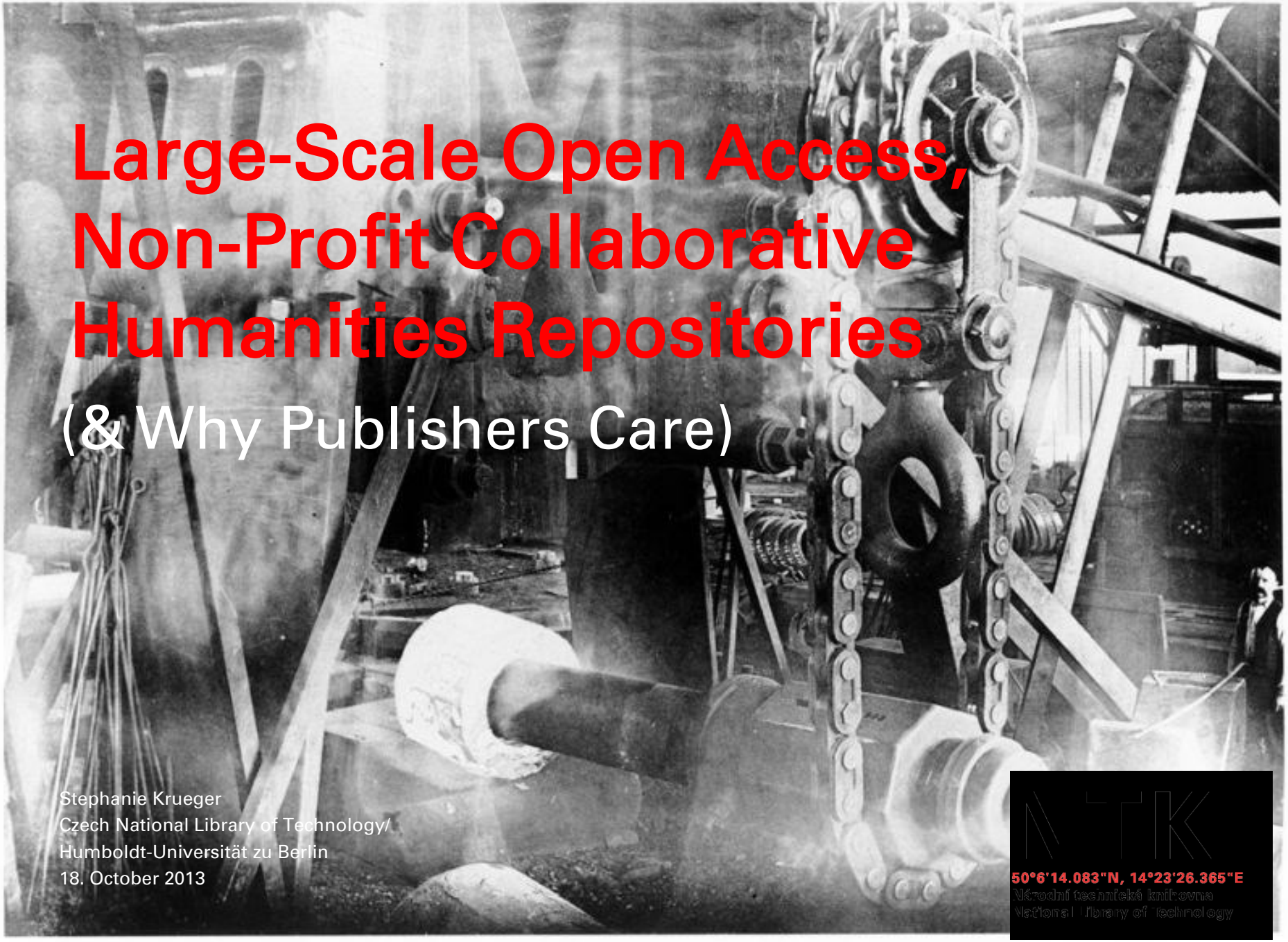


Large-Scale Open Access, Non-Profit Collaborative Humanities Repositories

(& Why Publishers Care)

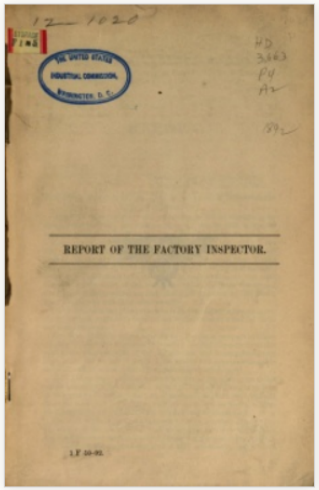
Stephanie Krueger
Czech National Library of Technology/
Humboldt-Universität zu Berlin
18. October 2013

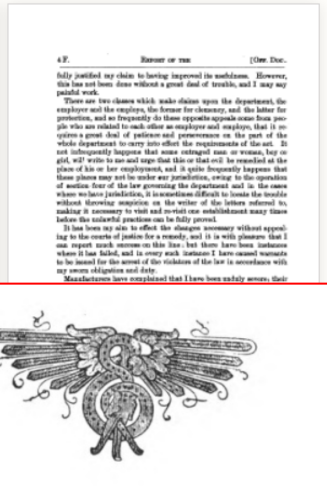
Overview

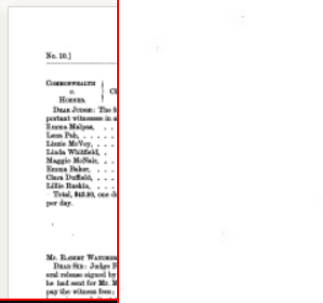
- Landscape of large (humanities) repositories in US with open access components
- Why now?
  - Selected drivers/environmental factors
- Concepts, infrastructure and funding models
  - Emphasis on institutional contributions, not national or individual user subsidization
  - Wealthy institutions take (some) responsibility for those less fortunate

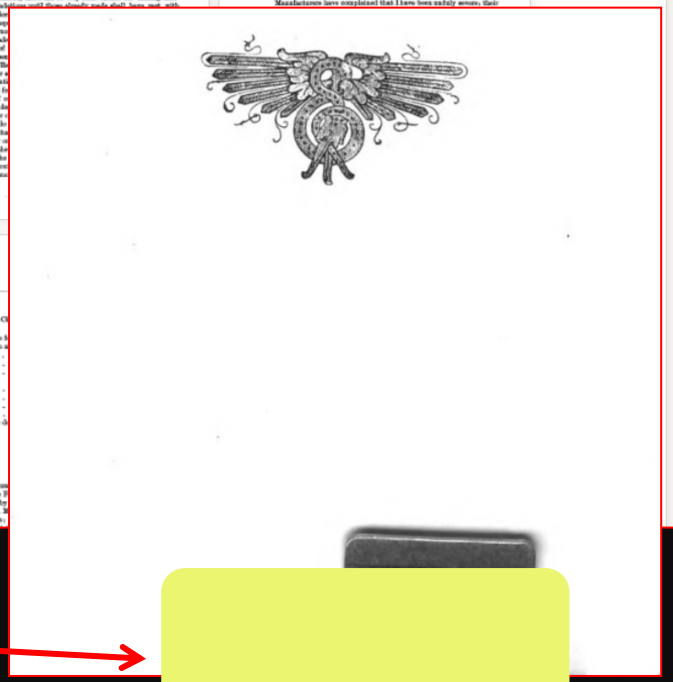
Landscape: Supplementing gated access with open access

- HathiTrust Digital Library
  - Public domain v. gated
- Former Mellon projects
  - JSTOR and ARTstor
  - Operating budgets no longer supported by Mellon but by member fees
  - Open v. gated examples
    - Community and publisher-driven
    - Digital Public Library of America/OpenARTstor
    - Shared Shelf Commons
    - Early Journal Content

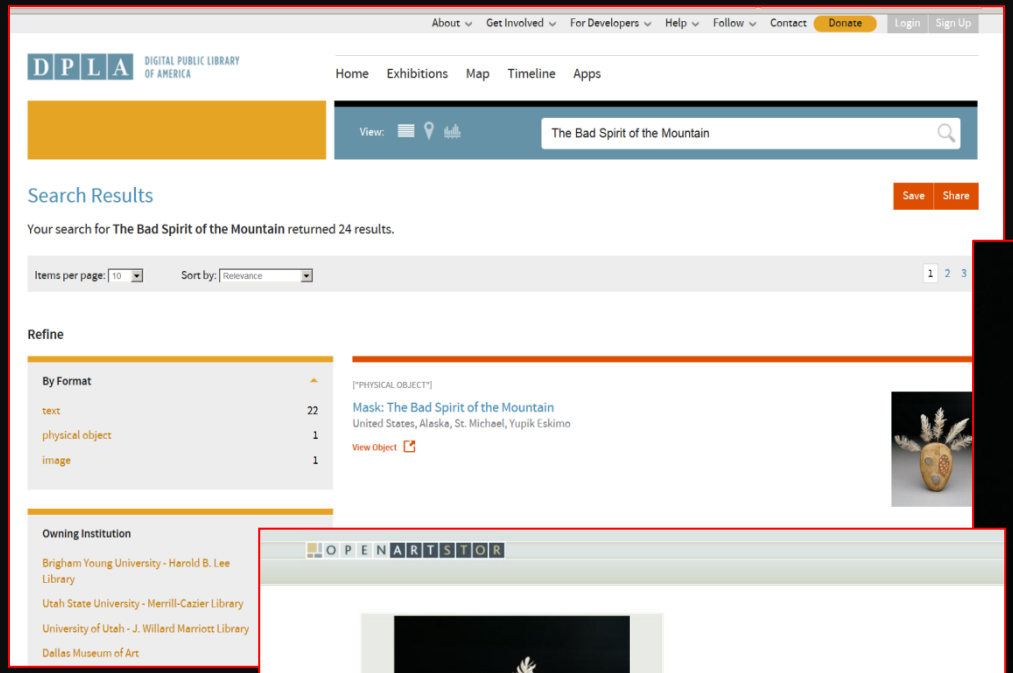


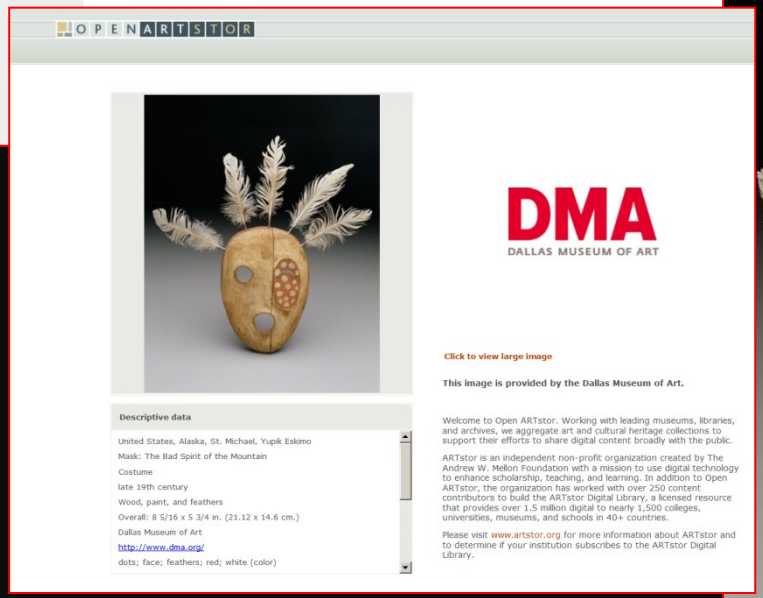




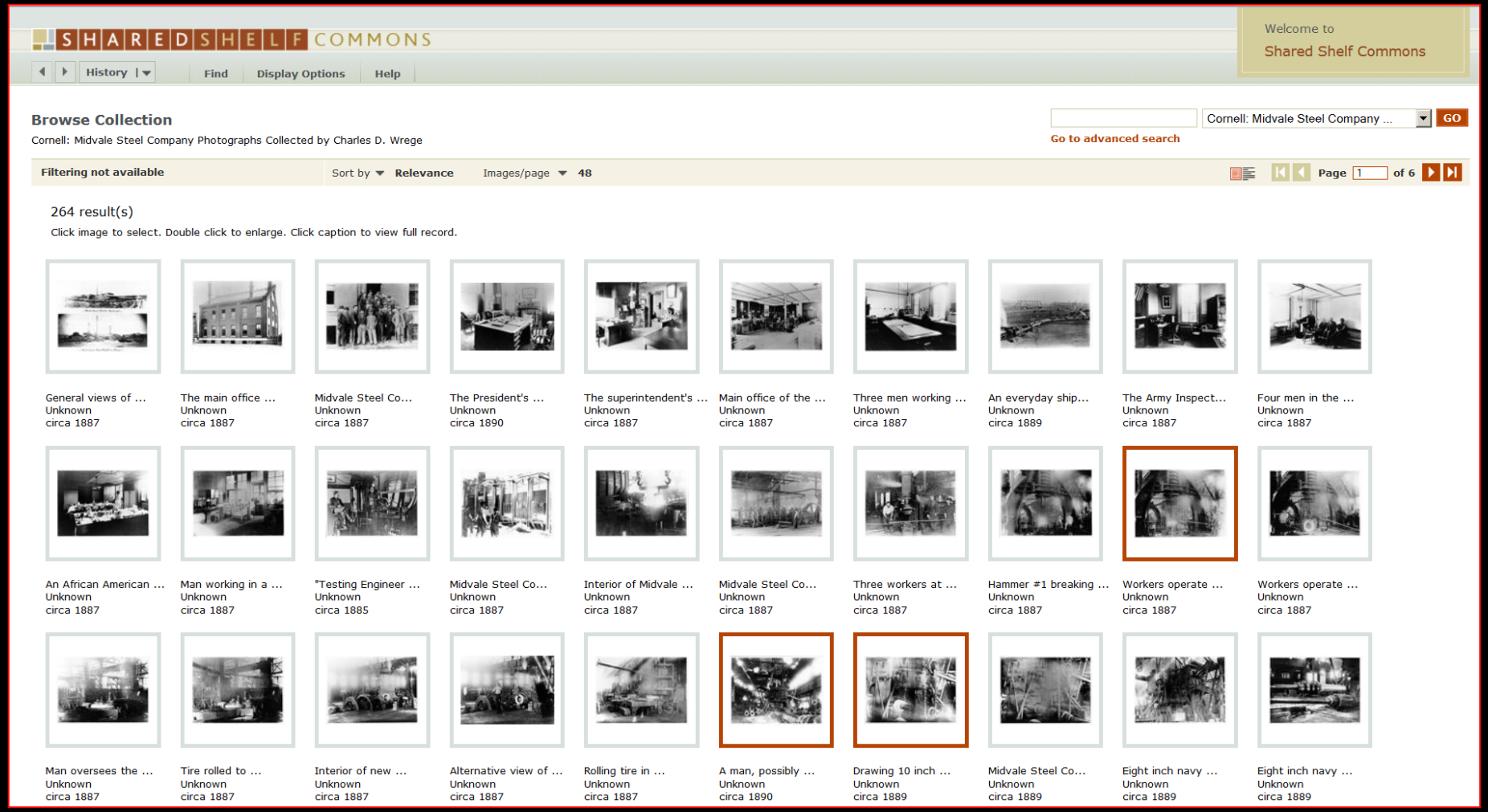


Digitized by Google; original from University of Michigan Collection; public domain



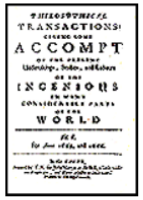


Digitization by the Dallas Museum of Art; metadata/platform by OpenARTstor; searchable via Digital Public Library of America platform



Digitization and metadata by Shared Shelf partner institution librarians (here: Cornell University); platform: Shared Shelf Commons

Philosophical Transactions (1665-1678) Publication Info



Vol. 12, 1677 - 1678
Published by: The Royal Society
Issue Stable URL: http://www.jstor.org/stable/i206877

TABLE OF CONTENTS

Digitization and metadata by JSTOR from microfilm; Publisher: The Royal Society of London; open access via JSTOR Early Journal content (1665-1922)

Why now? (or: necessity is the mother of invention)

- Driver #1: Google
  - Non-profit, community-based alternatives to access (only) via Google
  - "Our content on our platform"
- Driver #2: First wave of digitization of core objects in humanities "complete"; focus now: specialized research needs and interests
  - But…how does the user find information in over-saturated environment?
  - Hope: collaborative non-profit, open access platforms enhance findability (?)

Why now? (or: necessity is the mother of invention)

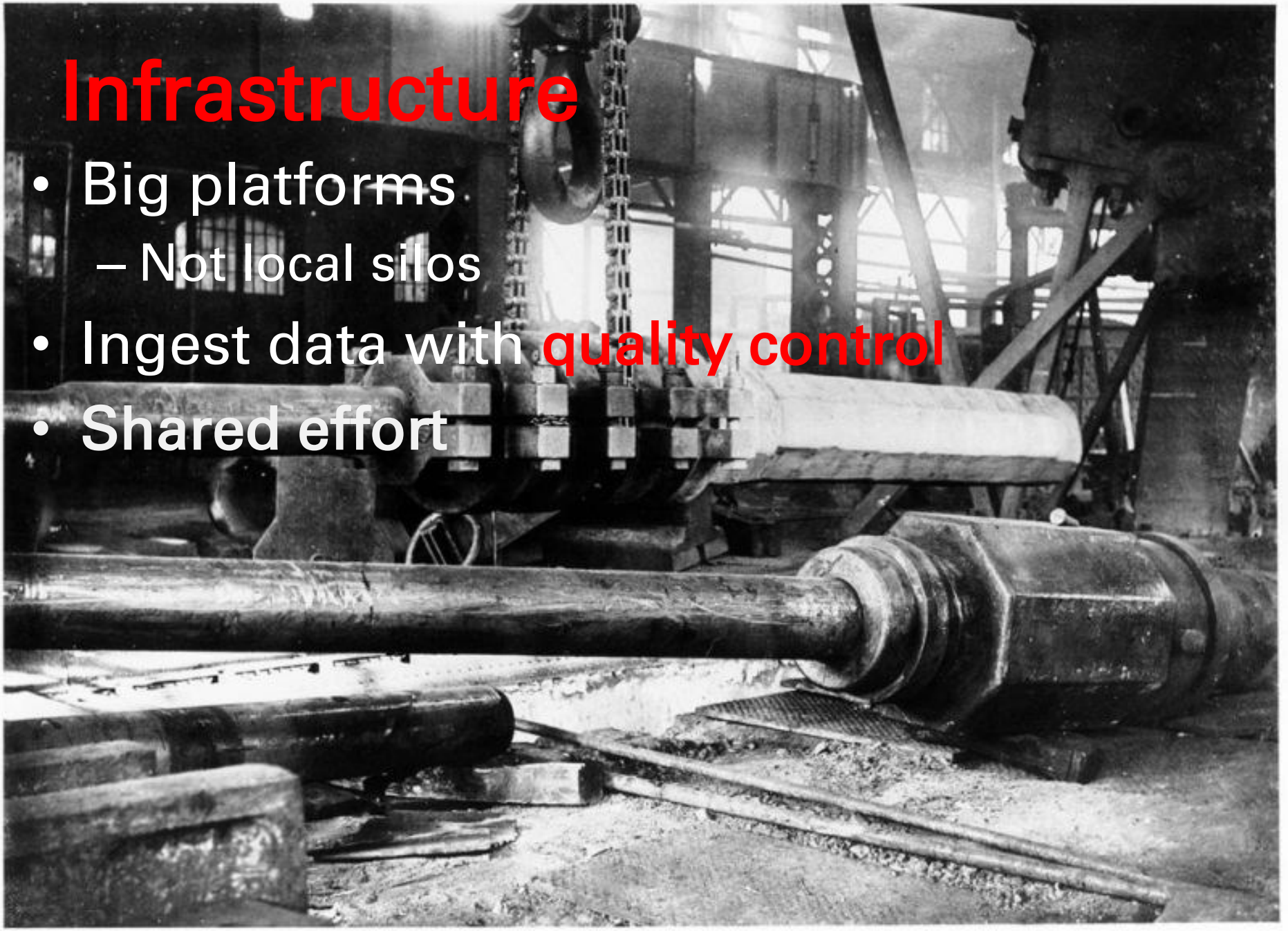
- Driver #3: Libraries/humanities in the US hit hard by 2008 economic crisis
  - Even many wealthy institutions funded library through endowments (investments) instead of normal operating budgets
  - Immediate impact on staffing, particularly arts/humanities librarians and IT
  - "Do more with less": collaborative platforms and simplified metadata/search structures theoretically reduce individual institutional IT and metadata creation investments

Why now? (or: necessity is the mother of invention)

- Driver #4: Publishers and vendors recognize they can't keep up the gate for public domain content
  - At least in the US, because of Fair Use
  - Google will provide public content domain anyway (at least free of charge for now; future unclear)
  - For publishers: open content leads to gated content (?)
  - **Last but not least:** desire by library community to provide open access **alternatives**

Concepts, infrastructure and funding models

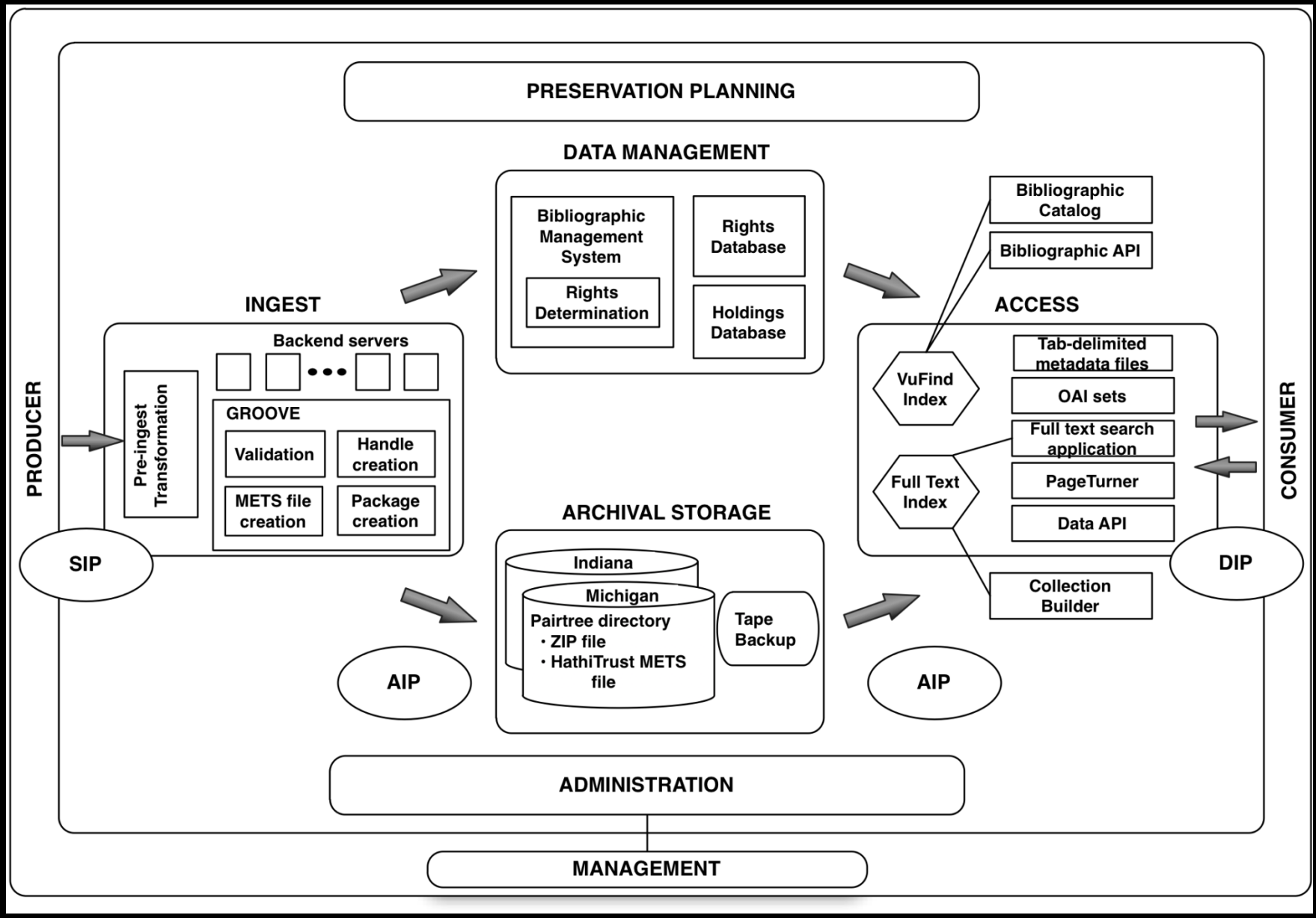
- Idea for these projects: independent non-profit organizations
  - May be hosted by institutions, but still have flexibility and independence in order to avoid intra-institutional politics; attempt to be as fair as possible to all
  - Therefore, transparence, governance and oversight key to trust within participating community
    - Trust is hard to win and easy to lose



Infrastructure

- Big platforms
  - Not local silos
- Ingest data with **quality control**
- **Shared effort**

HathiTrust


PRESERVATION PLANNING
DATA MANAGEMENT
Bibliographic Management System
Rights Determination
Rights Database
Holdings Database
Bibliographic Catalog
Bibliographic API
INGEST
Backend servers
Pre-ingest Transformation
GROOVE
Validation
Handle creation
METS file creation
Package creation
ACCESS
VuFind Index
Full Text Index
Tab-delimited metadata files
OAI sets
Full text search application
PageTurner
Data API
Collection Builder
PRODUCER
CONSUMER
SIP
DIP
ARCHIVAL STORAGE
Indiana
Michigan
Pairtree directory
• ZIP file
• HathiTrust METS file
Tape Backup
AIP
AIP
ADMINISTRATION
MANAGEMENT

Shared Shelf

- Developed with community-based steering group – 9 members
- Leverage ARTstor platform instead of multiple local platforms; example of Harvard:

*For decision-makers at Harvard, where an infrastructure for image management and use for 21 different departments was implemented in the 1990s, the partnership was attractive because of the joint investment that will update the cataloging systems and leverage protocols enabling interoperation with authority files, repository and discovery environments. "Images are becoming ever more important in both teaching and research. As a community we have lacked good tools for their management and discovery," **Dale Flecker, [former] Associate Director for Planning and Systems, Harvard University Library Office for Information Systems** noted. "Images present significant challenges. Having worked with ARTstor for years in this domain, we believe that combining forces and know-how offers the most promising approach to these challenges.“**

**http://www.artstor.org/shared-shelf/s-html/why-choose.shtml*

A word about JSTOR…

- University-developed (Michigan and Princeton)
  - Big, expensive servers on high-speed internet nodes
  - Geographically-distributed, load-balanced
- Outsourced to private platform mid-2000s
- Considering returning to own platform
  - More flexibility
  - Not trivial

Funding models

- Hathi, JSTOR, and ARTstor: Institutional annual subscriptions + new models
- DPLA: grants and donations
- Models developed with economists and the community
  - Increasingly complex

Funding example: Hathi + Shared Shelf

- Initial model for Hathi and Shared Shelf model: based on cost of basic infrastructure (cost of storage model)
  - Attempts to calculate cost per GB for storage, basic management, basic access
    - Hathi estimate (2010 baseline): $3.86
    - Shared Shelf initial cost was slightly higher
      - Different organizations may make different estimates based on internal politics and other factors; for example, regional costs of a programmer per hour may vary greatly

Funding example: Hathi + Shared Shelf

- Recent revisions to initial approach **Hathi: "benefits-based model"**
  - **Different costs for public domain v. in-copyright items**
    - **(Public domain volumes in all HathiTrust*average annual costs to support a volume*X [related to annual cost to generate surplus)/total number of partner libraries**
    - **Average annual cost to support a volume*X/number of partner libraries that hold a given print in-copyright volume**
- **Shared Shelf: annual platform fee based on amount of data**
  - **Lite: $26.7 (!) per GB (500GB or less)**
  - **Intensive: $8.6 per GB (over 4TB on platform)**

Funding example: JSTOR open access

- Digitization fees and platform operations/services already paid for with annual fees per collection and one-time collection contributions
  - Works as long as institutions pay annual fees
- Publisher royalties still paid for in-copyright materials
  - Annual review of revenue sharing paybacks
  - Usage part of the revenue sharing equation, but not 100%, so small humanities publishers still get a benefit from the in-copyright objects

Funding: Challenges

- Complexity of platform models
  - But not impossible
- Trust
  - Tricky business of calculating cost for storage/service seen in example of Hathi "X" variable
    - *"Initially, HathiTrust proposes using a value of two (2) for X, i.e., doubling the cost of maintenance in order to build a fund for the services. Thus, 50% of the funds collected will go to development and the other 50% will cover costs of storing content."**
- Sustainability over the long-term
- Ability to change course/ability to adapt to changes models

**http://www.hathitrust.org/documents/hathitrust-cost-rationale-2013.pdf*

What's in it for publishers/content providers?

- Goodwill within library and scholarly communities regarding open access progress
- Ability to stumble upon in-copyright/gated materials
- Small profit margins anyway; why not share more broadly with more potential readers
- Sometimes not a publisher but provider of objects
  - Metropolitan Museum of Art provides some open images in order to reduce staffing costs for high-quality image requests
  - Dallas Museum of Art wants to promote its collections more broadly
  - Shared Shelf libraries contributing to the commons want to supersede institutional content silos)
  - **etc.**

Summing it up

- Towards non-institutional based, collaborative collection and platform development
  - Because little or no profit margin in many humanities disciplines, non-profit institutions have room to lead in this space
- Someone has to pay in the end; models can vary greatly depending on societal, organizational, and larger macroeconomic factors
- Problem of institutional silos still exists
  - Not just library silos, but non-profit platform silos
  - Does the user care?

Questions + Discussion

Stephanie Krueger
Head, Communications and Fundraising
Czech National Library of Technology
Doctoral student, Humboldt-Universität zu Berlin
stephanie.krueger@techlib.cz