REIMAGINING MOVING IMAGE ARCHIVAL CATALOGING USING THE SEMANTIC WEB

Presented to Reimagining the Archive Seminar, November 12-14, 2010

BY

Martha M. Yee
Cataloging Supervisor
UCLA Film & Television Archive
myee@ucla.edu
http://myee.bol.ucla.edu

INTRODUCTION

1. How we demonstrate and share relationship information now

2. Current situation: duplicated efforts in silos

INTRODUCTION

3. Semantic web idea

- URIs for entities
- Data changes could be more efficient
- Multiple languages and scripts

INTRODUCTION

4. Yee Cataloging Rules

5. Hierarchy problems with current model

INTRODUCTION

6. RDF (semantic web) resists hierarchy?

7. RDF too binary for our data?

INTRODUCTION

8. RDF challenge

9. Read more about it

1. HOW WE DEMONSTRATE AND SHARE RELATIONSHIP INFORMATION NOW

Example of a relationship

All the films by the same filmmaker (Charlie Chaplin)

1. HOW WE DEMONSTRATE AND SHARE RELATIONSHIP INFORMATION NOW

Use of character string matching:

Chaplin, Charlie, 1889-1977

1. HOW WE DEMONSTRATE AND SHARE RELATIONSHIP INFORMATION NOW

When the Library of Congress catalogs a Chaplin film, it adds the heading:

Chaplin, Charlie, 1889-1977

1. HOW WE DEMONSTRATE AND SHARE RELATIONSHIP INFORMATION NOW

When UCLA Film & Television Archive catalogs a Chaplin film, it adds the heading:

Chaplin, Charlie, 1889-1977

1. HOW WE DEMONSTRATE AND SHARE RELATIONSHIP INFORMATION NOW

OCLC is a huge database of holdings from libraries and archives all over the world

1. HOW WE DEMONSTRATE AND SHARE RELATIONSHIP INFORMATION NOW

When the Chaplin films from UCLA and LC enter OCLC, the heading

Chaplin, Charlie, 1889-1977 tells the OCLC computer that these are all films by the same filmmaker

1. HOW WE DEMONSTRATE AND SHARE RELATIONSHIP INFORMATION NOW

If UCLA makes a typo:

Chaplin, Charlie, 1899-1977

not

Chaplin, Charlie, 1889-1977

the OCLC computer interprets that to mean that there are two different filmmakers named Charlie Chaplin

1. HOW WE DEMONSTRATE AND SHARE RELATIONSHIP INFORMATION NOW

If we catalog a new version or expression of a work, such as a DVD version with director commentary, we repeat all of the work information (cast and production credits, summary, genre and form, subject headings) in each record for a new version of that work

1. HOW WE DEMONSTRATE AND SHARE RELATIONSHIP INFORMATION NOW

For each new episode of a television series such as The Simpsons, we repeat all of the information about the series as a whole, such as series regulars, production company, network, etc.

2. DUPLICATED EFFORTS IN SILOS

The same films are described differently in:

UCLA/LOC catalogs

AFI catalog

Internet Movie Database

2. DUPLICATED EFFORTS IN SILOS

- If you Google *Gone With the Wind*, you will get the IMDb record, but not the UCLA/LOC catalog records and AFI catalog record, all of which must be searched inside databases located on the web; that's why UCLA/LOC/AFI are called "the deep web," or "a silo"

2. DUPLICATED EFFORTS IN SILOS

We at UCLA cannot cause data from IMDb or AFI records to display in our UCLA catalog, and vice versa; this is the “silo effect.” Instead we have to copy and paste information into our records

3. SEMANTIC WEB IDEA

The Web could be a shared database instead of a shared document store, using URIs instead of database records or documents with URLs

3. SEMANTIC WEB IDEA

URI stands for Uniform Resource Identifier

3. SEMANTIC WEB IDEA

Example of the URI for the Library of Congress subject heading *World Wide Web*:

http://id.loc.gov/authorities/sh95000541#concept

3. SEMANTIC WEB IDEA

A URI such as this could be a kind of hot link for pulling together everything about the *World Wide Web*:

http://id.loc.gov/authorities/sh95000541#concept

3. SEMANTIC WEB IDEA: URIs FOR ENTITIES

On the open Web, there could be a URI for a particular film **work** linked to all attributes of that film such as director, year of original release, summary, genre/form terms, etc.

3. SEMANTIC WEB IDEA : URIs FOR ENTITIES

- On the open Web there could also be a URI for a particular film version or expression linking only to data pertaining to that expression (e.g. a title change or director’s commentary)
- All work attributes such as director, year of original release, summary, genre/form terms, etc., could be linked to at the work URI
- There would be no need to repeat the work data on every expression

3. SEMANTIC WEB IDEA : URIs FOR ENTITIES

This could reduce duplicate effort on the part of catalogers

This could ensure that searchers got EVERYTHING concerning a particular film:

a. all different versions
b. all works about it
c. all works related to it (remakes and adaptations and the like)

3. SEMANTIC WEB IDEA : URIs FOR ENTITIES

On the open Web there could be a URI for a particular television series as a whole, linking to all series attributes that are true of every episode of that series, such as regular cast members, setting for the series as a whole (Portland, Oregon), genre/form terms for series as a whole (situation comedy), etc.

3. SEMANTIC WEB IDEA : URIs FOR ENTITIES

- On the open Web there could be a URI for each television episode linking only to data pertaining to that particular episode (such as guest stars
- All series attributes such as regular cast members, genre/form terms for the series as a whole, etc. could be linked to at the series URI.
- No need to repeat the series data on every episode

3. SEMANTIC WEB IDEA: CHANGES ARE MORE EFFICIENT

Change one URI not thousands of records

3. SEMANTIC WEB IDEA: CHANGES ARE MORE EFFICIENT

For example, if any data about a particular entity (film, filmmaker, etc.) needed to be changed, e.g. the addition of a death date to the heading for a particular actress, it would be changed once at the URI and immediately accessible to all users, archives and archive staff by means of links down to local data such as film traffic or shelf location data

3. SEMANTIC WEB IDEA: MULTIPLE LANGUAGES AND SCRIPTS

Currently a person who speaks Chinese must search our catalog in English

3. SEMANTIC WEB IDEA: MULTIPLE LANGUAGES AND SCRIPTS

The semantic web URI that stands for a particular entity is not language based

It could potentially be linked to its entity's "name" in every possible language and script

3. SEMANTIC WEB IDEA: MULTIPLE LANGUAGES AND SCRIPTS

This could enable systems that would allow the user to define a preferred language and script for searching, e.g. a Chinese person could search our catalog using Chinese characters and display records in Chinese characters

4. YEE CATALOGING RULES

You can find my cataloging rules and my data model, (including my RDF schema, and some RDF examples) at:

http://myee.bol.ucla.edu

4. YEE CATALOGING RULES

RDF is a family of languages used to develop the semantic web

Roughly equivalent to HTML and XML as the languages of the current Web (although RDF can be expressed in XML)

4. YEE CATALOGING RULES

- The most important part of the rules consists of examples of displays that show the value of hierarchical displays of our data
- See Section 15, p. 106-135 in the cataloging rules

4. YEE CATALOGING RULES

The second most important part of the rules are the rules themselves that target the bits of data necessary to build those displays (reverse engineering from the displays to the rules)

4. YEE CATALOGING RULES

The least finished part of the project is my amateurish attempt at an RDF model (probably full of errors)

…but it is valuable in demonstrating the problems I had getting RDF to do what I wanted it to do (reverse engineering)

5. HIERARCHY PROBLEMS WITH CURRENT MODEL

- I believe that the relational databases we are using now were always too binary for our data and that is why our current ILS systems perform so badly at indexing and displaying our data.
- Rows and columns in tables may be fine for business inventory work, but they don't seem to work for the complex and hierarchical relationships we need to demonstrate in our catalogs.

5. HIERARCHY PROBLEMS WITH CURRENT MODEL

Currently when you search for a film, you are given a jumble of records for that film, records for films and programs about that film, and records for films and programs related to that film (e.g. remakes, sequels and the like).

These records are in no particular order.

5. CURRENT SYSTEMS LACK HIERARCHICALLY STRUCTURED DISPLAYS

EXAMPLE OF WHAT WE WOULD LIKE:

Separate lists of:

All the versions of a film,

All the films related to that film (adaptations, remakes),

All the films about that film (e.g. a making of documentary).

5. CURRENT SYSTEMS LACK HIERARCHICALLY STRUCTURED DISPLAYS

EXAMPLE OF WHAT WE WOULD LIKE:

Lawrence of Arabia

- Versions
- Works about

5. CURRENT SYSTEMS LACK HIERARCHICALLY STRUCTURED DISPLAYS

EXAMPLE OF WHAT WE WOULD LIKE:

All the films by a particular filmmaker, with a break-down (at the user's request) by function performed.

5. CURRENT SYSTEMS LACK HIERARCHICALLY STRUCTURED DISPLAYS

EXAMPLE OF WHAT WE WOULD LIKE:

Polanski, Roman.

direction [i.e. director]

cast

5. CURRENT SYSTEMS LACK HIERARCHICALLY STRUCTURED DISPLAYS

EXAMPLE OF WHAT WE WOULD LIKE:

Polanski, Roman. cast

- Back in the U.S.S.R.
- Chinatown
- Quiet chaos
- The tenant

5. CURRENT SYSTEMS LACK HIERARCHICALLY STRUCTURED DISPLAYS

EXAMPLE OF WHAT WE WOULD LIKE:

All the television series set in a particular locale, and then, separately, all of the episodes of a selected television series.

5. CURRENT SYSTEMS LACK HIERARCHICALLY STRUCTURED DISPLAYS

EXAMPLE OF WHAT WE WOULD LIKE:

Manhattan (New York, N.Y.)

- Adventures of Ellery Queen
- Amos ‘n’ Andy
- Ann Sothern show
- Barney Miller
- Beauty and the beast

5. CURRENT SYSTEMS LACK HIERARCHICALLY STRUCTURED DISPLAYS

EXAMPLE OF WHAT WE WOULD LIKE:

Barney Miller

- Abduction
- Agent orange
- Altercation
- Ambush
- Appendicitis

5. CURRENT SYSTEMS LACK HIERARCHICALLY STRUCTURED DISPLAYS

EXAMPLE OF WHAT WE WOULD LIKE:

Separate listings of:

All the films on a subject

All the films on broader subjects

All the films on narrower subjects

All the films on related subjects.

5. CURRENT SYSTEMS LACK HIERARCHICALLY STRUCTURED DISPLAYS

EXAMPLE OF WHAT WE WOULD LIKE:

Women in television broadcasting

BROADER:

Television broadcasting

NARROWER:

Women television personalities

Women television producers and directors

6. RDF RESISTS HIERARCHY?

RDF seems to resist hierarchy even more than the data models underlying our current ILS systems. Every link is a one-to-one link rather than a one-to-many link.

Hierarchy is an essential tool for allowing users to navigate efficiently through hundreds of thousands, even millions, of records.

7. RDF TOO BINARY FOR OUR DATA?

In relational database systems, every piece of data must be designated as an entity or an attribute.

Similarly, in RDF, the language of the semantic web, every piece of data must be designated as a class or a property.

7. RDF TOO BINARY FOR OUR DATA?

In my attempt to model our data using RDF, I frequently felt I needed to create a property of a property yet this is discouraged in RDF modelling.

Instead, a single entity, such as a person, begins to be represented by more and more classes; current RDF models are up to 6-7 classes just to represent the entity person

7. RDF TOO BINARY FOR OUR DATA?

I would like to generalize from this observation to suggest that RDF, like our current data model for relational databases, may be too binary for our data.

8. RDF CHALLENGE

As the scientific method teaches us, I cannot prove that it is impossible to build a catalog using RDF.

All that can be proved is that it IS possible, and that can be proved by doing it.

8. RDF CHALLENGE

Therefore, I would like to challenge those who think that semantic web technology IS the way forward for us to build a demonstration system and then show us:

8. RDF CHALLENGE: SHOW US

how we can search for a known film using a variant of the director's name and a variant of the title

EXAMPLE: Charles Chaplin Modern Columbus

(variant title for *The Immigrant*)

8. RDF CHALLENGE: SHOW US

how we can search for a list of television series set in a particular locale and see the episodes of the desired series only after it is selected

EXAMPLE: Situation comedies set in Manhattan

8. RDF CHALLENGE: SHOW US

how multiple works are displayed when the user does a search, such as a subject search, that retrieves thousands of works

EXAMPLE: subject search on the Vietnam War

9. READ MORE ABOUT IT

Yee, Martha M. "Can Bibliographic Data be Put Directly Onto the Semantic Web?" *Information Technology and Libraries* 28:2 (June, 2009): 55-80. Also available on the Web at:

http://repositories.cdlib.org/postprints/3369