We Are Family: Negotiating the Needs of Multiple Units While Creating a User-Friendly Thesaurus

Laura Sullivan, Brad Kuennen, and Lindsey Hillgartner
So what is a thesaurus anyway?

[Image: Cartoon of a person being chased by a dinosaur with speech bubbles saying "GASP!", "INHALE, PANT, PUFF, RESPIRE....", and "Thesaurus."]

Walter Lim / Flickr
CC BY 2.0
So What Is a Thesaurus Anyway?

- Controlled vocabulary
- Required: relationships between terms (broader, narrower, related, preferred, etc.)
- Optional: explanatory material (scope notes, definitions, etc.)

from: https://hasset.ukdataservice.ac.uk
How Do Thesauri Differ from Subject Headings?

- **Similarities:**
  - Controlled vocabularies
  - Relationships between terms

- **Primary differences:**
  - Thesauri are often narrower in scope (National Agricultural Library's Agricultural Thesaurus, etc.)
  
  - Subject headings often combine several unique concepts in a single string (temporal, content, item type, etc.) while thesauri tend to more often use singular terms
Collaborating Across Library Units: Our Thesaurus Family

- Construction roles:
  - Special Collections & University Archives
  - Digital Initiatives
  - Metadata & Cataloging

- Consulting roles:
  - Digital Press & Repository
  - Research & Instruction
Thesaurus Project Goals

- Take advantage of archives & digital collections system migrations
- Collocate collections & objects with similar topical areas across units
- Create a vocabulary specific to collecting areas for library's rare and unique items and collections
- Focused, browsable vocabulary
- Eliminate need for archival collection subject guides
- Improve search precision and recall
- Eliminate frustration for staff & researchers!
Planning: Considerations

- Thesaurus versus more simple taxonomy
- Project management
- Relationship with Library of Congress
- Linked data compatibility
- Events
- Audience: staff-only versus online
Planning: Statement of Intent

Statement of Intent

Wednesday, August 30, 2017, 10:12 AM

Thesaurus Purpose
- The vocabulary should:
  - bring different types of resources together (e.g., "aboutness" of the archival and digital collections)
  - support indexing and retrieval of digital, archival, and manuscript collections at ISU
- The thesaurus is intended for general, collection-level description

AUDIENCE
- Internal audience
  - SCUA
  - Digital Initiatives
  - Metadata and Cataloging
  - Research and Instruction
- External audience
  - ISU community (undergrad, graduates, faculty, staff...)
  - Outside researchers (academic / non-academic)
  - Donors

INTENDED PLATFORMS FOR USE
- MARC records in: ALMA/PRIMO/WorldCat
- Finding aids in SKCA
- Artifact records in PastPerfect
- Digital collections in CONTENTdm
- Digital collections in Islandora

LEVEL OF SPECIFICITY AND DESCRIPTION
- The thesaurus will focus on terms for archival collections, and will later specifically look at more item-level description in digital collections once the thesaurus is in place.
- General goal to make the vocabulary useful, the terms need to be able to group a number of finding aids under a term, or move to another category in a number of finding aids.
PHASE 1 – deadline – beginning of October

- Analysis of the needs for a vocabulary
  - Define the audience
    - Internal audience
      - SCUA (archival collections – EAD, MARC, artifacts) - Platforms: SKCA, ALMA, PRIMO, CONTENTdm, Islandora
    - Digital Initiatives
  - External audience
    - ISU community (undergrad, graduates, faculty, staff...)
    - Outside researchers (academic / non-academic)
    - Donor – reinforce and add detail to the collecting to the collection scope

- Levels of description
  - Look at Digital Collections once the thesaurus is in place – start with collection level (focus on the FAs)

- Level of specificity
  - General – need to group a number of FAs under a term as opposed to just a couple FAs, so that the vocabulary is useful
  - This thesaurus should substitute the subject guides
Planning: What Went Missing

- Assigning individual roles & responsibilities: project manager, etc.
- Determining:
  - Smaller group assignments
  - All stakeholders from the beginning
- Establishing a timeline & schedule with milestones
First steps: Compiling Terms

Compiled from:

- SCUA collecting areas
- SCUA subject guides
- Digital Initiatives topical index

Reviewed:

- MARC subject headings for archival collections
- Digital Repository collections
- Data Repository categories
Organizing the Terms

Develop a basic hierarchy of terms

- Looked at major collecting areas (again)
- Explored Library of Congress classifications

Divide terms among several team members to review

- Several meetings to discuss and refine terms
Testing the Thesaurus

Transform the spreadsheet into SKOS
Testing the Thesaurus

- Visualizing the terms using SKOS Play
- Multiple views to browse and explore the hierarchy
Testing the Thesaurus

- Examine coverage of terms through the use of a survey (Google Forms)
- Looked at the organization of the terms
- Investigate overlap of terms or confusion in usage of terms
Findings from the initial survey

- Scope notes and definitions may be necessary
  - Ex.: aerospace engineering vs. aeronautics
- In some cases coverage was lacking
  - Ex. University-specific terms for administration and student life
- Analytics needed some clean-up
  - Ex. “Cooking” and “cooking” were counted as separate terms
Revise the Terms Again … and Again

More elimination and simplification:

* Events category eliminated
  * A few terms were retained and moved elsewhere
* Classes of People category eliminated
  * Not fully resolved
* Education category eliminated
* History and Local Interest category eliminated
  * Many were LC terms
Revise the terms again … and again

Identify gaps in the terms selected:

• Review archival record groups and subject guides
• Do we have enough collections to merit a new term?
The Results so Far

Settled on the following subject categories:

- Agriculture
- Arts & Culture
- Politics & Government
- Society & Human Interactions
- Science & Mathematics
- Engineering

Just under 200 terms and currently developing guidelines for adding new terms

University Library
What Comes Next?
Next Steps

1. Minting URIs for all terms compiled
2. Set up Server
3. Create SKOS model
   - Vocabularies built in Excel are turned into SKOS based RDF triples using Python’s RDFLib and the Protégé Ontology editor

What is an RDF triple?
- An RDF Triple is a statement which relates one object to another.
  - Subject----predicate---object

Why SKOS?
- SKOS is the accepted model for hierarchical linked data. Like all RDF, it can be made freely available, is easy to repurpose, and can be interoperable with other existing or future vocabularies.
Next Steps

4. Serialize and Upload to Triplestore
   - Once our vocabulary is turned into RDF, we serialize it with Python into the turtle format (.ttl) and store the RDF in a triplestore (a database for RDF)

5. Publish Vocabularies using SKMOS

This approach is the easiest way to standardize and centralize our vocabularies, thus potentially resolving discrepancies in how different library units apply terms.
Next Steps

- Skosmos is a SKOS browser and publisher built by the National Library of Finland, it offers a dynamic interface that pulls from our triplestore over HTTP using SPARQL.
Controlled Vocabulary Working Groups
Personal Names, Corporate Names, and ISU (Non-personal Names) Groups

- Names were compiled from:
  - SCUA MARC Records
  - Digital Collections creator, contributors, and authors fields (CDM)
Terms were compiled and then were reconciled with the Library of Congress Name Authority files using Open refine.

Separated Personal and Corporate Name

Created a reconciled and local list of names

Clean-up process

Guidelines were created
ISU (Non-personal Names) Group

These names were devised from the work of the personal names group

- Question: What level do we want to describe units on campus?

Controlled listed where created for:

- Department Level
- Centers and Institutions
- Student groups
- Athletics teams
- Campus administrative units

- A local format was devised
  - i.e. Iowa State University. Department of English
- Rules and guidelines were created
What Other Controlled Vocabulary list can we create?

- Places
- Format-Medium (AAT)
- Others?
Tips, Tricks and Lessons Learned

- Plan for plenty of time. Timelines = magic!
- Provide regular updates to departments/units involved
- Establish specific roles for team members within the group.
- Consider splitting into smaller groups with assigned tasks.
- Establish (semi)regular meetings for all groups to touch base
- Revisit goals & purpose of the project at regular intervals
- Focus on a hierarchy that works best for the library’s collections. Don’t automatically copy others.
- Share what you've done so units across the library have the opportunity to consider using the thesaurus in the future.
- File management
- In order to set up a solid linked data system you need someone who is well versed in coding languages.
Questions?

Laura Sullivan, lsulli12@iastate.edu
Brad Kuennen, kuennen@iastate.edu
Lindsey Hillgartner, lah1@iastate.edu