THE GEORGE WASHINGTON UNIVERSITY

LIBRARIES

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

THE GEORGE WASHINGTON UNIVERSITY
LIBRARIES

- George Washington University Libraries
- Lauinger Library at Georgetown University

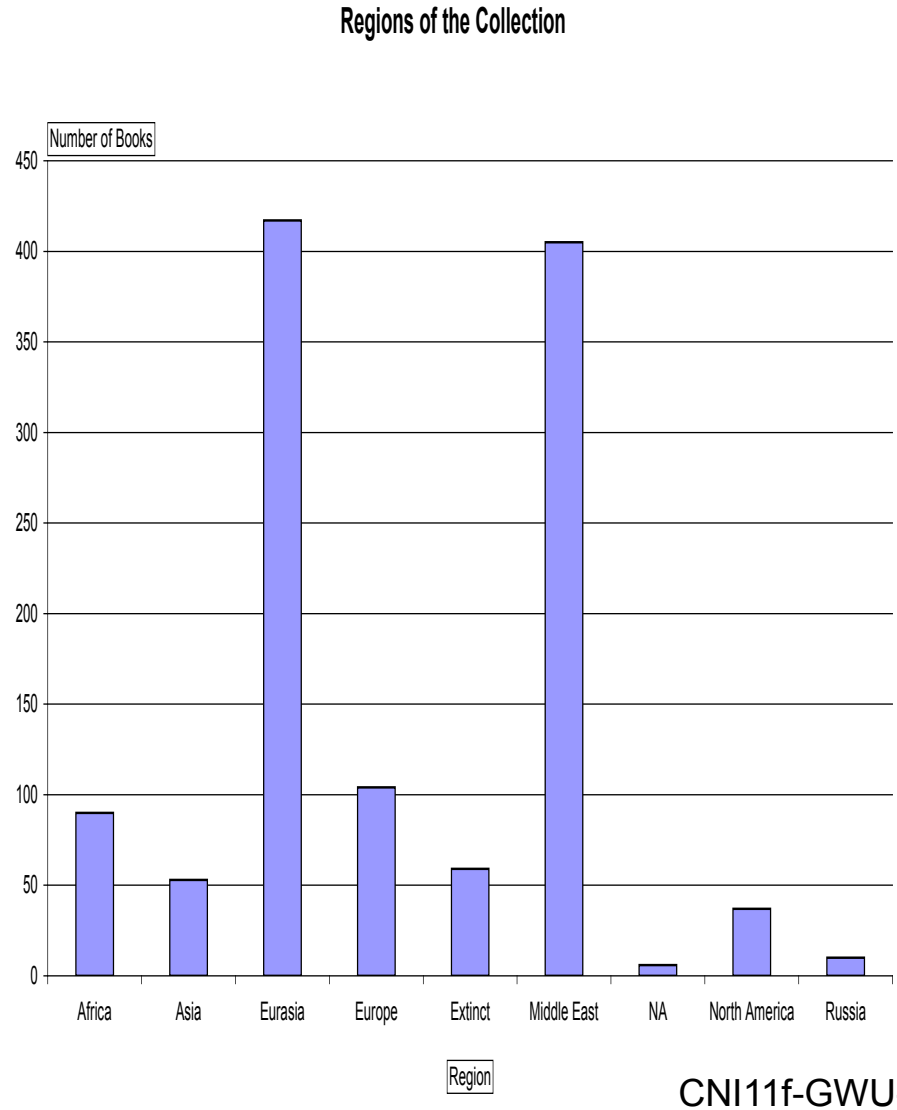
GEORGETOWN UNIVERSITY
LIBRARY

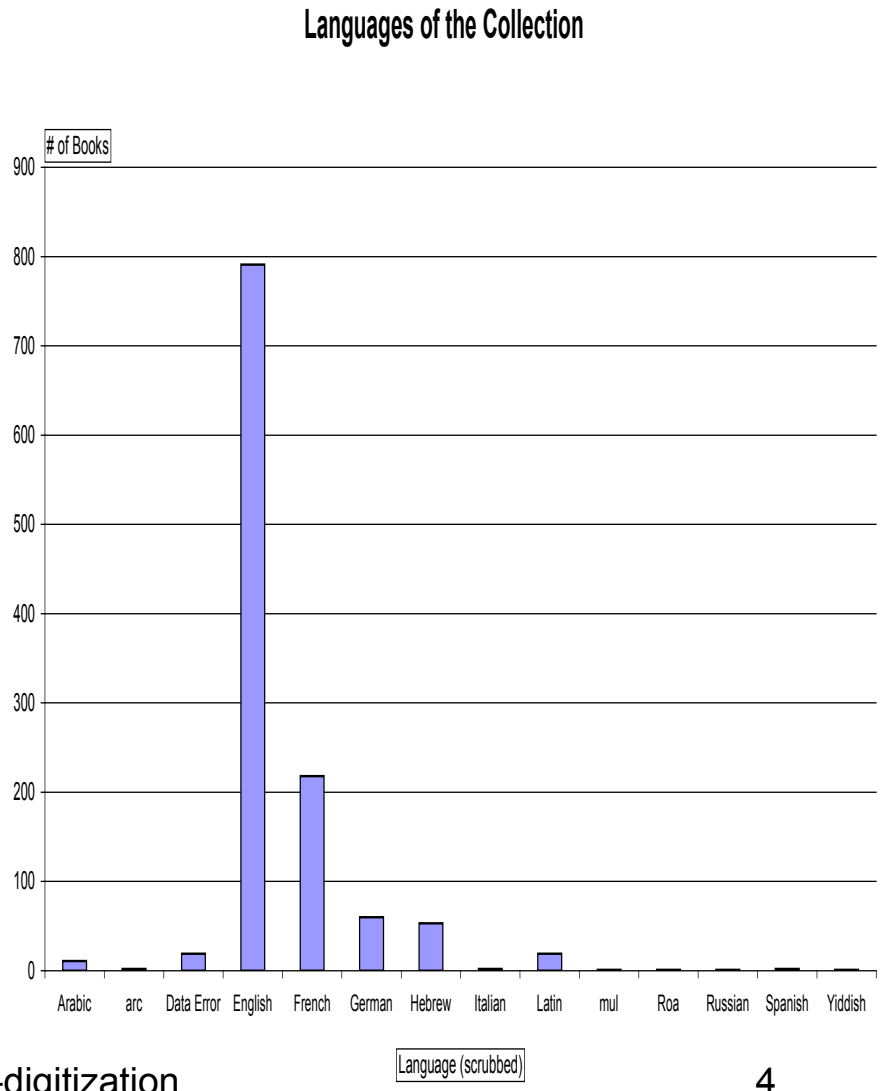
INSTITUTE of
Museum and Library
SERVICES

The Collections

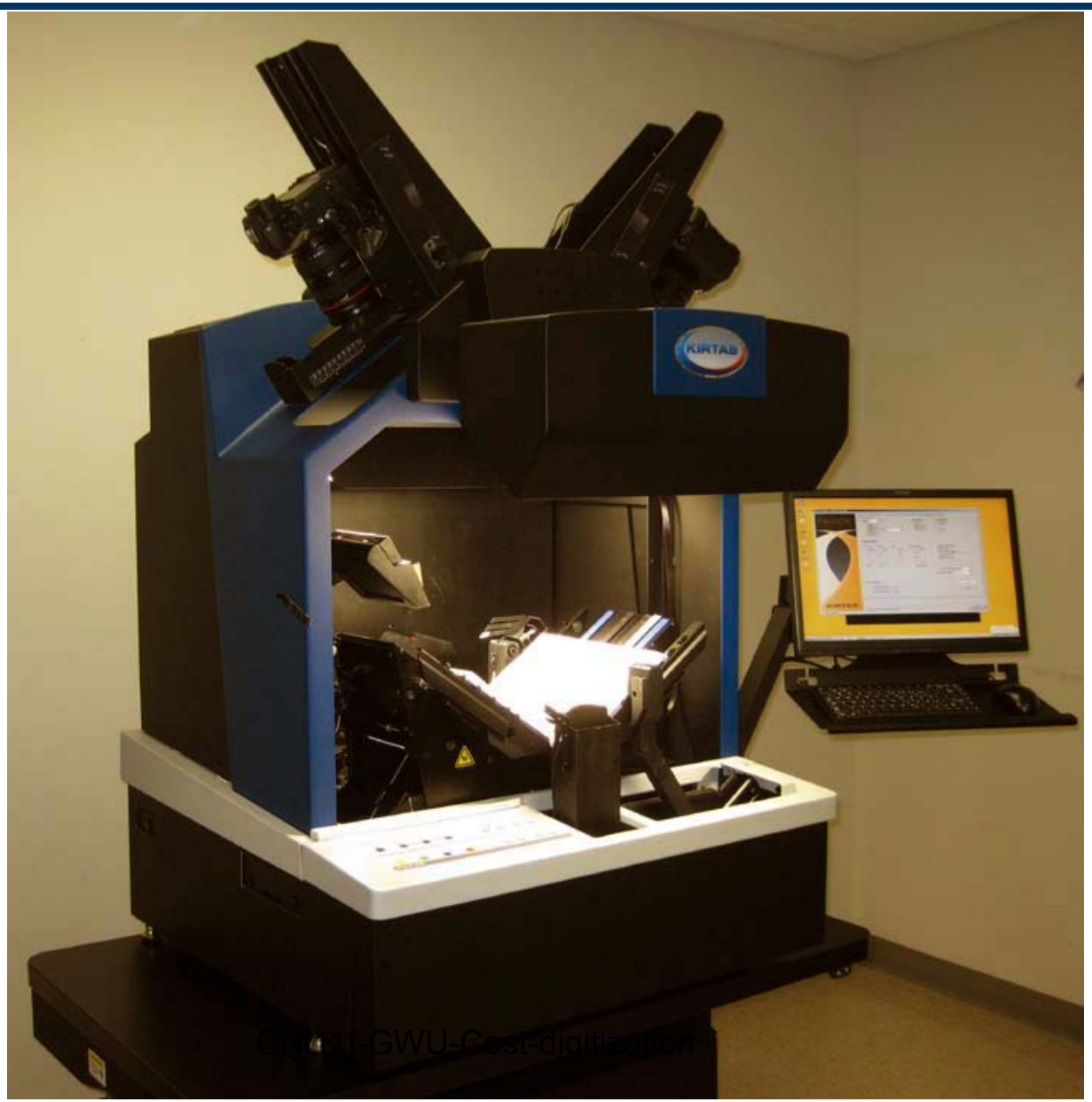
- George Washington University Libraries
  - Middle Eastern Institute Rare Book Collection
  - I. Edward Kiev Judaica Collection
  - Andrew Oliver Archaeology Collection
- Georgetown University Lauinger Library, Department of Special Collections
  - Orientalist accounts and images of Turkey and the Levant
  - Images and etchings of the Holy Land
  - Jesuitica relating to the Orient

Regions and Languages of the Cultural Imaginings Collection





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 Insitution 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

Best Quality	Good Quality	Fair Quality
• Preservation, Print on Demand • CR2, TIFF, jp2	• Preservation, print on demand at a lower level • CR2, TIFF (derived from jpegs), jp2	• Identification purposes • Jpegs, jp2

Research before we began…

- Steven Puglia's article, "The Costs of Digital Imaging Projects," in RLG DigiNews (October 15, 1999)
- 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 its $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/
- **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

Collections Trust Digitisation Cost Calculator

http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/annexes/digiti_report.pdf

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

British Library Lifecycle

http://www.life.ac.uk/

Cost model:

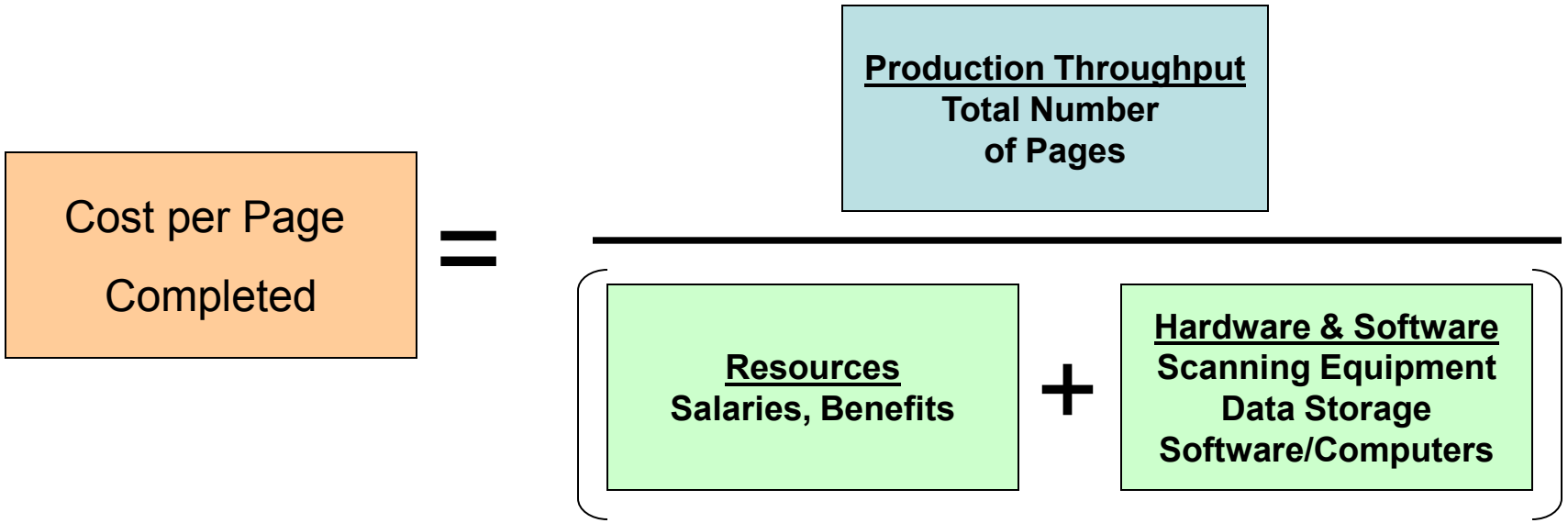
K(t) = s + ipr + cons + r + cap + q + m + acs(t) + p(t)

- 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



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)

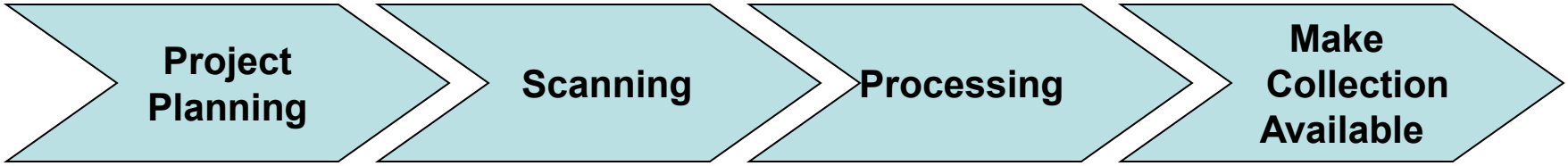
Collecting metrics

- 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



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

Scope:
Prep for Scanning
Scanning

Timelines: 17 Months
(August 2010 – Present)

Expenses & Resources:
- Scanning Equipment:
- # Unique Resources
- Total # Hours

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

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

Select

- Project theme
- Public domain
- Physical requirements
- Sturdy enough to scan

Prepare

- Pull books from stacks
- Vacuum/physically prepare them
- Conserve
- Prepare/check metadata
- Scan techs review

Set-up

- Set characteristics
  - Measure book width
  - Set speed settings (pgs per hr)
  - Choose image folder for daily images
  - Enter barcode #
  - Set scan settings (adjust pressure of clamps, speed of pages turned, align page, adjust/focus camera, center book, etc.)

Scan

- Take manual image of front and back covers
- Re-focus in middle of book's gutters
- Manually image first 20 pages
- Press start button for robotic arm to turn pages
- Scan color target
- Manually image last 20 pages

Process

- Create template, run test pages, and save book parameters
- Run QC for missing pgs, adjust templates, cleanup errors, etc.
- Run QA on the page scans to ensure quality results
- Run OCR (automated) for final outputs (PDF, Mets, etc.)

Access

- Dspace: digital repository that GWU uses for any type of digital documentation.
  - Using it to display PDF files
  - Not displaying jp2 yet
  - Still in development…

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 and Scan to color

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

3rd Workflow (September 2001 – present)

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

Optimize staff efficiency:

- Hired experienced f/t scanning tech = dramatic production increase

Optimize scanning

Use Jpeg for access copy.

Production improvement:

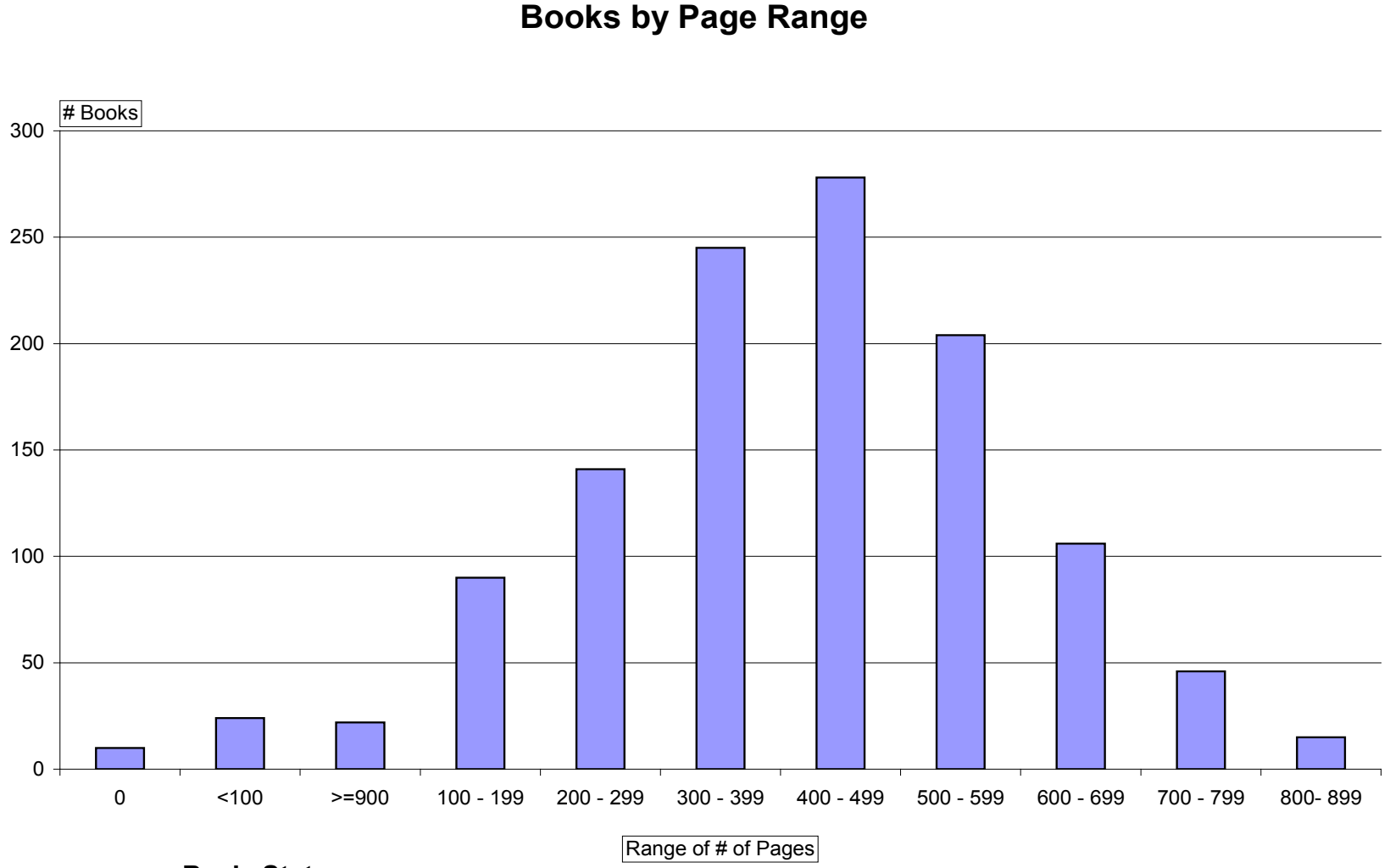
- 4 hours vs. 7 hrs to complete scanning a book

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



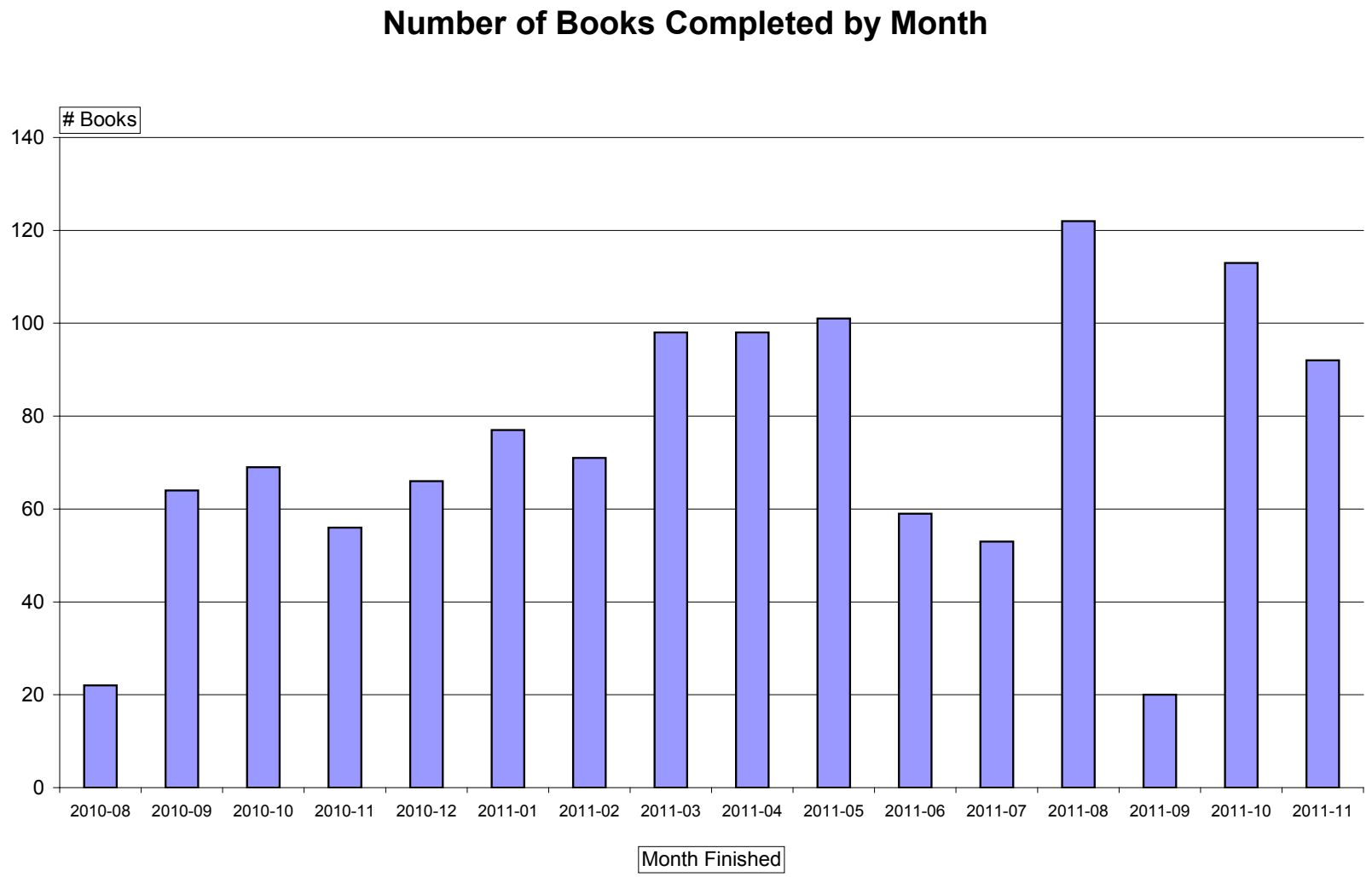
Basic Stats:

Total Number of Books: 1,181

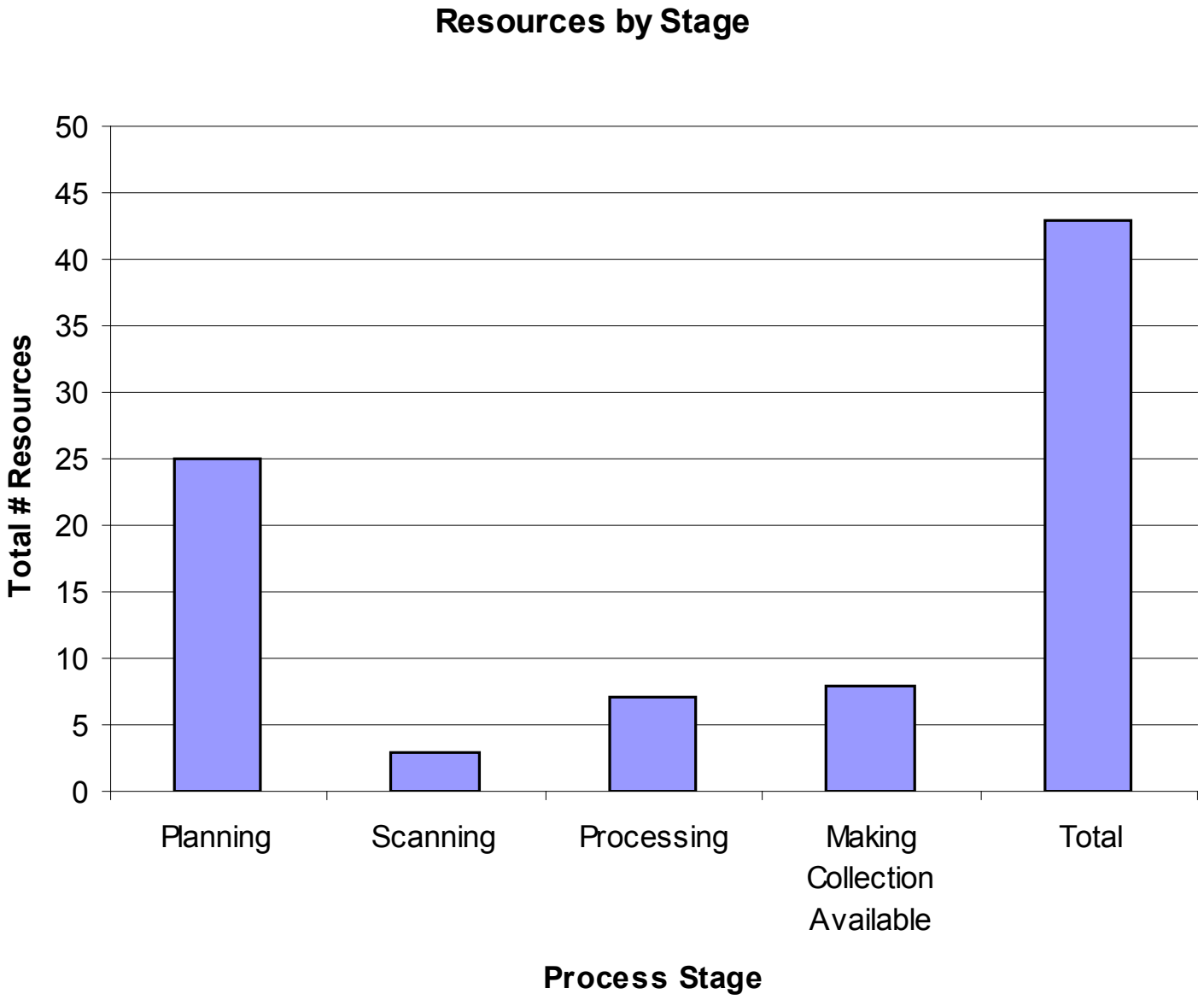
Average Number of Pages: 430

Highest Production Month: August 2011

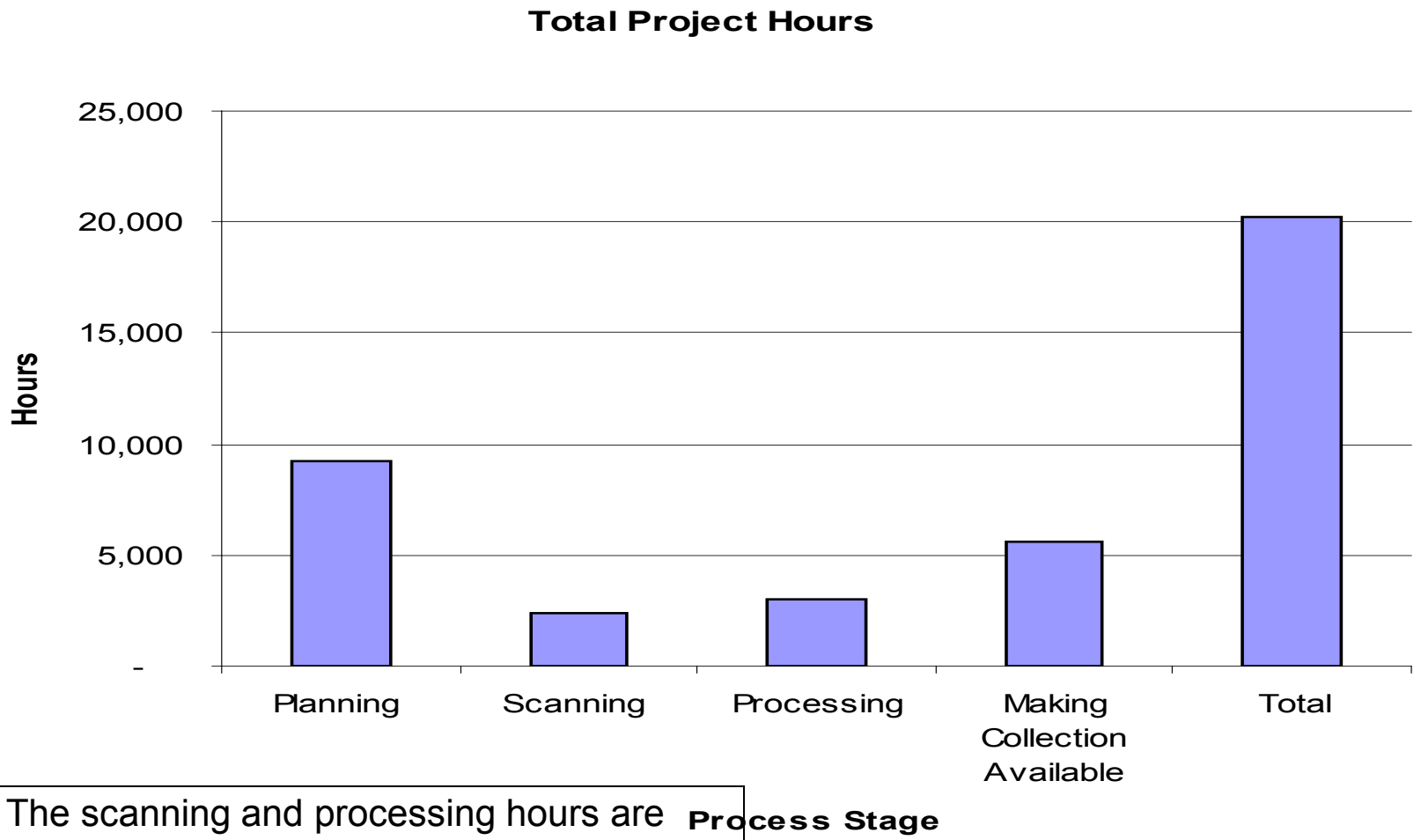
Overview of the Gelman Digital Collection: Books Completed By Month



The project planning phase had the most resources to set-up the infrastructure, processes, and standards for the project

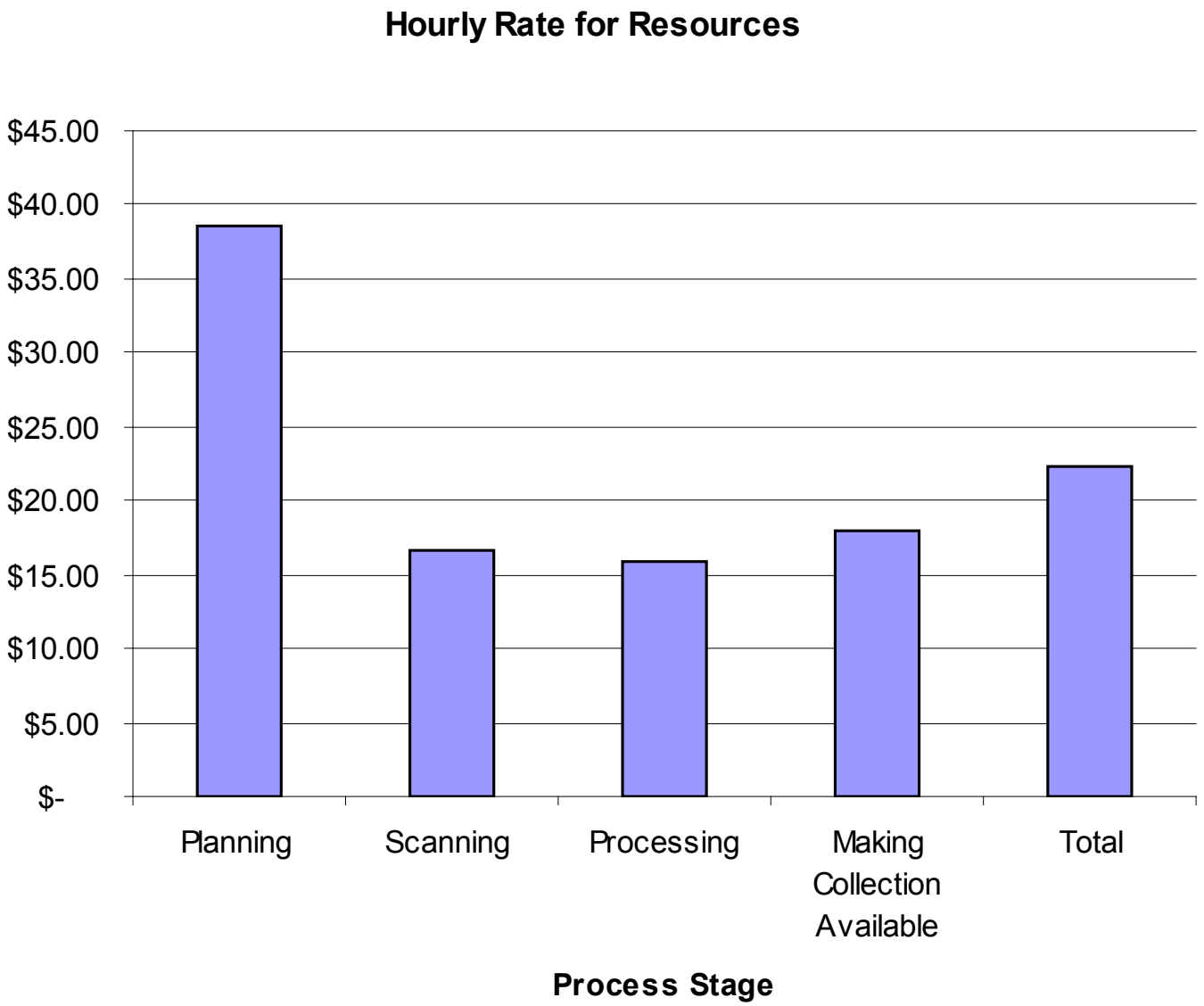


The project planning phase was also the most time consuming stage in the process

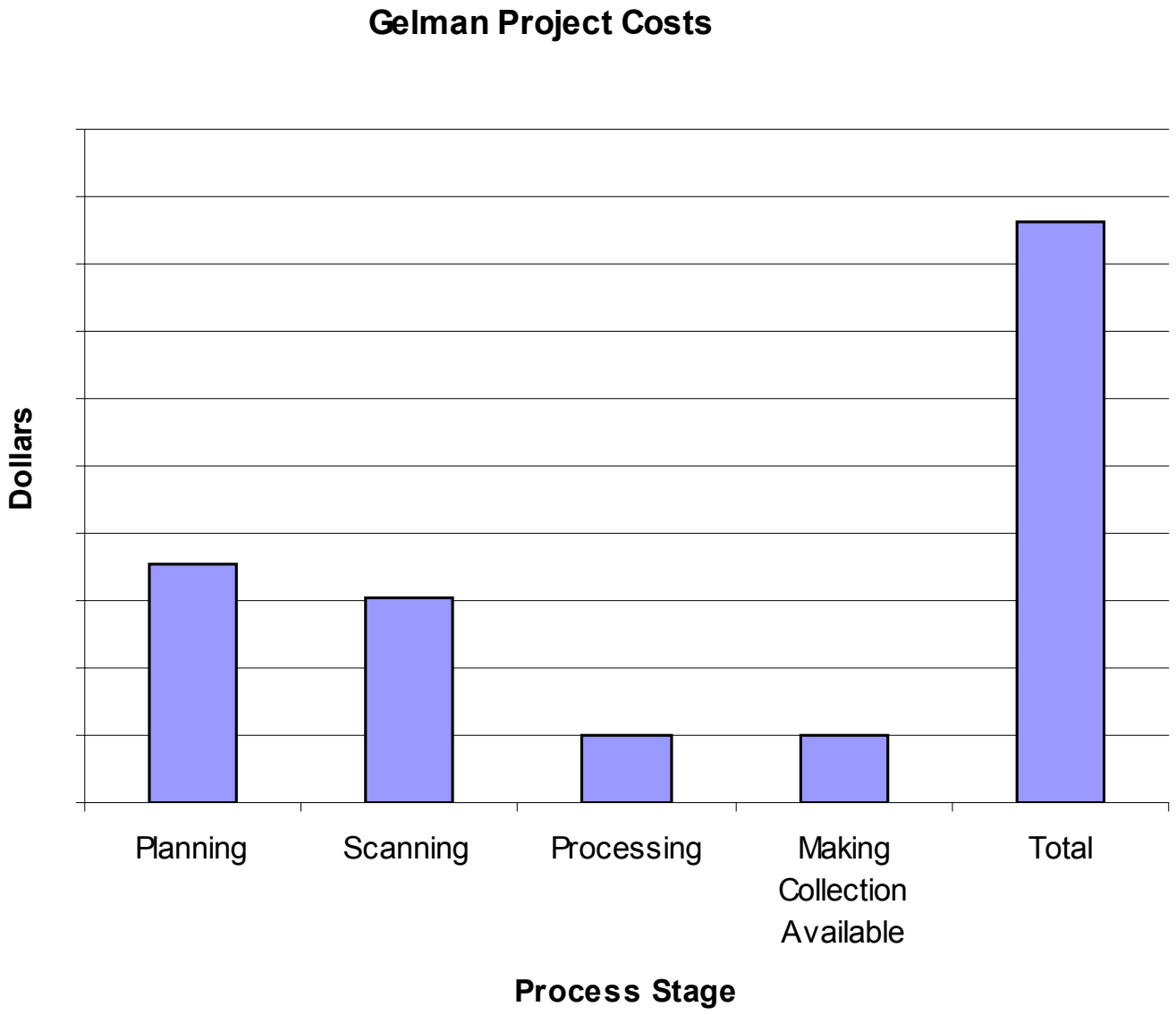


The scanning and processing hours are growing at a more rapid pace now that the operational processes have been implemented

The project planning phase resources tend to be professional staff with higher hourly rates

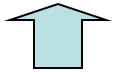


The project planning phase was the most expensive, and the scanning phase was second due to hardware costs



Gelman Project Cost Model

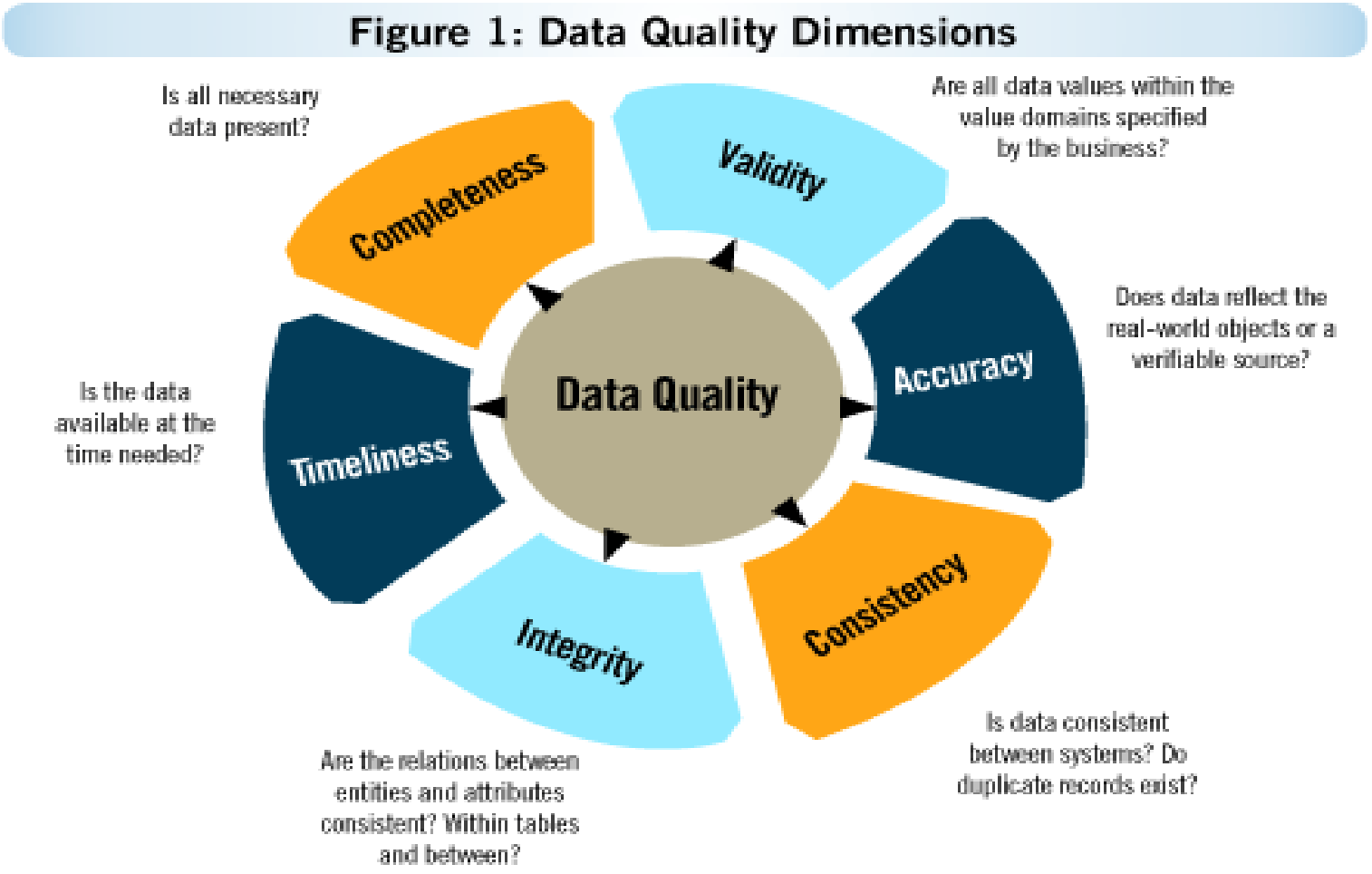
Metric	Project Planning	Scanning	Processing	Making the Collection Available	Total
Hardware Costs (Equipment, Servers, storage, etc)		$ 300,000.00	$ 40,000.00	$ 44,000.00	$ 384,000.00
Software Costs	$ -		$ 8,000.00	$ 20,000.00	$ 28,000.00
Number of Resources	23	3	7	8	41.0
Hourly Rate for Resources	$ 35.00	$ 17.00	$ 16.00	$ 18.00	$ 21.50
Total Hours per Month	1,221	480.00	416.00	483	2,600.00
Number of Months in Stage	28	17	17	17	79
Total Number of Hours	8,745	2,400	2,984	5,651	19,780.00
Resource Cost	$ 306,075.00	$ 40,800.00	$ 47,744.00	$ 101,718.00	$ 496,337.00
Number of Pages per Hour	NA	500.00	402.14	NA	NA
Cost for Project Planning	$ 306,075.00	$ 340,800.00	$ 95,744.00	$ 165,718.00	$ 908,337.00
Cost per Page	$ 0.60	$ 0.67	$ 0.19	$ 0.33	$ 1.79
Total Number of Pages	1,200,000				



The production throughput can be adjusted

The Cost per page updates based on the drivers

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?

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

Validity

Are the data values within specifications?

Recommendations:

- For each piece of capture data create a list of expect 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

Accuracy

Does the data reflect reality ?

Recommendations:

- 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)

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

COST PER PAGE

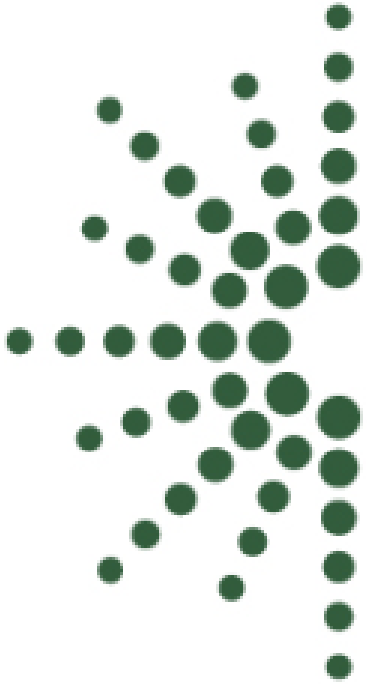
$ 1.70 per page

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



INSTITUTE of Museum and Library SERVICES