ALA ALCTS Cataloging and Classification Research Interest Group Meeting

Expanding the Metadata Librarian Horizon: Reflections on the Metadata Practices in the Web and Digital Repositories

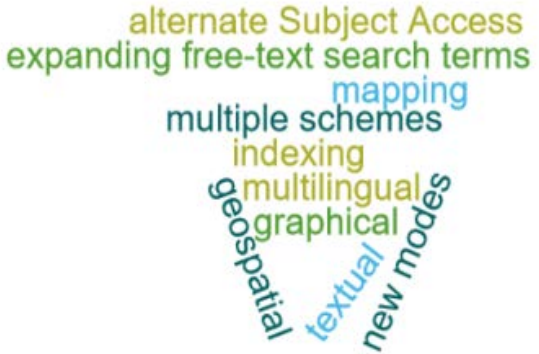
Sai Deng, University of Central Florida Libraries

Digital Repository, the Web and KOSs

- *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files,* by Gail Hodge. Copyright 2000 by the Council on Library and Information Resources.
- **Types of Knowledge Organization Systems (KOSs):**

Term Lists	Classifications and Categories	Relationship Lists
Authority files, Glossaries, Dictionaries, Gazetteers	Subject headings, Classification schemes, Taxonomies, Categorization schemes	Thesauri, Ontologies, Semantic networks

- **Linking Digital Library Resources to Related Resources**
- **Making Resources Accessible to Other Communities** (see graph)
- **Implementing Knowledge Organization Systems:**
  - **Making the Link (DL-KOS): metadata; text analysis**



Environmental Scanning of Authority Control in Digital Repositories

- Problems in name authority control in institutional repositories. (Salo, Dorothea. 2009)
- FAO's Survey Open Access Repositories in the Agricultural Domain (2009-2010): Only 40% is using some sort of authority control, especially for Journal titles.
- Survey of ARL-member institutions: Despite the system challenge, half use LCNAF in some capacity (Goslen, Anna S. 2011)
- Western Name Authority File Project: Authority Control for Digital Collections Survey, 2016.

Name Disambiguation:
Authority Control & Identity Management

- **Name Disambiguation**
  - Manual disambiguation (e.g., LCNAF);
  - Author registry site;
  - Automatic name disambiguation.

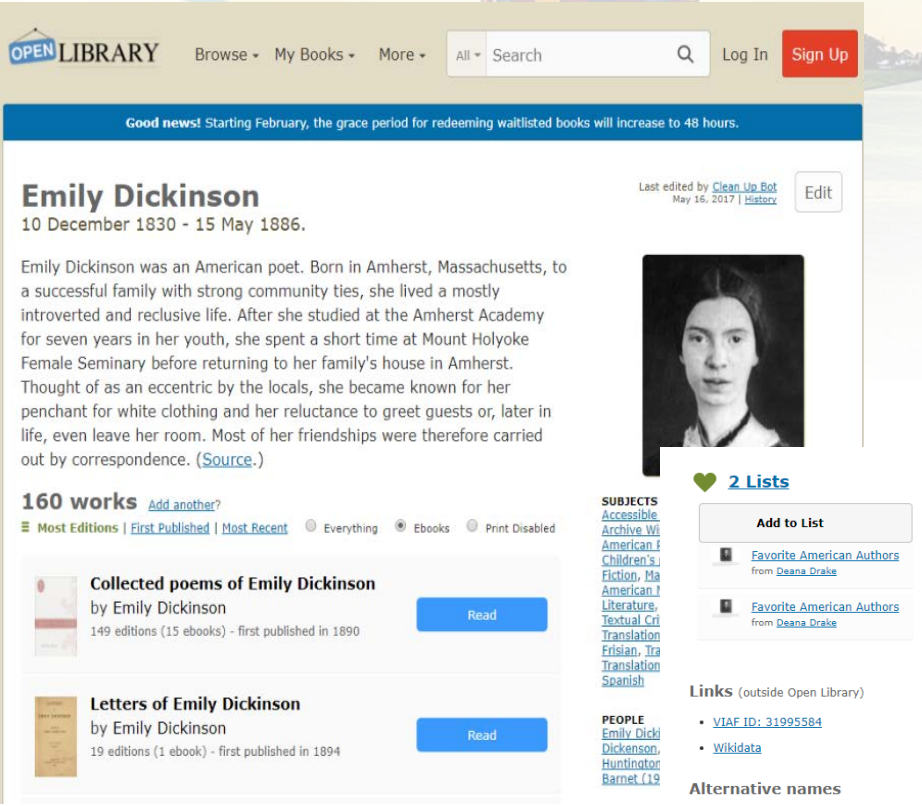
  (Elliott, Sarah. Survey of Author Name Disambiguation: 2004-2010)

- **Identities**: Scopus author identifier and profile, VIAF, ISNI, ResearcherID, ORCID, Worldcat identities…

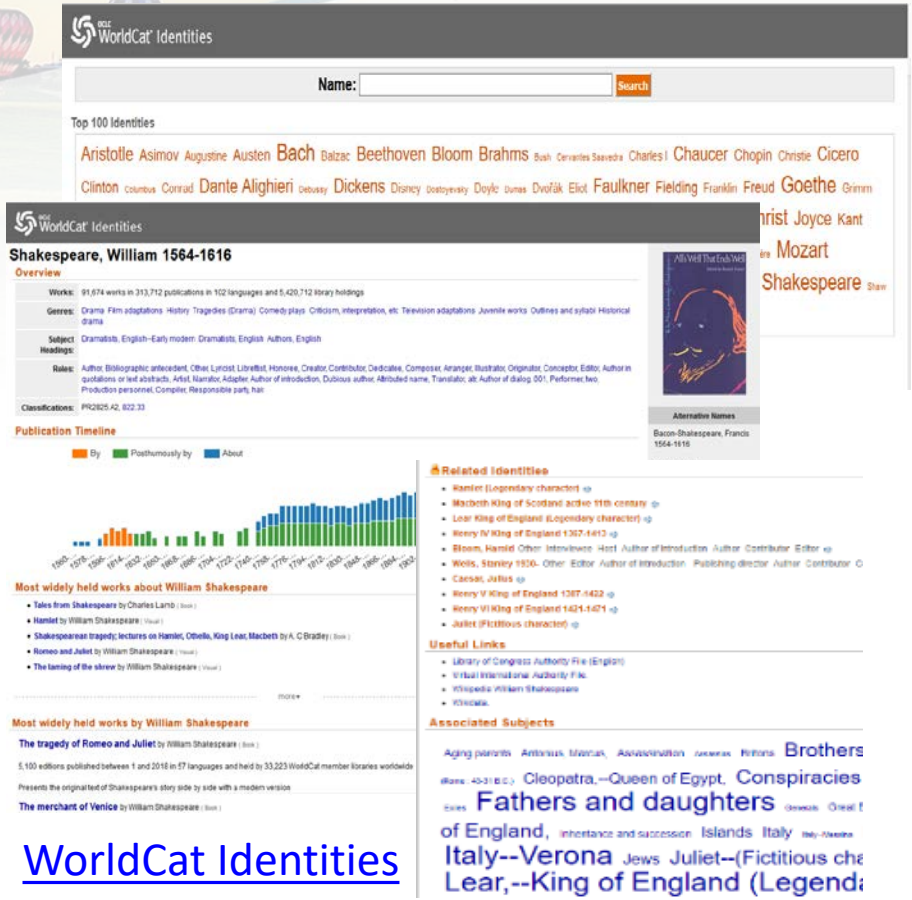
- **Shifting from Authority Control to Identity Management**

  (Identity Management or Authority Control? Liss, 2017 ALA Annual)

Engaging User Interfaces Using Author Information



[Open Libraries Author Page]



[WorldCat Identities]

Subjects & Keywords:
Debate on the Use of Controlled Vocabularies

- **Information retrieval research** shows that the use of CVs does not improve precision and recall;
- **Indexing researchers** show that CV usage can improve indexing consistency.
- **This study found that: LCSH works best for topical terms; Bio thesaurus ITIS works better for scientific terms.**
  (White, Hollie. Examining scientific vocabulary: Mapping controlled vocabularies with free text keywords, 2013.)

- **User search query compared with FAST terms and legacy keywords in the IR: FAST terms match better.**
  (Hanrath & Radio. 2016)

- **Researchers assign subject/keyword metadata to their own materials in IRs; Tagging**

- **Advantage of controlled vocabularies lies in linked data?!**

- **The web's impact to library practices and digital repositories: BIBFRAME! What about digital repositories**

What about the Web: Metadata & Text Analysis

- **Web archiving**: **Capturing E-Publications of Public Documents (CEP); SWISH-E based IGI search engine**
  - **Illinois Subject Tree, Metadata Generator available**; worked as RA;
  - **Author generated metadata is rare, supplement with**:
    - Inclusion of plain-text; extraction of noun phrases; knowledge of the ownership of the website; collection level metadata; classifier program for assigning subject classifications (Jackson, 2005)
  - **Automatic subject heading assignment** for online government publications using a semi-supervised machine learning approach. (Hu, X., Jackson, L., Deng, S. and Zhang, J., 2006.)

- **"Making the link":** metadata; machine learning, natural language processing…
- **Learn about web technology, web standards:** HTTP, RDF, URIs…
- **Search engine optimization and digital repositories**

What about the Web: IR Working w/ Search Engines

- **Google's ranking algorithm:** PageRank; RankBrain

- **Google Scholar:** "Rank documents the way researchers do"
  - **Author page**: created for authors
    - Name, position, email, research interest, **publications** (title, author, journal name, volume, issue, year), **cited by** (Citations, h-index, i10-index), Co-authors
  - Author has no unique alphanumeric identifiers; use **statistical model** to try to tell authors apart; author mediation

- **IR Content Discovery:** Provide metadata structured according to **Google Scholar guidelines** can greatly improve the indexing ratio of IR content; transform Dublin Core to **High Wire Press tags** and include as meta tags in html (Arlitsch & O'Brien)

What about the Web: Get Metadata to be Understood by Search Engines

- **Schema.org**: a set of extensible schemas that enable webmasters to embed **structured data** to web pages for use by search engines and other applications.
  - Created by search engines (Google, Bing, Yahoo)
  - **Schema.org vocabulary** (entities, actions, relationships)
  - **Extensions**
    - **Bibliographic Extension** (http://bib.schema.org/)
    - **Extend Schema.org to represent archives** (https://archival.github.io/schema-org/)

- **Schema.org, Libraries & Digital Repositories**: Get metadata to be understood by search engines
  - WorldCat, Europeana, Islandora…

Changing Digital Repositories

Islandora

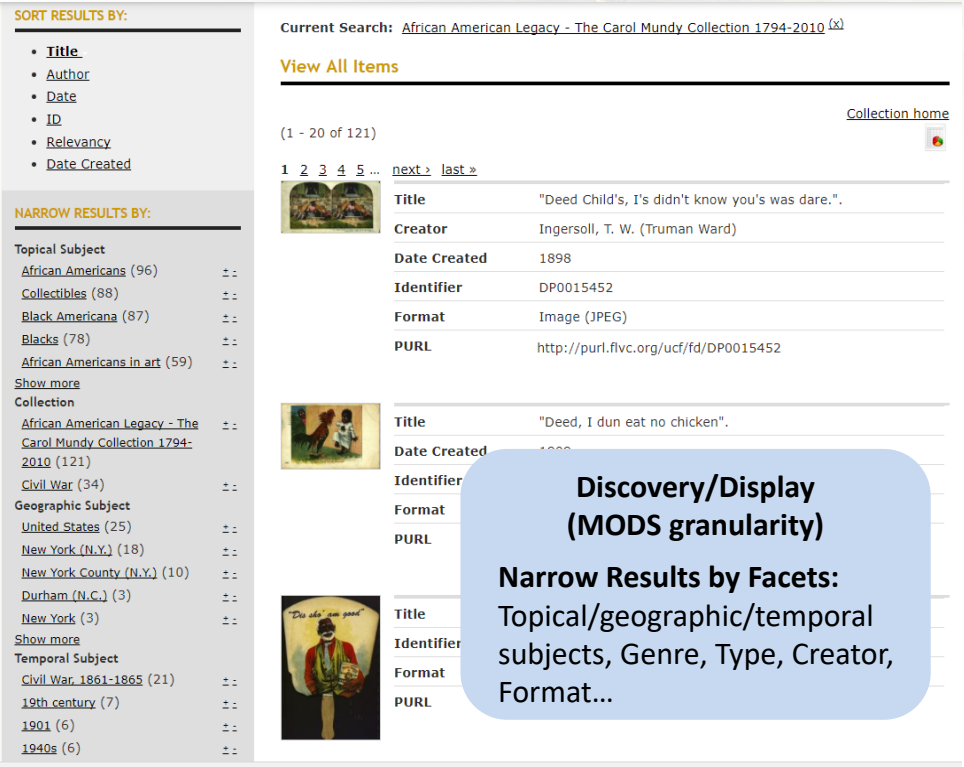
- Open source. Supports Dublin Core, MODSXML, MADS, METS, PREMIS…
- Authority Control in Islandora 7.x: **Entities Solution Pack**
  - **Entities:** Person, place, event, organization;
  - Entities created in **MADS** or **EAC-CPF**;
  - Build departmental directories and faculty profiles;
  - **Islandora FLVC:** Some institutions populated author and department names (w/ basic authority control)
- **Integrated doi** (batch ingest module), **ORCID**.
- **Subject:** authority can be defined, e.g.,<Subject authority="LCSH"…>
  - FSU DigiNole case: local script to reconcile and add subjects to MODS records (Miguez, 2017)
- **Google Scholar tags** added to citation object's page; **schema.org tags** added to metadata fields

CONTENTdm

- Proprietary. Supports DC, XML.
- **Support multiple controlled name& subject vocabularies and thesauri**:
  - AAT, TGN, TGM, ULAN…
  - LCSH, FAST not included;
  - CVs can only be chosen when first establishing a field; local terms can be added, but not separated from controlled lists; validation enforced

Changing Digital Repositories

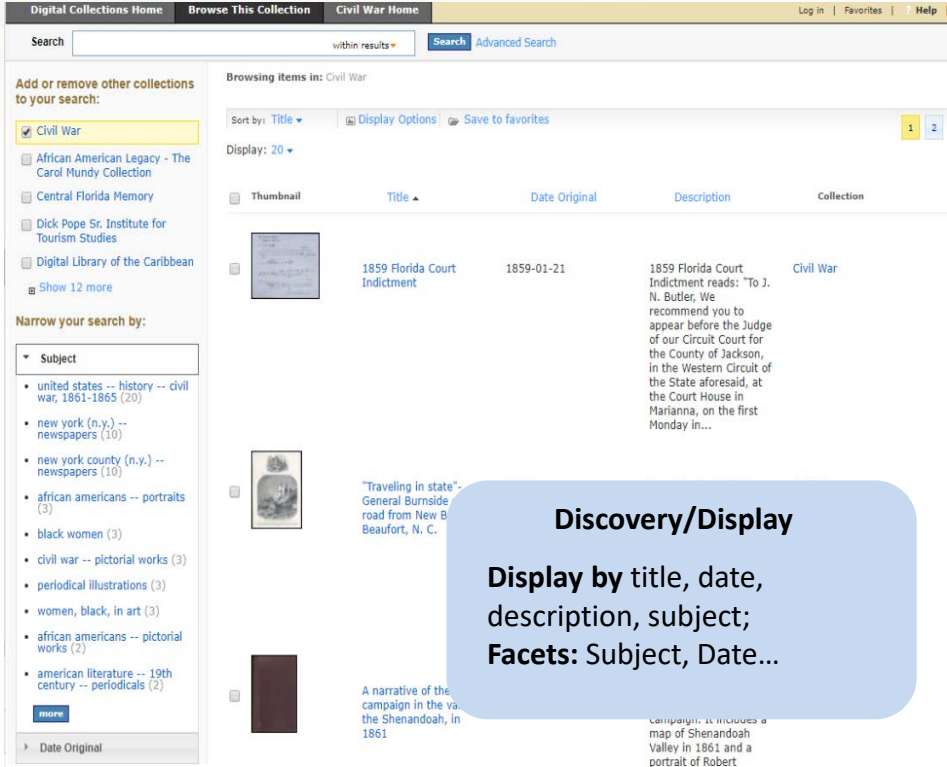
Islandora
Collection Example



Discovery/Display (MODS granularity)

Narrow Results by Facets: Topical/geographic/temporal subjects, Genre, Type, Creator, Format…

CONTENTdm
Collection Example



Discovery/Display

Display by title, date, description, subject;
Facets: Subject, Date…

Changing Digital Repositories

Islandora

- **Linked Data:** Effort in mapping MODS to RDF. Fedora 4, LDP implementation; Islandora CLAW
- **Cases:**
  - TU Delft repository example: Colonical Architecture (Delftdora module, Fedora 3)
  - A Lightweight Structured Data Implementation Using JSON-LD and Schema.org for Digital Repository (Mak et al.)
- **My own experience:**
  - Data mapping, transformation, Notepad++ text editor. Metadata migration to Islandora: Is there an easy way? 2014.
  - Standards implementation. Applying RDA to CONTENTdm & Islandora. 2015.

CONTENTdm

- **Linked data experiments:**
  - Not inside the system, but work on exported data.
  - **UIUC.** Challenges of Mapping Digital Collections Metadata to Schema.org (Lampron et. al. 2016.)
  - **OCLC& CONTENTdm**. Metadata Refinery, transforms CONTENTdm metadata to Linked Data, 2016. Not yet scalable. (Washburn et al. 2016.)
- **My own experience:**
  - Used a mixture of terms from id.loc.gov (text only) and local.
  - Reconcile certain fields (e.g., advisor) in exported datasheet against LC authorities in OpenRefine.

Changing Digital Repositories

DSpace

- Open source. Supports Qualified DC, can export to METS, MODS…
- DSpace 6.x, XMLUI (Manakin), JSPUI.
- **Metadata Registry**: Default registry (DC, DCTERMS); Local Metadata Registry.

- **Authority control & Identity Management**
  - Can be enabled in configuration;
  - CVs in form fields; in separate XML files;
  - Authority key; ORCID integration.

- dc.subjects doesn't enforce control; supports some specific vocabularies (e.g., SRSC); has dc.subject.lcsh field

- Usage analysis via plug-in, SEO

Digital Commons

- Proprietary. Supports qualified DC.
- **Authority Control & Identity Management**
  - **Author:** author names entered as they appear in email or source; include name, email and institution
  - **Author merge tool**
  - **DOI and ORCID integration**

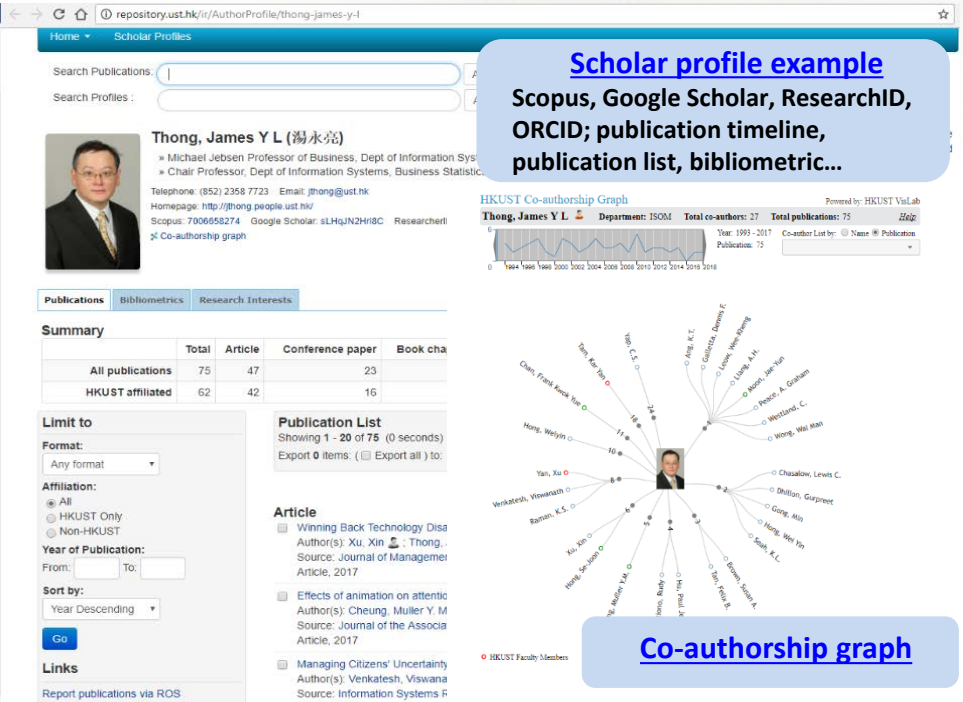
- **Subject:** built-in Discipline thesaurus; only has Keywords field, can request to add subject field, but not included in browsing/ facets

- **Some features:** Impact & Analytics map, improved SEO

Changing Digital Repositories

DSpace
Author Page Example

- **HKUST IR** (Powered by VuFind and DSpace)

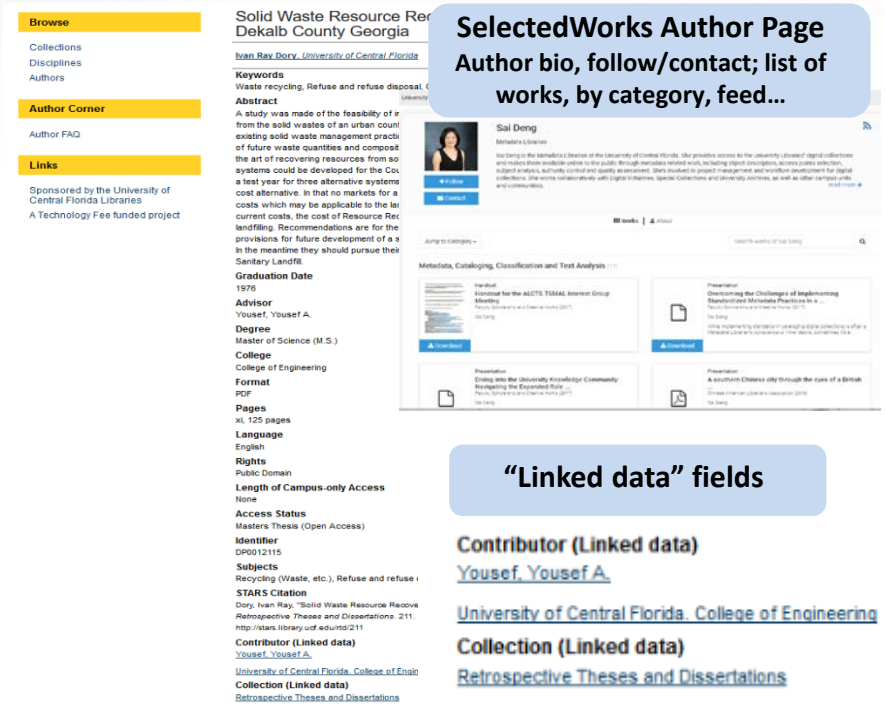


Scholar profile example
Scopus, Google Scholar, ResearchID, ORCID; publication timeline, publication list, bibliometric…

Co-authorship graph

Digital Commons
Author Page & Record Examples

- **UCF Author/contributor Linked Data**



SelectedWorks Author Page
Author bio, follow/contact; list of works, by category, feed…

“Linked data” fields

Changing Digital Repositories

DSpace

- **DSpace Linked Data support spans all three Layers:** **"**the storage layer with **a triple store**, the business logic with classes to **convert** stored contents into RDF, and the application layer with a module to **publish** RDF serializations." (Donohue, DSpace documentation)

- **Discovery/Display:**
  - **Browse by** communities, collections, date, authors, titles, **subjects**, type;
  - Author, subject, date **facets**.

- **My past experience at WSU** (in interface, data repurposing, transformation and transfer):
  - Building Collections in IRs from External Data Sources
  - Enhancing workflow through batch import from Excel to DSpace; Customized mapping and metadata transfer from DSpace/SOAR to OCLC and Voyager
  - Building New Interfaces for Shocker Open Access Repository

Digital Commons

- **Linked data**: Allow to add extra fields for links (not RDF)

- **Discovery/Display:**
  - **Browse by** collections, **disciplines**, authors;
  - Discipline, keyword, year, type **facets**;
  - **Exhibit-like presentation;**
  - Included in **Bepress' networks** (e.g., Digital Humanities Commons).

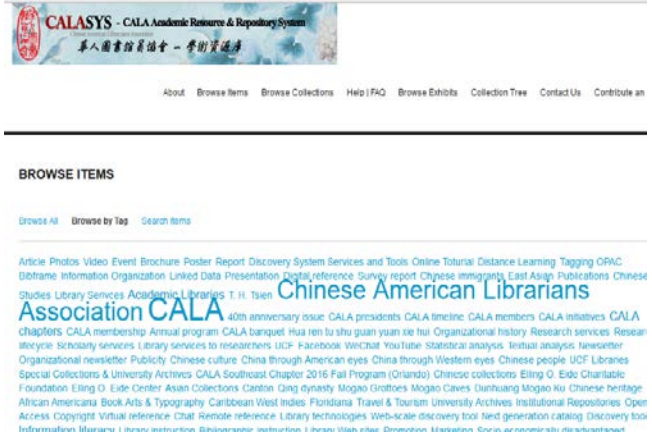
- **Our experience:**
  - Include links (id.loc.gov, viaf.org) in added additional fields for author, advisor, department names etc.;
    - Preparing for linked data in digital repositories.
  - Data (in spreadsheet) cleanup and reconciliation using OpenRefine;
  - Verifying author names (challenging): Google, university and departmental websites, publications, CV…

Changing Digital Repositories

Omeka

- Open source web publishing platform. Supports Dublin Core. Can export METS, dc-rdf, json etc.
- **Subject** field not enforced; has additional tag feature.
- **Browse/Discovery:**
  - Terms can be accessed via clicking tags in the record;
  - Browse by tag (word cloud);
  - Browse items/collections/exhibits;
  - Collection tree

- **Plug-ins:**
  - **Import from Dropbox, YouTube; CSV batch import**
  - **LC Suggest (plug-in)**
  - **User Profiles (plug-in)**
  - **Ozmeka (plug-ins)/ Omeka linked data**
- My own experience: CALASYS (as a CALASYS Group member)



Browse by Tag

Various Approaches in Authority Control/Identity Management/Discovery etc.

- **System approach:** Develop new functionality; develop extensions and plug-ins; collaboration
- **Institutional approach:**
  - e.g., Name Authority Control: An institutional repository approach (Waugh, 2013)
- **Regional approach,** e.g., Western Name Authority File Project
- **National and International approach,** e.g., WorldCat Identities, VIAF, ISNI…
- **Individual reflection:**
  - Being a cataloger means to be not only a guardian of principles and standards but also a knowledge organizer, collaborator, researcher and an innovative and curious-minded lifelong learner (Diao & Hernandez. 2014)

Some Discussions

- **A *Very Basic Understanding of the Big Environment (Broaden your horizon)**
  - **How to organize knowledge and information? How does search work in various platforms and the web?**
  - **Knowledge Organization Systems (KOSs); Every DL uses one or more KOS.**
    - **Primary applications:** Indexing, browsing, searching DLs
    - **Other applications:** Automatic annotation/indexing of contents/documents, automatic classification of contents/documents, developing Concept Space and ontologies for enhanced browsing/navigation, interoperability among DLs, linked data creation, publication venue authority control, visualization for browsing…
    - **Additional applications:** Automatic query expansion, Search Term Recommendation (STR) System

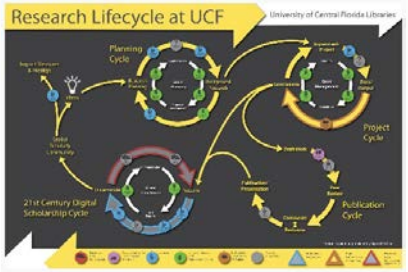
    (Sunny, TISS. Application of Traditional Knowledge Organization Systems in Digital Libraries: A Study on Current Status. Tata Institute of Social Sciences)

- **Understanding of Metadata and the Profession**
  - The Evolving Role of the Metadata Librarian: Competencies Found in Job Descriptions (Han& Hswe, 2010)
  - Metadata standards, schemas, vocabularies… “What then?”
  - While users and authors can contribute metadata, do Metadata Librarians have values to add? What are those values?

In addition…

- **Metadata Services is defined as:** Providing metadata consultation and assistance to university faculty and students in their research lifecycle.

- **The Research Lifecycle at UCF**
  - Sub-cycles: the Planning cycle, the Project cycle, the Publication cycle, the 21st century digital scholarship cycle
  - Dataset Metadata and Metadata Services



- **Challenges and Opportunities for Metadata Services**
  - Connecting the dots: Defining scholarly services in a research lifecycle model, 2013.
  - Diving into the University Knowledge Community, 2017.

Final Thoughts

- **Open mindset**
- **Learn new knowledge, information and skills**
  - Possible areas: vocabulary management, document encoding, data processing;
  - Learn some basic skills in working with data: XML, XSLT, RDF…
  - Learn to use tools, e.g., semi-automatic metadata editing tools: MarcEdit, Notepad++, OpenRefine…
  - Think out of the box: what can we learn from the web practices?

- **Be flexible and collaborative**
  - Flexible in working with different systems;
  - Collaborate with people beyond cataloging/metadata, Technical Services, your own library and the library field.

References and More Reading

- Archives and Linked Data Interest Group. (2017). Schema.org and Schema Architypes for linked archival description. Retrieved January 10, 2018 from https://archival.github.io/schema-org/
- Arlitsch, K. & O'Brien, P. (2013). Improving the visibility and use of digital repositories through SEO: A LITA guide. ALA Neal-Schuman. Retrieved January 10, 2018: https://dl.acm.org/citation.cfm?id=2509707
- Bepress. SelectedWorks guide for authors. Retrieved January 10, 2018: https://www.bepress.com/reference_guide_sw/selectedworks-guide-authors/
- Chinese American Librarians Association. CALASYS - CALA Academic Resources & Repository System. Retrieved January 10, 2018: http://ir.cala-web.org/
- CONTENTdm. Resources. Retrieved January 10, 2018: https://www.oclc.org/en/contentdm/resources.html
- Diao, J. L. & Hernandez, M. A. (2014). Transferring cataloging legacies into descriptive metadata creation in digital projects: Cataloger's perspective. *Journal of Library Metadata*, Volume 14, 2014 - Issue 2. Retrieved January 10, 2018: https://doi.org/10.1080/19386389.2014.909670
- Digital Commons: Three-Tiered Taxonomy of Academic Disciplines. Retrieved January 10, 2018: https://www.bepress.com/reference_guide_dc/disciplines/
- DSpace: Authority Control of Metadata Values. Retrieved January 10, 2018: https://wiki.duraspace.org/display/DSDOC6x/Authority+Control+of+Metadata+Values
- DSpace: Linked (Open) Data. Retrieved January 10, 2018: https://wiki.duraspace.org/display/DSDOC6x/Linked+%28Open%29+Data
- Elliott, S. (2010). Survey of author name disambiguation: 2004-2010. *Library Philosophy and Practice* 2010 November. Retrieved January 10, 2018: http://www.webpages.uidaho.edu/~mbolin/elliott.htm
- FAO. Survey open access repositories in the agricultural domain. Retrieved January 10, 2018: http://aims.fao.org/advice/open-access/survey
- Google Scholar. Retrieved January 10, 2018: https://scholar.google.com/
- Goslen, A. S. (2011). Name authority control in institutional repositories and its relationships to metadata quality. Retrieved January 10, 2018: https://cdr.lib.unc.edu/record/uuid:855b470a-dd7d-40a8-86fc-9dba25846297
- Han, M. J. & Hswe, P. (2010). The evolving role of the metadata librarian: Competencies found in job descriptions. *Library Resources & Technical Services*, Vol 54, No 3.
- Hanrath, S. & Radio, E. (2017). User search terms and controlled subject vocabularies in an institutional repository, *Library Hi Tech*, Vol. 35 Issue: 3, pp.360-367, https://doi.org/10.1108/LHT-11-2016-0133.
- Hodge, G. (2000). Systems of knowledge organization for digital libraries: Beyond traditional authority files. The Digital Library Federation. Retrieved January 10, 2018: https://www.clir.org/wp-content/uploads/sites/6/pub91.pdf.
- Hong Kong University of Science and Technology. Scholar Profiles: Thong, James Y L. Retrieved January 10, 2018: http://repository.ust.hk/ir/AuthorProfile/thong-james-y-l
- Hu, X., Jackson, L., Deng, S. and Zhang, J. (2006). Automatic subject heading assignment for online government publications using a semi-supervised machine learning approach. In *Proceedings of the American Society for Information Science and Technology*. Volume 42, Issue 1, 2006. Retrieved January 10, 2018: http://onlinelibrary.wiley.com/doi/10.1002/meet.14504201139/pdf
- International Standard Name Identifier (ISO 27729). Retrieved January 10, 2018: https://viaf.org/

References and More Reading

- Islandora CLAW. Retrieved January 10, 2018: https://islandora.ca/CLAW
- Islandora: Entities solution pack. DURASPACE. Retrieved January 10, 2018: https://wiki.duraspace.org/display/ISLANDORA/Entities+Solution+Pack
- Islandora FLVC. Retrieved January 10, 2018: https://islandora.pubwiki.fcla.edu/wiki/index.php/Islandora_FLVC
- Jackson, L. (2005). Difficulties in electronic publication archival processing for state governments. In: 1st International Conference on Universal Digital Library, ICUDL 2005, 175-185. Retrieved January 10, 2018: https://www.ideals.illinois.edu/handle/2142/16401
- Lampron, P., Mixter, J. & Han, M. J. (2016). Challenges of mapping digital collections metadata to Schema.org: working with CONTENTdm. Retrieved January 10, 2018: https://link.springer.com/chapter/10.1007/978-3-319-49157-8_15
- Liss, J. (2017). Identity management or authority control? The ALA ALCTS Cataloging Norms Interest Group Meeting, 2017 ALA Annual Conference.
- Mak, L., Lorenzo, L. & Smeltekop, N. (2017). A lightweight structured data implementation using JSON-LD and schema.org for digital repository. ALA ALCTS Cataloging Norms Interest Group Meeting, ALA Midwinter Conference, 2017. Retrieved January 10, 2018: http://connect.ala.org/files/2_ALAMidwinter2017_CNIG_Mak.pptx
- Miguez, M. (2017). Automating controlled subjects from IR keywords. Retrieved January 10, 2018: http://diginole.lib.fsu.edu/islandora/object/fsu:388899
- OMEKA Plugins. Retrieved January 10, 2018: https://omeka.org/classic/plugins/
- Open Libraries author page. Retrieved January 10, 2018: https://openlibrary.org/authors/OL19512A/Emily_Dickinson
- Salo, D. (2009). Name authority control in institutional repositories. *Cataloging and Classification Quarterly*, 47:3/4 (April 2009).
- Schema.org Vocabulary. Retrieved January 10, 2018: http://schema.org/docs/gs.html#schemaorg
- Sunny, S. K. (2017). Application of traditional knowledge organization systems in digital libraries: A study on current status. Tata Institute of Social Sciences. Retrieved January 10, 2018: http://rnd.tiss.edu/wp-content/uploads/2017/09/Abstract-of-Mr.-Sanjeev-Kumar-Sunny-for-Synopsis-Presentation.pdf
- TU Delft repository. Colonial architecture & town planning. Retrieved January 10, 2018: http://colonialarchitecture.eu/
- University of Central Florida Digital Library. African American Legacy - The Carol Mundy Collection 1794-2010. Retrieved January 10, 2018: https://ucf.digital.flvc.org/islandora/search/?type=edismax&collection=ucf%3Amundy
- University of Central Florida Digital Libraries. Overview: Research Lifecycle. Retrieved January 10, 2018: https://library.ucf.edu/about/departments/scholarly-communication/overview-research-lifecycle/
- University of Central Florida Digital Libraries. Retrospective Theses and Dissertations, 211. Retrieved January 10, 2018: http://stars.library.ucf.edu/rtd/211/
- VIAF: The Virtual International Authority File. Retrieved January 10, 2018: http://www.isni.org/
- Waugh, L., Tarver H., Phillips, M. D. (2014). Introducing name authority into an ETD collection, *Library Management*, Vol. 35 Issue: 4/5, pp.271-283, https://doi.org/10.1108/LM-08-2013-0074
- Western Name Authority File Project (2016). Authority control for digital collections survey. Retrieved January 10, 2018: https://sites.google.com/site/westernnameauthorityfile/survey
- White, H. (2013). Examining scientific vocabulary: Mapping controlled vocabularies with free text keywords. *Cataloging & Classification Quarterly*, Volume 51, 2013 - Issue 6.
- WorldCat identifiers. Retrieved January 10, 2018: www.worldcat.org/identities

Thank you!

Contact

Sai Deng, sai.deng@ucf.edu

Metadata Librarian and Associate Librarian

University of Central Florida Libraries