LINKED DATA

CONTROLLED VOCABULARIES AND BIBLIOGRAPHIC DATA

Presentation by

Ashleigh Faith, MLIS, PhD

Director, Platform Content and Data Management

In Association With EBSCO Information Industries

January 2019

What's the problem?

- 80% of researchers spend their time searching for content.
- Each user typically does three searches.
    - First to see what the general space looks like
    - Next to dig into a specific aspect of that space
    - Third to actually find content hey want to engage with
- **45% of digital library users express frustration in finding content.**
- 84% of ALL users start their search on Wikipedia or Google Scholar.
- 80% of searches on Google are now questions.

How can EBSCO help??

What is linked data?

Linked data has two meanings, the first is primary in data science and the second is used more in the library domain.

Defined in data science as:

Linked data in data science is primarily concerned with the sharing, linking, and interoperability of data. Most often connected with metadata schemes, linked data is used to create semantic webs, knowledge graphs, and ontologies. It is defined as "using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods. More specifically, Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF." (http://linkeddata.org/, 2018). The W3C goes on to define it as a manifestation of Tim Burners-Lee Semantic Web which is "a Web of Data — of dates and titles and part numbers and chemical properties and any other data one might conceive of. The collection of Semantic Web technologies (RDF, OWL, SKOS, SPARQL, etc.) provides an environment where applications can query that data, draw inferences using vocabularies, etc." (https://www.w3.org/standards/semanticweb/data, 2018).

Defined in libraries:

Linked data within the library domain borrows from the data science definition but in a defined library context. This context is usually connected with BIBFRAME and RDF/FRBR standards within the library metadata community. As Karen Coyle, one of the leading library researchers in this area states, libraries are moving toward the world of Linked Data with the Bibliographic Framework Initiative, known as BIBFRAME, which was announced by the Library of Congress in 2011. Since then, though BIBFRAME is still in development, rapid progress has been made that suggests that BIBFRAME may be the long-awaited replacement for the MARC format that could free library bibliographic information from its information silos and allow it to be integrated with the wider web of data. (Coyle, 2010; Coyle, 2012; Gonzales, 2014).

Why not use the search box and be done with it?
Here are some academic findings from the engineering community:

- Zhang, Ogletree, Greenberg, and Rowell (2015) analyzed the use taxonomies in scholarship and found that **73% of users preferred using free text and taxonomy together to search** and Cleverley and Burnett (2015), examining how a research team interacted with the taxonomy of a 13,000 volume repository, found that researchers' **"satisfaction with knowledge search grew after taxonomy was added" and increased "the propensity of a search UI [user interface] to facilitate unexpected, insightful, and serendipitous discoveries**."

- Szostak (2003) indicates that **scholars may be able to more effectively search for studies across multiple domains, find studies that will directly pertain to the methods they will use, and find content which are most suitable to their research through taxonomy knowledge graphs.** Additionally, he suggests that scholarly artcles lend themselves to term-relationship connections because of the emphasis on the scientific process and cause and effect.

- Du, Lau, Ma, and Xu (2013) investigate using ontologies to map terminologies across different domains and repositories for **better knowledge transfer and search**; Giess, Wild, and McMahon (2007), show that **researchers prefer the context rich knowledge search afforded by taxonomy**; Hjorland (2002), indicates terminological studies may **use taxonomies to further the researchers access to knowledge.**

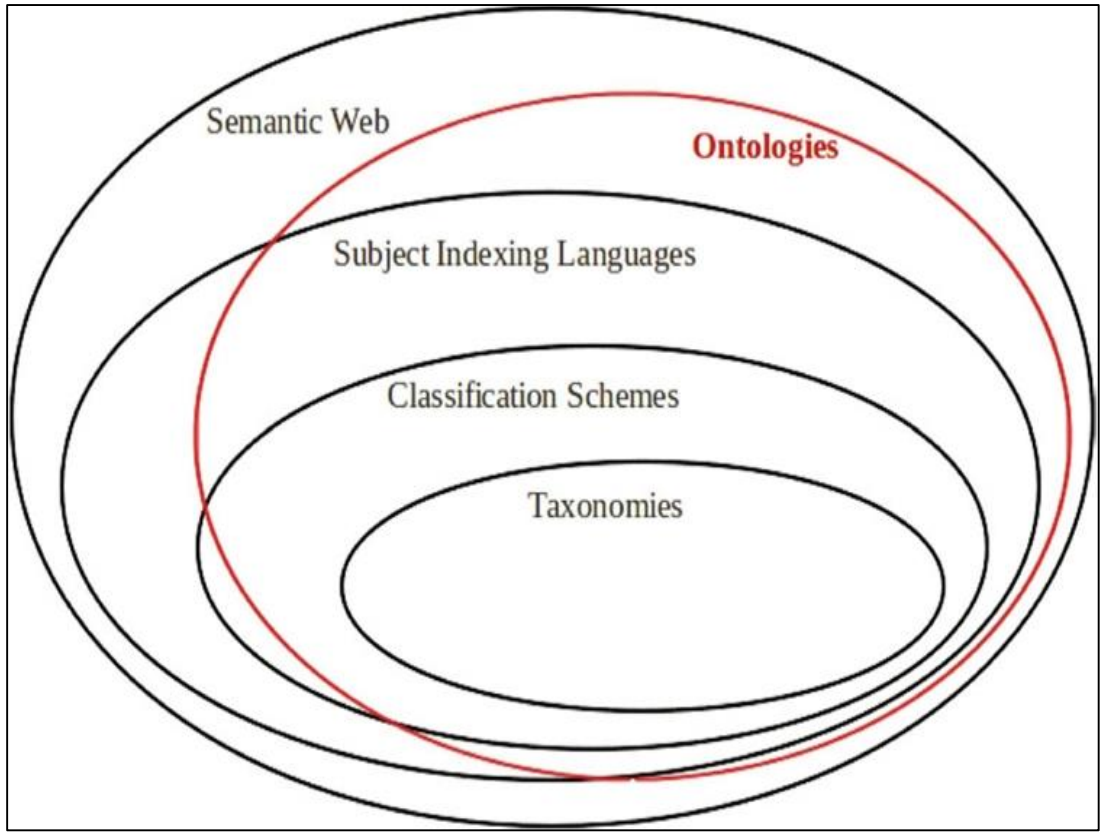
- According to a study conducted by the MIT Artificial Intelligence Lab, **knowledge seekers use keyword search less than 40% of the time, preferring to use browsing capabilities.**

- Cleverly and Burnett (2015) and Szostak (2003) explore how the research community searches for knowledge based on their knowledge needs. They found that a **topical knowledge graph, supported by taxonomy, may help engineers to more effectively meet their research needs and "stimulate new needs, improving a system's ability to facilitate serendipity**."

- Hjorland (1998), in agreement with Hislop (2013) and Oborn and Dawson (2010), state that **the most useful information to a scholarly user is being able to identify the different meanings of terms, especially terms which are shared across disciplines**. Hjorland finds that "**users need 'maps' of information structures…[which] uncover the more or less hidden meanings, interests, and goals in the document**" and that by **having a taxonomy that maps the distinctions between terms, more cross disciplinary research may occur.**

LINKED DATA FOR CONTROLLED VOCABULARIES

Where do ontologies (linked data for subject vocabularies) fall within other vocabularies within literature?



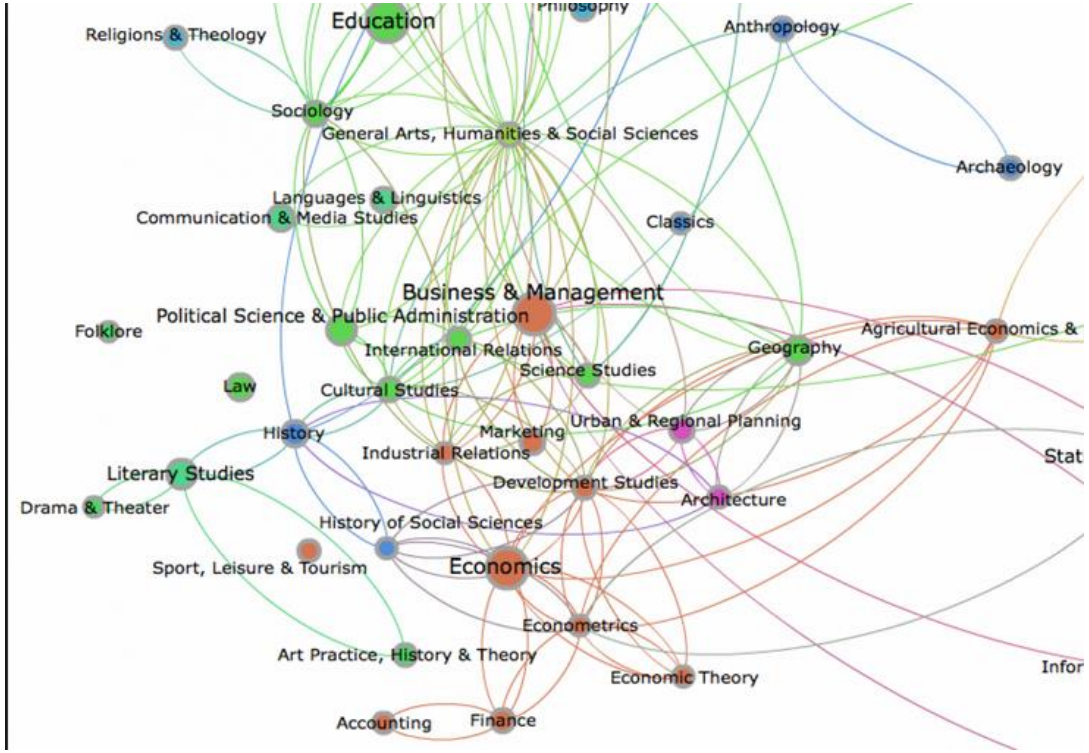
Knowledge Organization Literature- Structure saturation, analysis from Web of Science, conducted 2016

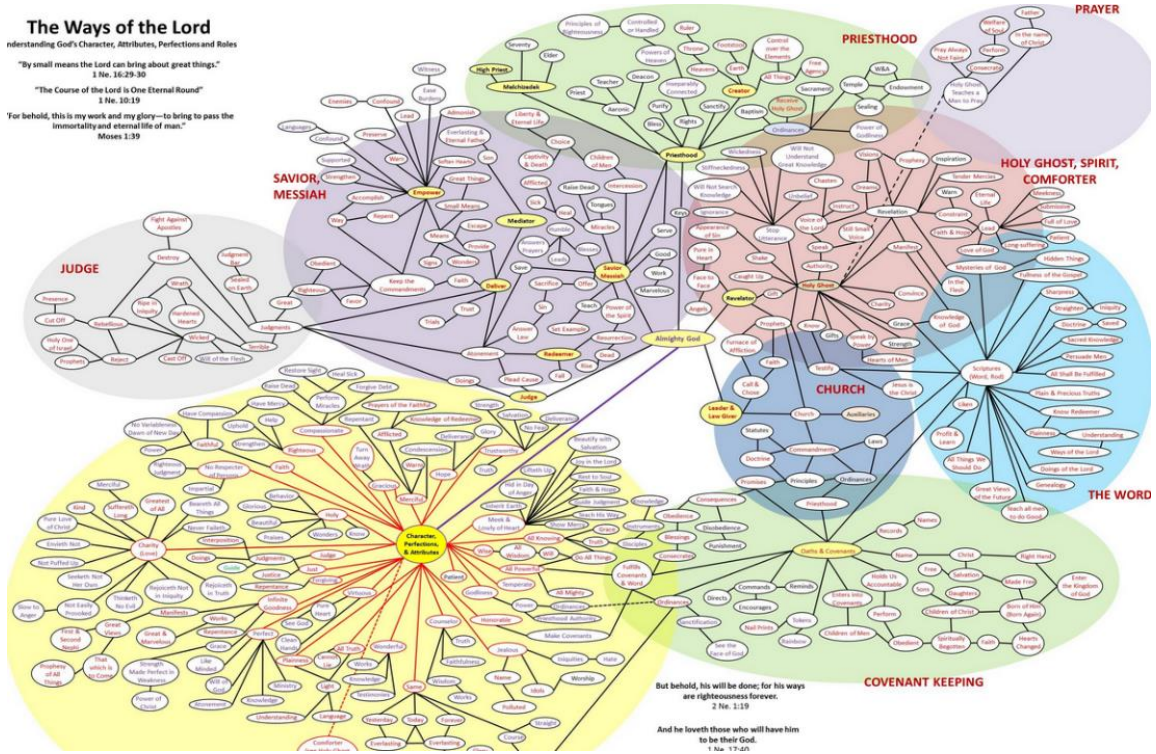
2016 Knowledge Organization Literature	n=	%
Ontology/ies	87	13%
Thesauri/thesaurus	70	11%
Taxonomy/taxonomies	30	5%
Subject heading/s	12	2%
Controlled vocabulary/ies	26	4%

Structure

- Taxonomy and Ontology are both hierarchical at a basic level.
- Folksonomy is unstructured and unorganized for the most part. When a visualization application is applied to folksonomic tag frequency can be represented in bigger font (like in a word cloud).
- Taxonomy and Ontology are similar.
  - Taxonomy is a vertical hierarchical organization.
  - Ontology cannot be created without first having a taxonomy.
- Folksonomy is not related to either, but words and terms in a folksonomy can be used to create a taxonomy or ontology. When used to create ontology (linked data) relationships it can be called folksontology.

Example of Ontology: Ontologies are forms of Linked Data Triples





Most often found organized in:

- Protégé (OWL, OBO, or SKOS file)
- Taxonomy/Ontology Management System such as Data Harmony, Semaphore, or Cogito

Semantics: A Triple Story

What are Triples?

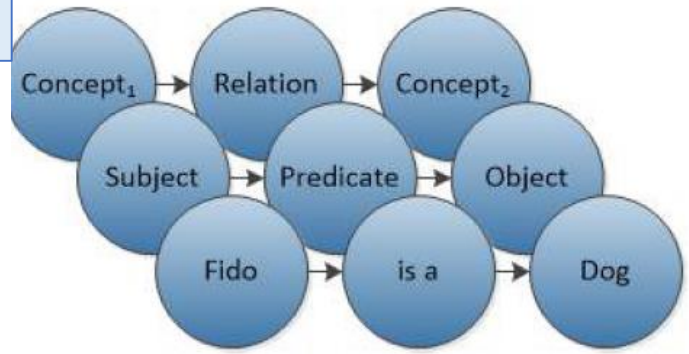
Crash Course

- Triples are what semantics are recorded as
- Each triple defines a specific type of relationship between two entities so a computer can understand how they relate to one another

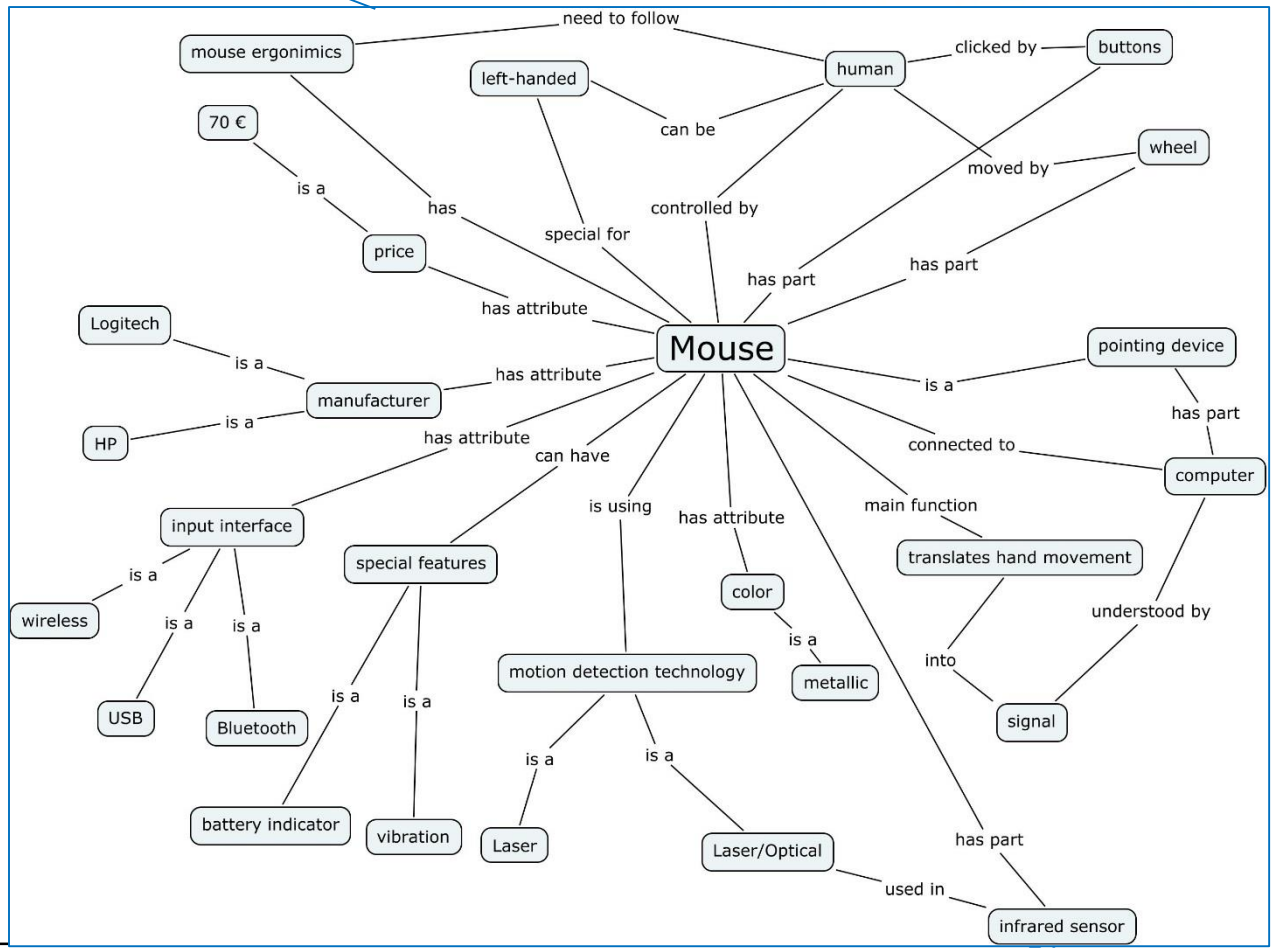
Entity 1	***Relationship***	***Entity 2***
Catalytic Converter	*isPartOf*	The Exhaust System
Identity protection	*isElementOf*	Cyber Security
John C. Doe	*isCoauthorWith*	Sam B. Smith
Boeing	*isA*	Manufacture

- Triples create computer logic which enhances search functions to facilitate more tacit information needs
- Sets of triples build knowledge graphs. A Triple store is a network of all triples
  - Not all knowledge graphs need to use RDF triples. There are other types of triples such as FOAF, OWL, SKOS, etc., that can be modified for specific use cases.

Example of Semantic Triples



Example of Semantic Triple Knowledge Graph



TECHNICAL ONTOLOGY

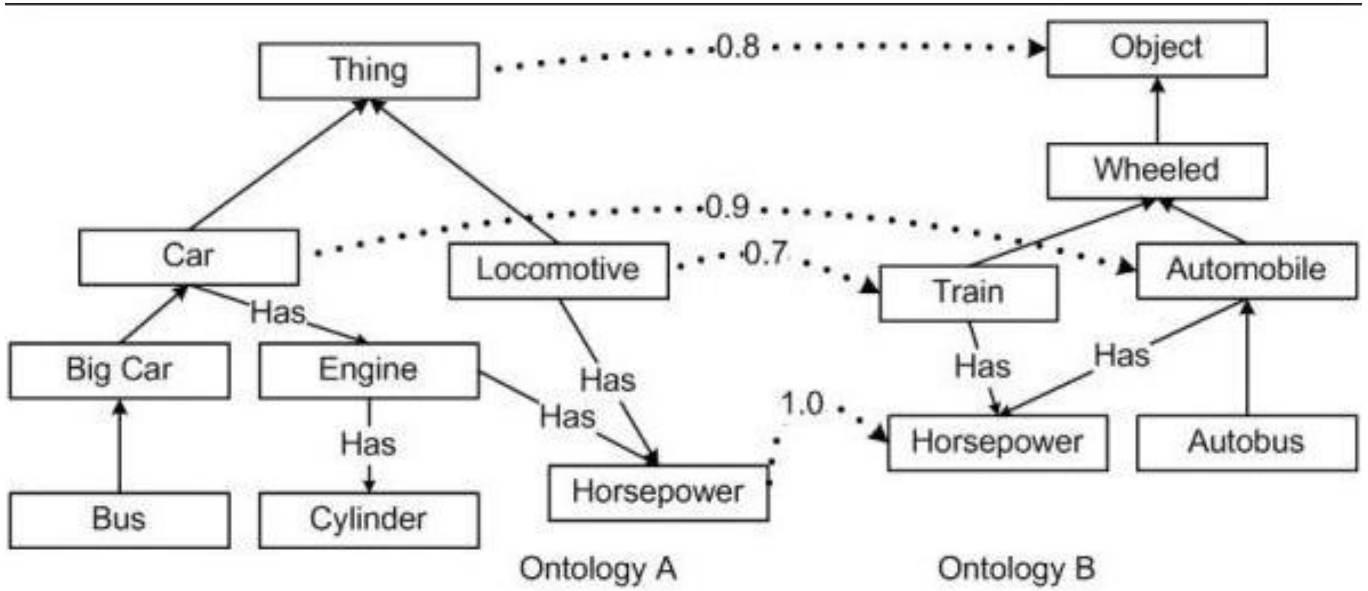
Levels of Ontology Complexity

Ontology Profile	Area	Formality of Logic	Interoperability Prospectus	Language
Upper-level schema	Philosophy	High	High	First-order logic (inference) OBO OWL
Heavyweight schema	Information science (inference)	High	Medium	OWL
Lightweight schema	Library and information scientists	Medium	Low	RDFs HTML JASON

What is an ontology used for?

- AS A KNOWLEDGE PROVIDER
  - Answering user questions
  - Smart recommendations
- CONNECTING KNOWLEDGE IN MEANINGFUL WAYS
  - Knowledge search rather than information search
  - Smart search- Mapping terminology to user's language
  - Semantic enrichment- Adding more value to content
  - Better search and retrieval
  - Connecting author/publisher data
  - Connecting institutional data
  - Connecting content to other types of resources
  - Data visualizations for enhanced knowledge search
- DATA MINING
  - Content data mining
  - Mining for knowledge representation
  - Mining for business/market intelligence
  - Mining for open access content
  - Mining & mapping terminology for systems

Mapping:



- Mapping is hugely important now because without it content cannot be aggregated and processed from multiple sources. This is also a part of linked data. EBSCO's qusai-ontology is the USI

Folksontology?

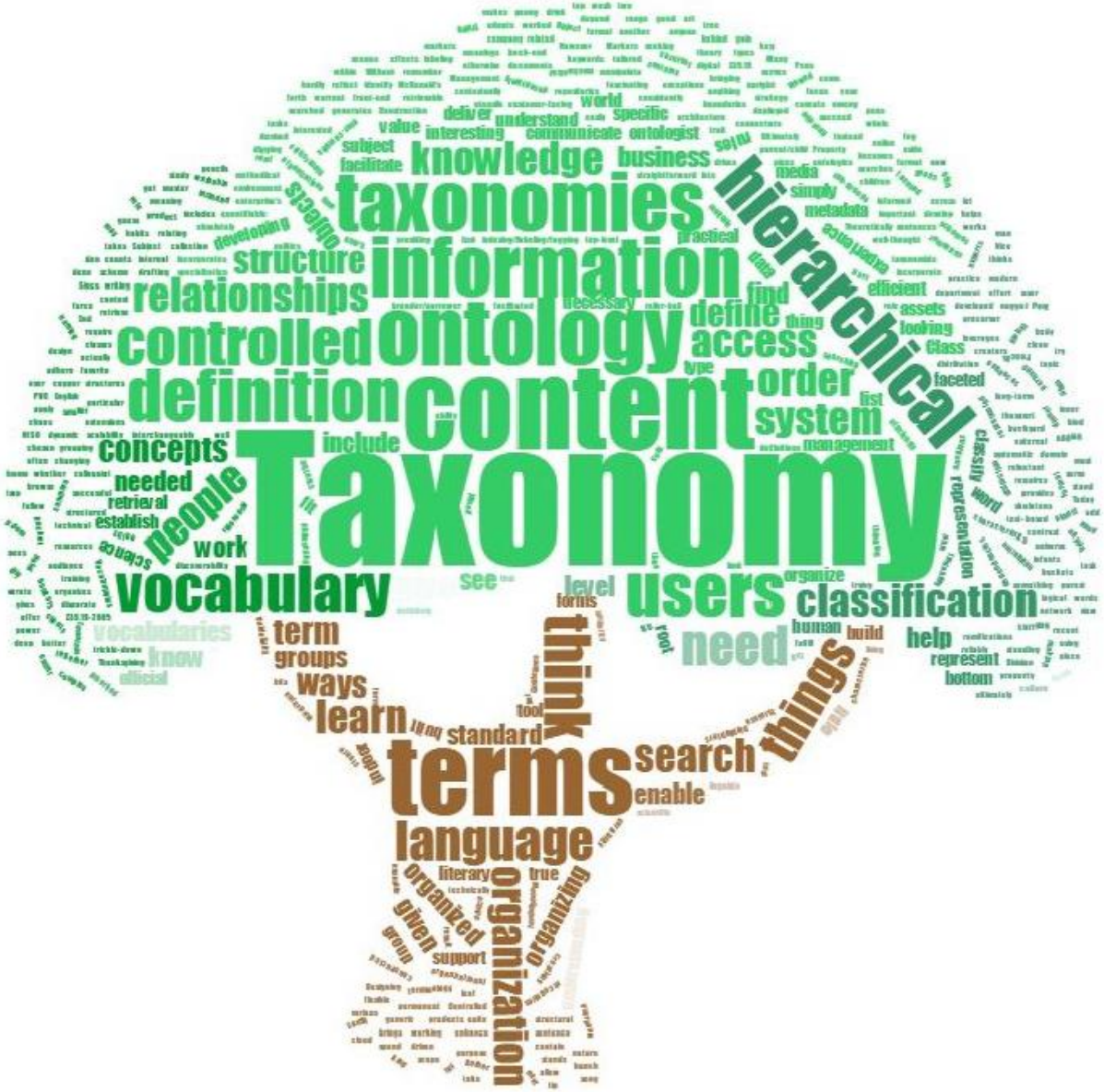
- Connecting users' natural language by way of an ontology increases user satisfaction by 55%.

Definition: Folksonomy

- User-generated, often crowdsourced, controlled or uncontrolled vocabulary (van Damme, Hepp, and Siorpaes, 2007). Flickr, Facebook photo tags, and Twitter tags are examples of a folksonomy.

Folksonomy

- They are good for assessing what terminology your users use and what they are interested in.
- Are not suitable for systematic processes.
- Fun to play with and use visualization tools for.
- This tag cloud was created from a sample of 80 taxonomists who were interviewed on what they think taxonomy is.



HOW TO MAP WITH LINKED DATA

Background

- *Why map data?*
- It is increasingly important to connect data from various databases to allow for linked sharable data (interoperable), collaboration between researches (especially across disciplines), evolve to a *just-in-time* model of access (medical and legal greatly benefit from this), and elevate information and knowledge access within repositories.

- *Here are some issues to contend with while linking controlled vocabularies:*

- Not many linked vocabulary quality standards
- Governance and provenance documentation needed
- Little understanding of what user's needs are
- Maintenance
- Terminology varies across databases (inconsistency of terms) requires more manual approach for granular vocabularies
- Poor quality and incomplete data
- Non-digitized content
- Restricted data
- Inconsistent terminology definitions/scope notes hinder access, reliability, and communication of data –notoriously so in gray literature

Linked vocabularies facilitate search in dispersed systems through vocabulary mapping

Linked vocabulary has the following characteristics:

✪	Data is made available in any format such as PDF. Open access under an open license is preferred.
✪✪	As above, but as structured data such as an Excel file.
✪✪✪	As above, but structured in a non-proprietary format like CSV or XML. Vocabularies at this stage can be mapped but without URIs the mapping decreases in effectiveness.
✪✪✪✪	As above, but use W3C open standards such as RDF as well as URIs so the vocabulary can be linked to. This includes linking within the vocabulary.
✪✪✪✪✪	As above, and also linking out to other vocabularies. This level represents the most robust linked vocabulary interoperability.

Table 1: Star system for linked subject heading vocabularies. Stars indicate level of interoperability. Modified from Berners-Lee (2009), Binding and Tudhope (2015) and Kim (2015).

Jargon note:
Source vocabulary= the terminology you are mapping from. Your "starting point."
Target vocabulary = the terminology which is analyzed for a match

Vocabulary Case Studies

Model Type	Use For	Case Study Name	Domain	Source
Derivation model	New vocab. from existing vocab.	North Atlantic Treaty Organization (NATO) Terminology	Military/International Relations	Panajotu, 2012
Application profile	Adding rules/protocols to existing vocab.	The Electronic Exchange of Social Security Information (EESSI)	Legal	Azuara, et al., 2013
Crosswalk model	Mapping between 1-3 vocabularies.	Evidence in Documents, Discovery, and Analysis (EDDA) Terminology	Medical	Bekhuis, et al., 2013
Switch or Hub model	Mapping between n vocabularies.	Dryad Digital Repository and White	STEM	Bedford, et al., 2010; White, 2015
Metadata Framework	New vocab. using framework's vocab.	Smithsonian American Art Museum (SAAM) Linked Data project	Art	Szekely, et al., 2013
Metadata Registry	Sharing existing vocab.	Linked Open Vocabularies (LOV)	General	Vandenbussche & Vatant, 2014

25

Most common mapping models:

*Can be recorded in a more simplistic format like Excel to more complex formats like OWL/SKOS

Type of analysis: Term-to-term *Source* term-to-*target* term

Vocabularies	Crosswalk (M2M)	Links (n^2-n)	Hub	Links (2n)
2		2		2
3		6		6
4		12		8
5		20		10

Mapping volume comparison between crosswalk and hub structures. Modified from Binding and Tudhope (2015).

Types of term mapping:

- *Partial* = syntax/semantic/contextual match
- *Exact* = matches exactly
- *Associative* = Matches as a related term
- No match = No match in target terminology

Considerations of term mapping:

- Hierarchy
- SMEs
- Automation/APIs
- Context/domain specificity

LINKING IT ALL TOGETHER

How library indexing fits in to linked data

This is an example of a triple in the library world

Use of RDA FRBR Example

* "Imagine that you have a patron who needs a copy of Heaney's translation of Beowulf. She doesn't care who published it or when, only that it's Heaney's translation. What if you (or your patron) could place an interlibrary loan call on that expression, instead of looking through multiple bibliographic records… for multiple manifestations and then judging which record is the best bet on which to place a request? Combine that with functionality that lets you specify 'not Braille, not large print,' and it could save you time. Now imagine a patron in want of a copy, any copy, in English, of Romeo and Juliet." (Gonzalez, Linda. "What Is FRBR?" Library Journal 130 [Spring 2005 NetConnect]: 12 – 14).

* Example (using an RDF triple- you won't get what an RDF triple is yet but in later weeks you will –for now just check out the example)

* Using this example, the work (i.e. the copy your library has) can be found by:
  - Translations of Beowulf
  - Heaney translations
  - Heaney AND Beowulf
  - The expression itself, Heaney's translation of Beowulf
  - Adding the attribute English translation can also be used

This is another example of a triple

Dublin Core and Linked Data

1. Dublin Core, or Dublin Core Metadata initiative (DCMI), started as a one of the first metadata standards used in digital libraries. It has since become a major standard used in linked data even outside the library. At the most recent DCMI conference in the states (Austin) I encountered more people using DCMI outside the library as I did within it.

* Converted to linked DCMI linked data, it would look like:
  * dc: contributor: Heaney
  * dc: title: Beowulf
  * dc: format: translation
* There are two types of linked data: bibliographic metadata and classification (indexing) data.
  * Bibliographic metadata is used to connect content; classification (indexing) data is used to connect subject heading-like terms for more consistent content tagging.

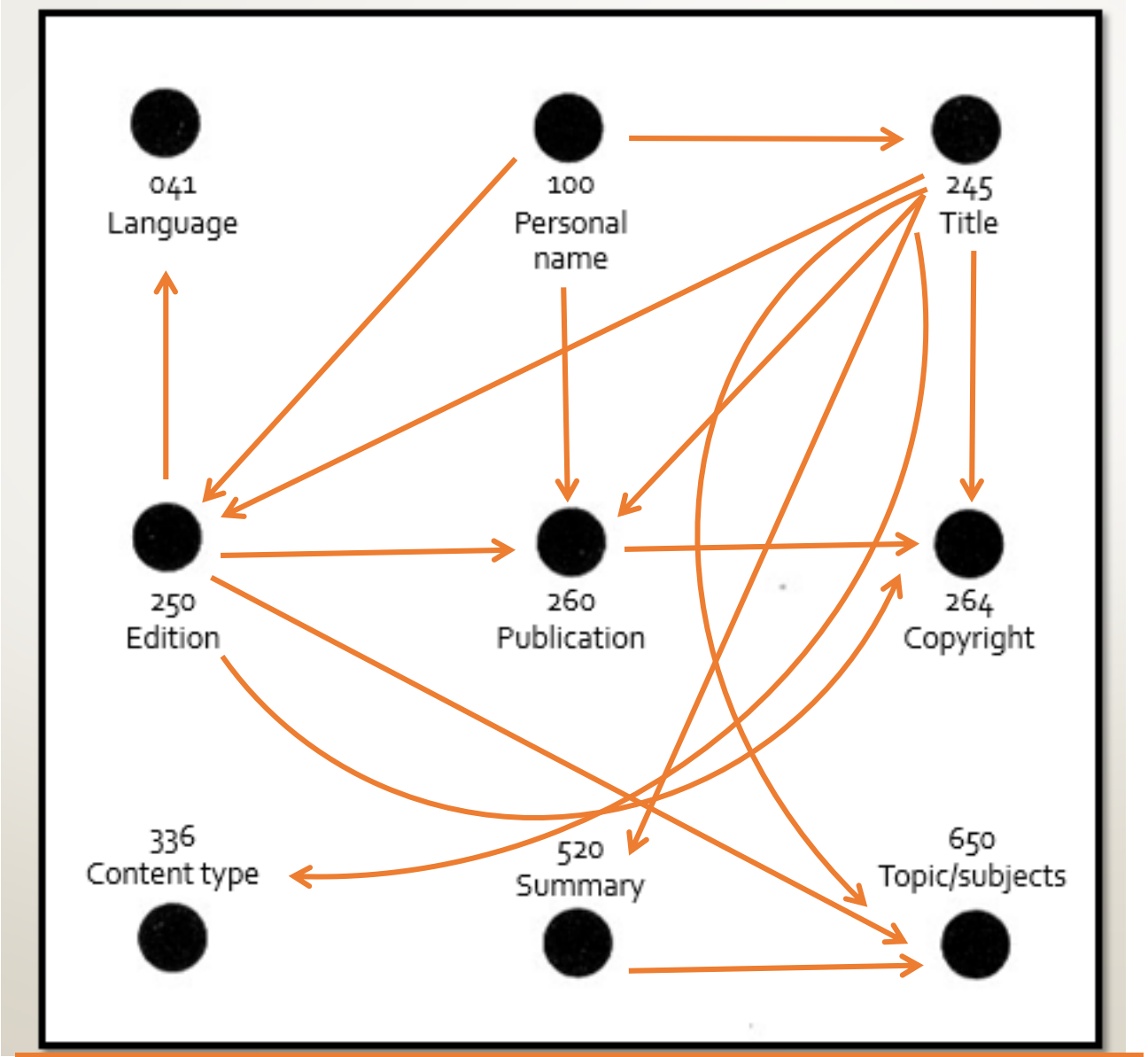
LINKED DATA/SEMANTIC WEB/RDF

* **Linked Data** is the foundation of the Semantic Web and means to connect pieces of data together with the ultimate goal of making things easier to find through multiple access points. The link structure is called RDA.
* The **Semantic Web** is defined as using the Web to connect data and "to lower the barriers to linking data" currently in use; specifically by "exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web (Heath, 2014)." By enabling data to be linked, usually through metadata, the data can be considered interoperable. *Interoperability is a key function of linked data because it allows systems and people to exchange information using similar access points such as metadata and search terminology.*
* As defined by the DCMI Glossary, **interoperability** is "the ability of different types of computers, networks, operating systems, [people] and applications to work together effectively, without prior communication, in order to exchange information in a useful and meaningful manner" (2011). This happens through linked data access points.

RDF and RDA are the same thing.

RDF is used in the technology world and RDA is used in the library world. They are essentially the same thing.

	Subject	**Predicate**	**Object**
	Bibliographic Linked Data:		
	dc	contributor	Heaney
Explanatory Notes	This is to indicate which schema it is from	This is the schema element	This is the "individual" or the information access point –i.e. the object of the link.
	Classification (indexing) Linked Data:		
	Cancer	isSynonymOf	Melanoma
Explanatory Notes	This is the source term. A source terminology is the terminology used on your own content.	Link/relationship is used to indicate how these two terms are related (or linked).	This is the target term. A target terminology is the terminology used to find alternative terms to your own – i.e. synonyms.



Each relationship is an additional access point.

Some example free-text (i.e. not coded) links within a MARC document record:

- The 4th edition is in Spanish
- Maria Luguio wrote the book
- Maria Luguio wrote the books current edition
- Mario Luguio worked with the publisher XIX
- The title is in its 4th edition
- The title was published by XIX Inc.
- The title is under this copyright
- The title is a textual content type
- The title is about the summery (abstract)
- The 4th edition was produced in Italy
- XIX Inc. filed and received copyright
- The 4th edition was produced and in copyright by 1980
- The book is about horses
- The edition is about horse healthcare
- The summary describes horses

How to code human understanding into linked data:

Free Text Linked Data	Coded Linked Data	Coded Linked Data	Coded Linked Data
The 4th edition is in Spanish	4th edition isTranslatedAs Spanish	Spanish isSpokenIn Italy	
Maria Luguio wrote the book	Maria Luguio isAuthor of book	Maria Luguio isA author	Maria Luguio isA female
Maria Luguio wrote the books current edition	Maria Luguio isAuthor of 4th edition		
Mario Luguio worked with the publisher XIX	Maria Luguio workedWith XIX	XIX Inc. isA publisher	
The title is in its 4th edition	title version 4th edition		
The title was published by XIX Inc.	title publishedBy XIX Inc.		
The title is under this copyright	title inCopyright 1980		
The title is a textual content type	title contentType textual		
The title is about the summery (abstract)	title hasA summery		
The 4th edition was produced in Italy	4th edition wasProducedIn Italy		
XIX Inc. filed and received copyright	XIX Inc. filedFor copyright in 1980		
The 4th edition was produced and in copyright by 1980	4th edition inCopyright 1980		
The book is about horses	title isAbout horses		
The edition is about horse healthcare	4th edition isAbout horse healthcare	horse halthcase isPartOf horses as topic	
The summary describes horses	summary isAbout horses		

Notice how the relationships are consistent when the relationship is the same.

Relationships should not include spaces. Underscores or capitalization should distinguish the individual words.

Are all relationships manually created? No.

- Other places to find Linked Data:
  - Schema online
  - Content in your collection
  - Bibliographic records
  - Online resources like Wikipedia and DbPedia
  - user preference and trends, i.e. user activity

Are there rules to relationships? Not really.

- Using schema like DCMI or other metadata field tags is one way to remain consistent. More often than not relationships are built for the need of the collection. Unique relationships however, do not facilitate interoperability so try to stick with a known schema if possible.
- Schema and relationship tags can be from multiple standards, schema, and lists.

How are triples used?

- Linked Data:
  - Enables computers to do more with your content and data
  - Aggregators, like librarians and library databases, can understand and pull more specialized information from their collection.
  - Users have a much wider variety of search options which helps them find the most applicable content. Also allows for users to customize their search more leading to more discovery.
  - Usually entered in a content management system, knowledge graph software, or an ontology program.
- Allows information to be connected to the outside world, like DbPedia or Wikipedia.
- Triples can define ***knowledge*** about your content. Knowledge and information are different things. Information is the raw data and the knowledge is the relationships librarians create and the tagging (as in MARC fields) that add value to data.
- Specific search, such as users asking for all books about horses written by women, are facilitated by triples.

Conclusion

By connecting data and controlled vocabularies across different publishers and databases, users benefit by way of easier cross database and cross disciplinary research.

This allows for a seamless search experience so researchers can do what they do best. Research.

For more information contact Ashleigh Faith at afaith@ebsco.com

36