



College of Information Studies

University of Maryland Hornbake Library Building College Park, MD 20742-4345

Description

Week 5

LBSC 671

Creating Information Infrastructures

Metadata Capture: User Behavior

Minimum Scope

Segment Object Class

Behavior Category

Examine	View Listen	Select	
Retain	Print	Bookmark Save Purchase Delete	Subscribe
Reference	Copy / paste Quote	Forward Reply Link Cite	
Annotate	Mark up	Tag Publish	Organize
Create	Type Edit		

Exploiting Behavioral Metadata

When you visit a website ...

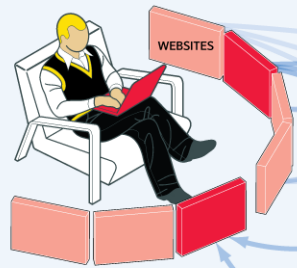
... tiny tracking files watch what you do online ...

... and develop a profile of your behavior.

Some sell your data on an exchange ...

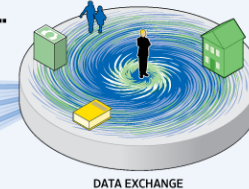
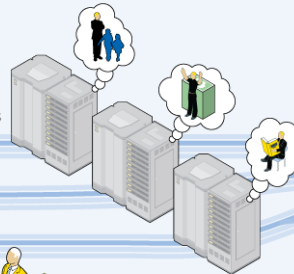
... which can combine it with other sources of personal data ...

... to be sold to advertisers looking for consumers like you.



PARENTING INTERESTS
SHOPPING ONLINE
BROWSING BOOKS

TRACKING COMPANIES



DATA EXCHANGE



OFFLINE DATA
Census figures, real estate records, car registration, etc.

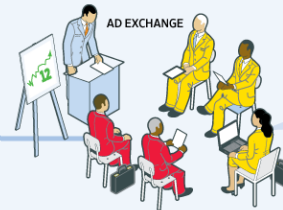
You might like this book!

You might like this car!



Often, a tracking company sells this information directly to advertisers.

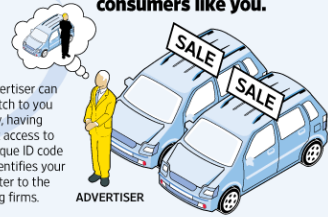
Advertisers buy ad space from websites at auctions.



AD EXCHANGE

An advertiser can now pitch to you directly, having bought access to the unique ID code that identifies your computer to the tracking firms.

ADVERTISER



BACK TO YOU

The websites you visit show you ads or other content based on the description of you in the dossiers they've built and analyzed.

Metadata Extraction: Named Entity “Tagging”

- Machine learning techniques can find:
 - Location
 - Extent
 - Type
- Two types of features are useful
 - Orthography
 - e.g., Paired or non-initial capitalization
 - Trigger words
 - e.g., Mr., Professor, said, ...

Your query has finished



Search	Topic		Person	
Clear	Organization		Location	
OR	Speaker		Text	
AND	Story	Jewish-Arab relations : Politics and government : Palestinian Arabs : Middle East : Israel : Terrorism		

- 5 stories about: Jewish-Arab relations : Politics and government : Palestinian Arabs : Middle East : Israel : Terr
- Jewish-Arab relations : Politics and government : Palestinian Arabs : Middle East : Israel : Terrorism : Pale
- Jewish-Arab relations : Israel : Middle East : Middle East peace negotiations : Politics and government : P

male 5

Well as all work during president Clinton's trip to New York tonight and he enjoys the performance of the opera Carmen at Lincoln Center and see the scene there is a lot of Broadway. Now earlier today Mr. Clinton announced that the UN united Nations general assembly that he plans to send a nuclear test ban treaty to the Senale the treaty bans all nuclear test explosions and is regarded as a milestone in the arms control. Two israeli security guards were wounded in an early morning shooting in Jordan a government official says three men and a car opened fire on the guard's car wounding both before Skipping guards were treated at a hospital and released it is real several West Bank villages were sealed by israeli soldiers who search for the islamic militants behind two recent suicide bombings in Jerusalem palestinian leader Yasser Arafat says that he believes those was counsel for the bombing case and abroad.

- Jewish-Arab relations
- Middle East peace negotiations
- Middle East
- Palestinian self-rule areas
- Israel
- Politics and government
- Arafat, Yasir
- Palestinian Arabs

Metadata Sources

- Automated
 - Capture
 - Extraction
 - Classification
- Manual
 - Professional
 - Community
 - Personal

Community Metadata: “Folksonomies”



del.icio.us / tag / radio

[popular](#) | [recent](#)

[login](#) | [register](#) | [help](#)

All items tagged **radio** ([create tag description](#)) → view **popular**

del.icio.us

[« earlier](#) | [later »](#)

[Playbill Radio](#) [save this](#)

by [wheelmaker2](#) to [music radio](#) [broadway playbill](#) [Entertainment ...](#) [saved by 12 other people](#) ... 2 mins ago

[Rhapsody](#) [save this](#)

by [srminnton](#) to [music rhapsody radio](#) [streaming entertainment mp3 ...](#) [saved by 515 other people](#) ... 3 mins ago

[Kasper Hauser's "This American Life" Parody: Episode 1](#) [save this](#)

Sounding like This American Life.

by [hansenn](#) to [comedy radio thisamericanlife ...](#) [saved by 27 other people](#) ... 5 mins ago

[Breaking News | Latest News | Current News - FOXNews.com](#) [save this](#)

by [parcley](#) to [radio news ...](#) [saved by 2839 other people](#) ... 7 mins ago

[Family.org](#) [save this](#)

by [bastian_balthasar_bux](#) to [Family christian Christianity radio news RELIGION reference ...](#) [saved by 311 other people](#) ... 16 mins ago

[BBC - 1Xtra - Homepage](#) [save this](#)

by [okajun](#) to [reggae radio ...](#) [saved by 135 other people](#) ... 17 mins ago

[Sound & Spirit](#) [save this](#)

by [dragonjazz](#) to [radio ...](#) [saved by 19 other people](#) ... 19 mins ago

<http://www.pandora.com/?tc=x-036821-0035-1149> [save this](#)

[music](#)

by [sarah.bierman](#) to [radio ...](#) [saved by 4 other people](#) ... 20 mins ago

▼ related tags

[music](#)

[media](#)

[audio](#)

[scanner](#)

[streaming](#)

[radiocator](#)

[frequencies](#)

[ham](#)

[musik](#)

[journalism](#)

[imported](#)

Community Metadata: Games With a Purpose

The ESP Game - Netscape 6

0:11
Time Left

The ESP Game

2100
score



Taboo Words
MAN
BEARD

Your Guesses
HAT

Type your next guess:

Guess



© 2000-2003 Carnegie Mellon University. All rights reserved. Patent Pending

Community Metadata: Crowdsourcing

ANCIENT LIVES

TRANScribe MEASURE Save!

The screenshot shows a digital interface for transcribing ancient Greek text. The main area displays a fragment of a papyrus scroll with several lines of text. The visible text includes: ΩΓΓΕ, ΚΡΑΤΗ, ΩΓΓΕ, ΣΩΛΕΤΕ, and ΠΑΤΟ. A red circle highlights the letter Ω in the third line, and a blue circle highlights the letter Γ in the same line. The interface includes a top bar with 'TRANScribe' and 'MEASURE' tabs, a 'Save!' button, and a 'Key' section at the bottom with a grid of Greek characters for selection. The 'Key' section includes characters like Σ, Εε, Ρρ, Ττ, Υυ, Θθ, Ιι, Οο, Ππ, Αα, Κκ, Λλ, Ζζ, Χχ, Ψψ, Ωω, Ββ, Νν, Μμ, and symbols for punctuation and navigation.

- Colour
- Map
- Match
- Talk
- Next





Sources of File Type Metadata

- Capture:
 - MyDocument.xls
 - Attachment MIME type
- Extraction
 - “Magic bytes”
- Classification
 - Machine learning on byte sequences
- Manual
 - Mechanical Turk

Metadata Challenges

- Balancing cost and benefit
- Accommodating dynamic factors
 - Content
 - Location
- Reuse for unanticipated purposes
- Remaining interpretable in the far future

Putting It All Together






	<i>Material Culture Libraries Archives Museums</i>	<i>Bibliographic Libraries Archives Museums</i>	<i>Archival Libraries Archives Museums</i>
 Data Structure	CDWA	MARC	EAD
 Data Content	CCO	AACR2 (RDA)	DACS
 Data Format	XML	XML/ISO2709	XML
 Data Exchange	OAI	OAI Z39.50 SRU/SRW	OAI

Some Types of “Metadata”

- Descriptive
 - Content, creation process, relationships
- Technical
 - Format, system requirements
- Administrative
 - Acquisition, authentication, access rights
- Preservation
 - Media migration
- Usage
 - Display, derivative works

Adapted from
Introduction to Metadata,
Getty Information Institute (2000)

Aspects of Metadata

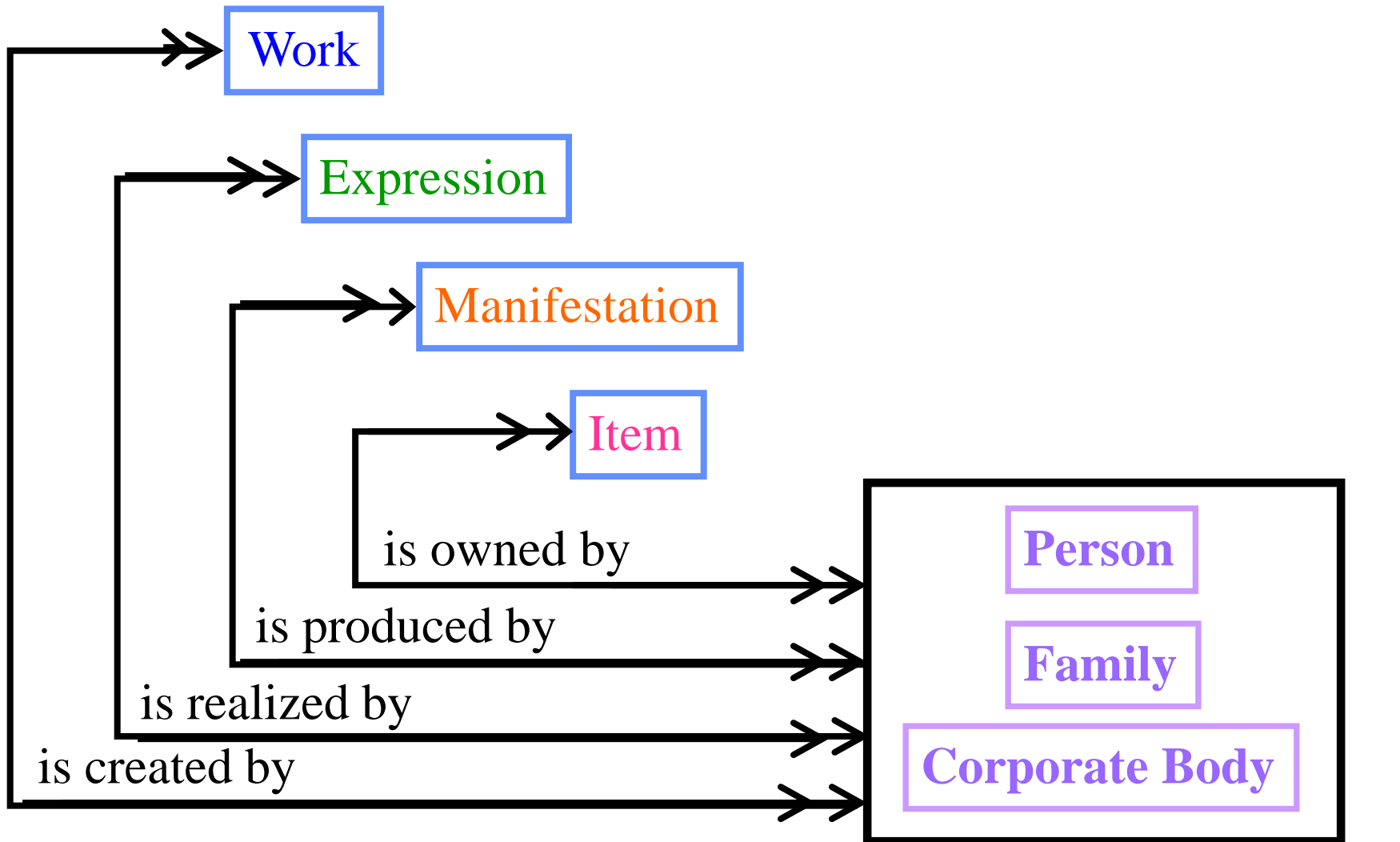
- Framework 
 - Functional Requirements for Bibliographic Records (FRBR)
- Schema (“Data Fields and Structure”) 
 - Dublin Core
- Guidelines (“Data Content and Values”) 
 - Resource Description and Access (RDA)
 - Library of Congress Subject Headings (LCSH)
- Representation (abstract “Data Format”) 
 - Resource Description Framework (RDF)
- Serialization (“Data Format”) 
 - RDF in eXtensible Markup Language (RDF/XML)

Fostering Consistency

- Content Standards
 - Resource Description and Access (RDA)
 - Describing Archives: a Content Standard (DACS)
- Authority Control
 - Subject Authority
 - Name authority

FRBR Entity Types

- Subject-Only Entities
 - (abstract) Concepts
 - (tangible) Objects
 - (any kind of) Places
 - Events
- Subject or Responsibility Entities
 - Persons
 - (any kind of) “Corporate” Bodies
 - Families (technically, only in FRAD)
- Product Entities
 - Works, Expressions, Manifestations, Items



Work

- The idea or impression in the mind of its creator
 - Completely abstract, no physical form
- What all forms, presentations, publications, or performances of a work have in common
 - *Romeo & Juliet*
 - Homer's *Odyssey*
 - Debussy's *Syrinx*

Expression (Realization)

- A work formulated into an ordered presentation
- When a work takes a *form*
 - Can be notational, aural, kinetic, etc.
- Excludes aspects of form not integral to the work
 - Font, layout, etc. (with some exceptions)
- Attributes: Form, Language

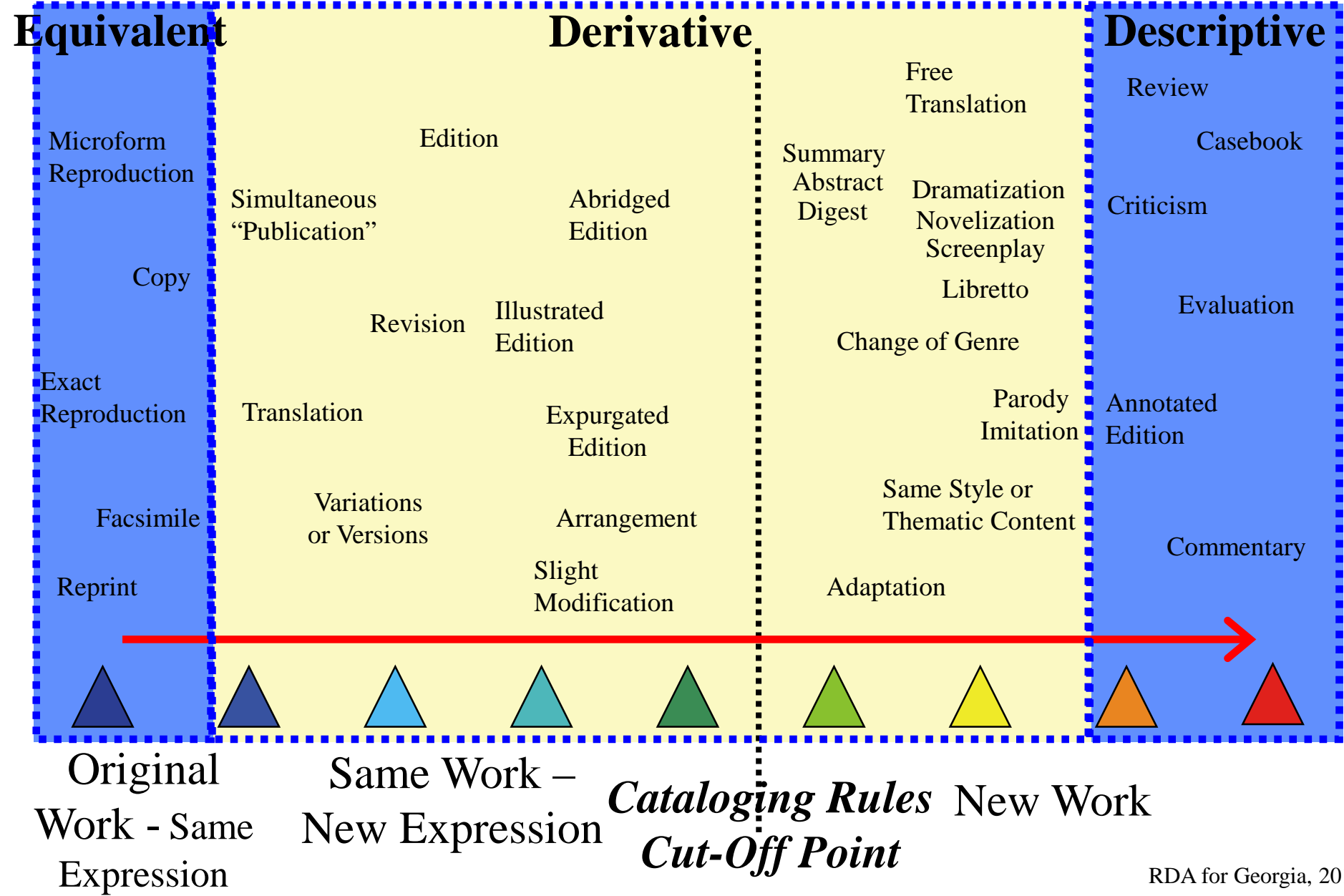
Manifestation

- Physical embodiment of an expression
 - The level usually described via cataloging
- **Set** of physical objects that bear the same:
 - *intellectual content* (expression), and
 - *physical form* (item)
- May have one or many items
 - Mona Lisa, Gone with the Wind, ...
- Attributes
 - Format, Physical medium, Manufacturer

Item

- Instance of a manifestation
 - *A thing!*
- Attributes:
 - Owned by, Location, Condition

Family of Works



FRBR Bibliographic User Tasks

- Find it
 - Search (“to find”)
 - Recognize (“to identify”)
 - Choose (“to select”)
- Serve it
 - Location (“to obtain”)

Resource Description & Access (RDA)

- RDA metadata describes entities *associated with* a resource to help users perform the following tasks:
 - **Find** information on that entity and on resources associated with the entity
 - **Identify**: confirm that the entity described corresponds to the entity sought, or to distinguish between two or more entities with similar names, etc.
 - **Clarify** the relationship between two or more such entities, or to clarify the relationship between the entity described and a name by which that entity is known
 - **Understand** why a particular name or title, or form of name or title, has been chosen as the preferred name or title for the entity

Components of RDA

- “Elements” (Attributes)
 1. Of manifestations and items
 2. Of works and expressions
 3. Of persons and corporate bodies
 4. Of concepts
- Relationships
 5. Among product entities
 - Content entities: work, expression, manifestation, item
 6. Between product and responsibility entities
 - Responsibility entities: person, family, corporate body
 7. Between works and subject entities
 - Subject entities: concepts, objects, places, events

Bibliographic Relationships

- **Equivalence:** exact (or nearly exact) copies
 - mp3 recording burned from a CD, ...
- **Derivative:** work based on/derived from another
 - Updated edition, adaptation, ...
- **Descriptive:** work that describes another work
 - Criticism, commentary, summary (e.g., Cliffs Notes), ...

More Bibliographic Relationships

- Whole-part: One work is part of another work
 - Volume in an encyclopedia, chapter in a book, ...
- Accompanying: A work meant to go with another work
 - Math workbook w/ textbook, index, documentation, ...
- Sequential: Work precedes/continues an existing work
 - Issues of a publication, sequels/prequels, ...
- Shared characteristic: Something in common
 - Author, title, language, subject, ...

Some RDA Elements for Products

- Work
 - ID
 - Title
 - Date
 - etc.
- Expression
 - ID
 - Form
 - Date
 - Language
 - etc.
- Manifestation
 - ID
 - Title
 - Statement of responsibility
 - Edition
 - Imprint (place, publisher, date)
 - Form/extent of carrier
 - Terms of availability
 - Mode of access
 - etc.
- Item
 - ID
 - Provenance
 - Location
 - etc.

RDA: Person

- “An individual or an identity established by an individual (either alone or in collaboration with one or more other individuals)”
- Includes fictitious entities
 - Miss Piggy, Snoopy, etc. in scope if presented as having responsibility in some way for a work, expression, manifestation, or item
- Also includes real non-humans
 - Only in US RDA test

RDA Person Examples

```
100 0# $a Miss Piggy.  
245 10 $a Miss Piggy's guide to life / $c  
      by Miss Piggy as told to Henry Beard.  
700 1# $a Beard, Henry.
```

```
100 0# $a Lassie.  
245 1# $a Stories of Hollywood / $c told  
      by Lassie.
```

RDA: Language and Script

- Names:
 - USA: In authorized and variant access points, apply the alternative to give a romanized form.
 - For some languages, can also give variant access points in original language/script
- Other elements:
 - If RDA instructions don't specify language, give element in English

RDA: Preferred Name

- Used as the “authorized” (i.e., canonical) access point
- Choose the form most commonly known
- Variant spellings:
 - Choose the form found on the first resource received
- If individual has more than one identity
 - Construct a preferred name for each identity

RDA: Additions to Preferred Name

- title or other designation associated with person
- date of birth and/or death * ^
- fuller form of name * ^
- period of activity of person * ^
- profession or occupation *
- field of activity of person *

* = if need to distinguish; ^ = option to add even if not needed

RDA: Surnames Indicating Relationships

- Include words, etc., (e.g., Jr., Sr., IV) in preferred name – not just to break conflict

```
100 1# $a Rogers, Roy, $c Jr., $d 1946-  
670 ## $a Growing up with Roy and Dale, 1986:  
      $b t.p.(Roy Rogers, Jr.) p. 16 (born  
      1946)
```

RDA: Terms of Address When Needed

- When the name consists only of the surname
 - (Seuss, **Dr.**)
- For a married person identified only by a partner's name and a term of address
 - (Davis, Maxwell, **Mrs.**)
- If part of a phrase consisting of a forename(s) preceded by a term of address
 - (Sam, **Cousin**)

RDA: Profession or Occupation

- Core:
 - for a person whose name consists of a phrase or appellation not conveying the idea of a person, **or**
 - if needed to distinguish one person from another with the same name
- Overlap with “field of activity”

```
100 1# $a Watt, James $c (Gardener)
```

RDA: Field of Activity of Person

- Field of endeavor, area of expertise, etc., in which a person is or was engaged
- Core:
 - For a person whose name consists of a phrase or appellation not conveying the idea of a person, or
 - If needed to distinguish one person from another with the same name

```
100 0# $a Spotted Horse $c (Crow Indian chief)
```

RDA: Associated Date for Person

- Three dates:
 - Date of birth
 - Date of death
 - Period of activity of the person
- Guidelines for probable dates are in RDA 9.3.1

RDA: Associated Place for Person

- Place of birth
- Place of death
- Country associated with the person
- Place of residence

DACS Principles

1. Records in archives possess unique characteristics.
2. The principle of respect des fonds is the basis of archival arrangement and description.
3. Arrangement involves identification of groupings within material.
4. Description reflects arrangement.
5. The rules of description apply to all archival materials regardless of form or medium.
6. The principles of archival description apply equally to records created by corporate bodies, individuals, or families.
7. Archival descriptions may be presented at varying levels of detail to produce a variety of outputs.
8. The creators of archival materials, as well as the materials themselves, must be described.

(Single-Level) DACS Elements

Required

- Reference code
- Name+location of repository
- Title
- Date
- Extent
- Name of creator(s)
- Scope and content
- Conditions governing access
- Languages and scripts
- Plus, for “Optimal”
 - Administrative/biographical history
 - Access points

Optional

- System of arrangement
- Physical access
- Technical access
- Conditions for reproduction and use
- (other) Finding aids
- Custodial history
- Immediate source of acquisition
- Appraisal, destruction, scheduling
- Accruals (anticipated additions)
- Existence+location of originals
- Existence+location of copies
- Related archival materials
- Publication note
- Notes
- Description control

Authority Control

- Unify references to the same entity (synonyms)
 - Samuel Clemens, Mark Twain
- Distinguish references to different entities (homonyms)
 - Michael Jordan (basketball), Michael Jordan (computers)
- Establish “access points”
 - Canonical and variant forms, to better support “find it” tasks

Access Points

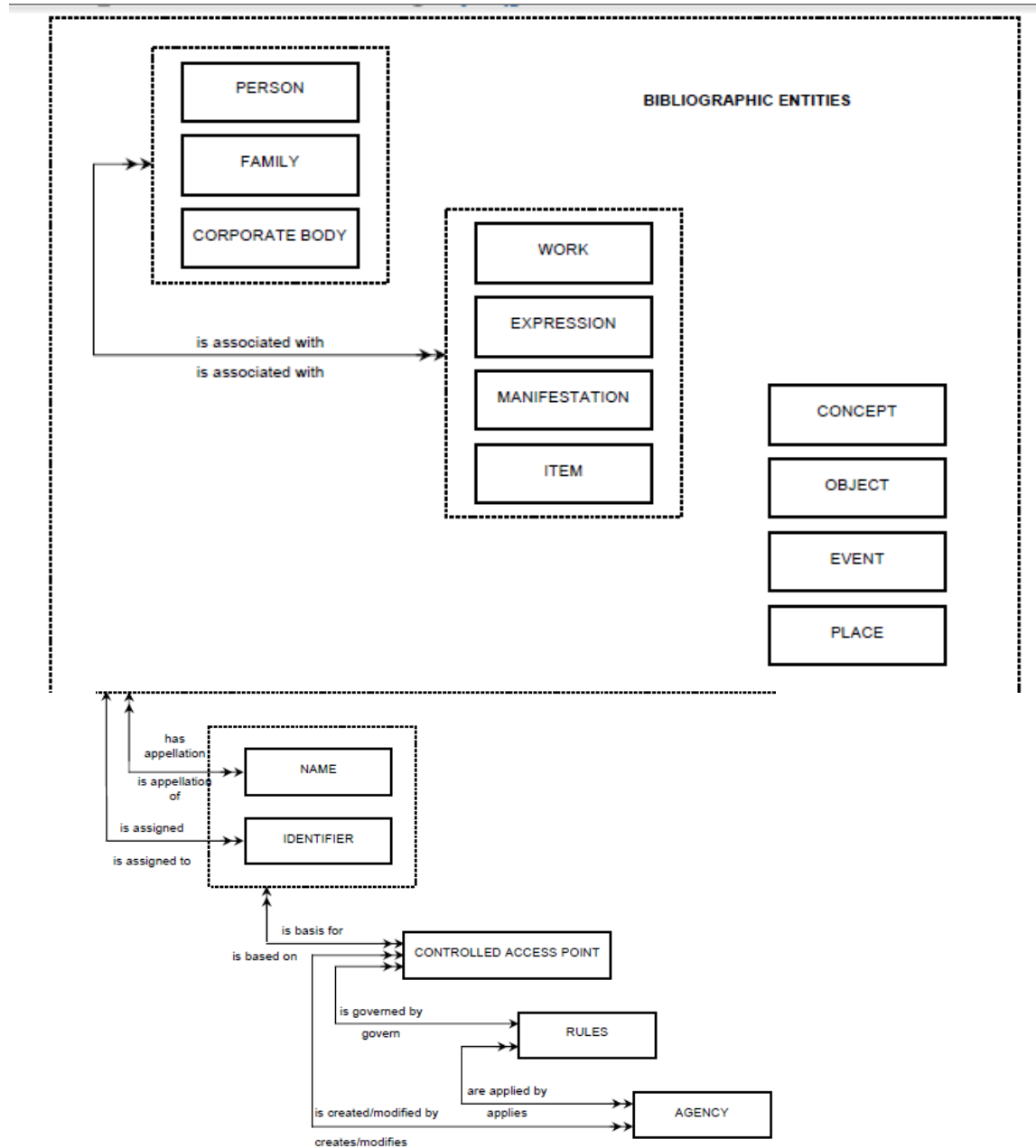
- Originally designed for card catalogs
 - One card for every “authorized” access point
- Four types “dictionary” catalog access points
 - Title (uniform titles)
 - Author (name authority)
 - Subject (controlled vocabulary)
 - Series
- Other things can serve a similar purpose
 - Call number (shelf order)
 - “Keywords” (full-text search)



Functional Requirements for Authority Data (FRAD)

- Name
 - Canonical form for display to users
- Identifier
 - Canonical form for use by systems
- Controlled access points
 - Forms that can be used as a basis for access
- Rules
 - For creating access points
- Agency
 - Organization responsible for creating access points

Functional Requirements for Authority Data



FRBR Bibliographic User Tasks

- Find it
 - Search (“to find”)
 - Recognize (“to identify”)
 - Choose (“to select”)
- Serve it
 - Location (“to obtain”)

FRAD Authority Control User Tasks

- Searcher tasks
 - Find
 - Identify
- Authority control tasks
 - Contextualize
 - Justify



LIBRARY OF CONGRESS AUTHORITIES

[Help](#) [New Search](#)[Search History](#)[Headings List](#)[Start Over](#)

SOURCE OF HEADINGS: Library of Congress Online Catalog

INFORMATION FOR: Oard, Doug

Please note: Broader Terms are not currently available

Select a Link Below to Continue...

[Authority Record](#)

See: [Oard, Douglas W.](#)

LC control no.: no 97043761

LCCN permalink: <http://lccn.loc.gov/no97043761>

Personal name heading: Oard, Douglas W.

Variant(s): Oard, Doug

Found in: A survey of information retrieval and filtering methods, 1995: t.p. (Douglas W. Oard) p. 1 (Electrical Engineering Dept., University of Maryland, College Park, MD)

Cross-language text & speech retrieval, c1997: t.p. (Doug Oard)

Hands On

- Find the authoritative LC name for one of ...
 - <http://ischool.umd.edu/faculty-staff/jennifer-j-preece>
 - <http://www.umiacs.umd.edu/~jimmylin/>
 - <http://terpconnect.umd.edu/~pwang/>
 - http://en.wikipedia.org/wiki/Robert_S._Taylor
 - http://en.wikipedia.org/wiki/Hans_Peter_Luhn

Classification

- Classification
 - A system for organizing knowledge
- Notation
 - Expressing the classification in a systematic way

Library of Congress Subject Headings

- Controlled vocabulary for subject access points
 - Most commonly applied to books and serials
- Used when a subject describes $\geq 20\%$ of the work
- Choose the most specific appropriate headings
 - But if more than 3 subtopics, choose a broader heading

LCSH Subdivisions

- Topical

 - Archaeology – Methodology

- Form

 - Archaeology – Fiction

- Chronological

 - Archaeology – History – 18th century

- Geographic

 - Archaeology – Egypt

Hands On

- Find the LCSH for one of:
 - <http://www.mayoclinic.com/health/heart-attack/DS00094>
 - <http://en.wikipedia.org/wiki/AS-204>
 - <http://www.apollotheater.org/>
 - <http://www.flickr.com/photos/usnationalarchives/4153755504/>
 - http://en.wikipedia.org/wiki/Operation_Entebbe

Before You Go!

- On a sheet of paper (no names), answer the following question:

What was the muddiest point in today's class?