Access

Cal Lee

School of Information and Library Science

University of North Carolina, Chapel Hill

SERI Institute

July 8-12, 2013

Indianapolis, Indiana

SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

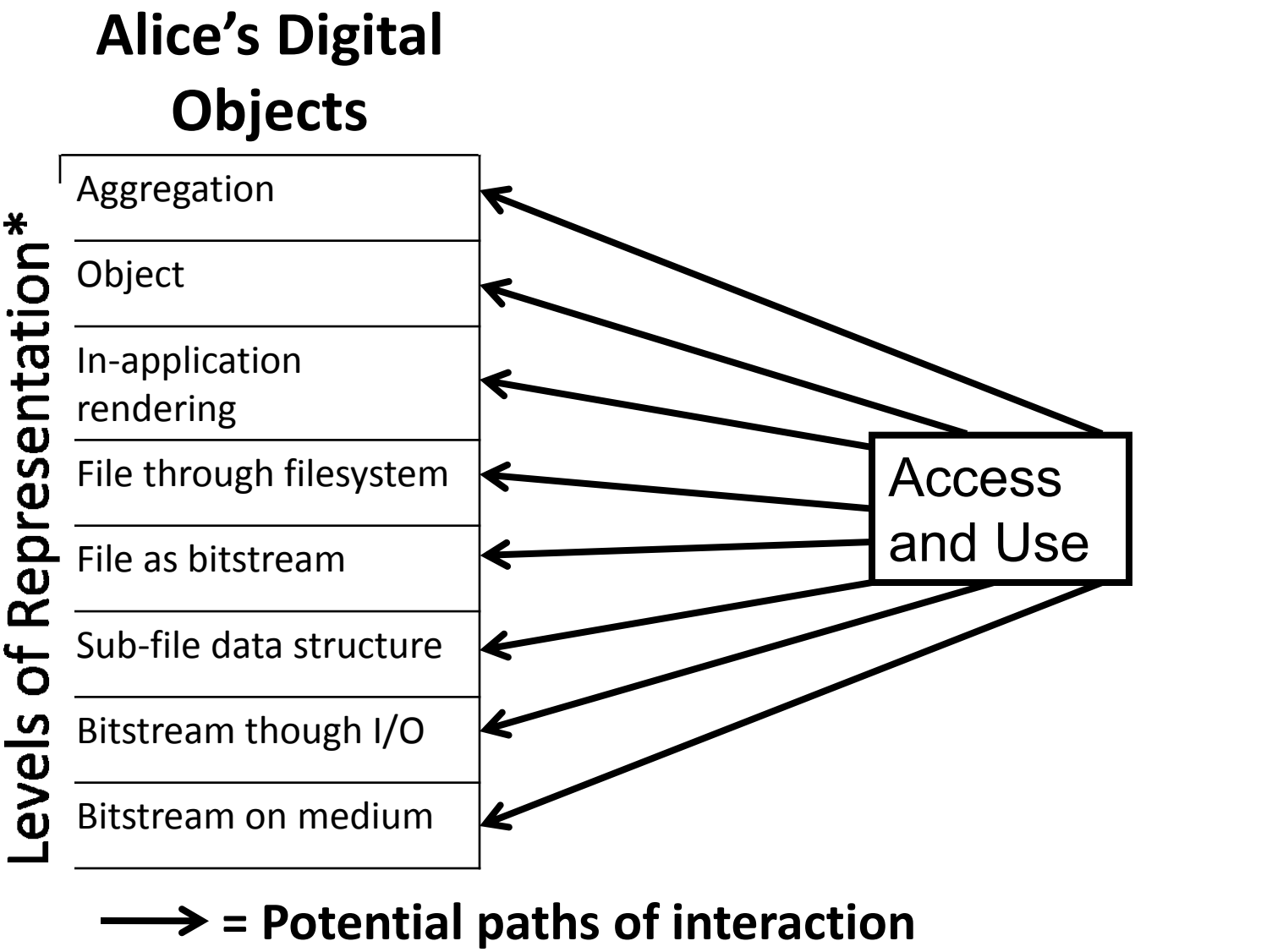
UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Some Differences with Provision of Access to Digital Collections

- Access permissions/control more complicated – overlaps between circulation & publishing
- Requirements (& thus policies, rules) may be more driven by remote, rather than local user needs
- Often treated at collection level (similar to existing archives & special collections practices)

Many More Decisions about Levels of Investment (Connection to Preservation Session)

- How much to pre-process content vs. simply giving the bits to the users
- For digitized materials:
  - OCR or not
  - Page turner vs. page-level access
  - Quality control for access copies
- For data sets:
  - “FTP-style” download of whole files vs. specialized API for manipulation/analysis vs. hosting and providing manipulation/analysis services vs. one unified user interface
- For collections of web pages:
  - Static or dynamic rewriting of links vs. presenting as collected
  - Full-text search vs. only URL-based access
  - User interface that clearly distinguishes archived from live content vs. presenting as collected



**Recall these layers from earlier session about digital preservation*

A Useful Resource: “Opening Gateways”*

- Assessment Tool
- Diagnostic Tool
- Program Design Tool
- Cost Estimation Tool

*Pardo, Theresa A., Sharon S. Dawes, and Anthony M. Cresswell. "Opening Gateways: A Practical Guide for Designing Electronic Records Access Programs." Albany, NY: Center for Technology in Government, 2002. http://www.ctg.albany.edu/publications/guides/gateways/gateways.pdf

Appendix 1. Assessment Tool

low cost/low risk — *high cost/high risk*

Characteristics of users

homogeneous and predictable ⟷ highly variable and unpredictable

some variation in characteristics

Predictability of uses

same purpose(s) for which the data were collected ⟷ unrelated to original use

some relationship to original purpose

Sensitivity of content

not context, politically or personally sensitive ⟷ legally sealed

some context, political or personal sentivity

Appendix 2. Diagnostic Tool

The Diagnostic Tool - Users, uses, suppliers and content dimensions				
	Dimension	**Nature of dimension**		**Source and nature of constraint or flexibility**
		Key Constraint	***Adjustable***	
Users, uses, suppliers and content dimensions	Characteristics of users			
	Predictability of uses			
	Sensitivity of content			
	Frame of reference needed to interpret and use content			
	Status of meta data			

Appendix 3. Program Design Tool

Program Design Tool, Part 1 - Features and functionality			
Features and functionalities	**Modest**	**Moderate**	**Elaborate**
Users, uses, suppliers and content			
Who are your customers?			
What will customers be able to do?			
What information sources will be included and what are their characteristics?			
How extensively will the sources be integrated?			
What meta data will be provided?			
What security and confidentiality measures must be implemented?			

Appendix 4. Cost Estimation Tool

Cost Estimation Worksheet						
	MODEST		**MODERATE**		**ELABORATE**	
	First Year	**Annual**	**First Year**	**Annual**	**First Year**	**Annual**
Project Leadership						
Human Resources						
Develop program policy structure						
Full program design						
Project sponsorship activities						
Governance board						
Other						
Project Management						
Human Resources						
Overall project manager						
Develop program management procedures						
Develop program implementation plan						
Support Staff						
Other						

What is an Identifier?*

- “A stated **association** between a symbol and a thing; that the symbol may be used to **unambiguously** refer to the thing within a **given context**."
- “...an identifier will only exist as long as anyone **remembers the declaration of association**. Persistence of identifiers is not so much about remembering the identifier itself, but what it is associated with."

*Campbell, Douglas. "Identifying the Identifiers." Paper presented at the International Conference on Dublin Core and Metadata Applications, Singapore, August 27-31, 2007. (emphasis mine)

Uses for Identifiers

- Discovery – finding the object identified
- Retrieval – getting the object identified
- Citation – telling others about the object identified
- As surrogate for object itself - store the identifier, not the whole object

3 Ways to Access Information in Files

- Where it is
- What it contains
- What is said about it

Where it is – Going to specific location

- Pros:
  - fast retrieval
  - easy to implement
  - path to location can often provide hints about meaning (e.g. “finance-committee/2008/minutes”)
- Cons:
  - pointers often must be updated to reflect location changes
  - access often dependent on specific way that storage is implemented
  - hierarchical systems usually place each object in only one place, thus reifying a single path to the object that may lose relevance over time

What it contains – Searching across contents of files themselves

- Pros:
  - can serve as “stop-gap” approach when other mechanisms aren’t available
  - can often reveal information that wouldn’t have been visible otherwise (e.g. search term that never appeared in an index)
- Cons:
  - very slow to do a serial search over all content
  - often very low precision of results
  - no intellectual control (e.g. authority control)

What is said about it – Use of surrogates

- Various data values serve as access points to the object (e.g. title, checksum, file name)
- Finding an object requires either (1) a single access point or (2) combination of access points, that uniquely identify the object
- Access point or set of access points that are unique in one context may not be unique in another context, e.g.
  - calling someone only by her first name usually works within your family but doesn't work in a government database
  - file name of Untitled.doc on your computer will be unique within a given folder, because operating system prevents you from using the same name twice within a folder, but there could be many Untitled.doc files in other folders
- Implications of this...

Relying on External Surrogates to Get to Objects

- Pros:
  - Object can have an arbitrarily large number of identifiers, which can be optimized for different purposes
  - Massive performance gains, e.g. searching over index of a database, rather than serial search over data directly
  - Integrity & access controls better managed when finding doesn't require reading object's data directly
- Cons:
  - Surrogate creation, management & exposure to the outside world has to be done right! This is the theme of the week.

The Reality of Getting to Digital Objects

- Usually involves combination or hybrid of all 3 types of access
- Examples:
  - Checksum as access point – can be stored within file or outside of it, but in both cases is based on content of file itself (as opposed to externally imposed descriptor)

Think Globally, Name Locally

- Many existing mechanisms for assigning persistent IDs, generally all based on:
  - Conventions for generating strings of text that are unique within a given domain
  - Syntax for identifying the ID system used and domain
  - System for resolving persistent ID to more direct path within a storage or file system
- Can use whatever system is most appropriate in local context, but must have ability to qualify that name for use outside the local context
  - Namespaces – context within which an ID/name is applied to a given object (e.g. dc:title indicates that title is being used as that element is defined within Dublin Core namespace)

Factors in ID Persistence

- **Sustainability of the system and/or its administering organisation(s):** an identifier issued by an organisation which has questionable sustainability has limited credibility as a persistent identifier.
- **Popularity of the system**: if a PID service, and the technologies and rules underpinning it, are widely adopted and understood, then a community with an interest in the long-term sustainability of the PID scheme will be formed.
- **Quality of system documentation**: the PID system must be well documented if it is to be understood and implemented over time.
- **Standards compliance**: compliance with web standards, such as URI, can bring interoperability and transparency.
- **Low cost or free**: a repository responsible for preserving digital archives and manuscripts will use a great many PIDs, it is therefore highly desirable that any PID system is economic to administer.
- **Independent of, but interoperable with, other systems**: PIDs must outlast all systems. When repository systems and storage technologies are upgraded, PIDs must remain consistent.
- **Ability to incorporate existing identification schemes**: if there is a long-established persistent identifier scheme already in use within an institution or particular sector, it might be useful to incorporate these identifiers into the namespace of whatever PID system is adopted for the electronic environment (an example might be ISBN identifiers for books).

Source: Thomas, Susan, Renhart Gittens, Janette Martin, and Fran Baker. "Workbook on Digital Private Papers." 2007. p.52 [PARADIGM Workbook]

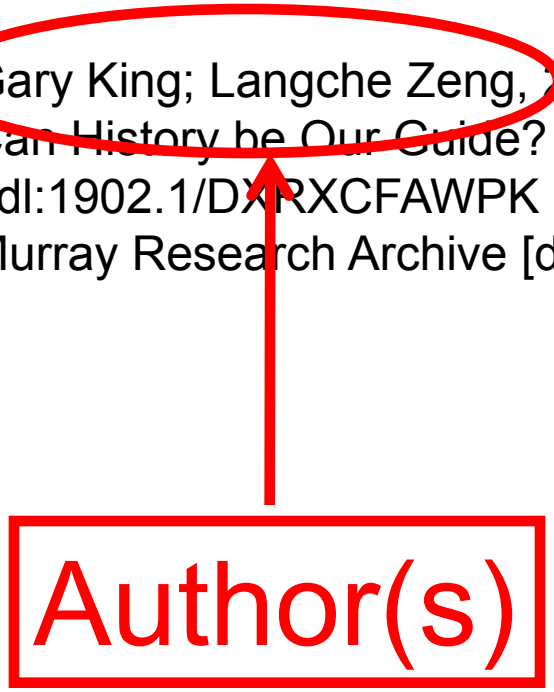
Citation Standard for Data Sets (Dataverse Network Project)*

- Replication data-set citation example (six components):

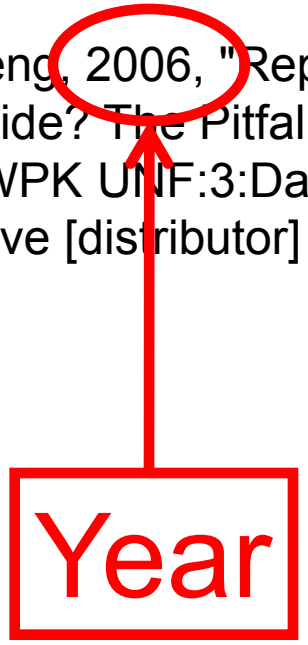
Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History be Our Guide? The Pitfalls of Counterfactual Inference'" hdl:1902.1/DXRXCFAWPK UNF:3:DaYlT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor]

*http://thedata.org/citation

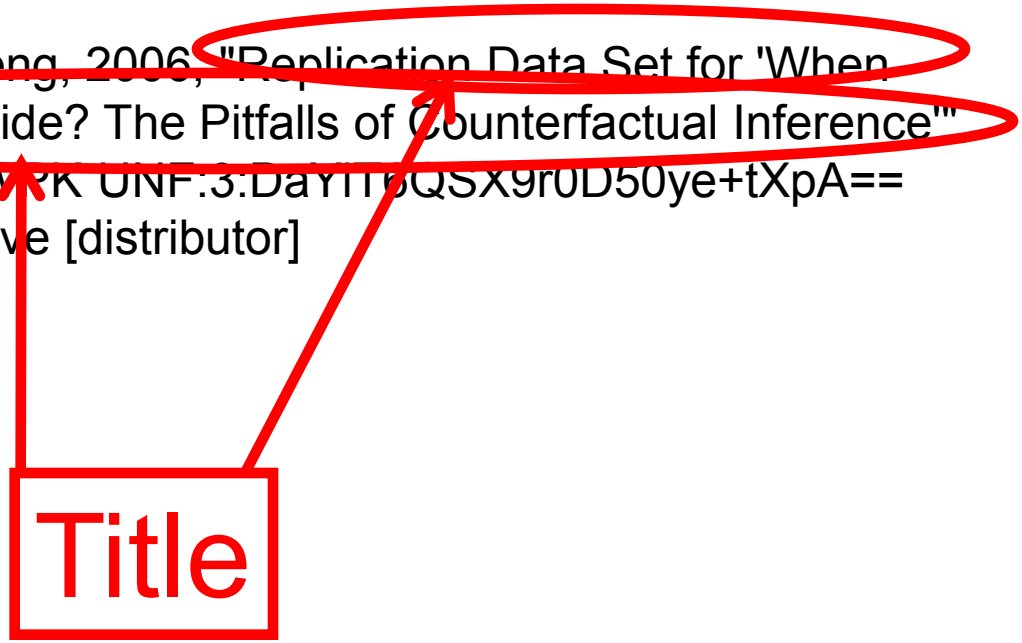
Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History be Our Guide? The Pitfalls of Counterfactual Inference'" hdl:1902.1/DXRXCFAWPK UNF:3:DaYIT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor]



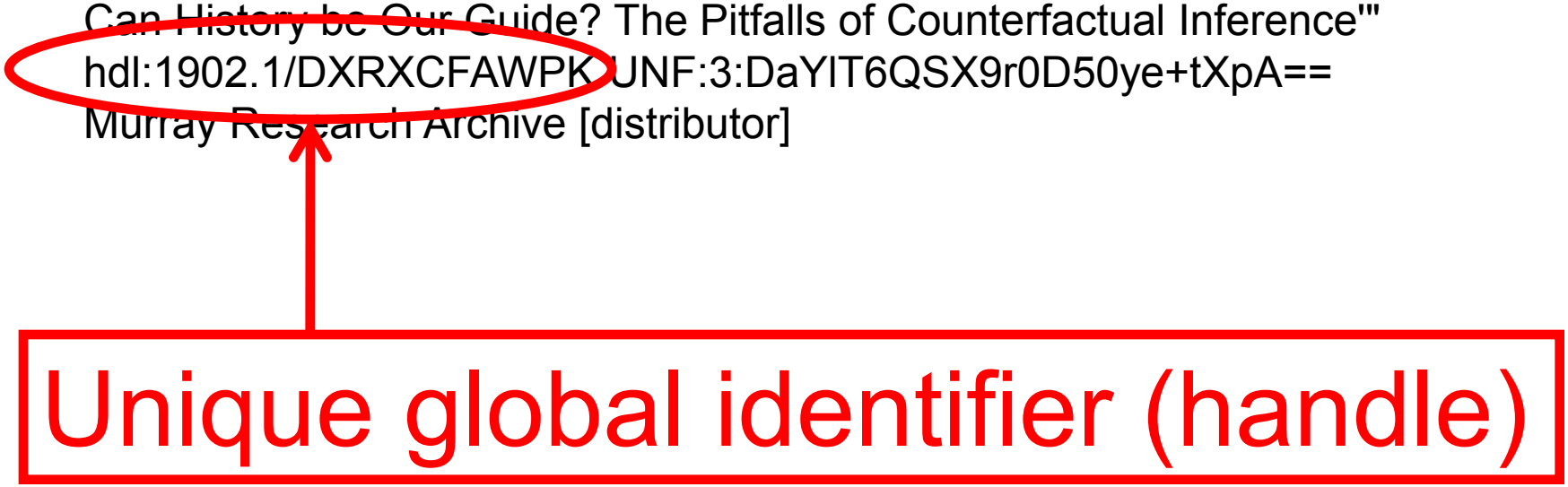
Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History be Our Guide? The Pitfalls of Counterfactual Inference'" hdl:1902.1/DXRXCFAWPK UNF:3:DaYIT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor]



Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History be Our Guide? The Pitfalls of Counterfactual Inference'", hdl:1902.1/DXRXCFAWRK UNF:3:DaYIT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor]



Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History be Our Guide? The Pitfalls of Counterfactual Inference'"
hdl:1902.1/DXRXCFAWPK UNF:3:DaYlT6QSX9r0D50ye+tXpA==
Murray Research Archive [distributor]



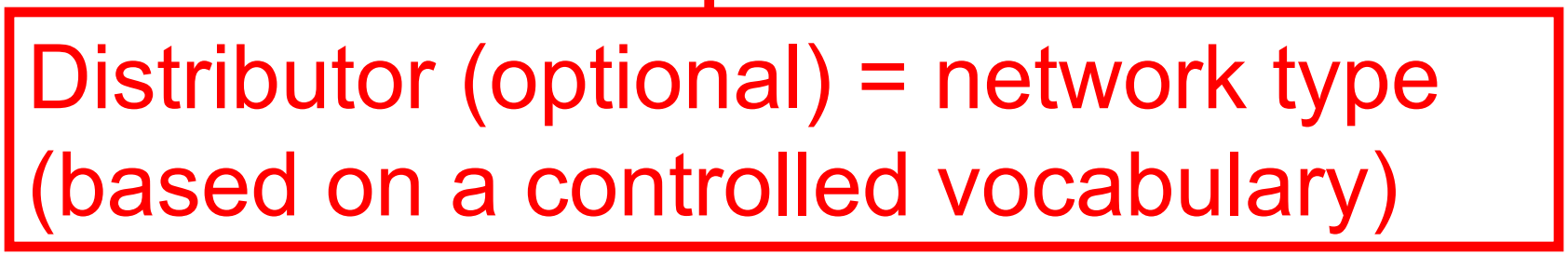
Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History be Our Guide? The Pitfalls of Counterfactual Inference'" hdl:1902.1/DXRXCFAWPK UNF:3:DaYIT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor]

Universal Numerical Fingerprint (UNF)

Based on a one-way algorithm applied to the data content*, designed to be independent of file format of the data

*Normalize values in various ways, encode as UTF-8 strings, then apply SHA256 hashing

Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History be Our Guide? The Pitfalls of Counterfactual Inference'" hdl:1902.1/DXRXCFAWPK UNF:3:DaYIT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor]



Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History be Our Guide? The Pitfalls of Counterfactual Inference'" hdl:1902.1/DXRXCFAWPK UNF:3:DaYlT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor]

When in print, also provide URL

Discovery

- Search and delivering records can be expensive, particularly when there is limited intellectual control
- Delete doesn't mean delete
- The Sedona Principles[1]
- Risk Profiler Self-Assessment for E-Discovery – ARMA and NetDiligence (2006)[2]
- Companies provide specialized training and certification

1. The Sedona Principles Addressing Electronic Document Production, Second Edition, June 2007. https://thesedonaconference.org/download-pub/81
2. http://www.arma.org/profiler/ediscovery.cfm

What does it mean for an electronic record to be “accessible”?

The Pragmatism of US Discovery Law

- Federal Rules of Civil Procedure (emphasis mine):
  - "The person responding need not provide discovery of electronically stored information from sources that the person identifies as **not reasonably accessible because of undue burden or cost**." (Rule 45 (d)(1)(D))
  - "A party must produce documents as they are kept in the **usual course of business or** must organize and label them to correspond to the **categories in the request**" (Rule 34 (b)(2)(E)(i))
  - "If a request does not specify a form for producing electronically stored information, a party must produce it in a form or forms in which it is **ordinarily maintained or in a reasonably usable form** or forms" (Rule 34 (b)(2)(E)(ii))
- Client-attorney privilege: "…disclosure does not operate as a waiver in a federal or state proceeding if:
  - (1) the disclosure is inadvertent;
  - (2) the holder of the privilege or protection took reasonable steps to prevent disclosure; and
  - (3) the holder promptly took reasonable steps to rectify the error…" (Rule 502(b) of the Federal Rules of Evidence - added in 2008)

Judge Shira Scheindlin:
“[t]he more information there is to discover, the more expensive it is to discover all the relevant information until, in the end, ‘discovery is not just about uncovering the truth, but also about how much of the truth the parties can afford to disinter.’ ” (Zubulake I, 217 F.R.D. at 311 (quoting Rowe Entm’t, Inc. v. William Morris Agency, Inc., 205 F.R.D. 421, 423 (S.D.N.Y. 2002)).

Seven-Factor Test from *Zubulake v. UBS Warburg*

1. extent to which the request is specifically tailored to discover relevant information
2. availability of such information from other sources
3. total cost of production, compared to the amount in controversy
4. total cost of production, compared to the resources available to each party
5. relative ability of each party to control costs and its incentive to do so
6. importance of the issues at stake
7. relative benefits to the parties of obtaining the information

(217 F.R.D. at 322)

Zubulake's Five Categories of ESI (Most to Least Accessible)*

- Active, online data
- Near-line data
- Offline storage
- Backup tapes
- Erased, fragmented or damaged data

*See: Lange, Michele C. S., and Kristin M. Nimsger. *Electronic Evidence and Discovery: What Every Lawyer Should Know Now*. 2nd ed. Chicago, IL: Section of Science & Technology Law American Bar Association, 2009. p.75.

Magistrate Judge John Facciola:

"…I am anything but certain that I should permit a party who has failed to preserve accessible information without cause to then complain about the inaccessibility of the only electronically stored information that remains" (*Disability Rights Council of Greater Wash. v. Wash. Metro. Transit Auth.,* 242 F.R.D. 139 (D.D.C. 2007)).

What are the main factors that determine whether a record is “accessible”?

Which factors are most likely to pose accessibility risks for files stored on your computer?

Rights to Control Information

- Most frequently discussed in library lit is copyright
- Claims can extend far beyond intellectual property rights, as defined by US law
- Cultural property, replevin and repatriation
- Right to privacy
- Protection of human subjects in research
- Privileged or protected information (e.g. client-attorney, healthcare, social services, library circulation, source-journalist)
- Right to publicity - individual's protection from unauthorized commercial use of her name, persona or likeness
- Prevention of misappropriation (including plagiarism)

Some Types of Documents that May Contain Sensitive Information

- Server, circulation and use logs
- Personnel files
- Research data sets
- Oral histories
- Medical records (e.g. clinical trials)
- Correspondence within donated collections of papers
- Cultural materials from indigenous or foreign populations
- Captured web pages
- Courseware

Example of a “Dark Archive” - Portico

- $3 Million from LC and matching funds from Mellon to Ithaka (nonprofit organization in New York City and Princeton, NJ)
- Aims to build sustainable technical infrastructure and business model for long-term preservation of electronic journals
- Participating libraries pay fee and sign license agreement

Preserved Archival Units (articles, books, etc.)	25,258,365
Preserved E-Journal Titles	13,315
Preserved E-Book Titles	93,664
Preserved Files	380,534,401
Preserved Images	224,694,807
Preserved Repository created archival files	77,328,064
Preserved Supplied text files	56,137,484
Preserved Application Specific Files	20,687,047
Preserved Multi-file Packages	845,979
Preserved Video Files	44,607
Preserved Audio Files	1,200
Preserved Executable Files	5

Portico Terms of Participation

- Payment based on Library Materials Expenditure (LME), with savings for early adopters (Archive Founders) and consortia
- Five-year term with automatic one-year renewal
- Trigger Events (not immediate and then "may be printed or saved only for educational, research or noncommercial use":
  - Publisher no longer in business
  - Publisher stops publishing title or no longer offers back issues
  - "Catastrophic failure" (technical or business) for more than 90 days
- Can't "copy, download, or attempt to download an entire issue or issues of a publication from the Archive or substantial portions of the Archive"

Portico Journal Archive License Agreement, version 4.1. http://www.portico.org/digital-preservation/wp-content/uploads/2011/10/2012-Portico-Library-Agreement.doc

Publisher Copyright Agreements - SHERPA/RoMEO Database (http://www.sherpa.ac.uk/romeo.php)

ROMEO Color	Archiving policy
green	can archive pre-print and post-print or publisher's version/PDF
blue	can archive post-print (final draft post-refereeing) or publisher's version/PDF
yellow	can archive pre-print (pre-refereeing)
white	archiving not formally supported

Shared E-Resource Understanding (SERU)

- “SERU offers publishers and libraries the opportunity to save both the time and the costs associated with a negotiated and signed license agreement by agreeing to operate within a framework of shared understanding and good faith”
- Provides “a set of common understandings for publishers and libraries to reference as an alternative to a formal license when conducting business”
- NISO Recommended Practice document issued in February 2008*

*http://www.niso.org/publications/rp/RP-7-2008.pdf

Section 108 Study Group

http://www.section108.gov/

- “select committee of copyright experts, convened by the Library of Congress, and charged with updating for the digital world the Copyright Act's balance between the rights of creators and copyright owners and the needs of libraries and archives”
- Report issued on March 31, 2008*

*http://www.section108.gov/docs/Sec108StudyGroupReport.pdf

Some Section 108 Study Group Recommendations

- Include museums for Section 108 eligibility
- “Permit certain qualified libraries and archives to make preservation copies of at-risk published works prior to any damage or loss” which would have limited access
- “Permit libraries and archives to capture and reproduce publicly available Web sites and other online content for preservation purposes and to make those copies accessible to users for private study, research or scholarship” with ability for rights holders to opt out
- Allow libraries & archives “to make a limited number of copies, as reasonably necessary, to create and maintain a single replacement or preservation copy” (in contrast to current 3-copy limit)

Nonexpressive Works*

"…**nonexpressive uses** of copyrighted works—i.e., acts of copying that **do not communicate the author's original expression to the public**—should not generally be regarded as infringing.

The legal status of actual copying for nonexpressive uses was not a burning issue before digital technology: there simply was no commercially relevant total literal copying directed towards a nonexpressive end. …it would be both uncommon and nonsensical to photocopy *Gone With The Wind* and then to use it to light a fire. … However, **digital technology** and the increasing value of **metadata** have combined to make the legality of **nonexpressive copying** arguably the **most significant issue** in copyright law today." (1625)

"However, given the significant role of nonexpressive copying in Internet search engines and other **copy-reliant technologies**, the legality of nonexpressive copying is an issue that copyright doctrine must now address." (1625, emphasis mine)

*Sag, Matthew. "Copyright and Copy-Reliant Technology." *Northwestern University Law Review 103, no. 4 (2009): 1607-82.*

Major Issue – Orphaned Works

- Carnegie Mellon feasibility study (1999-2001)*
  - To determine likelihood of publishers granting nonexclusive permission to digitize & provide open Web access to copyrighted books
  - 21% of publishers (19% of titles) couldn't be located
  - 27% of publishers granted permission to digitize & provide web access (24% of the copyrighted books)
  - 68% of publishers that granted permission applied some restriction
- General lesson from several investigations: often a decent approval rate for digitizing, but tracking down the "approver" can be hugely expensive

*Source for above data: Denise Troll, Convey, "Acquiring Copyright Permission to Digitize and Provide Open Access to Books," Digital Library Federation & Council on Library & Information Resources, October 2005, http://www.diglib.org/pubs/dlf105/dlf105.htm

See also: Heather Brison, Mark Allen Greene, Cathy Henderson, Peter Hirtle, Peter Jaszi, William Maher, Aprille Cooke McKay, Richard Pearce-Moses, and Merrilee Proffitt. "Orphan Works: Statement of Best Practices." Society of American Archivists. January 12, 2009 (Revised June 17, 2009). http://www.archivists.org/standards/OWBP-V4.pdf

Hidden Data within Files

- Lots of data in many files that you don't always see with the naked eye. For example:
  - Comments within the code
  - Stored rules and styles
  - Change tracking information
  - Metadata stored in file headers and elsewhere
  - Viruses

Examples of Hidden Data in MS Office Documents

- Application used to create document
- Authors, user names, organizational affiliations & author history
- Comments
- Custom properties
- Database queries
- Embedded objects (OLE) – elements not immediately visible (e.g. spreadsheet)
- Fast save – change history appended to end of file, rather than applied to body of document
- GUID – globally unique identifier for computer (see Leach et al, 2005)
- Hidden cells, slides, text – purposely hidden but then possibly forgotten
- Outlook (email) properties & routing slips
- Path information – audio & video paths, author history, linked objects, printers, hyperlinks, include fields, template
- Presentation notes
- Printer driver information
- RSID – Revision save ID (differentiates changes from different editing sessions)
- Tracked changes (added to PPT and Excel in Office XP)
- Versions
- Visual Basic code – including macros & viruses (and identity of code creators)
- Web server information
- White text (on white background)

Hidden Image Data

- Content outside crop area
- Layered objects (hidden from view)
- Pixel information in resized and embedded image
- Metadata:
  - GIF – comment extensions and application extensions
  - JPEG - camera use, date/time, distance settings, location, thumbnail image

E-Discovery and Forensics Impact on Computer Industry

- Changes to Microsoft Office, e.g.
  - Document Investigator (Office 2007), Prepare for Distribution (Office 2010 Backstage View)
  - Appearance of comments and tracked changes by default when opening document (Office 2007)
  - “Fast save” turned off by default in Word 2000 and disabled in Word 2003
  - Rise & fall of the embedded PID GUID (introduced in Office 97, abandoned in Office 2000)
- Some other Windows changes
  - No more accidental dumping of RAM slack (now writes zeros)
  - In Index.dat, deleted entries are now (since IE7 and Vista) reportedly now zeroed out
- Macintosh – encryption and safe delete
- Large market for software designed specifically for managing e-discovery, e.g. EnCase & Neutrino (Guidance Software), Discovery Partner (Electronic Evidence Discovery), iScrub (Esquire Innovations)

Difficulties of Determining Use of Digital Resources

Who is the User?

Technical layers between user & document

- Dynamic allocation of IP address to user's computer
- Proxy servers – server through which a user can connect to network services, rather than connecting directly
- Cached copies – saved closer to the user to reduce bandwidth demand & download time
- Aggregators – host, deliver & get paid for content
- Gateways – doesn't store, but refers users to content or submits requests on their behalf
- Access controls – can result in "turnaways" (rejected sessions)
- User name – not always traceable to a person
- User must click on something, which can result in double-clicks

Organizational and institutional layers

- Consortium (expressed as range of IP addresses)
- Consortium member (generally an organization and expressed as subset of the above range of IP addresses)
- Organizational units & individuals within consortium member

Ways to Identify a Persistent User

- IP address or address range
- Host name
- Session cookie
- User cookie
- User name (log in)

How Does this Fit into the Information Seeking Process?

- Server stats show only a tiny slice into the information seeking process
- Can fail to capture:
  - Determination of original information need
  - Interactions with other people
  - Use of other documents
  - Use of search engines
  - Navigation to the library's site
  - Use of the document (e.g. viewing, printing, sharing, copying, annotating)

What is the Documentary Unit?

- Functional Requirements for Bibliographic Records (FRBR): work, expression, manifestation, item
- COUNTER
  - Defines: article, book, chapter, collection, database, database record, entry, full-text article, item, journal, section, title, volume
  - HTML & PDF copies of a single work treated as two different uses
  - Files associated with a web version (e.g. GIF images, style sheets) not counted

Some Future Use Projections

- More macro-level tools & methods
  - Federated search
  - Data mining in large data sets
  - Network analyses
- Access at point of interest (mobile computing)
- Visualizations
- Computer-supported redaction
- Parallel moves **toward** and **away from** item-level focus
- Analysis of archives as reflection of ideology and power relationships
- Revisiting authenticity issues, but within more public discussion (why should I trust your repository more than theirs?)