Creating a Digital Microfilm Library

Dan R. Olsen Jr

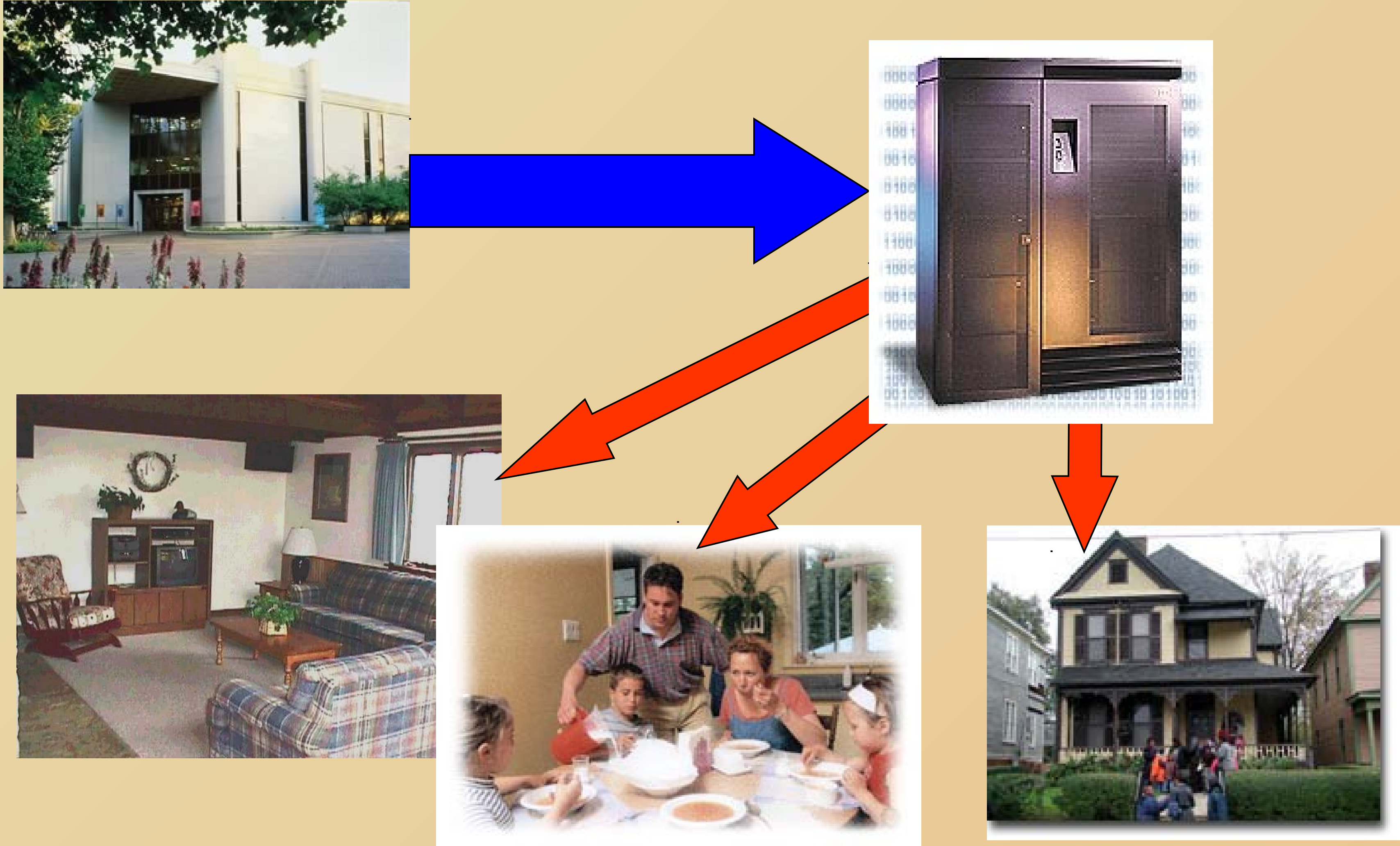
Computer Science Dept

Brigham Young University

The meaning of family history

Getting in touch with your ancestors

Family history out of the library and into the home



How big is the problem?

- 2.5 million films and growing
- 1000 images per film = 2.5 billion images
- 600 KB per image =

1,500,000 Gigabytes

25,000 laptop hard disks :-)

What will it cost to store it?

- 1,500,000 GB
- $30 per GB - real servers not PCs
- Total library cost
  - Today - $45,000,000
  - 5 years - $4,500,000
  - 10 years - $450,000

Producing the Image Library

- Scanning rate - 100 frames per minute (optimistic)
- Images to scan - 2.5 Billion
- Scanner time per year - 2000 hours
- To complete the library
  - 208 scanner years

Cost of production

- 20 scanners = $1,000,000
  - 10 years to finish
  - replacement costs $1,000,000
- Worker costs = 10x$100,000 = $1,000,000
- 10 year plan
  - GB per year - 150,000
  - Cost - $3,000,000

Creating the Digital Microfilm Library

- 10 years

- $8,000,000

Delivering images to the home

- Image size = 500K
- Dialup data rate = 5K Bytes / sec
  - (on a good day)
- Time per image = 1.6 minutes

- You cannot scan digital images the way you scan through microfilm

- The library must be indexed at the image level

Extracting data from images

- Extraction 12/hour
  - Assumes one record per image
- hours to extract the entire library
  - 208 million hours
  - cost at $5/hour = $1 billion
- 20,000 extractors - 100 hours per year
  - 104 years to complete
- 2 million extractors to complete in 10 years

To build the library in 10 years

- $5,000,000 to store
- $3,000,000 to scan
- must be indexed
- $1,000,000,000 to extract using current approach

- Need a new indexing plan

Extract for index

- Ordered collections
    - by name
    - by date
  - Parish records
  - Death records
  - Main archive group sheets
- Unordered collections
  - Wills
  - Deeds
  - Other records where image order is not helpful

Ordered collections

- Extract top date or name from each image
  - 100/hour
  - 12 years to complete - 20,000 volunteers
- Sample extract every 10 images then interpolate
  - 1000/hour
  - 1.2 years to complete - 20,000 volunteers

Unordered collections

- Extract only essential name, date and place info

- Let the image carry most of the data
  - Eliminate interpretation errors in extraction
- Extractors map extracted data to image fragments
- Auto extraction methods - OCR and Handwriting
  - Use extraction as a training set for new algorithms

What library should we build?

Lessons from the past

- Microfilm library
  - Make the raw data available in a uniform way
- WWW/GEDCOM
  - Make the library open
  - Base collection
    - Tools on top but not in place
    - Support both people and software
- Guaranteed archiving
  - **Digital Stone**
- Guaranteed naming

The vision

- Out of the library into the home
- Scan it all in 10 years
  - $8,000,000
- Index not extract
  - High level index (beat microfilm)
  - Deeper indices on important collections
- Open library architecture
  - Raw data and raw indexes publicly available
- Library of last resort - Digital Stone
  - Guaranteed archive, guaranteed naming

Family history out of the library and into the home