Cost forecasting model for new digitization projects

CNI December 13th, 2011

Karim Boughida, AUL for Digital Initiatives & Content Management
Martha Whittaker, Dir of Content Management
Linda Colet, Project Consultant / President, DaoPoint, Digital LLC
Dan Chudnov, Dir of Scholarly Technology
Cultural Imaginings: the Creation of the Arab World in the Western Mind

From the collections of

- George Washington University Libraries
- Lauinger Library at Georgetown University
The Collections

• George Washington University Libraries
  – Middle Eastern Institute Rare Book Collection
  – I. Edward Kiev Judaica Collection
  – Andrew Oliver Archaeology Collection

• Georgetown University Lauinger Library, Department of Special Collections
  – Orientalist accounts and images of Turkey and the Levant
  – Images and etchings of the Holy Land
  – Jesuitica relating to the Orient
Regions and Languages of the Cultural Imaginings Collection

Regions of the Collection

Languages of the Collection

CNI11f-GWU-Cost-digitization
Kabis III Robotic Arm Scanner (KIRTAS)
Linda Colet

• President of DaoPoint Digital, LLC
  specializing in digital planning for the arts
  www.daopoint.com/digital/

• Clients include:
  • The Smithsonian
    – Smithsonian Institution Archives
    – Cooper Hewitt National Design Museum
  • George Washington Universities
  • Harvey B. Gantt Center for African American Art
  • Arts Council of Fairfax County, VA
  • Whitney Museum of Art
  • Wadsworth Atheneum Museum of Art
Perspective I bring…

How to…

– Implement cataloguing and digital projects for museums and archives
– Develop small and large-scale workflow solutions to improve standards and processes
– Budget for digitization projects at both the institutional and pan-institutional level
Cost Model Concept

• Prepare a forecasting cost model for institutions to plan out (predict) digitization budgets to determine cost per page and total project costs.
Defining the cost model

• Our cost model
  – What it is:
    • A case study that tracks variables and associated costs of the GWU/Georgetown project.
    • A model that offers institutions a way to predict costs of their project to help budget and apply for grants.
  – What it is not:
    • A broad calculation of every possible variable that exists
    • A model that takes into account every type of book or collection that exists
Quality standards

Cost model will provide categories that institutions can fit in... small, medium, and large budgets

<table>
<thead>
<tr>
<th>Best Quality</th>
<th>Good Quality</th>
<th>Fair Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Preservation, Print on Demand</td>
<td>• Preservation, print on demand at a lower level</td>
<td></td>
</tr>
<tr>
<td>• CR2, TIFF, jp2</td>
<td>• CR2, TIFF (derived from jpegs), jp2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Identification purposes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Jpegs, jp2</td>
<td></td>
</tr>
</tbody>
</table>
Research before we began…

- Besser, Howard, Bonn, Maria, et al. NINCH SYMPOSIUM: April 8, 2003, New York City The Price of Digitization
- Good Practices in Cost Reduction for Digitisation www.minervaeurope.org
Cost models we reviewed

- **Internet Archive**
  - Cost $0.10 a page to digitize a book (300-page book is $30 a book).

- **ENUMERATE**
  - $1.30 per page
  - Cost calculator provided

- **British Library Lifecycle**
  - Cost model based on variables
Cost Model: Enumerate

- [http://www.enumerate.eu/](http://www.enumerate.eu/)
- **ENUMERATE** is a EC-funded project, led by Collections Trust in the UK. The primary objective of ENUMERATE is to create a reliable baseline of statistical data about digitization, digital preservation and **online access** to cultural heritage in Europe.

Great reference but cost calculator does not break down variables for our needs
According to this cost calculator, digitizing 2500 books should cost at the range of $507,253 - $697,473.

Collections Trust Digitisation Cost Calculator


| LIBRARY |
|------------------|------------------|------------------|------------------|
| My project is a: | In-house digitisation programme | (select project type) |
| I am in a:      | National Library | (select library type) |
| I am digitising | 2500 BOOKS      | at a cost of BETWEEN 310000 AND 426250 EURO |
| I am digitising | 0 RARE BOOKS    | at a cost of BETWEEN 0 AND 0 EURO |
| I am digitising | 0 pages of ARCHIVAL MATERIAL | at a cost of BETWEEN 0 AND 0 EURO |
Cost model:
\[ K(t) = s + ipr + cons + r + cap + q + m + acs(t) + p(t) \]

- Based on what its costs long-term to manage digitization project
- Identifies important variables that affect costs but doesn’t tell you how to calculate them.
How GWU cost model is different

• How we differentiate from these studies..
  – A way to track variables at the project planning level to prepare for budget forecasting and grant applications.
  – Focus is on 3-5 yr costs of a project.
The major inputs to the cost model are resources, hardware, software, and production throughput.

\[
\text{Cost per Page Completed} = \frac{\text{Production Throughput Total Number of Pages}}{\text{Resources Salaries, Benefits} + \text{Hardware & Software Scanning Equipment Data Storage Software/Computers}}
\]
GWU Cost Model Approach

- Learn about current workflow and processes
- Identify variables and bottlenecks affecting production
- Budget review (grant amount / actual expenditures)
- Research cost models on digitization projects
- Work with staff to collect metrics
- Create cost model concept (a forecasting tool)
- Build cost model template in Excel
- Create web interface (pending)
Collecting metrics

• Budgets
• Software reports re: workflow
• IT storage cost analysis
• Time study:
  – Log time it takes for each workflow step:
    • Selecting books for scanning
    • Scanning
    • Post-processing steps
    • Archiving and Access
  – Additional tests as needed
Metrics / Variables affecting cost

- Project Planning
- Scanning
- Processing
- Making Collection Available
- Within these categories…
  - Hardware, software, server costs
  - Staff costs
  - Production speed
The cost model is based on the Gelman Library Digital process

**Project Planning**
- **Scope:**
  - Book Selection & Prep
  - Project Management
  - Manager Oversight
  - Grant & Budget Mgt
  - Process Design
  - Metadata Creation
  - Archival standards
  - Procedures for managing digital content
- **Timelines:** 28 Months (September 2009 – Present)
- **Expenses & Resources:**
  - No major expenses outside resources
  - # Unique Resources
  - Total # Hours

**Scanning**
- **Scope:**
  - Prep for Scanning
  - Scanning
- **Timelines:** 17 Months (August 2010 – Present)
- **Expenses & Resources:**
  - Scanning Equipment:
  - # Unique Resources
  - Total # Hours

**Processing**
- **Scope:**
  - Post Processing
  - QA/QC
  - Book Tagging
  - OCR
- **Timelines:** 17 Months (August 2010 – Present)
- **Expenses & Resources:**
  - Data Storage Costs:
  - Additional hardware purchased to handle the workflow needs
  - # Unique Resources
  - Total # Hours

**Make Collection Available**
- **Scope:**
  - Website Design
  - Web Development
  - OCR
  - Storage Design
  - Technology Management
- **Timelines:** 17 Months (August 2010 – Present)
- **Expenses & Resources:**
  - No major expenses outside resources
  - # Unique Resources
  - Total # Hours
<table>
<thead>
<tr>
<th>Select</th>
<th>Prepare</th>
<th>Set-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Project theme</td>
<td>• Pull books from stacks</td>
<td>• Set characteristics</td>
</tr>
<tr>
<td>• Public domain</td>
<td>• Vacuum/physically prepare them</td>
<td>• Measure book width</td>
</tr>
<tr>
<td>• Physical requirements</td>
<td>• Conserve</td>
<td>• Set speed settings (pgs per hr)</td>
</tr>
<tr>
<td>• Sturdy enough to scan</td>
<td>• Prepare/check metadata</td>
<td>• Choose image folder for daily images</td>
</tr>
<tr>
<td></td>
<td>• Scan techs review</td>
<td>• Enter barcode #</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Set scan settings (adjust pressure of clamps, speed of pages turned, align page, adjust/focus camera, center book, etc.)</td>
</tr>
<tr>
<td>Scan</td>
<td>Process</td>
<td>Access</td>
</tr>
<tr>
<td>------------------------------</td>
<td>---------------------------------------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>• Take manual image of front and back covers</td>
<td>• Create template, run test pages, and save book parameters</td>
<td>• Dspace: digital repository that GWU uses for any type of digital documentation.</td>
</tr>
<tr>
<td>• Re-focus in middle of book’s gutters</td>
<td>• Run QC for missing pgs, adjust templates, cleanup errors, etc.</td>
<td>• Using it to display PDF files</td>
</tr>
<tr>
<td>• Manually image first 20 pages</td>
<td>• Run QA on the page scans to ensure quality results</td>
<td>• Not displaying jp2 yet</td>
</tr>
<tr>
<td>• Press start button for robotic arm to turn pages</td>
<td>• Run OCR (automated) for final outputs (PDF, Mets, etc.)</td>
<td>• Still in development…</td>
</tr>
<tr>
<td>• Scan color target</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Manually image last 20 pages</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Improved production workflow procedures throughout the project

Workflow #1
(Aug 2010 – Feb 2011)
Ramp up and begin scanning in CR2 and JPEG Small format. Then processed in color.

Workflow #2
(Aug 2010 – Feb 2011)
Scan in CR2 and JPEG Small. Then processed in grayscale.

Workflow #3
(Feb 2011 – Dec 2011)
Scan in CR2 and Jpeg Large and use jpeg for access copy to speed up production. Images are processed in grayscale.
First Workflow (August 2010 – Feb 2011)

Ramp Up
• Identify workflow standards
• Set up workflow
• Test scans
• Train staff
• Create scripts

Begin Scanning
• Scan file and save in raw format (CR2)
• Convert CR2 to uncompressed TIFFs
Second Workflow (Feb – September 2011)

Scan to color and process in grayscale

Revise Scanning
- Scan in color and process in grayscale; instead of color, since file sizes were too large to produce PDFs needed for access

Optimized scanning process
- Scan file and save in raw format (CR2)
- Convert CR2 to uncompressed TIFFs
- Saving text in grayscale = reasonable sized PDFs
Optimize staff efficiency:
- Hired experienced f/t scanning tech = dramatic production increase

Scan to color and process in grayscale but save in jpeg instead of TIFF

Optimize scanning
Use Jpeg for access copy.

Production improvement:
- 4 hours vs. 7 hrs to complete scanning a book

3rd Workflow (September 2001 – present)
<table>
<thead>
<tr>
<th>Task (based on a 350 page book)</th>
<th>Old workflow</th>
<th>New workflow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scan book in CR2 and jpeg</td>
<td>20-30 min</td>
<td>20-30 min</td>
</tr>
<tr>
<td>Convert CR2 to uncompressed unprocessed TIFF</td>
<td>30-40 min</td>
<td>*</td>
</tr>
<tr>
<td>Complete a template and process the book</td>
<td>60-80 min</td>
<td>30-40 min</td>
</tr>
<tr>
<td>Perform QC</td>
<td>1-2 hrs</td>
<td>1-2 hrs</td>
</tr>
<tr>
<td>Complete OCR</td>
<td>40-60 min</td>
<td>40-60 min</td>
</tr>
<tr>
<td>Create the preservation and access bags</td>
<td>1 hr</td>
<td>1 hr</td>
</tr>
<tr>
<td>Run program script #1 to convert CR2 to uncompressed unprocessed TIFF, and then, moves CR2 off production server.</td>
<td>Automated: no time recorded</td>
<td>Automated: no time recorded</td>
</tr>
<tr>
<td>Run program script #3 to create set of jp2's</td>
<td>Automated: no time recorded</td>
<td>Automated: no time recorded</td>
</tr>
<tr>
<td><strong>Total time</strong></td>
<td><strong>4 1/2 - 6 1/2 hrs</strong></td>
<td><strong>3 1/2 - 5 hrs</strong></td>
</tr>
</tbody>
</table>

Preservation bag will have jp2's from the original scanned CR2.

Access bag will have jp2's from the processed uncompressed TIFF.
Bottlenecks encountered

• Barcode of book scanned not matching metadata record.
• Waiting for materials to scan
• Server going down
• Post-processing took significant time and server couldn’t handle the load so some processes had to wait in a queue and be run at night (not during day as it would take up too much memory).
Overview of the Gelman Digital Collection: Books by Page Range

Basic Stats:
Total Number of Books: 1,181
Average Number of Pages: 430
Highest Production Month: August 2011
Overview of the Gelman Digital Collection: Books Completed By Month

Number of Books Completed by Month

<table>
<thead>
<tr>
<th>Month Finished</th>
<th># Books</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010-08</td>
<td>0</td>
</tr>
<tr>
<td>2010-09</td>
<td>60</td>
</tr>
<tr>
<td>2010-10</td>
<td>70</td>
</tr>
<tr>
<td>2010-11</td>
<td>70</td>
</tr>
<tr>
<td>2010-12</td>
<td>70</td>
</tr>
<tr>
<td>2011-01</td>
<td>100</td>
</tr>
<tr>
<td>2011-02</td>
<td>100</td>
</tr>
<tr>
<td>2011-03</td>
<td>100</td>
</tr>
<tr>
<td>2011-04</td>
<td>100</td>
</tr>
<tr>
<td>2011-05</td>
<td>100</td>
</tr>
<tr>
<td>2011-06</td>
<td>100</td>
</tr>
<tr>
<td>2011-07</td>
<td>100</td>
</tr>
<tr>
<td>2011-08</td>
<td>120</td>
</tr>
<tr>
<td>2011-09</td>
<td>90</td>
</tr>
<tr>
<td>2011-10</td>
<td>110</td>
</tr>
<tr>
<td>2011-11</td>
<td>100</td>
</tr>
</tbody>
</table>
The project planning phase had the most resources to set-up the infrastructure, processes, and standards for the project.

Resources by Stage

- Planning: 25
- Scanning: 5
- Processing: 7
- Making Collection Available: 6
- Total: 43

Process Stage
The project planning phase was also the most time consuming stage in the process.

The scanning and processing hours are growing at a more rapid pace now that the operational processes have been implemented.
The project planning phase resources tend to be professional staff with higher hourly rates.
The project planning phase was the most expensive, and the scanning phase was second due to hardware costs.
The production throughput can be adjusted based on the drivers.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Project Planning</th>
<th>Scanning</th>
<th>Processing</th>
<th>Making the Collection Available</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware Costs (Equipment, Servers, storage, etc)</td>
<td>$300,000.00</td>
<td>$40,000.00</td>
<td>$44,000.00</td>
<td>$384,000.00</td>
<td></td>
</tr>
<tr>
<td>Software Costs</td>
<td>$23</td>
<td>23</td>
<td>7</td>
<td>8</td>
<td>41.0</td>
</tr>
<tr>
<td>Number of Resources</td>
<td>35.00</td>
<td>3</td>
<td>16.00</td>
<td>$21.50</td>
<td></td>
</tr>
<tr>
<td>Total Hours per Month</td>
<td>1,221</td>
<td>480.00</td>
<td>416.00</td>
<td>483</td>
<td>2,600.00</td>
</tr>
<tr>
<td>Number of Months in Stage</td>
<td>8</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>79</td>
</tr>
<tr>
<td>Total Number of Hours</td>
<td>8,745</td>
<td>2,400</td>
<td>2,984</td>
<td>5,651</td>
<td>19,780.00</td>
</tr>
<tr>
<td>Resource Cost</td>
<td>$306,075.00</td>
<td>$40,800.00</td>
<td>$47,744.00</td>
<td>$101,718.00</td>
<td>$496,337.00</td>
</tr>
<tr>
<td>Number of Pages per Hour</td>
<td>NA</td>
<td>500.00</td>
<td>402.14</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Cost for Project Planning</td>
<td>$306,075.00</td>
<td>$340,800.00</td>
<td>$95,744.00</td>
<td>$165,718.00</td>
<td>$908,337.00</td>
</tr>
<tr>
<td>Cost per Page</td>
<td>$0.60</td>
<td>$0.67</td>
<td>$0.19</td>
<td>$0.33</td>
<td>$1.79</td>
</tr>
<tr>
<td>Total Number of Pages</td>
<td>1,200,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Quality is a measure of the accuracy, completeness, and validity of data in comparison to defined business requirements.

Data Quality is typically monitored and measured to ensure the reliability and effectiveness of data for a particular use in the fulfillment of business processes, decision making, planning and/or reporting.
Completeness, Validity and Accuracy are probably the key dimensions the project should focus

**Completeness**

Is all the necessary data present?

**Validity**

Are the data values within specifications?

**Accuracy**

Does the data reflect reality?

**Recommendations:**

• Develop a list of operational metrics to measure the business process. Examples: Number of pages complete, time to complete a page, cost per page

• For each metric determine the data needed and how often

• Build into the process the capture and reporting of the data and metrics
  – Data Capture: Manual logs; automated logs; customize application to collect the data
  – Metrics Reporting: Monthly review meetings

• For each piece of capture data create a list of expected valid values

• If data is manual captured – train people on the valid values and definitions

• Data Entry – only allow data entry for the valid values

• Build a custom application that records data automatically

• Build in data quality checks into the process to alert the process owner when data is not valid

• For all data that is built on an assumption – think about ways to capture data in the process to remove the assumption

• Build in data quality checks into the process to alert the process owner when data is not aligned with expectations (which does mean it is wrong)

Is all the necessary data present?

Are the data values within specifications?

Does the data reflect reality?
Dan Chudnov, Director of Scholarly Technology

From Project To Program
Cost analysis lessons

• Cost planning
• What to track
• Bottlenecks
• Effects of quality
• Infrastructure needs
Tracking

- Time per stage
- Per-item attributes
- Processing transitions
- Storage “float”
- Server usage
Bottlenecks

• Quality
• Storage
• Backups
• Access
• Communications
Quality

- Big, slow, hard
- Expensive
- Spectrum: from necessary to frivolous
Program goals

• Small-run reformatting
• Reliable
• Predictable
• Quality service
• Communication
• Free, or affordable
Infrastructure needs

- Lots of small projects
- Useful tracking
- Service focus
- Cross-trained staff
- Value from “float”
Infrastructure pieces

- Discrete apps
- Focus on functions: id, operations, inventory, storage, troubleshooting
- HTTP access to everything
- Web UI w/raw content access
Outcomes

• Composable
• Immediate use
• Iterate over apps
• Scale as needed
• Content / status visibility
$ 1.70 per page
Presenters

Linda Colet, lcolet@prodigy.net
Martha Whittaker, marthaw@gwu.edu
Dan Chudnov, @dchud, dchud@gwu.edu
Karim Boughida, @kboughida, boughida@gwu.edu
Thank You

• Grant US-IMLS NLG 2008