From the Big Picture to Those Pesky Details

Starting a Digitization Project in Your Library

Mark Jordan
BC Libraries Conference, 2006

The process

1. Define goals and scope of the collection
2. Evaluate and select source material
3. Clear permission to use the source material
4. Define project objectives and preliminary milestones
5. Determine technical specifications
   A. Metadata
   B. Search and display
   C. File formats
   D. Content Management System
6. Develop workflows
7. Determine preliminary procedures based on workflows; begin project documentation
8. Determine what resources you need (hardware, software, staff)
9. Decide if you will outsource
10. Develop budget
11. Identify and acquire necessary resources
12. Finalize milestones
13. Finish project documentation
14. Hire and train staff, if necessary
15. Execute the project
    A. Create the content and metadata
    B. Add content and metadata to CMS
16. Evaluate the project
17. Evaluate the collection

Projects vs. a program

- Projects are discrete, with a defined start and end
- A program integrates digitization into the library's daily services
  - Users with disabilities
  - Course reserves
  - Distance learners
  - Theses, projects, papers
  - Institutional Repositories
  - Local historical collections

1. Goals and scope of collection

- Goals: A statement of the reasons for digitizing the material, identifying primary and secondary audiences for the digital collection, and determining how the material will be organized and presented.
- Scope: Describes the amount of material that will be in the collection, and can be expressed in terms of numbers of items, geographic coverage, temporal coverage, or any other aspect of the collection that is appropriate.

Sample goals statement

“The Southeastern Regional News Collection contains selected issues of the *County Herald* and *Crighton Daily* newspapers published between 1900 and 1920. The collection, which will be freely available to everyone over the World Wide Web, will be of interest to local historians, to genealogists, and to the students seeking primary source material from the early part of the 20th century. Access to the major articles and in each issue will be aided by the addition of subject keywords. Each newspaper issue will be presented as a single Adobe Acrobat file for easy printing.”

2. Evaluate source material

- Criteria for research collections
  - □ *Selecting Research Collections for Digitization*, Dan Hazen, Jeffrey Horrell, and Jan Merrill-Oldham
- Systematic approaches
  - □ Diane Vogt-O’Connor, “Selection of Materials for Scanning.” In Sitts, Maxine K., ed. *Handbook for Digital Projects: A Management Tool for Preservation and Access*.
- Flexible approaches
  - □ Columbia University Library’s “Selection Criteria for Digital Imaging”

Columbia's evaluation criteria

- Collection Development Criteria
  - Value
  - Demand
  - Non duplication
- Added Value Criteria
- Intellectual Property Rights Criteria
- Preservation Criteria
- Technical Feasibility Criteria
- Intellectual Control Criteria

http://www.columbia.edu/cu/libraries/digital/criteria.html

3. Clear permissions

- What Libraries Can Put Online
  - Fair Dealing (or other exemptions)
  - Material in the Public Domain
  - Material Covered by a Written Agreement
  - Material Distributed with an Explicit License

Managing permissions

1. Determine the copyright status of the item (either public domain or covered by copyright)
2. If the item is covered by copyright, determine whether fair dealing applies
3. If fair dealing does not apply, determine the rights holder and contact him or her
4. If the copyright owner grants permission, the item can be included in an online collection.

4. Define project objectives and milestones

- Project vs. collection
- Objectives describe the desired outcomes of the project, specifically, the deliverables expected at the end of the project
- Milestones describe desired outcomes at specific points in time throughout the project and are used to assist in monitoring the progress of the work being completed.

Sample objectives and milestones

Objectives

To digitize approximately 760 newspaper issues (9800 pages), with issue-level metadata.

Milestones

End of month	Number of pages to be digitized
1	984
2	1968
3	2952
4	3936
5	4920
6	5904
7	6888
8	7872
9	8856
10	9840

5. Determine technical specifications

- Files
- Metadata
- Search and retrieval
- Preservation

Files

- Create high quality masters and web-enabled derivatives
  - TIFF → JPEG, TIFF → PDF
- Follow best practices for images, text, audio, and video
  - NDIIP (http://www.digitalpreservation.gov/)

Metadata

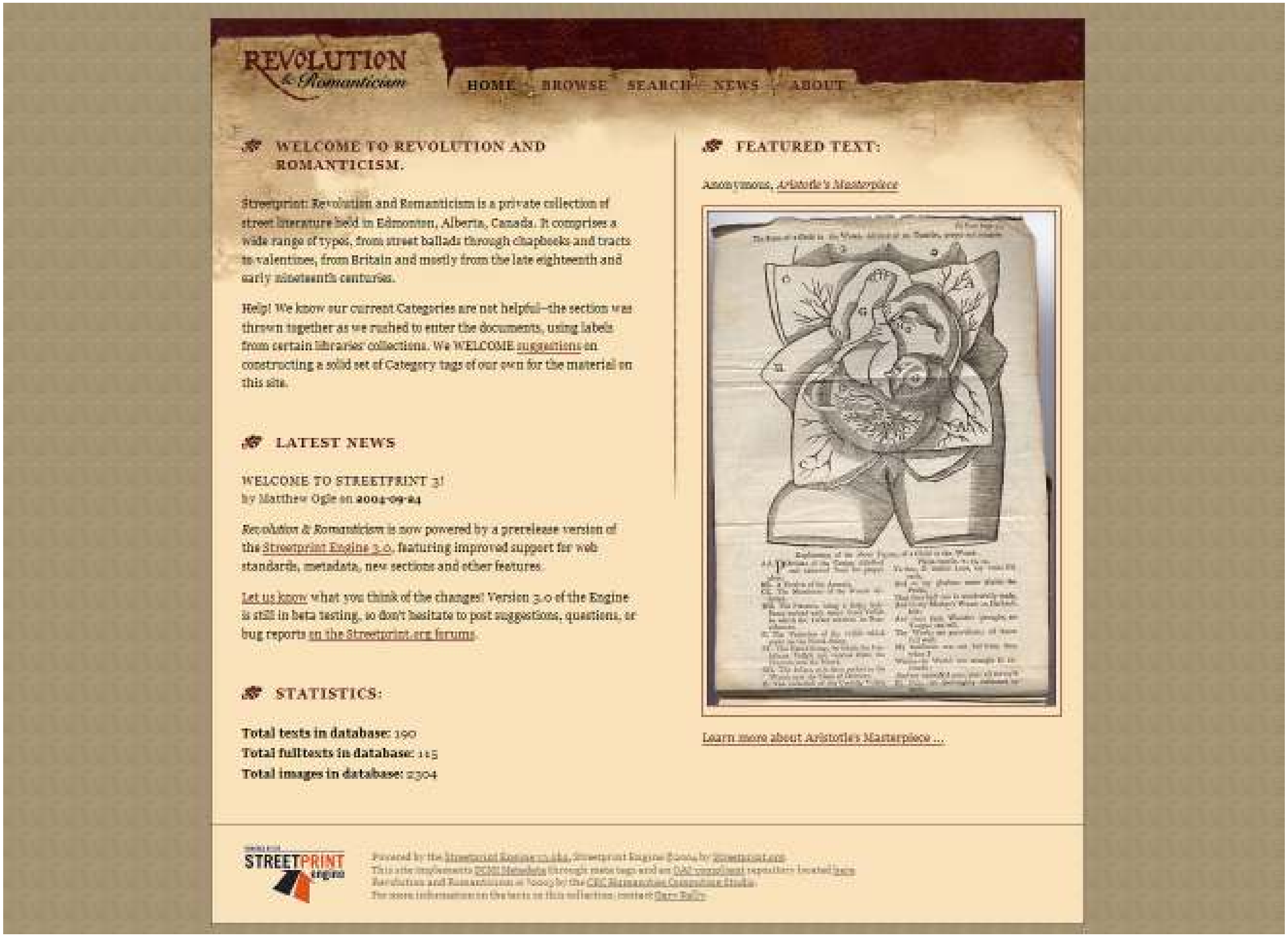
- Resist NIH (Not Invented Here) Syndrome
- Being like everyone else is a virtue
- Native vs. derived
- Best practice for descriptive: use Dublin Core but focus on vocabularies
- Structural, administrative, preservation metadata are all vital

Filenames and identifiers

- File names must be systematic and consistent
- Identifiers
  - Opaque (car4806) vs. transparent (1-1967-12-04)
  - Internal (5183) vs. public (http://urlofyourresource.ca/1-1967-12-04/)

Search and retrieval

- How your collection “works” online
- Related to metadata, file types, content management systems
- Images, text, audio, video, data all have their own requirements
- Mixed collections have their own too
- Requirements will change over time



WELCOME TO REVOLUTION AND ROMANTICISM.

Streetprint: Revolution and Romanticism is a private collection of street literature held in Edmonton, Alberta, Canada. It comprises a wide range of types, from street ballads through chapbooks and tracts to valentines, from Britain and mostly from the late eighteenth and early nineteenth centuries.

Help! We know our current Categories are not helpful–the section was thrown together as we rushed to enter the documents, using labels from certain libraries' collections. We WELCOME suggestions on constructing a solid set of Category tags of our own for the material on this site.

LATEST NEWS

WELCOME TO STREETPRINT 3!
by Matthew Ogle on **2004-09-24**

Revolution & Romanticism is now powered by a prerelease version of the Streetprint Engine 3.0, featuring improved support for web standards, metadata, new sections and other features.

Let us know what you think of the changes! Version 3.0 of the Engine is still in beta testing, so don't hesitate to post suggestions, questions, or bug reports on the Streetprint.org forums.

STATISTICS:

Total texts in database: 190
Total fulltexts in database: 115
Total images in database: 2304

FEATURED TEXT:

Anonymous, *Aristotle's Masterpiece*

Learn more about Aristotle's Masterpiece ...

Powered by the Streetprint Engine v3.0a. Streetprint Engine ©2004 by Streetprint.org
This site implements DCMI Metadata through meta tags and an OAI-compliant repository located here
Revolution and Romanticism © 2003 by the CRC Humanities Computing Studio
For more information on the texts in this collection, contact Gary Kelly

http://streetprint.org

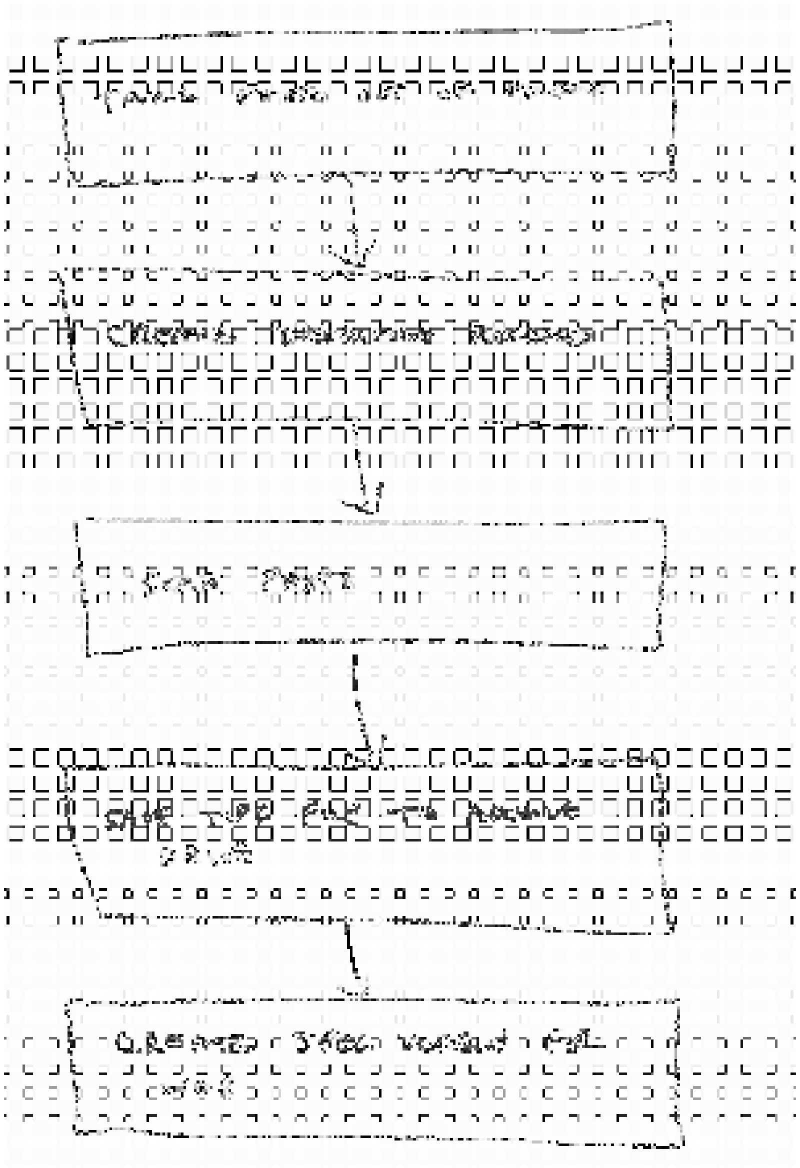
Preservation

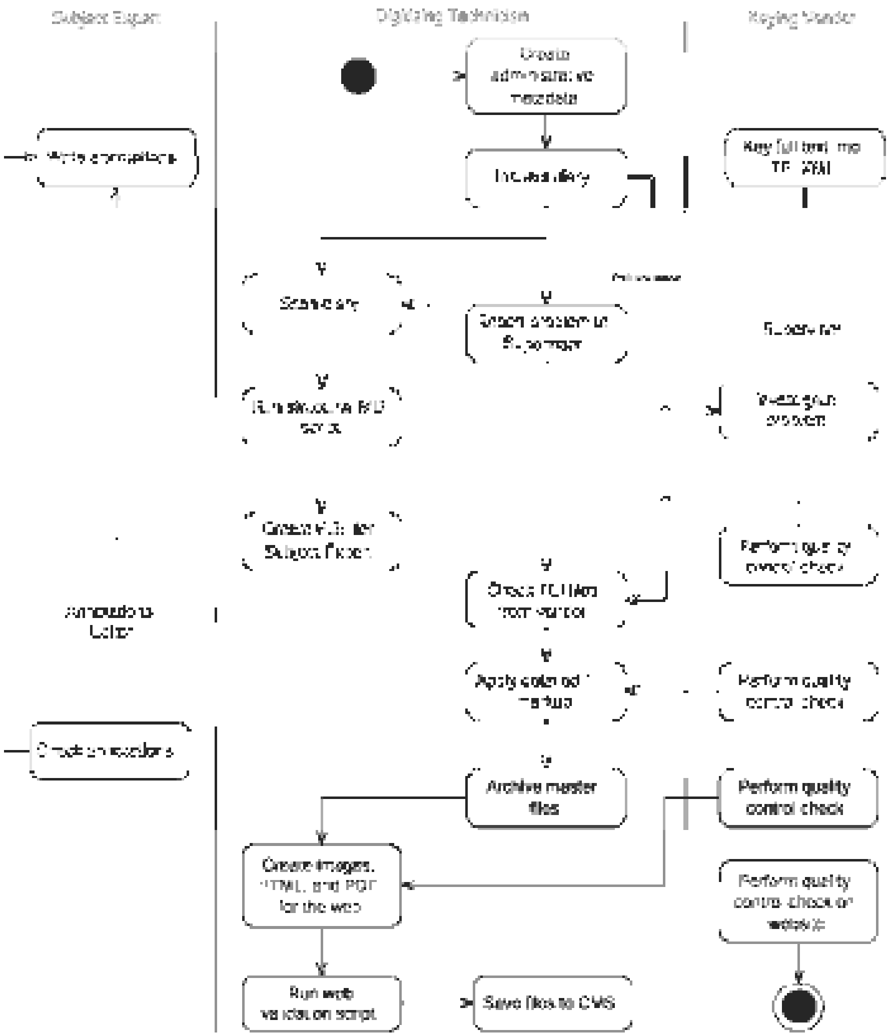
- If you’re not doing anything about it, you’re not alone…
- Trusted Digital Repositories
- First step: develop a written policy
- Metadata schemas and practical tools are becoming available
  - PREMIS, JHOVE
- Not all masters are worth keeping

6. Develop workflows

- Define and order tasks necessary to create deliverables
- Can assist in costing, clarify roles
- Promote process reusability
- Assist in creating procedural documentation
- Methods: Outlining, diagramming

Workflow stage	Required activity
Create metadata	1. Convert MARC records into required structure, supplement them if necessary 2. Create administrative metadata 3. Create unique identifiers and filename scheme
Prepare originals	1. Inspect diaries to see whether they are suitable for scanning 2. Have subject experts nominate candidates for annotations 3. Write annotations
Capture / convert	1. Scan diaries 2. Vendor: Key texts 3. Create structural metadata
Process	1. Vendor: Create basic TEI markup from digital images 2. Apply detailed markup, inc. add element attributes to enable linking to annotations 3. Create web versions from master images
Store	1. Save master images and XML to archive drive 2. Save web images and XML to web server




Subject Expert
Digitizing Technician
Keying Vendor
Create administrative metadata
Write annotations
Key full text into TEI XML
Run structural MD script
Create MD for Subject Expert
Check TEI files from vendor
Investigate problem
Perform quality control check
Annotations Editor
Apply editorial markup
Check annotations
Archive master files
Create images, HTML, and PDF for the web
Run web validation script
Save files to CMS
Perform quality control check on website

7. Determine preliminary procedures

- Workflow is high level, procedures are low level
- Preliminary vs. final
- Refine later – we’re still in planning phase

Intermission

The process (so far)

1. Define goals and scope of the collection
2. Evaluate and select source material
3. Clear permission to use the source material
4. Define project objectives and preliminary milestones
5. Determine technical specifications
   A. Metadata
   B. Search and display
   C. File formats
   D. Content Management System
6. Develop workflows
7. Determine preliminary procedures based on workflows; begin project documentation
8. Determine what resources you need (hardware, software, staff)
9. Decide if you will outsource
10. Develop budget
11. Identify and acquire necessary resources
12. Finalize milestones
13. Finish project documentation
14. Hire and train staff, if necessary
15. Execute the project
    A. Create the content and metadata
    B. Add content and metadata to CMS
16. Evaluate the project
17. Evaluate the collection

Project documentation

- Necessary for efficient project operations
- Vital for consistency
- Allows institution to develop expertise
- Allows for more efficient planning and budgeting in future

8. Determine required resources

- Determined by workflows
  - Hardware
  - Software
  - Staff
  - Space

9. Decide if you will collaborate and/or outsource

- Possible roles
  - Content ownership, evaluation
  - Metadata creation
  - Digitization
  - Hosting
  - Preservation

Multi-institution projects

- Defining roles, deliverables, and timelines is essential
- Example: Our Roots
  - □ Funded by CCOP
  - □ U of Calgary, U Laval lead partners
  - □ Regional "nodes"
  - □ Partners

Outsourcing

- Benefits
  - Institution need not invest in hardware, space, etc.
  - Staffing requirements
  - High economies of scale
  - Predictable cost
- Drawbacks
  - Institution loses some control
  - Source material must be sent off site
  - Risk may be high

Janet Gertz, "Vendor Relations" in *Handbook for Digital Projects: A Management Tool for Preservation and Access*. Ed. Maxine K. Sitts. Northeast Document Conservation Center, 2000.

Outsourcing: example vendors

- Micro Com Systems
  - http://microcomsys.com/
- OCLC
  - http://www.oclc.ca
- E-BookServices
  - http://www.e-bookservices.com/
- Discount Document Scanning
  - http://discountdocumentscanning.com/
- Academic Imaging Associates
  - http://www.academicimaging.com/

10. Develop a budget

- Extremely difficult to generalize
- Every project is different
- Should be based on similar projects or on samples

Cost variability

Projects funded by the Central New York Library Resources Council

Reported Cost per image	Details
$3.74	Broadside collection; 306 titles/2673 pages; HTML versions of broadsides; scan, OCR, correct.
$36.87/$12.63 w/o equipment	29 architectural drawings, 7 pamhlets/169 pages; HTML versions of pamphlets; scan, OCR, correct.

Cost Comparisons for Digitization Projects. Prepared for the CLRC Automation Committee, September, 2002. http://clrc.org/lstadigital/CostComparDigitizProjRev.pdf

Cost components

- 1/3 cost is conversion (i.e., scanning)
- Slightly less than 1/3 cost is metadata creation
- Slightly more than 1/3 cost is administration and quality control

Puglia, Steven. “The Costs of Digital Imaging Projects” RLG DigiNews 3.5 (1999): http://www.rlg.org/preserv/diginews/diginews3-5.html#feature

Evidence-based budgeting

- Uses trials and samples to determine estimated costs
- See handout for example

11. Acquire necessary resources

- Determined initially by workflows, refined at this point by collaboration, outsourcing decisions
    - Hardware
    - Software
    - Space
    - Staff

12 & 13. Finalize production details

- Milestones
- Documentation
  - Procedural
  - Finances
  - Staffing
  - Monitoring
  - Meetings

14. Hire and train staff

- Roles, job descriptions finalized by now
- Workflows, tools, documentation now in place
- Space must be finalized by this time as well

Staff roles

- Project manager
- Selector
- Conservator
- Cataloguer
- Scanning tech
- Quality control tech
- Data entry tech
- Programmer/CMS tech
- Systems admin
- User interface designer

Stephen Chapman, "Considerations for Project Management" in *Handbook for Digital Projects: A Management Tool for Preservation and Access*. Ed. Maxine K. Sitts. Northeast Document Conservation Center, 2000.

15. Execute the project

- Opening day
- Supervisor(s) must...
  - Schedule staff
  - Queue work
  - Perform (or delegate) quality control
  - Monitor project
  - Handle exceptions, problems

Monitoring

- Staff supervision
- Are we on track?
  - Time
  - Cost
- Reporting

Quality Control

- Don’t ignore it
- Integrate into workflow
- Establish benchmarks and exception procedures before production begins
- Should be effective but efficient

Create content and metadata

- Rely on procedural documentation
- Be prepared for exception / problem handling
- Aren’t you glad you spent so much time on workflows?

Put it online and offline

- Creation vs. publishing
- Derivative files vs. master versions
- Final testing and approval

16. Evaluate the project

- Production goals and milestones
- Budget
- Quality benchmarks
- Operational aspects
- Learn from mistakes

17. Evaluate the collection

- Principles: NISO *Framework of Guidance for Building Good Digital Collections*
- Mechanisms for user evaluation: NINCH *Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials*
- IMLS's Outcome Based Evaluation

SFU Cartoons Collection Logs

- Number of searches / clicks on subject headings per day
  - 120
- Keywords searched more than 80 times
  - Canada (126), Trudeau (113), Canadian (109), money (89), Quebec (89), Gordon (81)
- Searches that return no hits
  - 39%
- Searches comprised of a single keyword
  - 72%

Wrap up: 45 slides in 5 points

- Clear collection goals and scope lead to concrete project objectives
- Killers: Copyright, lack of institutional support, unclear roles, sloppy monitoring
- Invest in technical specifications, workflows, and budgeting
- Integrate evaluation into collections and projects
- Successful projects lead to successful programs