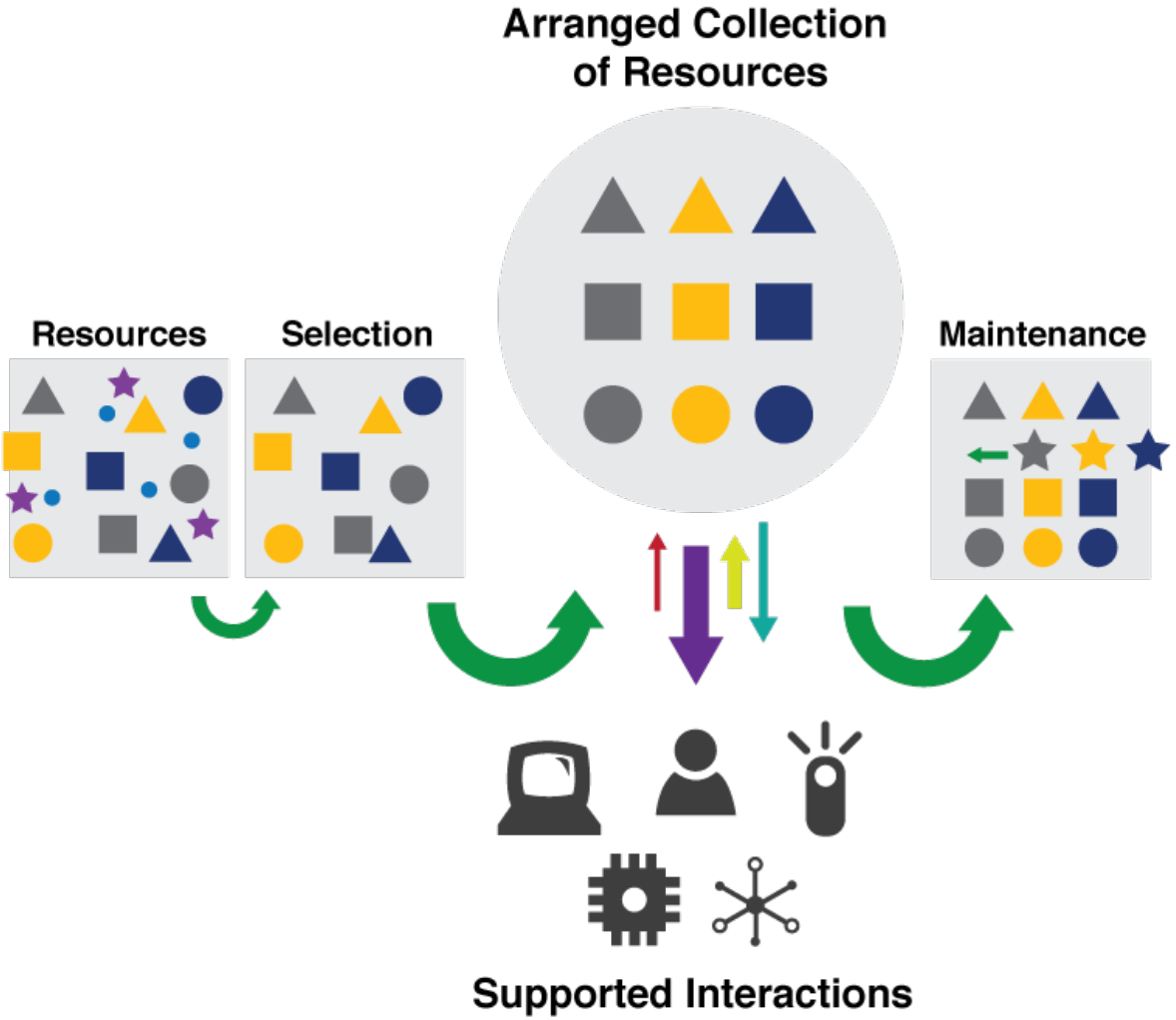
UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

5 September 2013
Lecture 3.1 – Activities in Organizing Systems

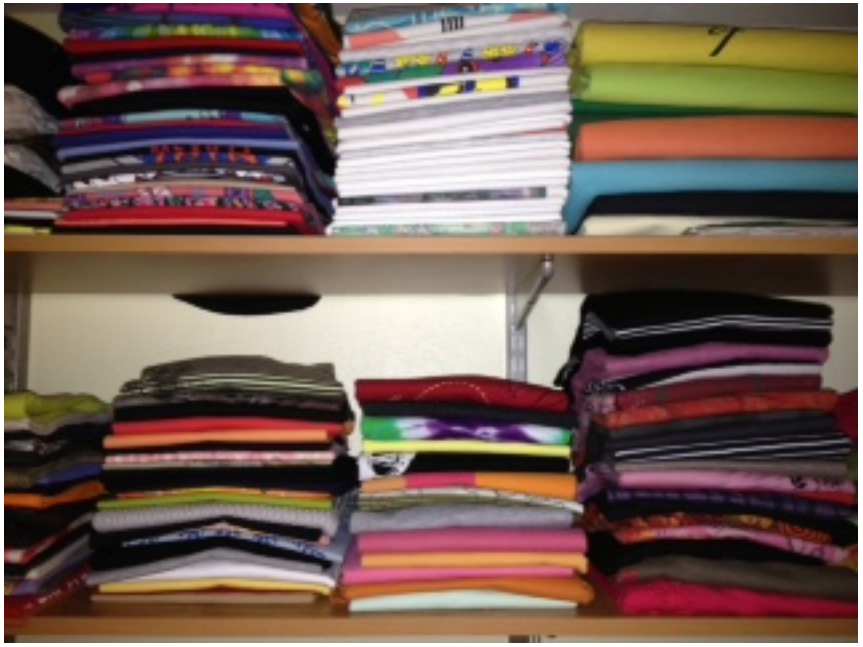
The Activities in Organizing Systems



The Activities in Organizing Systems

- We can identify four activities in the lifecycle of every organizing system:
  - Selecting resources
  - Organizing resources
  - Supporting resource-based interactions and services
  - Maintaining resources

Closet Organizing Systems



For Your Closet Organizing System…

- **Selecting**: *Should I hang up my sweaters in the closet or put them in a drawer?*
- **Organizing**: *Should I sort my shirts by color, sleeve type, or season?*
- **Supporting Interactions**: *Do I need separate places for laundry or dry cleaning?*
- **Maintaining:** *Should I toss out my clothes based only on stains and tears, based on how long I've owned them, or based on whether I'm tired of them?*

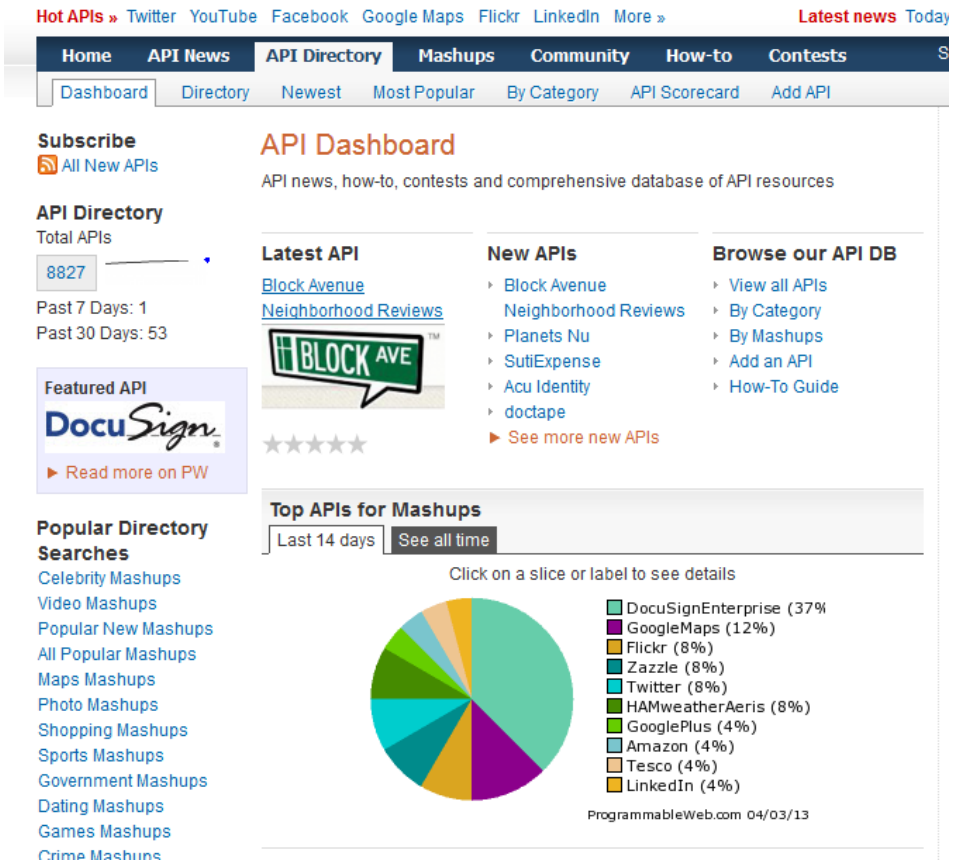
Stop and Reflect…

• Does the definition of "Organizing System" apply to all of the types of collections, documents, and information services we've seen?

• What are the consequences of using abstract terms like "resource," "interaction," and "maintenance" instead of more specific and domain-specific terms?

- Collection development, Appraisal vs. Selecting
- Acquisition, Accession, Ingesting vs. Adding to Collection
- Cataloging, Indexing vs. Organizing
- Curation, Governance vs. Maintenance …

And Your Closet Has "Interfaces" But They're Not Listed Here



ProgrammableWeb.com

Stop and Reflect Again…

- Why are there FOUR activities and not 3 or 5 or some other number?
- Why are the activities given generic names?

What about “Creating” Resources?

- Does "intentional arrangement" imply the existence of resources?
- Can "intentional arrangement" be a pre-requisite for resource creation?

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

5 September 2013
Lecture 3.2 – Selection in Organizing Systems

Selecting Resources

- SELECTION is the process by which resources are identified, evaluated, and added to a collection
- Selection is by definition an intentional process
- Selection methods and criteria vary across domains (as do the words that describe them); these can be subjective or objective

Selecting {and,or,vs.} Organizing

- SELECT everything and keep it all --> often defaults to weak organizing principles
- Selection PRECEDES Organizing --> Unique or rare resources are naturally evaluated one-at-a-time
- Selection and Organizing are more CONCURRENT --> homogeneous or predictable resources
- Selection FOLLOWS Organizing --> Resources created according to schema

Selection Principles

- Utility or relevance
- Intrinsic value
- Scarcity or uniqueness
- To support social or symbolic goals
- … many others

"Path Dependence"

- Does the selection principle apply to each candidate resource in isolation?
- ... or does the selection decision consider the items already in the collection?
- ... or does the selection decision consider possible future selections?

Selection in Different Domains

- Libraries select, collect, organize, conserve, preserve and provide access to information on behalf of a community of users
- Museum collections, especially of special or rare items, bring status or prestige whether or not the items have any contemporary use
- In "work systems" or "business models" selection is very task dependent and highly data-intensive

"My Life Bits"

- In 1998 Gordon Bell (MSFT research) began digitally recording as much of his life as possible
  - Articles, books, cards, CDs, letters, memos, papers, photos, pictures, presentations, home movies, videotaped lectures, and voice recordings...
  - Printed resources were scanned
  - Everything he did on his computer was captured
  - He wore a "sensecam" that took photos and logged biometric and ambient data
- Little or no concurrent organization but "The more that is captured, the more correlation is possible to help find things"
- 2009 Book: Your Life, Uploaded http://totalrecallbook.com/about-the-book/

The SenseCam

http://www.youtube.com/watch?v=3NjZxCll9Z8

http://www.youtube.com/watch?feature=player_embedded&v=ZuzByAsbDIE

Digitization and Selection

- Digitization separates selection from storage concerns because access need not imply actual "holding" of the resources
- Digital capture and copying of photos, videos, music and documents reduces their distinctiveness
- The "digital torrent" of scientific and sensor data makes collection easy and selection more important

Selection on the Web

- The web has no centralized and authoritative directory or catalog
- The web can be treated as a combination of a very large number of domain- or topic-specific resource collections that are created by individuals or identified by computational analysis

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

5 September 2013
Lecture 3.3 – Organizing in Organizing Systems

Organizing Resources

- Almost any property of resources can be used to organize them, but the most appropriate or effective properties differ across resource types and tasks
- Are you organizing the resources you have, or do you need to create an organizing system that can apply to resources that you have not yet collected or that might not even exist yet?
- Selection is easier if there are clear organizing principles because they establish specifications or locations for the resources to be collected

Matching Organizing Principles to Resource Types and Interactions

- Almost any property of resources can be used to organize them, but the most appropriate or effective properties differ across resource types and tasks

Resource Organization in Libraries

- Bibliographic classification
- We'll discuss this a lot in chapter 4 on Resource Description and chapter 7 on classification

Resource Organization in Corporate Knowledge Management

- Corporate or specialized libraries are narrowly focused organizing systems that support the knowledge management and reuse needs of a particular group of experts
- They need a narrower and deeper system that will often have proprietary aspects
- Access control mechanisms are often implicit in the organizing system

Resource Organization in Museums

- Museum collections often have a very distinctive or idiosyncratic character because many of them began as or grew through the acquisition of private collections
- Because non-text objects are more semantically opaque than text ones, there is greater need for description resources or metadata

Resource Organization in Archives

- Archives often collect sets of documents ("fonds") that have themselves been previously organized as a result of their generation and use
- This "original order" embodies the implicit or explicit organizing system of the person or entity that created the documents
- As a result, archival documents are often kept in this original order
- The "provenance" is an important set of information about the documents and influences how much they can be trusted

“Respect pour les fonds” in Museums

- The “fonds” principle of archives sometimes has an analogy in museums, when the museum is housed in a building donated by the original collector and the resources are kept in their original locations

- This is the Frick Museum in Pittsburgh; see also TDO case study about Barnes Collection (10.5.2)

Digitization and Organization

- For digital resources the distinction between primary resources and description resources is less clear and less important because they are often organized and stored together
- Being apply to apply "computational organization" in search engines or other NLP applications reduces the need for organization work by people

Webification and Organization

- Can we define a logical or technical boundary around a collection of web resources?
- What are the implications of doing so?

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

5 September 2013
Lecture 3.4 – Interaction Design in Organizing Systems

Interactions in Organizing Systems

- INTERACTIONS include any activity, function, or service supported by or enabled with respect to the resources in a collection or with respect the collection as a whole
- ACCESS is the fundamental interaction in all organizing systems – you wouldn't organize anything you didn't expect to access again – but most organizing systems have additional interactions to make access more efficient and to create additional value with the resources

The Four Big Ones for Libraries…

- FINDING resources
- IDENTIFYING resources to confirm that that a found resource is the intended one by distinguishing it from similar ones
- SELECTING a resource from a collection when more than one might seem to fit the search criteria
- OBTAINING the resource if it isn't immediately available

Designing Resource-Based Interactions

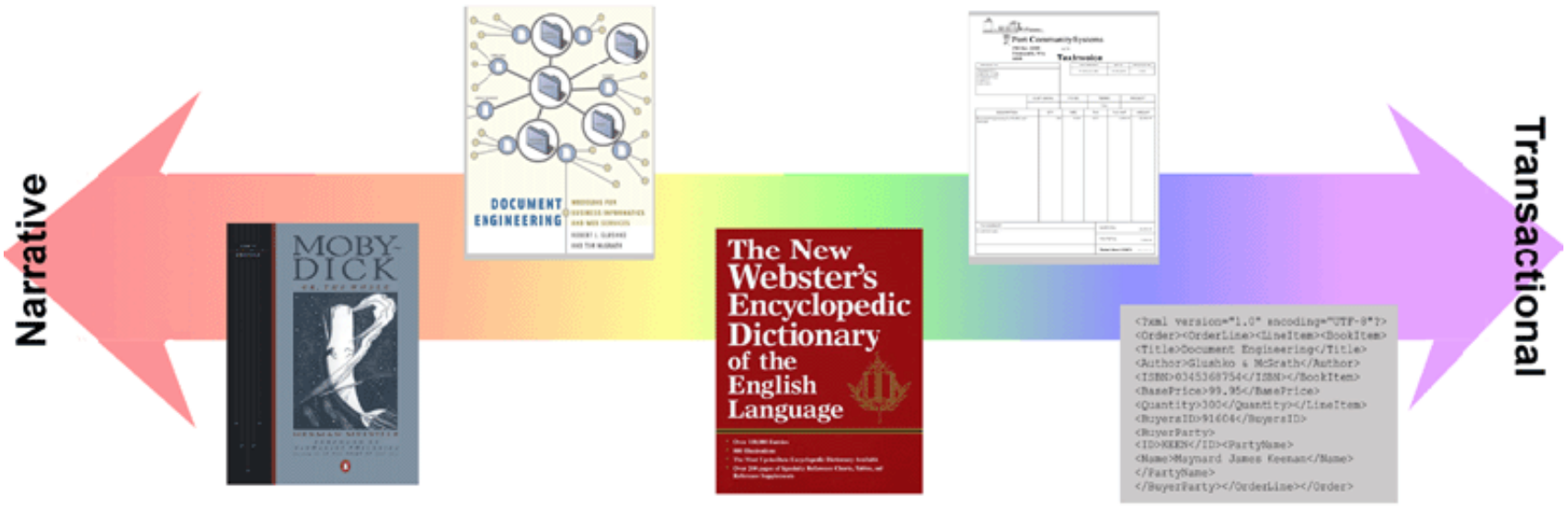
- Important distinction between interactions that are directly afforded by the resources from those that must be designed
- The nature of interactions depends on how we identify and define "resource"
- Selecting and organizing have no purpose if you don't allow interactions to the resource (to the appropriate "agents")
- Who we consider "appropriate" differs a great deal; contrast libraries, museums, archives, scientific and business data

Mediated Interactions (in Paris)

Digitization and Interactions

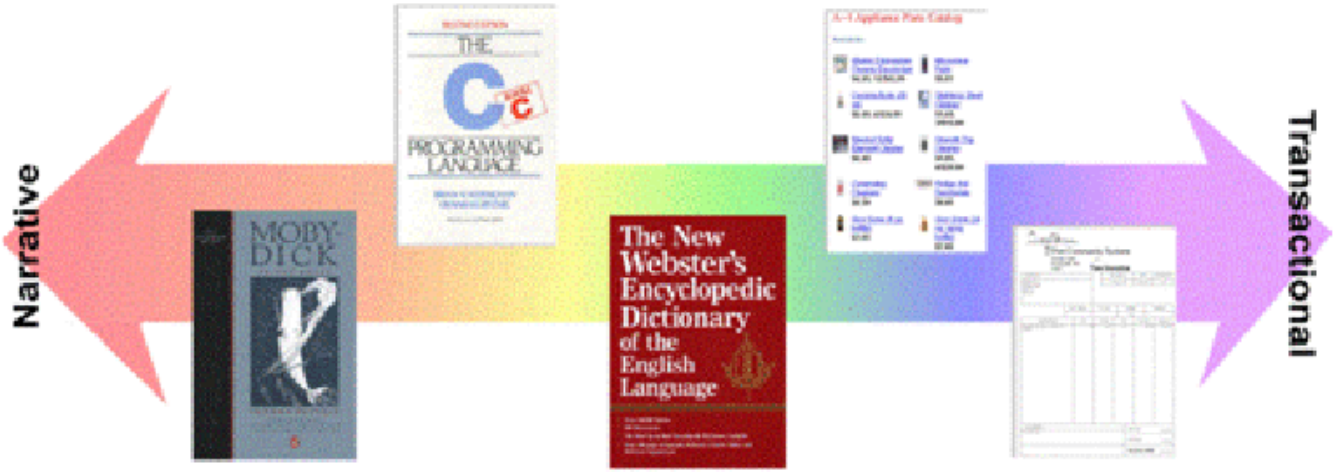
- The variety and functions of interactions with digital resources are determined by the amount of structure and semantics in their encoding
- Interactions can also be supported with digital resources that describe or are otherwise associated with the primary or original digital resource
- Access controls for the interactions involving digital resources can be very precisely defined and enforced

The Document Type Spectrum



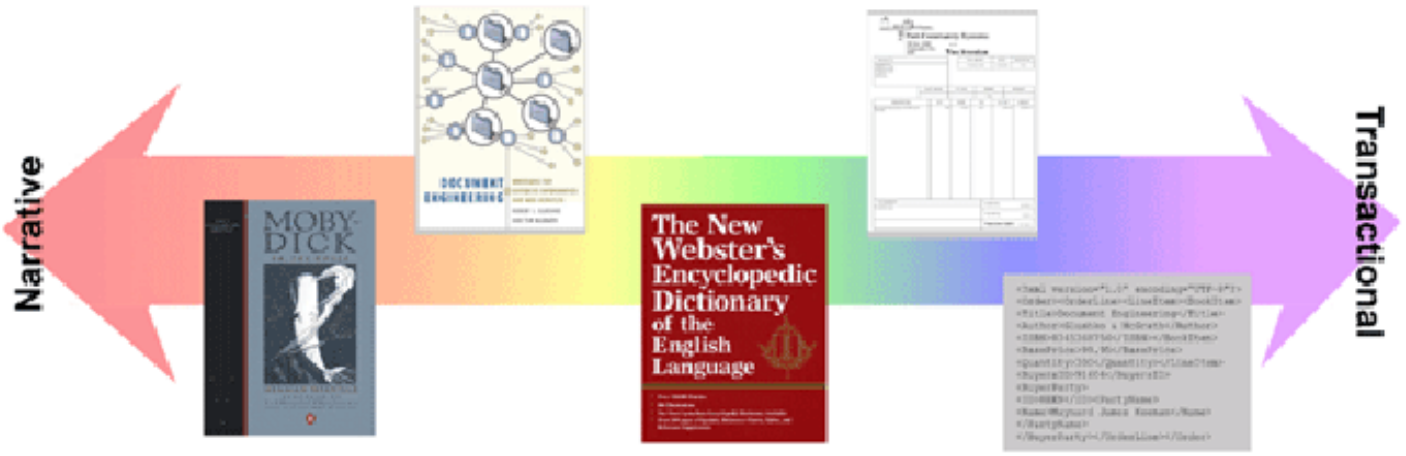
From Glushko & McGrath, DOCUMENT ENGINEERING, MIT Press 2005

Organizing System Types & the DTS



Knowledge Management | **Content Management** | **Data Management**

Information Retrieval Models & the DTS



FULL TEXT RETRIEVAL

STRUCTURE-BASED RETRIEVAL

DATABASE RETRIEVAL

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

5 September 2013
Lecture 3.5 – Maintenance in Organizing Systems

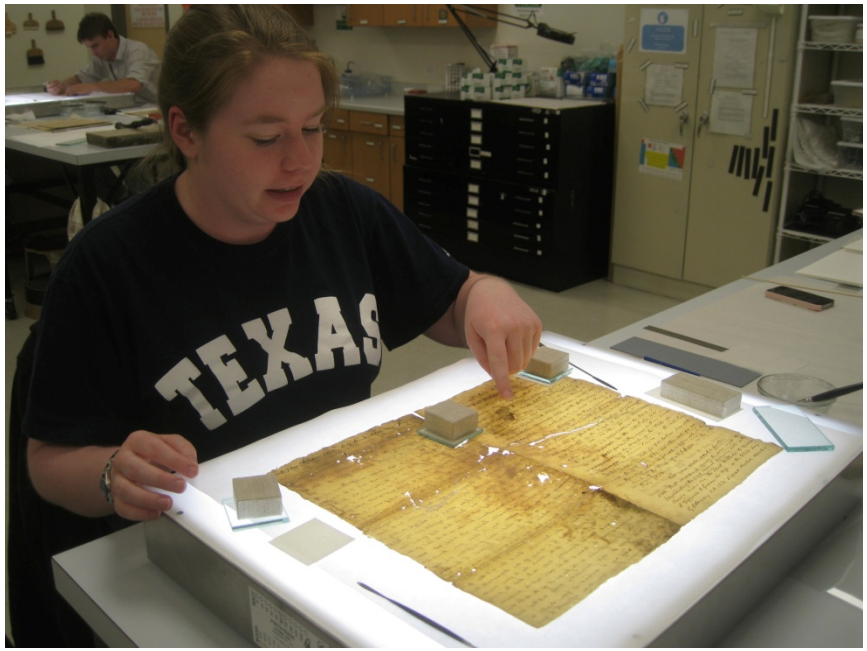
Maintaining Resources

- Maintaining includes any activity whose purpose is to ensure that a resource will be available at some future time for use or reuse
- We can predict future use for some kinds of resources but not for others
- Note that the intended future "user" of a resource might be a computational process or service
- For many kinds of resources, you need to preserve more than the "primary resource" for it to have future use
- Maintenance has many synonyms or near-synonyms across domains: Preservation, curation, governance, storage…

Preservation

- Preservation of physical resources requires that they be kept in conditions that prevent their deterioration
- It might also include actions to improve access or to restore the resource
- Preservation of digital copies might seem easier, but other issues arise:
  - How many copies do we need to ensure preservation?
  - Do we care about which copy is the original?
  - How do we prevent or at least detect tampering?
- Preservation (or "non-preservation") might be required by law or regulation

Restoration



Using a light table to see the tears and losses in a 19th century document

Repairing a similar document with conservation tissue and wheat starch paste

Records Management

- "Records management" and "compliance" are old terms but the concerns about what information a government or business needs to preserve are very important today
  - Retention requirements
  - Non-retention requirements
  - Access control
  - Prohibitions about keeping personally identifiable information
  - Purging requirements and authority

Backing Up Your Computer

"No, I'm not backing up our files—I'm just assuming that the F.B.I. is making copies."

Curation

- CURATION a common term for maintenance activities in "memory institutions“
- It includes work to improve access (through better organization or description) or to restore / transform the resource
- Curation decisions are often on an item-by-item basis, or for a set of closely related ones
- Some people call "Crowdsourcing" and similar activities "community curation" but professional curators deprecate this usage

Cook and the "Contribution Revolution"

- Scott Cook, "The Contribution Revolution: Letting Volunteers Build Your Business", Harvard Business Review, October 2008
- Much design thinking focuses on individual customers or users treated as independent actors or whose interactions with each other are not viewed as important
- More recently, designers of information systems and services have started to notice and exploit the collective actions of people who implicitly or explicitly contribute content or preference information
- Putting a "user contribution system“ into place to encourage and exploit this work is challenging (hard to get incentives right)

Cook's "Contribution Taxonomy"

User Contribution Systems

Active		Passive	
Aggregates content	Aggregates stuff for sale	Aggregates behavioral data	Aggregates resources
Opinion & ratings: Zagat guides	**Goods:** eBay online marketplace	**Buying behaviors:** Amazon's product recommendations	**Computing capacity:** Skype internet-based phone system
Expertise: Wikipedia encyclopedia	**Advertising:** Google's AdWords advertising placement system	**Web-linking behavior:** Google's search engine algorithm	**Computer sensing capabilities:** Honda's InterNavi traffic information service
Software code: Firefox web browser	**Services (and goods):** Craigslist online marketplace	**Company behavior:** Westlaw's PeerMonitor law firm database	
Creative expression: YouTube video-sharing site			
Social connections & personal information: Facebook social networking site			

Governance

- GOVERNANCE and Curation overlap, but the former term is more associated with corporate or enterprise data management and focuses more on policy than on process
- Governance often has an economic basis or metaphor at its foundation
- "IT Governance" has broad scope, applying to the management of the complete technological environment in which information resources are maintained

Assigned Readings for Next Lecture

- Hearst, Marti. Search User Interfaces, 2009. Chapter 3 through 3.4
- Trant, Jennifer. “Studying social tagging and folksonomy: A review and framework.” Journal of Digital Information 10, no. 1 (2009).
- Morville, Peter and Rosenfeld, Louis – “Information Architecture for the World Wide Web, Third Edition”, Sections 5.3 and 5.4
- Berman, Francine, and Vint Cerf. “Who Will Pay for Public Access to Research Data?.” Science 341, no. 6146 (2013): 616-617.
- Keller, Erasing History in the Internet Era, NY Times 28 April 2013