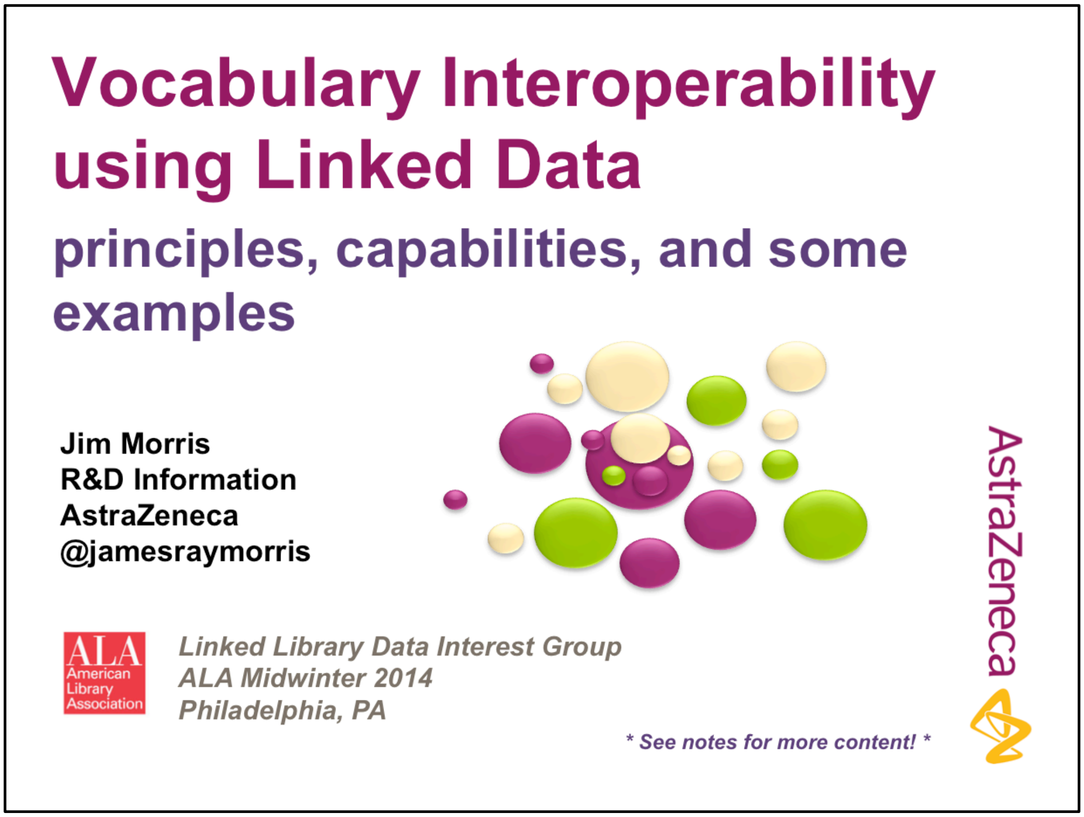


Vocabulary Interoperability using Linked Data

principles, capabilities, and some examples

Jim Morris
R&D Information
AstraZeneca
@jamesraymorris

ALA American Library Association

Linked Library Data Interest Group
ALA Midwinter 2014
Philadelphia, PA

** See notes for more content! **

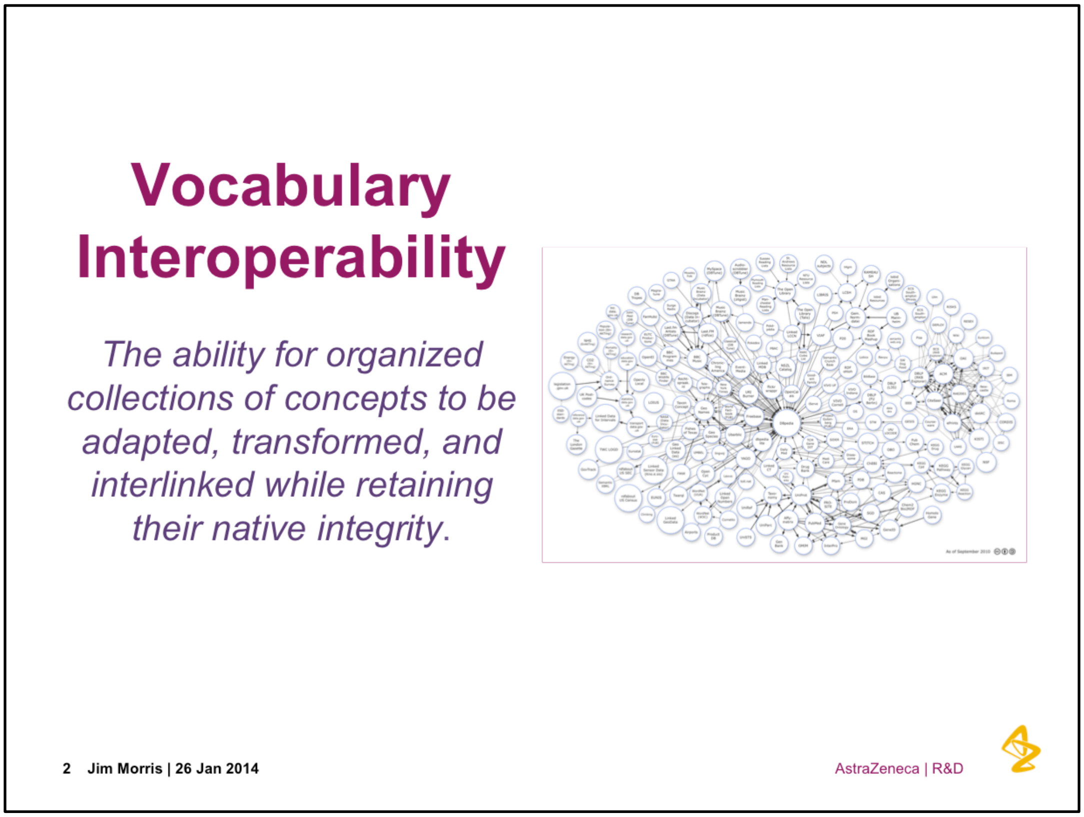
AstraZeneca

By “Vocabularies” I mean any controlled and structured set of terms or concepts – taxonomies, thesauri, ontologies, glossaries…

Vocabularies (taxonomies, thesauri, etc.), as every taxonomist knows, are essential for finding, navigating, and connecting the information or data we need. But just like the rest of the information we need, we don’t control those either (unless we own them). We can’t import them all and we can’t buy them all. And we can’t recreate them all--which is often what we end up doing.

To continue to build our profession's capability in building, managing, and using taxonomies, thesauri and other vocabularies, we need an approach that recognizes the reality of today’s highly networked information world. This approach includes Linked Data principles and the technology of the Semantic Web, which are designed to harness the inherent open-endedness of the World Wide Web. It's the right platform on which to continue building our professional expertise in vocabulary management.

The opportunities that this field presents to our profession are very exciting. We need more discussions about Linked Data vocabulary management among librarians and information scientists. To an equal extent, the values, skills, and experiences of librarians and information scientists are needed to make this emerging field successful.



Vocabulary Interoperability

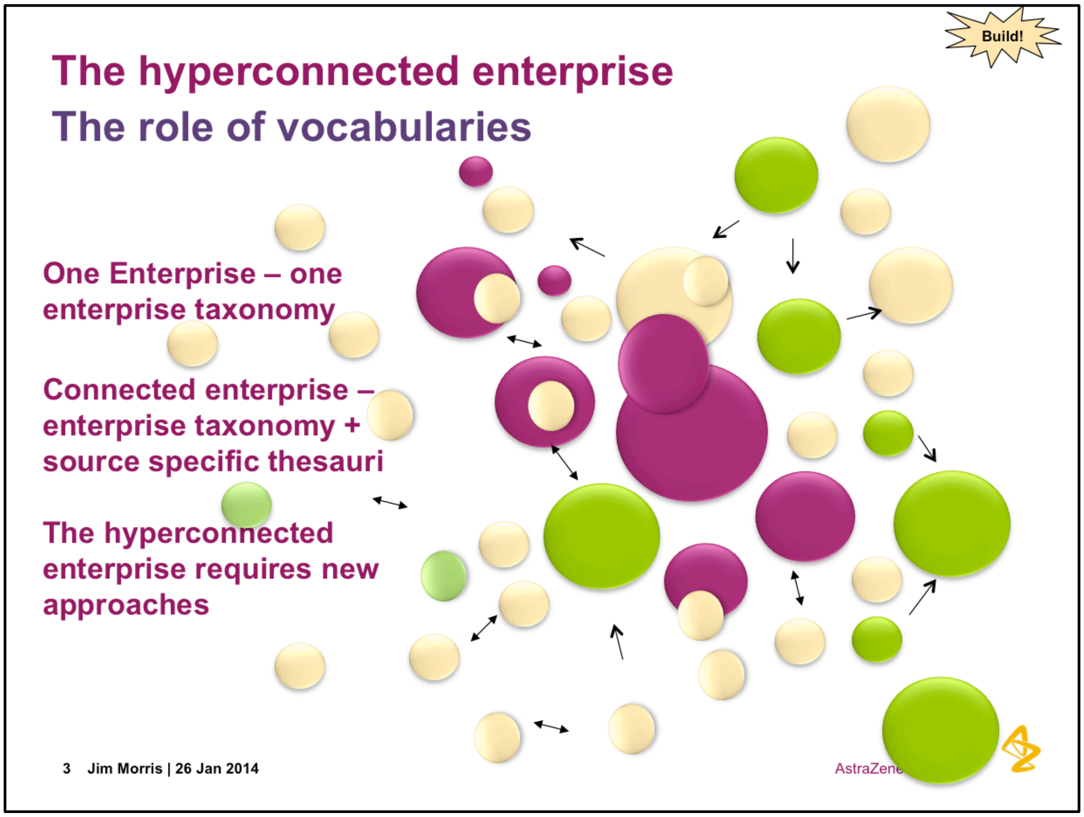
The ability for organized collections of concepts to be adapted, transformed, and interlinked while retaining their native integrity.

2 Jim Morris | 26 Jan 2014

AstraZeneca | R&D

This concept is at the heart of my talk.

(As I thought about this idea it occurred to me that it is not revolutionary at all – especially to a librarian. I have a habit of relating every new field I become involved in back to traditional librarianship. Very few advances in information management are truly revolutionary to a librarian. Take "vocabulary interoperability" for example. Consider a traditional "library". Isn't it about "book interoperability"? Isn't it about the ability for books to be adapted (selected, purchased, circulated), transformed (read, used to create new knowledge), and interlinked (arranged and made accessible according to common subjects and classifications), while--and this is key--never actually changing the books themselves?)



The hyperconnected enterprise

The role of vocabularies

One Enterprise – one enterprise taxonomy

Connected enterprise – enterprise taxonomy + source specific thesauri

The hyperconnected enterprise requires new approaches

3 Jim Morris | 26 Jan 2014

AstraZene

The purpose of this slide is to position the Semantic Web and Linked Data against an imperative of living in networked world. Too often Linked Data appears as a solution looking for a problem. No—the problem, the opportunity, is here. Linked Data is a solution designed to address the issues facing every information professional, including librarians and taxonomists.

Organizations have been evolving from being very self-contained to becoming, what Gartner calls, the "hyperconnected enterprise"—a highly fluid, highly connected collection of organizations, where coordination is key—not control. What does a hyperconnected enterprise use to manage this network? Information of course. How do controlled vocabularies fit into this picture?

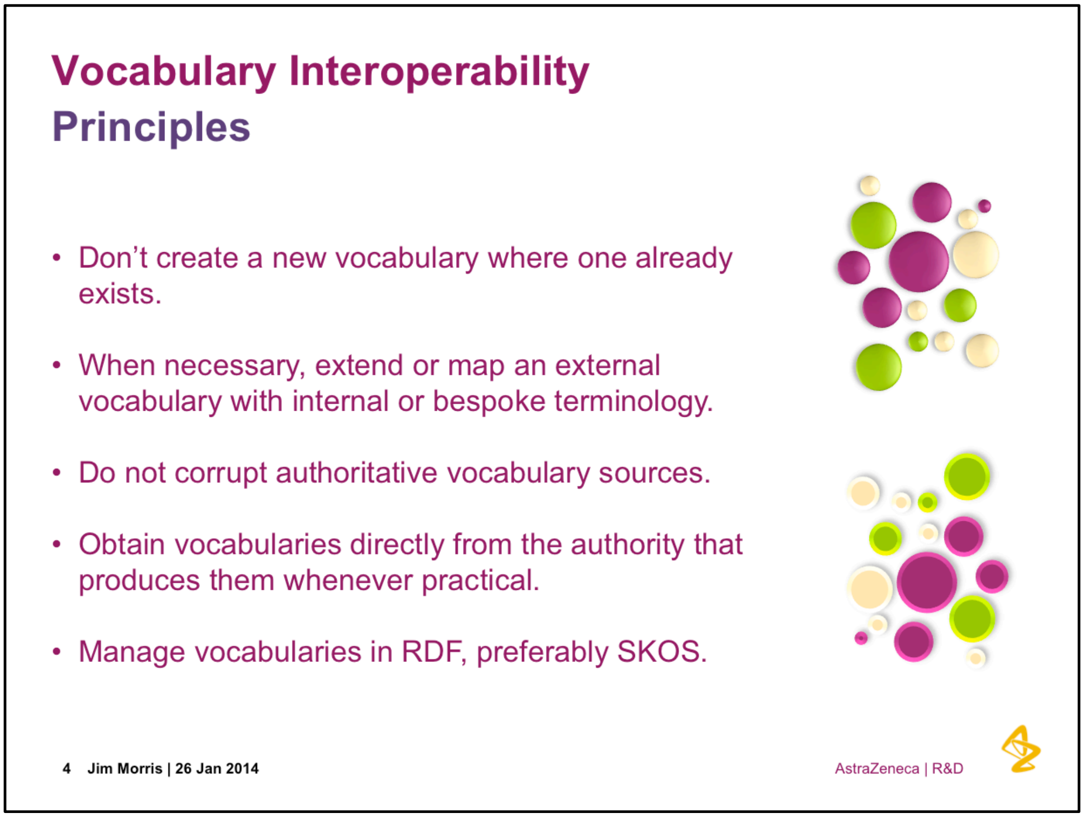
1. One Enterprise: Simple if you're one organisation, with good control of your information. Create an enterprise taxonomy, or set of enterprise vocabularies, and have everyone use it; establish a governance structure for maintaining it. As I mentioned before I represent R&D on the AstraZeneca "enterprise taxonomy" – it's a valuable resource used in our enterprise systems like SharePoint, Search, and digital asset management systems. We still need to manage our own information well.

2. Connected Enterprise: Of course, you need to use external information sources, which use their own custom vocabularies
And you'll need to deliver information to regulatory agencies that require the use of certain vocabularies (e.g. MedDRA, ICD-9/10, and other standards)
Maybe another company has an enterprise taxonomy or product thesaurus – we can integrate those; give us a year.
One of my first big projects at AZ was intergrating three product literature databases – each with their own thesaurus. We, of course, built an entirely new one.

3. Hyperconnected enterprise: But these other partners, universities, companies, divisions, subsidiaries… spread across the globe….partners come can partners go
•Small biotech – "Enterprise vocabulary? We're too small – we don't just use terms that we need; it's not complicated."
•Big pharma partner or regulatory agency – "Over here, yes we follow standard vocabularies…they're not the same as your standards, but… "
• Information supplier, database provider -- "Our proprietary databases use specially designed vocabs optimized for our software – you should really be using our software, you know. We have all the information need".

4, 5. And this is not going to stop

This is why can't keep recreating vocabularies, or going through massive mapping efforts. It's unsustainable. It's the same driver that led to the world wide web – coordination trumps control. We need a way to for our vocabularies to be "interoperable", where they can be linked to each other, while still retaining their native integrity.



Vocabulary Interoperability
Principles

- Don’t create a new vocabulary where one already exists.
- When necessary, extend or map an external vocabulary with internal or bespoke terminology.
- Do not corrupt authoritative vocabulary sources.
- Obtain vocabularies directly from the authority that produces them whenever practical.
- Manage vocabularies in RDF, preferably SKOS.

4 Jim Morris | 26 Jan 2014

AstraZeneca | R&D

How will we accomplish “vocabulary interoperability”? We need some basic capabilities. These were brainstormed at my organization.

We don’t create a new vocabulary where one already exists. Our preference is for authoritative, trustworthy, well-managed sources.

- Standard criteria used to evaluate resources – authority, accuracy, currency, point of view (objectivity), coverage, relevance, format. http://library.uwb.edu/guides/eval.html , http://www.noblenet.org/merrimack/FYS_eval.pdf
- We can extend those sources if we have to, but even then we want to extend them by making associations with other authoritative sources. Whenever possible we will leverage cross-vocabulary mappings already available publically.

When necessary, we will extend or map an external vocabulary with internal or bespoke terminology. This could include mapping obvious synonyms between vocabularies, or could entail more nuanced relationships.

- For example – mapping a published disease vocabulary to proprietary database value set, or list of values from internal database.

We will not corrupt authoritative vocabulary sources. It will always be possible to identify the source vocabulary, including what version, in the network.

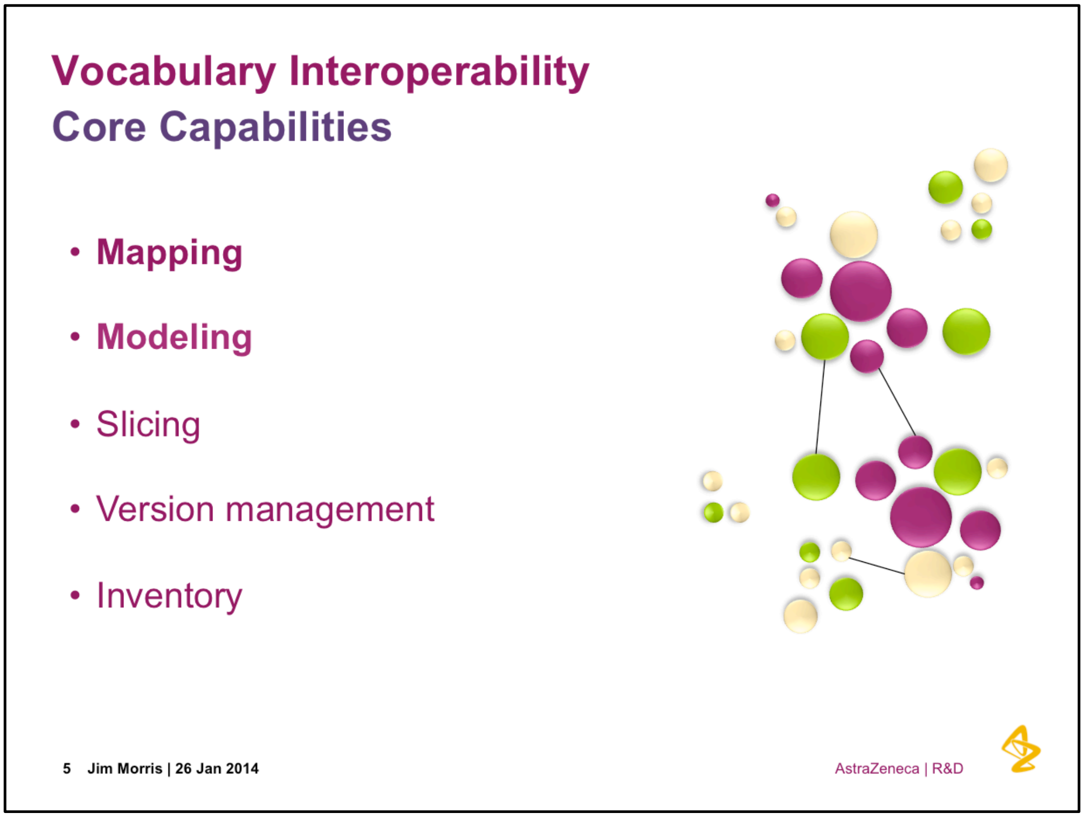
- We don’t take a page out of this book and paste it into another, cross out a section here, add one there to the point where we can’t put the books back together again – we don’t do that with vocabularies either.
- But we do repurpose these vocabularies – so in a way we might need to pull them apart. Perhaps a deeply hierarchical taxonomy just needs to be rendered as flat list. But the semantic web promises us to enable this without corrupting the source.

We will obtain vocabularies *directly* from the authority that produces them whenever practical.

- “directly” is the key word here. We’ll talk about this later.

Vocabularies will be managed in RDF. Preference is for this “SKOS” standard I’ve mentioned before. SKOS is a lightweight standard that can be applied to any controlled vocabulary. Focusing on a simple standard allows us to standardize our curation, management, and delivery processes. [In many cases, this requires doing our own SKOS conversion.]

[aside: in some ways the nature of “published information” is becoming more like the early days of copying, commenting, annotating, all at the same time.]



Vocabulary Interoperability
Core Capabilities

- **Mapping**
- **Modeling**
- Slicing
- Version management
- Inventory

5 **Jim Morris | 26 Jan 2014**

AstraZeneca | R&D

In addition to principles, we need capabilities. Again, these were brainstormed at my organization:

Vocabulary Mapping.

Reusing, creating, and maintaining the relationships between vocabularies holds the power of the "vocabulary network". We need to develop the approach, processes and tech for light-weight, sustainable, semantics-based vocabulary mapping.

- How to map two or more vocabularies together using Semantic tech
- Facilitating the reviewing of automatic mapping - approve, reject, adjust, amend mapping
- Quality measures (e.g. single synonyms can't be mapped to more than one parent).
- Leverage mapping already done in published vocabs.
- Publish mapped "bridging" vocabularies to customers.

Vocabulary modeling and conversions.

The vocabularies in the network need to be accessible as "Linked Data". And, until external suppliers deliver or make their vocabs available as linked data, we will need to be extremely proficient in converting them from their native formats.

Vocabulary Slicing.

Sometimes a vocabulary consumer – system or project – needs only a subset of a vocabulary that is relevant to their needs. For example, only the diseases that are of interest to a particular project. Rather than create a duplicate – we need to identify terms or branches of relevance, and make those available--all while retaining their link back to the authority, so that updates to the authority can continue to be leveraged (new synonyms, new mappings, new child terms, new notes).

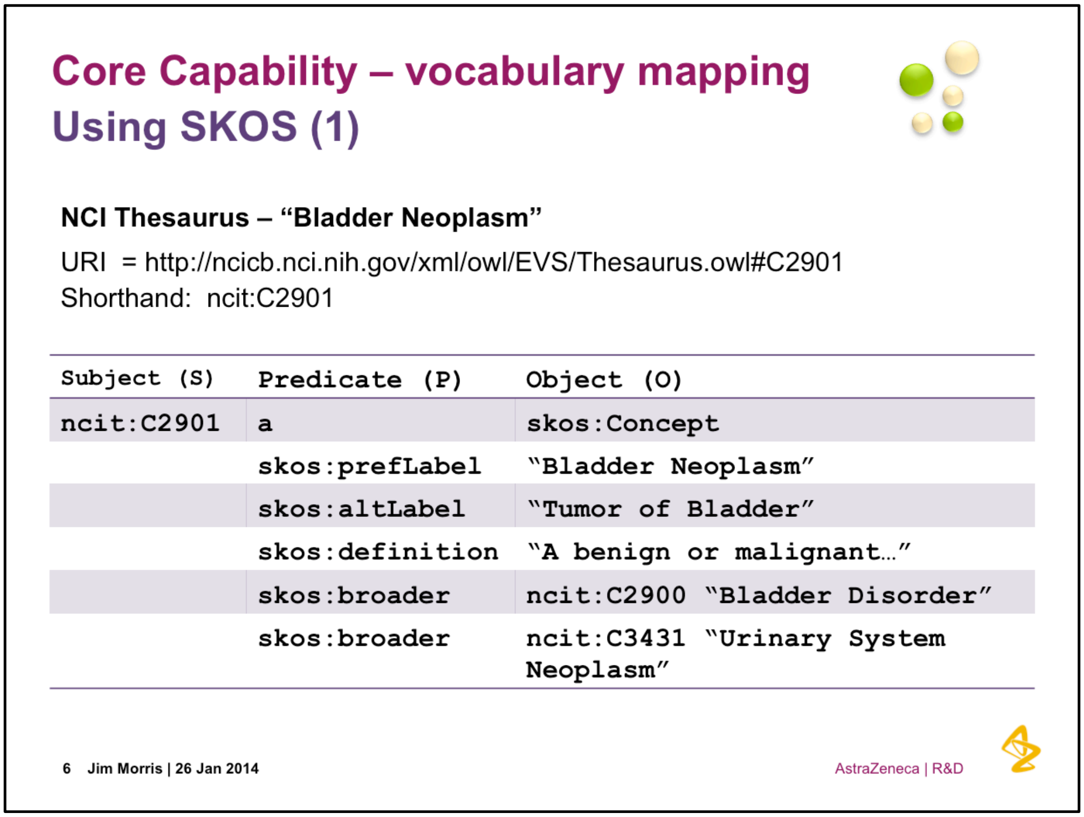
Vocabulary Version Management.

As our focus is the reuse of already published vocabularies, so successfully managing the impact of new versions on the network is critical.

- Comparing new version to existing version
- Identifying impact of vocab changes to mapped vocabs, consuming systems or related datasets.
- Once we identify the impact – how do we addressing impact.

Vocabulary Inventory.

Managing a newtwork of vocabularies requires a sophisticated inventory. Ideally managed as linked-data itself, the inventory can directly inform queries using the vocabulary network. For example, tell me all the vocabularies that cover concept X, the datasets that reference those vocabularies, and how to access them. There are developing standards in the semantic web specifically for managing these types of inventories.



Core Capability – vocabulary mapping
Using SKOS (1)

NCI Thesaurus – "Bladder Neoplasm"

URI = http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C2901
Shorthand: ncit:C2901

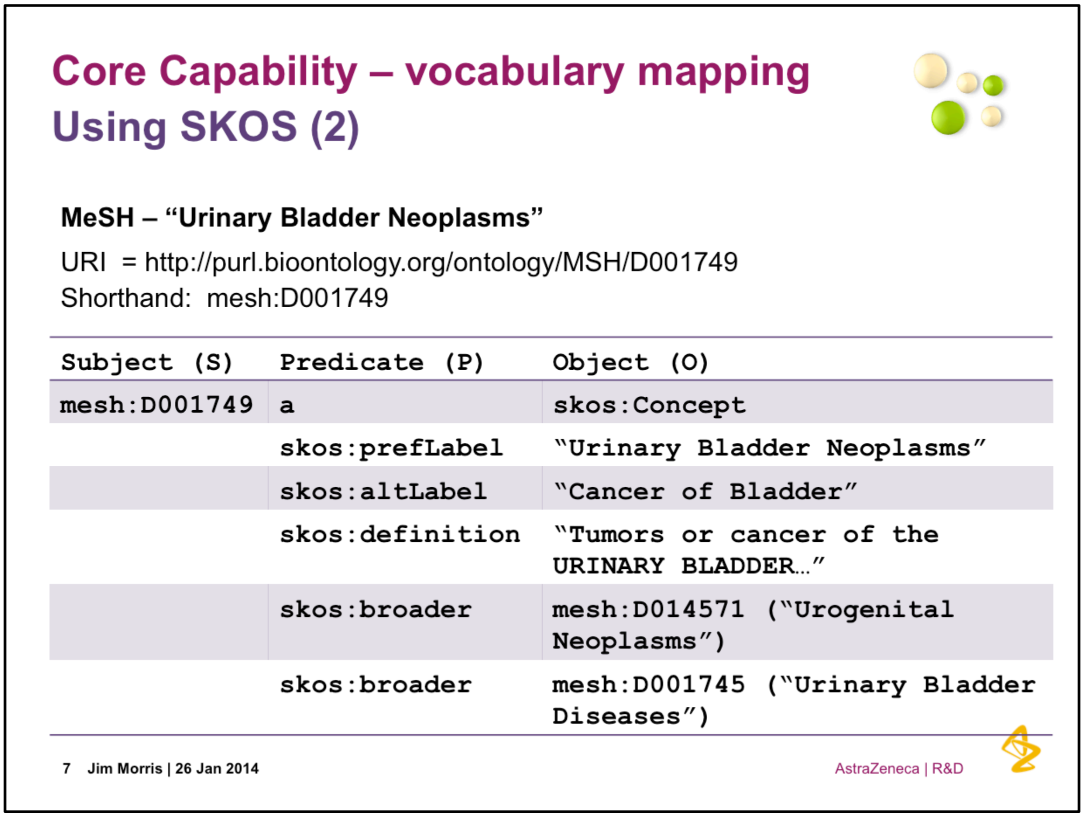
Subject (S)	Predicate (P)	Object (O)
ncit:C2901	a	skos:Concept
	skos:prefLabel	"Bladder Neoplasm"
	skos:altLabel	"Tumor of Bladder"
	skos:definition	"A benign or malignant…"
	skos:broader	ncit:C2900 "Bladder Disorder"
	skos:broader	ncit:C3431 "Urinary System Neoplasm"

These examples assume a basic knowledge of Linked Data standards and principles, RDF, etc., and a basic knowledge of SKOS. See extra slides for some additional background on these standards.

NCIT has a unique URI minted by the owing authority.

Once we can uniquely identify the concept we can start to record some basic things about it:

- It's a concept
- It has a preferred label
- It has synonyms – obviously many more than than this one.
- In has a definition.
- It has one or more parents – listing two here to show how a polyhierarchy would work.



Core Capability – vocabulary mapping
Using SKOS (2)

MeSH – “Urinary Bladder Neoplasms”

URI = http://purl.bioontology.org/ontology/MSH/D001749
Shorthand: mesh:D001749

Subject (S)	Predicate (P)	Object (O)
mesh:D001749	a	skos:Concept
	skos:prefLabel	“Urinary Bladder Neoplasms”
	skos:altLabel	“Cancer of Bladder”
	skos:definition	“Tumors or cancer of the URINARY BLADDER…”
	skos:broader	mesh:D014571 (“Urogenital Neoplasms”)
	skos:broader	mesh:D001745 (“Urinary Bladder Diseases”)

7 Jim Morris | 26 Jan 2014

AstraZeneca | R&D

Here’s the same concept in MeSH – D001749. Here we’re using an ID assigned in the NCBO’s BioPortal (http://bioportal.bioontology.org)

Once we can uniquely identify the concept we can start to record some basic things about it:

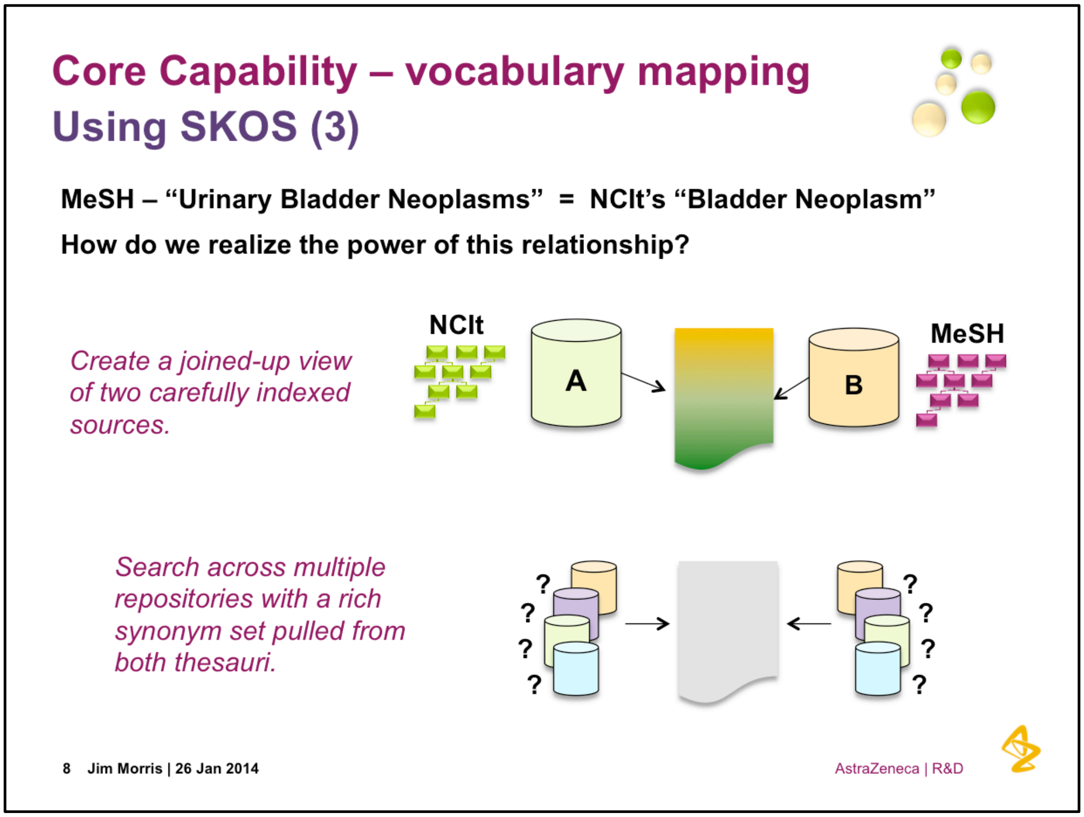
- It’s a concept
- It has a preferred label
- It has synonyms (“Entry Terms” in MeSH) – obviously many more than this one.
- In has a definition.

It has one or more parents. Listing two here.

A couple critical points here:

- If you look at the full record for this concept in Mesh, it does in fact have the NCIT preferred label as a synonym.
- But NCIt does not have the Mesh preferred name as a synonym.
- Note especially that alternate labels of “Cancer of the Bladder” in MeSH and “Tumor of the Bladder” in NCIt are unique to each vocabulary.

But – what if we know that mesh:D001749 is the same concept as ncit:C2901!



Core Capability – vocabulary mapping Using SKOS (3)

MeSH – "Urinary Bladder Neoplasms" = NCIt's "Bladder Neoplasm"

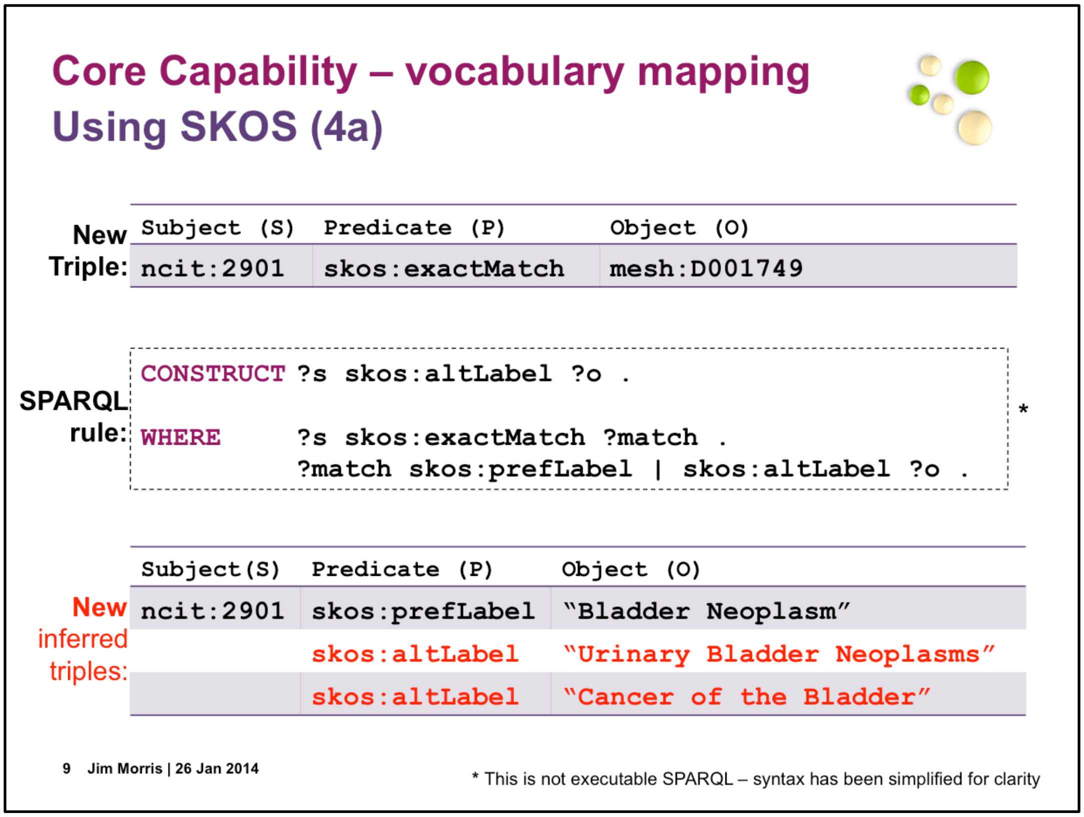
How do we realize the power of this relationship?

Create a joined-up view of two carefully indexed sources.

Search across multiple repositories with a rich synonym set pulled from both thesauri.

What if you're searching two carefully indexed repositories – one indexed using MeSH and another using NCIt. You want to produce a joined up view of these two repositories, leveraging the power of the two separate indexing methods?

What if you're searching across multiple repositories and you want as rich a synonym set for "Bladder Cancer" as you can get. As we just saw – by using only one, you could end up missing things.



Core Capability – vocabulary mapping Using SKOS (4a)

New Triple:

Subject (S)	Predicate (P)	Object (O)
ncit:2901	skos:exactMatch	mesh:D001749

SPARQL rule:

```
CONSTRUCT ?s skos:altLabel ?o .

WHERE      ?s skos:exactMatch ?match .
           ?match skos:prefLabel | skos:altLabel ?o .
```
*

New inferred triples:

Subject(S)	Predicate (P)	Object (O)
ncit:2901	skos:prefLabel	“Bladder Neoplasm”
	skos:altLabel	“Urinary Bladder Neoplasms”
	skos:altLabel	“Cancer of the Bladder”

9 Jim Morris | 26 Jan 2014

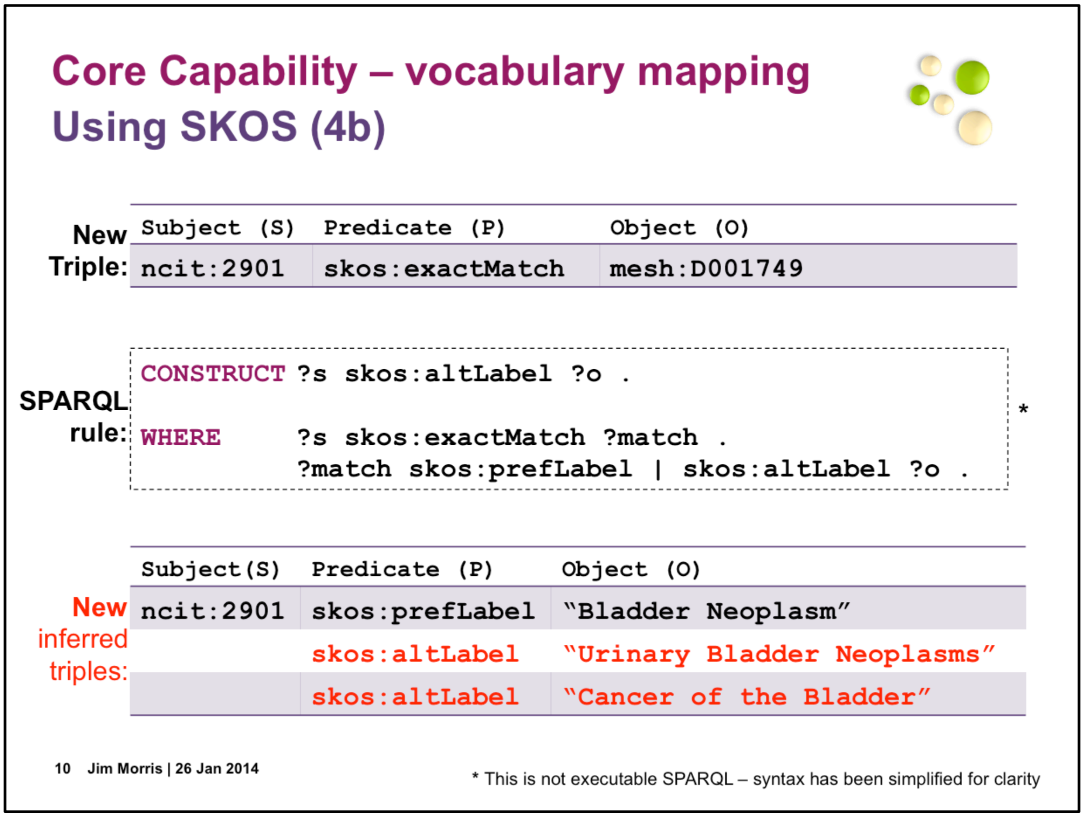
* This is not executable SPARQL – syntax has been simplified for clarity

Here’s how we use Linked Data to do really interesting things. This is a key slide.

New Triple: ncit:2901 skos:exactMatch mesh:D001749
By creating one additional “triple” that identifies the match between the two concept URI, we’re saying that the concept called “Bladder Cancer in NCIt is *the same thing, the same concept,* as the concept called “Urinary Bladder Neoplasms” in MeSH.

We can then exploit that new relationship by creating, essentially, a rule:
SPARQL rule:
We haven’t talked about SPARQL – but it is akin to SQL, in that it is the language with which to query and manipulate Linked Data in the semantic web

New Inferred Triples:
This three line SPARQL command takes that new relationship and *reasons* that the preferred labels or synonyms in the matched MeSH concept can be *inferred* as additional synonyms (in red) in the NCIt concept. It would do this for all the concepts in the NCIt that have been matched to Mesh terms. Those terms “reasoning” and “inferred” are very important. First you’re making a rule that a computer can use to infer additional information. Second--that’s all this has to be--an inference. We are not necessarily hard-coding additional data. We are not manipulating the source vocabularies at all. Every inferred triple can be managed completely separately from either source vocabulary!



Core Capability – vocabulary mapping
Using SKOS (4b)

New Triple:

Subject (S)	Predicate (P)	Object (O)
ncit:2901	skos:exactMatch	mesh:D001749

SPARQL rule:

```
CONSTRUCT ?s skos:altLabel ?o .

WHERE      ?s skos:exactMatch ?match .
           ?match skos:prefLabel | skos:altLabel ?o .
```

*

New inferred triples:

Subject(S)	Predicate (P)	Object (O)
ncit:2901	skos:prefLabel	"Bladder Neoplasm"
	skos:altLabel	"Urinary Bladder Neoplasms"
	skos:altLabel	"Cancer of the Bladder"

10 Jim Morris | 26 Jan 2014

* This is not executable SPARQL – syntax has been simplified for clarity

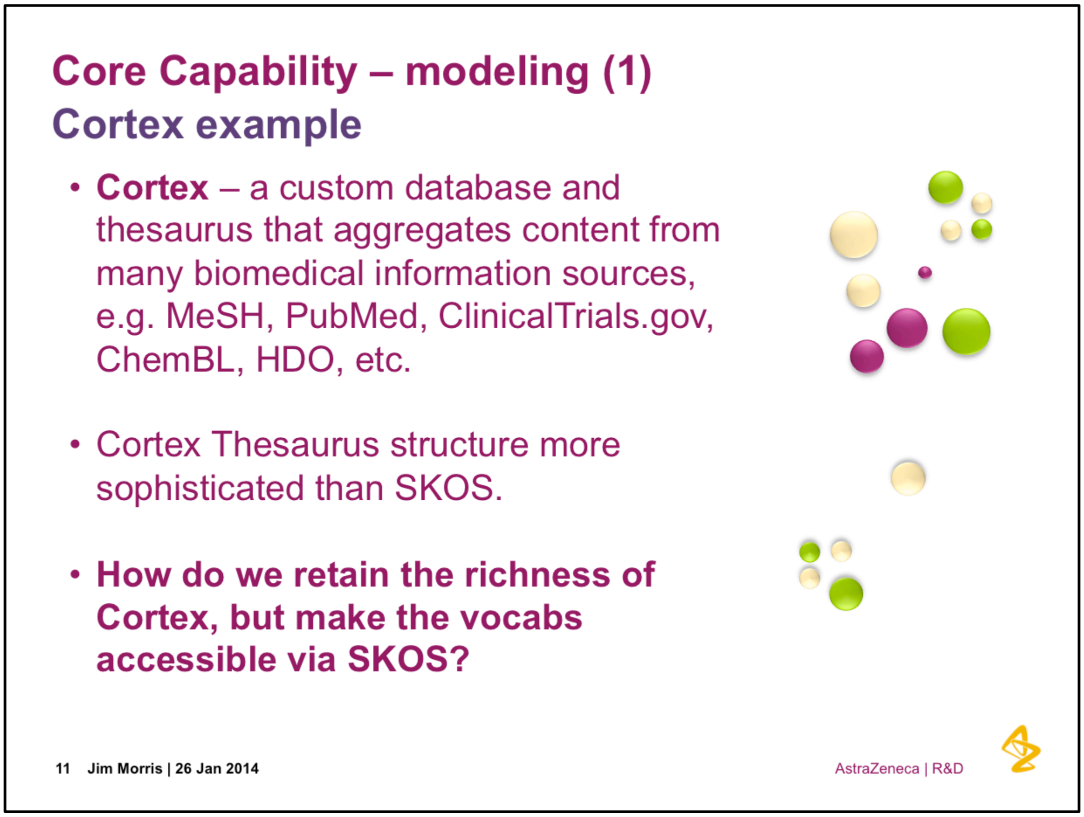
More about this slide, the power of mapping using SKOS and building inference rules that exploit it…

What about UMLS (unified medical language system) or NCI Metathesaurus and other large mapping initiatives? Aren't they solving this problem? Yes: But…

1. Mesh and NCIt are just two familiar examples. What if you trying to map vocabularies or search across repositories that use less standard vocabularies, proprietary, internal, or bespoke vendor vocabs? Or if you needed to search, and view information in an interface using the richness of NCIt, but against an repository that uses something different – a folksonomy perhaps, or just values selected from a drop-down list.

2. What if the mapping you wanted to do was proprietary to your project, your company? How concepts are arranged and mapped to each other has power. What if that mapping was relevant today – but not tomorrow? How can this be accomplished by leveraging these large published vocabularies – without corrupting them, or worse -- building *yet another one*?

Now I want to be clear. Identifying those initial matches--and they are not always as straightforward as an "exactMatch"--is the hard intellectual work of vocabulary management that is, and will continue to be a hallmark of the taxonomy profession. The semantic web doesn't solve the problem of knowing which terms to match. **But** what it does is--unlike any other taxonomy, thesaurus, vocabulary management system--is give you the tools to encode those matches and create rules that computers can use to make additional inferences. And very importantly, it does not require you to MERGE or INTEGRATE or otherwise CORRUPT original authoritative sources.



Core Capability – modeling (1)
Cortex example

- **Cortex** – a custom database and thesaurus that aggregates content from many biomedical information sources, e.g. MeSH, PubMed, ClinicalTrials.gov, ChemBL, HDO, etc.
- Cortex Thesaurus structure more sophisticated than SKOS.
- **How do we retain the richness of Cortex, but make the vocabs accessible via SKOS?**

11 **Jim Morris | 26 Jan 2014**

AstraZeneca | R&D

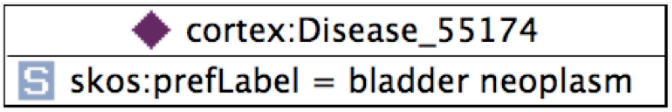
This next example is looking at the "Modeling" capability necessary for interoperable vocabularies. In order for vocabularies to interoperable they need to conform to a standard data format—at least at some higher level.

This is the key: we want to be able to retain the native richness and integrity of native sources; but they need to conform to a common data standard to be interoperable. Semantic Web standards and the SKOS ontology are designed to allow for this.

Think of this more broadly as well. Beyond the domain of vocabularies, imagine how this would apply to datasets of your own particular domain. How would this apply to bibliographic standards used in different library systems, institutional repositories, union catalogs, and even the web itself? This is a very hot topic in libraries these days.

Core Capability – modeling (2)

Using one disease term as an example…



◆ cortex:Disease_55174

S skos:prefLabel = bladder neoplasm

(screen shots are from TopBraid Composer)

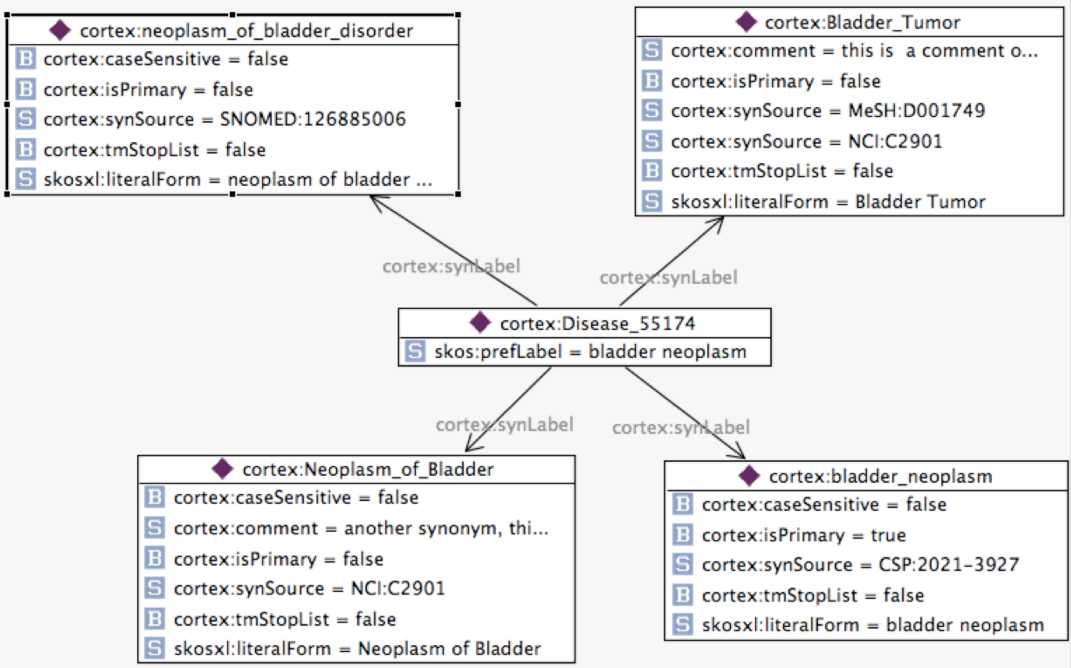
TopBraid Composer™

These screen shots are from TopBraid Composer – Standard Edition.

This screen shot is stating that:

- there is a concept with the identifier, “cortex:Disease_55174”.
- The concept has a preferred label of “bladder neoplasm”.

Core Capability – modeling (3)



◆ cortex:neoplasm_of_bladder_disorder
B cortex:caseSensitive = false
B cortex:isPrimary = false
S cortex:synSource = SNOMED:126885006
B cortex:tmStopList = false
S skosxl:literalForm = neoplasm of bladder ...

◆ cortex:Bladder_Tumor
S cortex:comment = this is a comment o...
B cortex:isPrimary = false
S cortex:synSource = MeSH:D001749
S cortex:synSource = NCI:C2901
B cortex:tmStopList = false
S skosxl:literalForm = Bladder Tumor

◆ cortex:Disease_55174
S skos:prefLabel = bladder neoplasm

◆ cortex:Neoplasm_of_Bladder
B cortex:caseSensitive = false
S cortex:comment = another synonym, thi...
B cortex:isPrimary = false
S cortex:synSource = NCI:C2901
B cortex:tmStopList = false
S skosxl:literalForm = Neoplasm of Bladder

◆ cortex:bladder_neoplasm
B cortex:caseSensitive = false
B cortex:isPrimary = true
S cortex:synSource = CSP:2021-3927
B cortex:tmStopList = false
S skosxl:literalForm = bladder neoplasm

This screen shot shows the relationship of the core concept to properties that are kept in other related records.

Like many of our most important published vocabularies, the structure of Cortex is very complex. Because of that, you can’t run the SKOS query we saw in the previous example against this thesaurus.

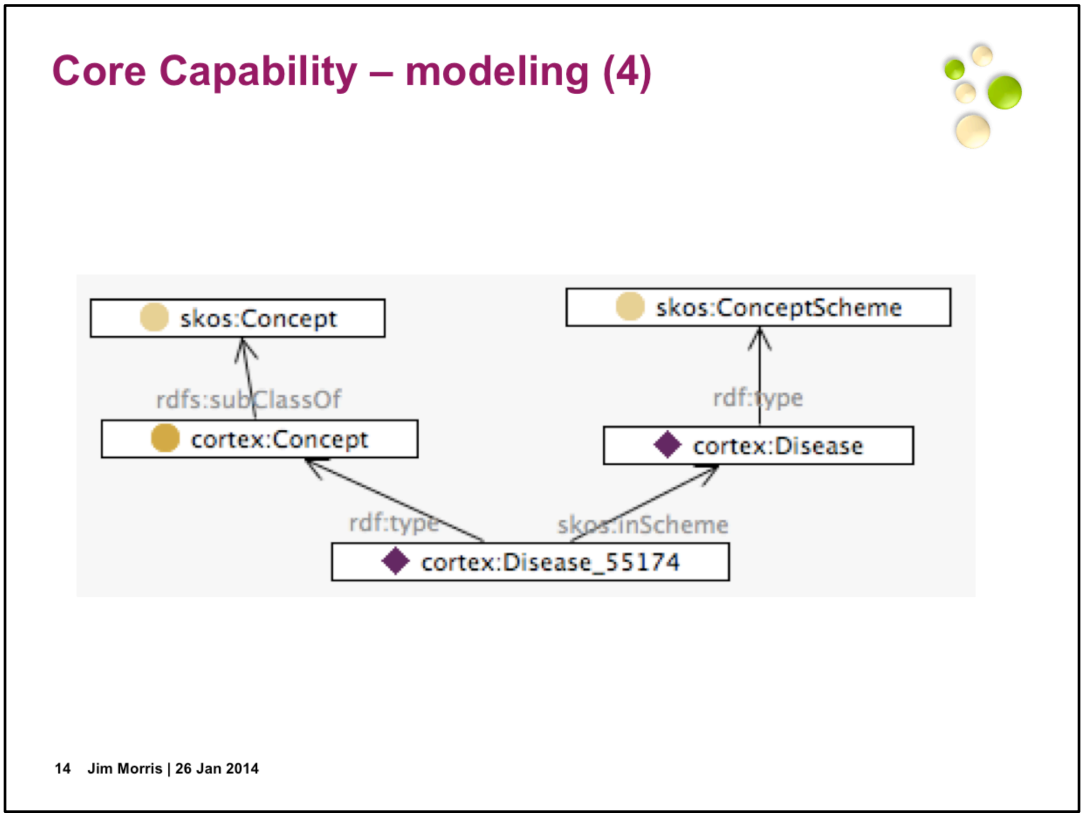
The main reason for this is because the synonyms or “altLabel” properties are stored in separate records, instead of as properties of the main concept. This was done for a good reason: so that additional properties could be added to each synonym. For example, some synonyms may be inappropriate for text-mining applications.

Following the relationships in this picture you can see that the concept cortex:Disease_55174 (“bladder neoplasm”) has four related records. Each of these records holds one synonym. And for each synonym there are a set of properties. For example, in the upper right you can see the synonym labeled “Bladder Tumor”. One of the properties is “tmStopList” and the value is “false”. This means that “Bladder Neoplasm” is a not a “stop term”, and is valid for text-mining. You might want this value to be true if the synonym was a common word that would mislead the indexing engine. This is common with certain acronyms.

This is also is an example of a mapped vocabulary. Looking at the “synSource” property you can see that the synonyms were derived from common biomedical vocabularies like MeSH, NCI, and SNOMED.

So there is a lot of added value in Cortex; value we don’t want to lose. And that is precisely the benefit of using the linked data approach. We don’t want to corrupt Cortex. We want to make it more *interoperable* by modeling it in SKOS.

On the next slides we’ll see how we can make all the synonyms to appear as “altLabels” of the primary concept, instead of as separate records. But modeling this vocabulary in SKOS does not “dumb it down”. Instead it will retain it’s native integrity, while making it more widely useful.



Core Capability – modeling (4)

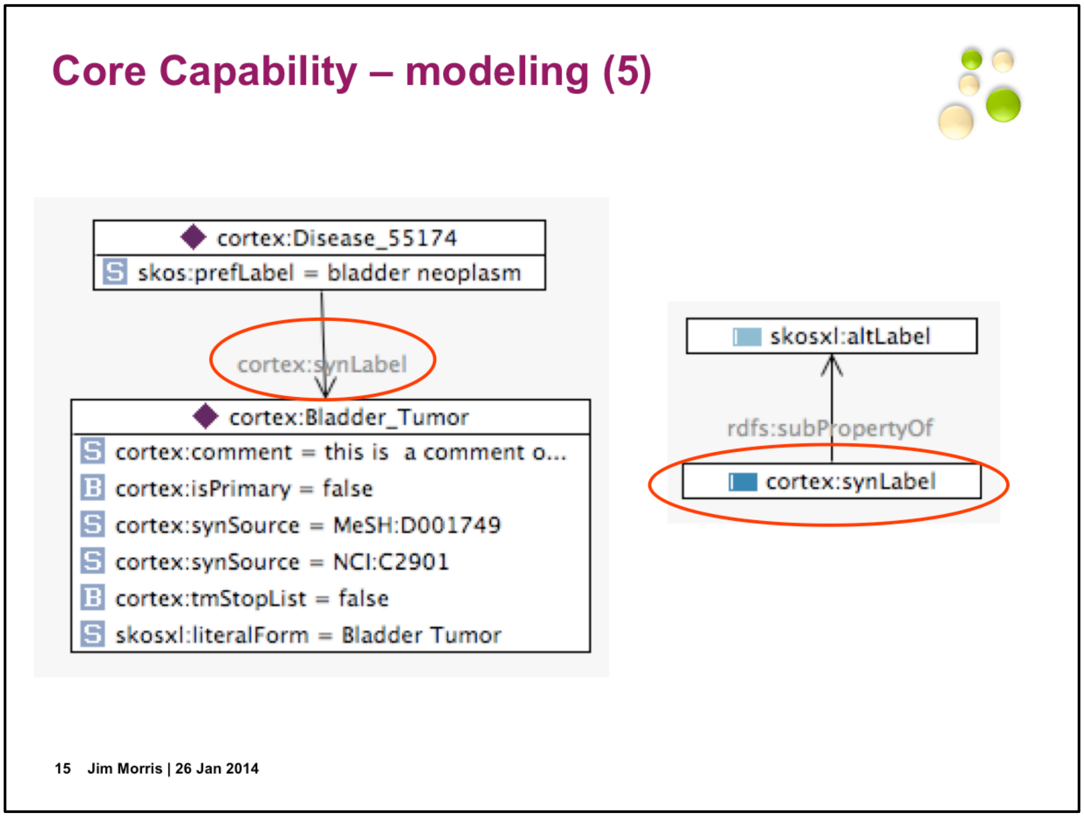
14 Jim Morris | 26 Jan 2014

Before dealing specifically with the synonyms we want to make the vocabulary structure align with SKOS properties or classes right away. On this screen:

- We've said that that concept cortex:Disease_55174 is a cortex-specific concept class we've called cortex:Concept.
- Next we've made cortex:Concept a subClass of skos:Concept. This means that all cortex concepts can be found by looking for basic skos:Concepts. They'll also inherit aspects of the parent skos:Concept class.
- We've also set up a concept scheme, cortex:Disease and identified it as a valid skos:ConceptScheme. SKOS ConceptSchemes are generally used to define a particular vocabulary.

Since the cortex concepts are now skos:Concepts we can use basic properties like skos:prefLabel to encode the primary label.

So now, just with these high-level modeling decisions, we would be able to produce a list of cortex concepts using a standard, simple, SKOS query. Otherwise we'd be forced to learn the proprietary query language of the host system. This alone is a big step forward towards interoperability!



Core Capability – modeling (5)

◆ cortex:Disease_55174

S skos:prefLabel = bladder neoplasm

◆ cortex:Bladder_Tumor

S cortex:comment = this is a comment o...

B cortex:isPrimary = false

S cortex:synSource = MeSH:D001749

S cortex:synSource = NCI:C2901

B cortex:tmStopList = false

S skosxl:literalForm = Bladder Tumor

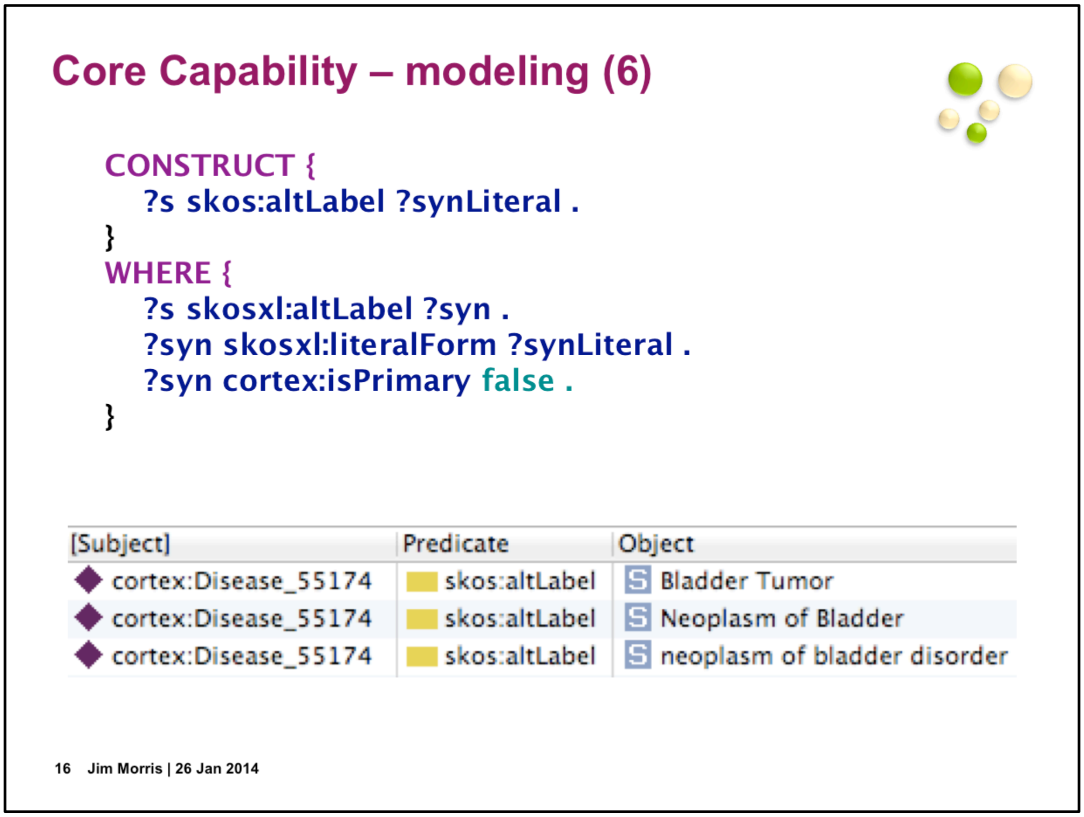
15 Jim Morris | 26 Jan 2014

Here we are exploiting some of the properties of SKOS-XL, the SKOS “extension for labels”.

SKOS-XL was created to support essentially what Cortex is doing: managing synonyms and other non-preferred labels as distinct entities so that additional properties can be added to them. So it makes sense to see if any of the SKOS-XL properties or classes can be applied here.

On the left is just an example of the core concept with one of its related synonym records. You can see that we’ve already used the SKOS-XL property “literalForm” for the label of the synonym.

On the right we’ve said that the cortex-specific property that manages the relationship between the synonym and the core concept (cortex:synLabel) is a subproperty of the SKOS-XL property skosxl:altLabel. Again we’re only doing this so that standard SKOS queries, using standard SKOS or SKOS-XL properties can be used to access this vocabulary.



Core Capability – modeling (6)

```
CONSTRUCT {
   ?s skos:altLabel ?synLiteral .
}
WHERE {
   ?s skosxl:altLabel ?syn .
   ?syn skosxl:literalForm ?synLiteral .
   ?syn cortex:isPrimary false .
}
```

[Subject]	Predicate	Object
cortex:Disease_55174	skos:altLabel	Bladder Tumor
cortex:Disease_55174	skos:altLabel	Neoplasm of Bladder
cortex:Disease_55174	skos:altLabel	neoplasm of bladder disorder

Now that we've mapped as many Cortex properties and classes to standard SKOS properties and classes, we want to construct a rule that does what we set out to do: make all the synonyms that are in related records actually properties of the core concept.

To figure out what the rule should be, we can experiment with queries that will create the necessary new triples, in this case creating the skos:altLabels on the main concept.)

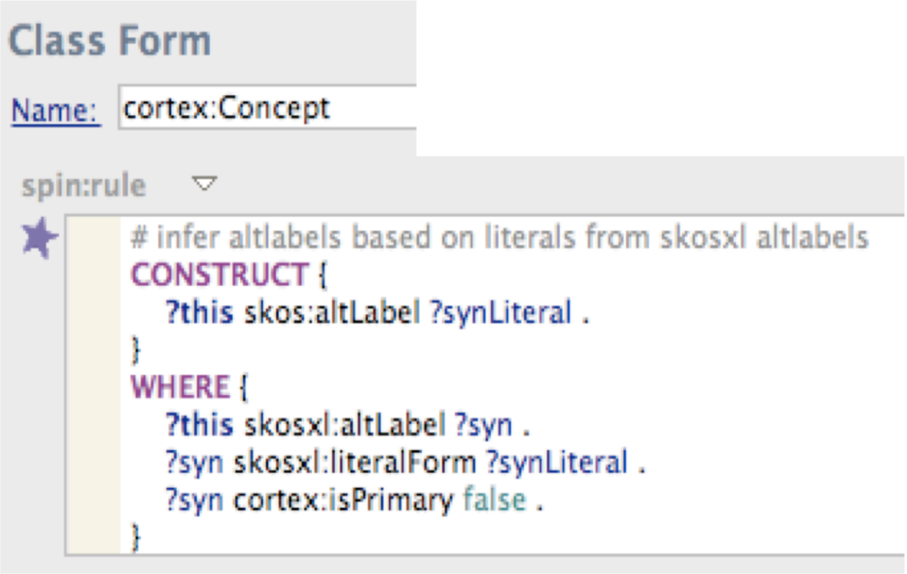
This SPARQL query is saying:

1. look for any related synonym records (?s skosxl:altLabel ?syn.)
2. Get the literal values from those records (?syn skosxl:literalForm ?synLiteral .)
3. Only get those values if the "isPrimary" flag is false.

The objective here isn't to teach SPARQL, but the key point here is that, *as much as possible*, we are using standard SKOS or SKOS-XL terminology in our query. There is only one part of this query that is specific to Cortex. The original modeling we did on the previous slides is what allows us to do this.

The results of this query are exactly what we want: we want those labels to be pulled back as properties of the core concept. Here in the resulting triples we see each of the values now as altLabels of the core concept.

Core Capability – modeling (7)



Class Form

Name: cortex:Concept

spin:rule

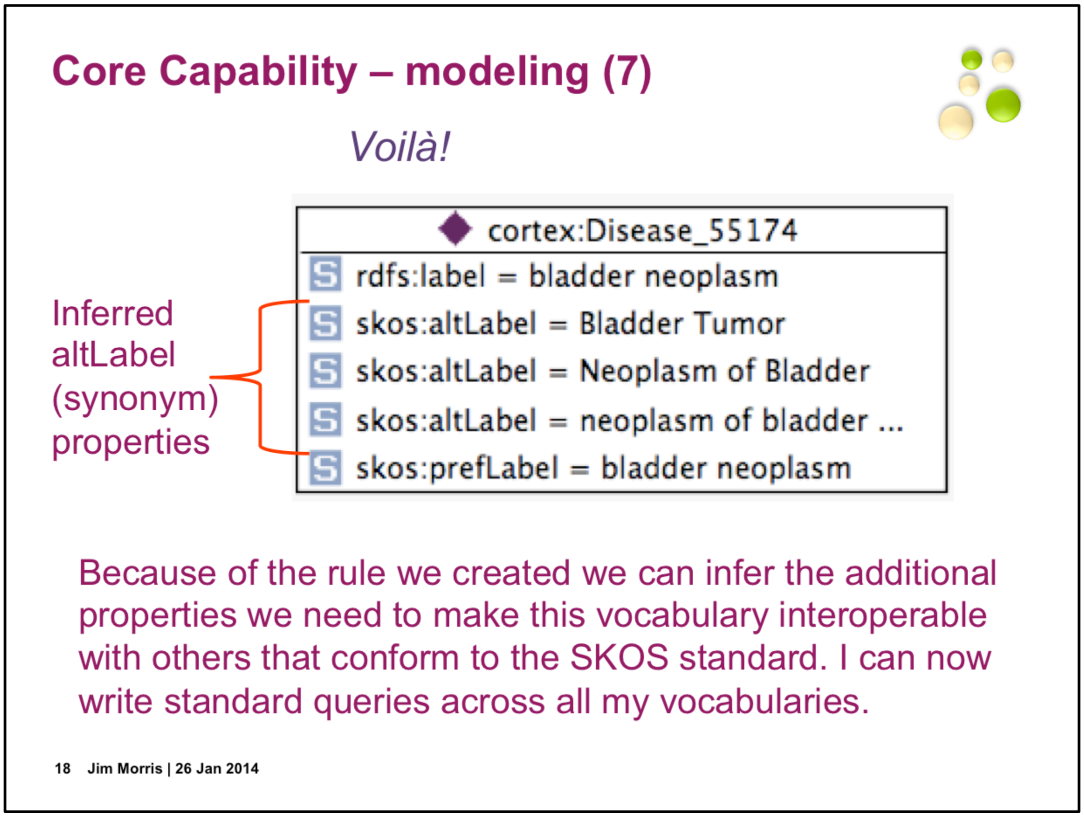
```
# infer altlabels based on literals from skosxl altlabels
CONSTRUCT {
   ?this skos:altLabel ?synLiteral .
}
WHERE {
   ?this skosxl:altLabel ?syn .
   ?syn skosxl:literalForm ?synLiteral .
   ?syn cortex:isPrimary false .
}
```

But it’s not enough to just create the query. Now, just as we did in the previous example, we will turn that query into rule that can be stored and used when needed.

Using SPIN, which stands for SPARQL Inferencing Notation, we can turn SPARQL queries into rules.

SPIN rules can be associated with RDF resources. In this case we’ve associated the rule with the cortex:Concept class. This is why we modeled the cortex concepts as their own class and then made them subClasses of the skos:Concept class. We want to associate this rule only with Cortex concepts – not every SKOS concept in the world! This is a key point—we can have our cake and eat it, too. We can manage rules and processes specific to a vocabulary, but through effective modeling, we can also make those vocabularies more useful.

SPIN rules can be turned on and off as part of Inferencing. If we turn this rule on, the new triples will be *inferred.* And then we can choose to do what we want with those inferred triples.



Core Capability – modeling (7)

Voilà!

cortex:Disease_55174

S rdfs:label = bladder neoplasm

S skos:altLabel = Bladder Tumor

S skos:altLabel = Neoplasm of Bladder

S skos:altLabel = neoplasm of bladder ...

S skos:prefLabel = bladder neoplasm

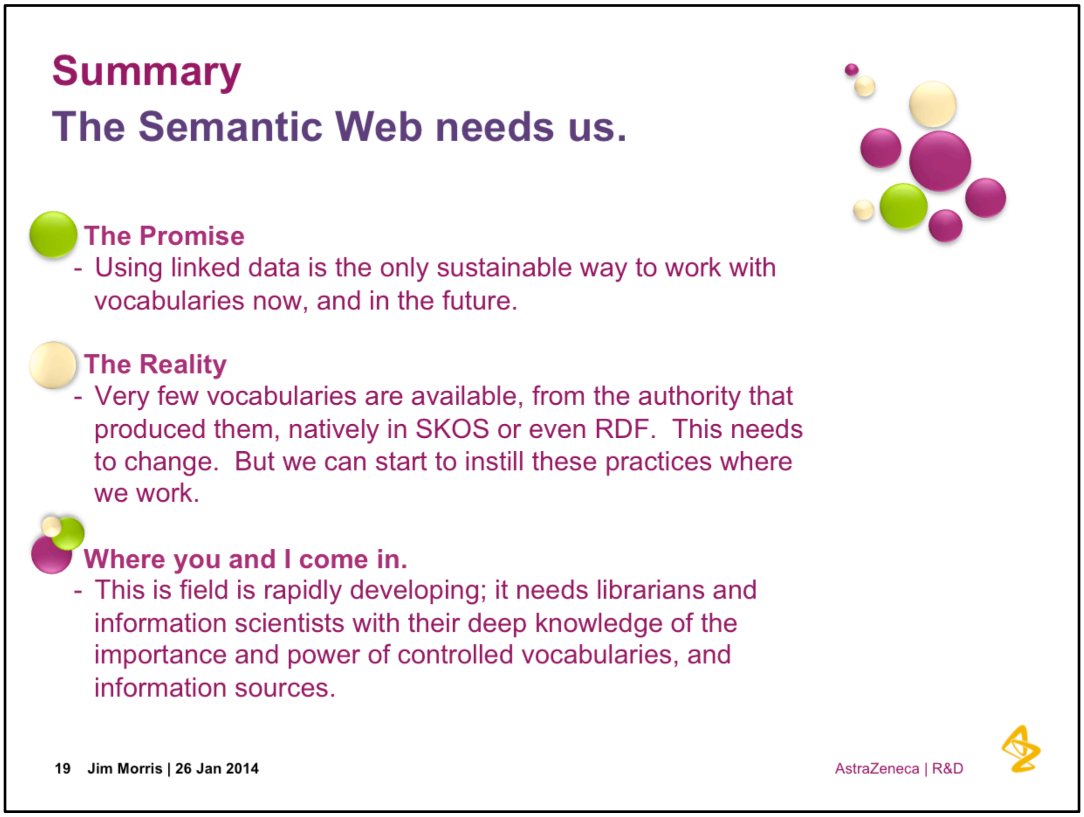
Inferred altLabel (synonym) properties

Because of the rule we created we can infer the additional properties we need to make this vocabulary interoperable with others that conform to the SKOS standard. I can now write standard queries across all my vocabularies.

Here we see the SPIN rule in action. As soon as it is active, we see additional skos:altLabel properties of our original concept appear before our eyes.

Now, our custom vocabulary can be seen as a standardized SKOS vocabulary--interoperable with other vocabularies in our network. I can now write the same queries across all my vocabularies! When I design new systems that use vocabularies, I can just write standard interfaces using SKOS, and plug-and-play whichever vocabulary is appropriate.

Note especially that I have not altered my original vocabulary structure in any way. Everything that you could do with Cortex in it's native application, you can do here. But in addition, you can now use a standard language to make it interoperable with other vocabularies and systems.



Summary
The Semantic Web needs us.

The Promise

- Using linked data is the only sustainable way to work with vocabularies now, and in the future.

The Reality

- Very few vocabularies are available, from the authority that produced them, natively in SKOS or even RDF. This needs to change. But we can start to instill these practices where we work.

Where you and I come in.

- This is field is rapidly developing; it needs librarians and information scientists with their deep knowledge of the importance and power of controlled vocabularies, and information sources.

19 Jim Morris | 26 Jan 2014

AstraZeneca | R&D

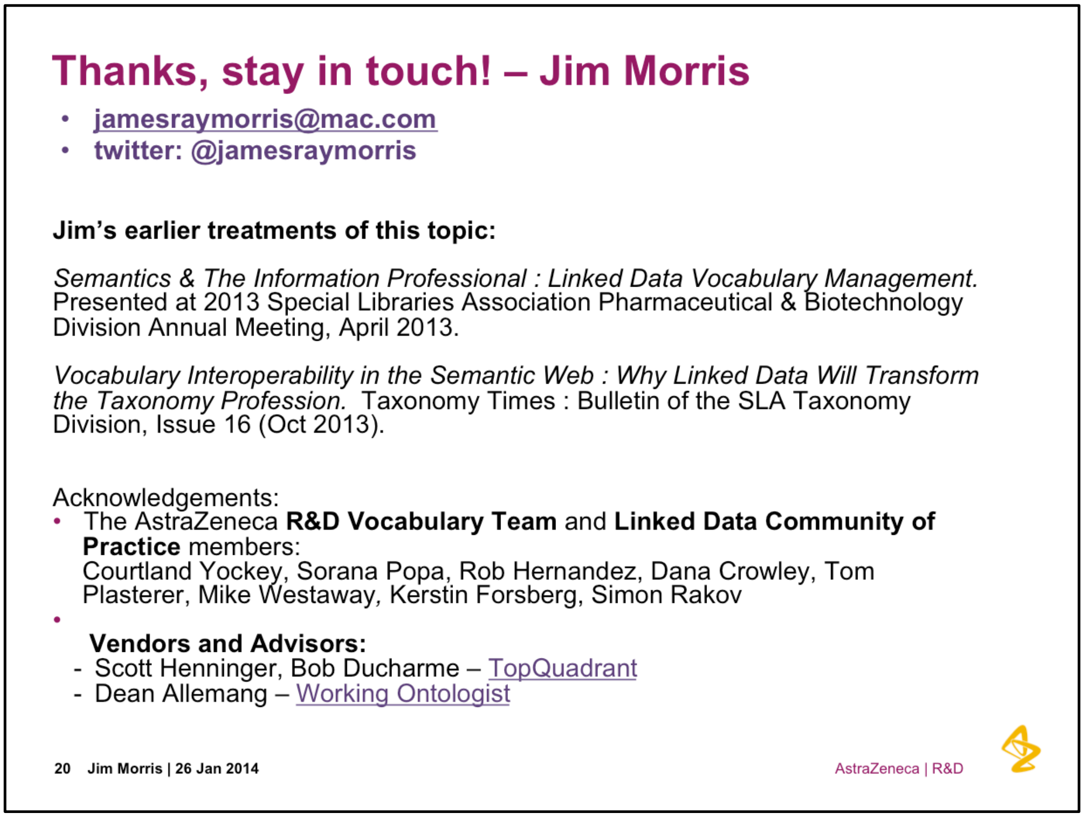
The Promise and the Reality. Linked Data approaches are the sustainable way to work with vocabularies in an increasingly federated and linked information world. The reality is that not enough vocabularies are available, from the authority that produced them, natively in SKOS or another form of RDF. This needs to change. But we can start to instill these practices where we work. This is field is rapidly developing; it needs librarians and information scientists with their deep knowledge of the importance and power of controlled vocabularies and the information sources that use them.

This is not the future--it's now. Leading organizations in the US, like OCLC and the Library of Congress, are adopting these standards. This even more true in Europe. The field of biomedical ontologies is very exciting right now, and firmly moving toward a Linked Data paradigm. The nature of business and of information demands new approaches. Even taking the Semantic Web's promise of universally sharing data across the world out of the picture, SKOS and RDF--even just applied to vocabularies--is opportunity alone. These vocabularies can be published in whatever form people or systems need them in. However, to manage a web of vocabularies requires tools and techniques designed for the web.

This is not easy. Breaking down information, including vocabularies into their most basic constructs, is what RDF does. It is elegant in design, but can quickly become very complex in execution. Writing the SPARQL queries that do complete transformations of vocabularies, manage versions, etc. can become very sophisticated. But isn't it a challenge worth tackling? Aren't librarians and information scientists poised to make this happen?

SKOS is a great example of a simplified model with practical application. RDF can be used to model ontologies in linguistically accurate, but extraordinarily complex, ways as well. The discussions at a Bioontology or Semantic Web conference can get very deep. But the beauty of RDF is that those complex models can be transformed, again without corrupting the source, into a simple model like SKOS. Guarding against unnecessary complexity is something that librarians are especially good at.

Managing unconnected silos of information, including taxonomies, will not harness the power of a networked, collaborative world. Vocabulary interoperability, enabled by the Linked Data principles of the Semantic Web will--if further developed as a discipline within our profession.



Thanks, stay in touch! – Jim Morris

- jamesraymorris@mac.com
- twitter: @jamesraymorris

Jim’s earlier treatments of this topic:

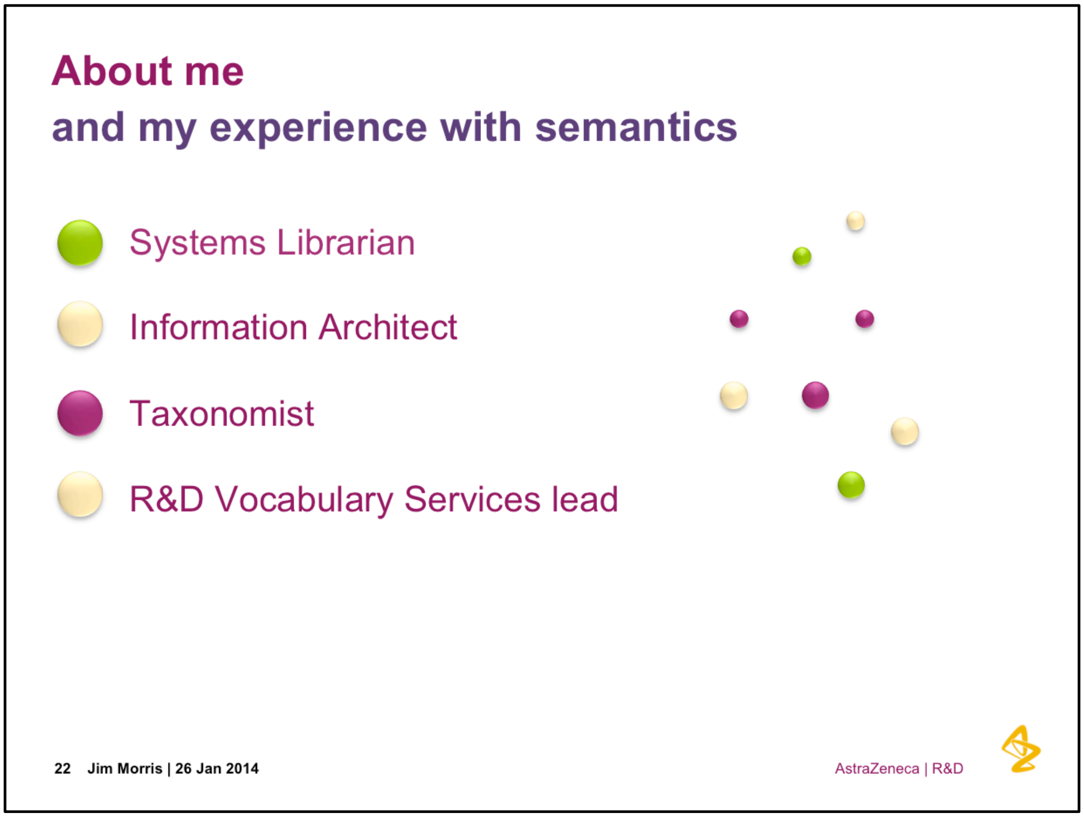
Semantics & The Information Professional : Linked Data Vocabulary Management. Presented at 2013 Special Libraries Association Pharmaceutical & Biotechnology Division Annual Meeting, April 2013.

Vocabulary Interoperability in the Semantic Web : Why Linked Data Will Transform the Taxonomy Profession. Taxonomy Times : Bulletin of the SLA Taxonomy Division, Issue 16 (Oct 2013).

Acknowledgements:

- The AstraZeneca **R&D Vocabulary Team** and **Linked Data Community of Practice** members:
  Courtland Yockey, Sorana Popa, Rob Hernandez, Dana Crowley, Tom Plasterer, Mike Westaway*,* Kerstin Forsberg, Simon Rakov
- **Vendors and Advisors:**
  - Scott Henninger, Bob Ducharme – TopQuadrant
  - Dean Allemang – Working Ontologist

Extra slides…



About me
and my experience with semantics

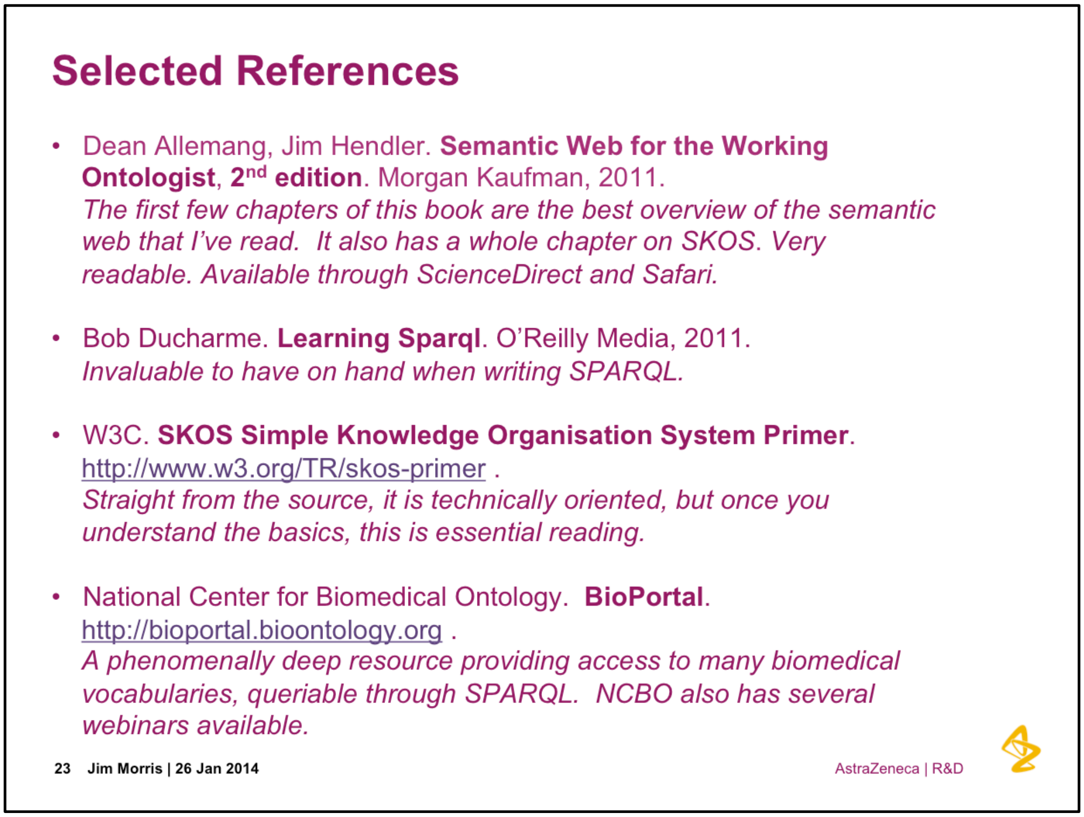
- Systems Librarian
- Information Architect
- Taxonomist
- R&D Vocabulary Services lead

Systems Librarian
Starting around 1989 was responsible for and helped implement several library-oriented systems. While working at OCLC I had my first detailed experience with a thesaurus management system. I helped write a module that integrated MARC authority records with the thesaurus manager.

Information Architect
At AZ, around 2006, I led a team of librarian/webmasters that managed all the webcontent for the global library organisation.

Taxonomist
While working with enterprise content, I became involved with the global AZ team responsible for the AstraZeneca Enterprise Taxonomy.

R&D Vocabulary Capability lead
I first took on this role of vocabulary capability lead within the “R&D Information” department, working on taking data interoperability to the next level. It was clear that leveraging semantic web technology was the most sustainable way forward for vocabulary management.



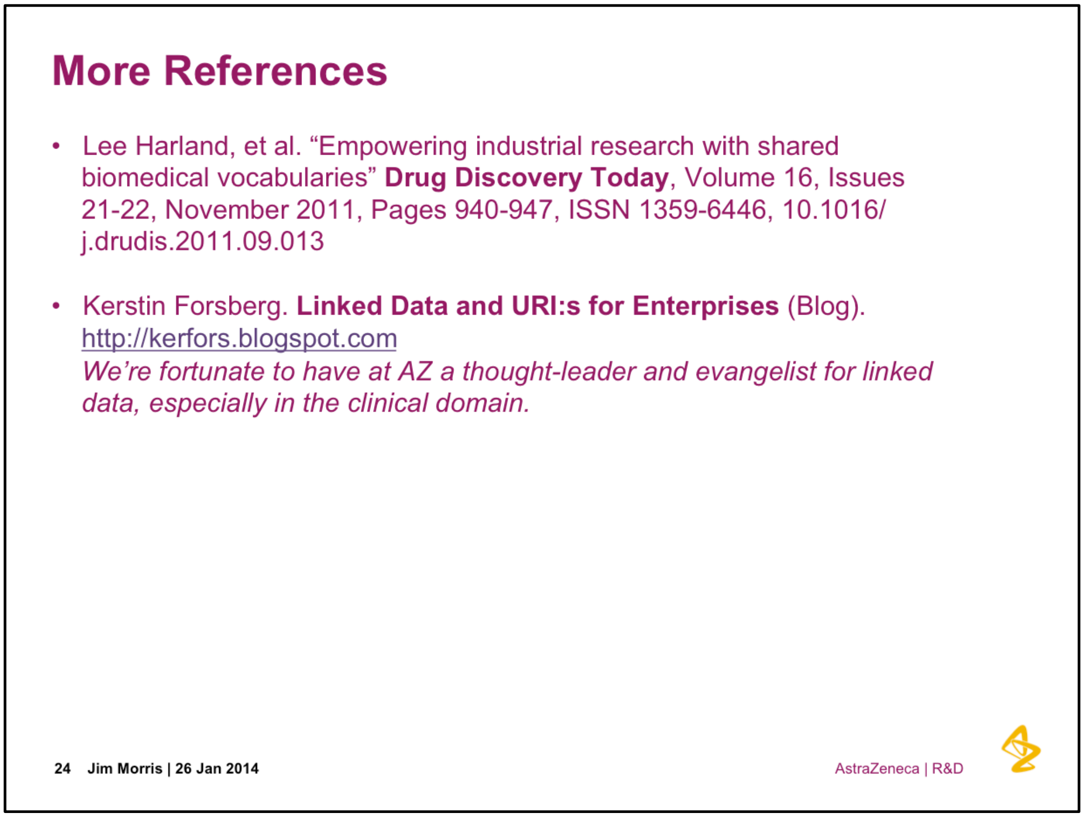
Selected References

- Dean Allemang, Jim Hendler. **Semantic Web for the Working Ontologist**, **2nd edition**. Morgan Kaufman, 2011.
  The first few chapters of this book are the best overview of the semantic web that I've read. It also has a whole chapter on SKOS. Very readable. Available through ScienceDirect and Safari.

- Bob Ducharme. **Learning Sparql**. O'Reilly Media, 2011.
  Invaluable to have on hand when writing SPARQL.

- W3C. **SKOS Simple Knowledge Organisation System Primer**.
  http://www.w3.org/TR/skos-primer .
  Straight from the source, it is technically oriented, but once you understand the basics, this is essential reading.

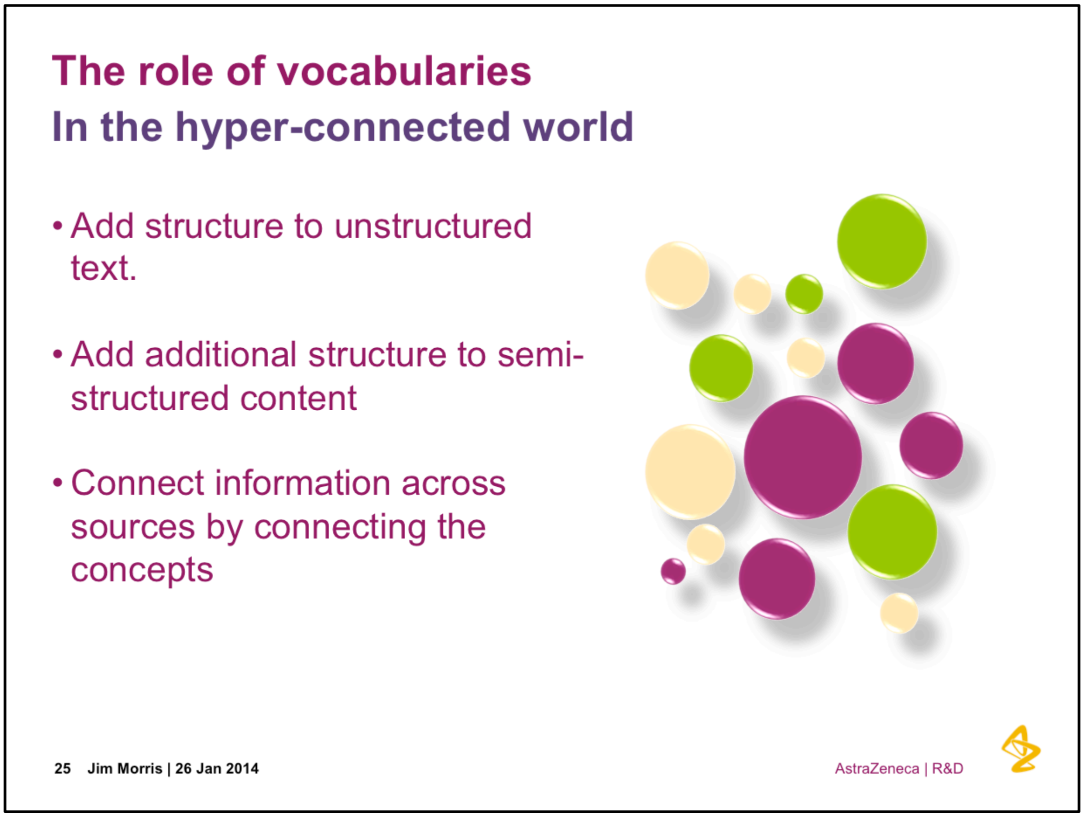
- National Center for Biomedical Ontology. **BioPortal**.
  http://bioportal.bioontology.org .
  A phenomenally deep resource providing access to many biomedical vocabularies, queriable through SPARQL. NCBO also has several webinars available.



More References

- Lee Harland, et al. “Empowering industrial research with shared biomedical vocabularies” **Drug Discovery Today**, Volume 16, Issues 21-22, November 2011, Pages 940-947, ISSN 1359-6446, 10.1016/j.drudis.2011.09.013

- Kerstin Forsberg. **Linked Data and URI:s for Enterprises** (Blog). http://kerfors.blogspot.com
  We’re fortunate to have at AZ a thought-leader and evangelist for linked data, especially in the clinical domain.



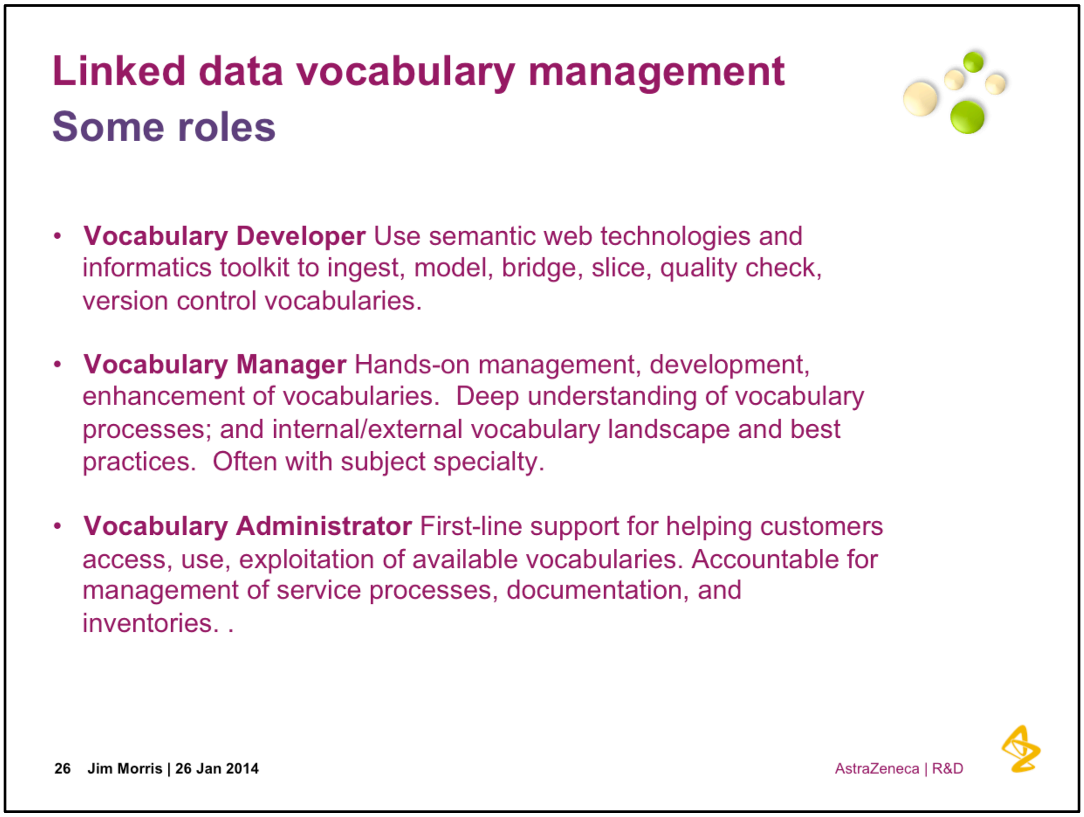
The role of vocabularies
In the hyper-connected world

- Add structure to unstructured text.
- Add additional structure to semi-structured content
- Connect information across sources by connecting the concepts

Adding structure to unstructured text; essentially tagging unstructured content with vocabulary term identifiers, by looking for strings that match a wide-array of synonyms, and by other more sophisticated algorithms.

Adding additional structure to already indexed content so that we can organize information in ways specific to our requirements.

Adding vocabulary metadata to the content or to a search index to connect information across repositories of information by connecting the concepts embedded in that information.



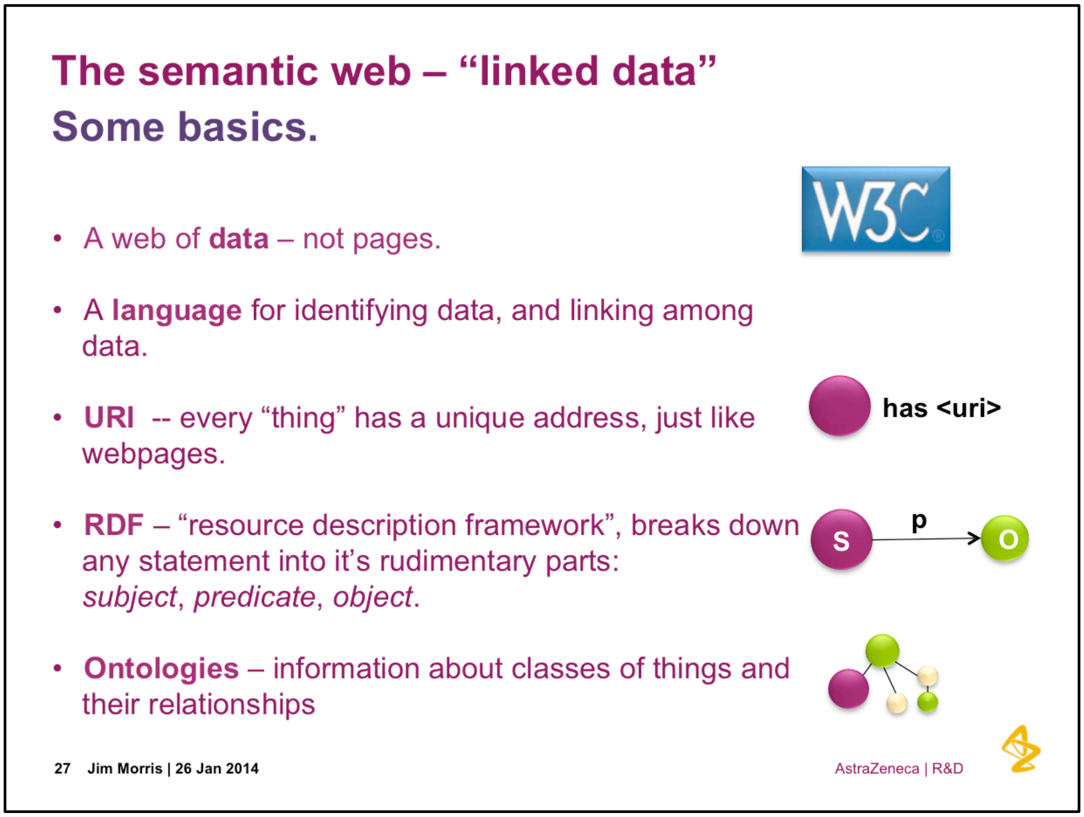
Linked data vocabulary management
Some roles

- **Vocabulary Developer** Use semantic web technologies and informatics toolkit to ingest, model, bridge, slice, quality check, version control vocabularies.
- **Vocabulary Manager** Hands-on management, development, enhancement of vocabularies. Deep understanding of vocabulary processes; and internal/external vocabulary landscape and best practices. Often with subject specialty.
- **Vocabulary Administrator** First-line support for helping customers access, use, exploitation of available vocabularies. Accountable for management of service processes, documentation, and inventories. .

26 Jim Morris | 26 Jan 2014

AstraZeneca | R&D

In thinking about how to develop and deliver this service based on these new approaches, we've begun drafting some role descriptions.



The semantic web – “linked data”
Some basics.

- A web of **data** – not pages.
- A **language** for identifying data, and linking among data.
- **URI** -- every “thing” has a unique address, just like webpages.
- **RDF** – “resource description framework”, breaks down any statement into it’s rudimentary parts: *subject*, *predicate*, *object*.
- **Ontologies** – information about classes of things and their relationships

27 Jim Morris | 26 Jan 2014

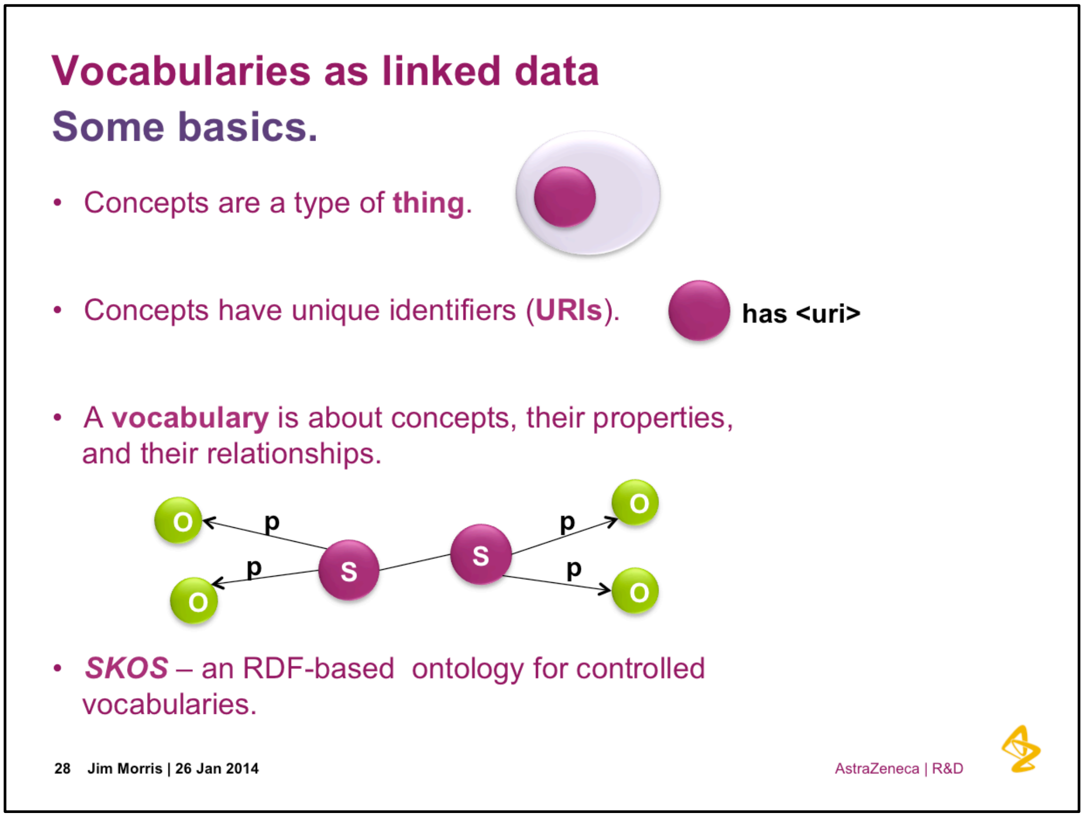
AstraZeneca | R&D

I’ve been using the terms “semantic web” and “linked data” interchangably. The term “semantic web” has fallen out of favor among it’s advocates who instead refer more often to “linked data” – which sounds much more practical and obtainable than the futuristic vision touted in the seminal 2001 Scientific American article by Tim Berners-Lee, Jim Hendler, and Ora Lassila.

1. A web of data – not pages.
Pages need to be read by a person, or theoretically an AI agent mimicking a person, in order to obtain elements of data. For example, I want to learn the times that local bars offer “happy hour” – I need to find the web page for each bar, scan pages, follow links, and finally read content that says something like “we have happy hour M-F from 4-7”. It doesn’t matter that that information might originally be in a database that builds the page. Because to interpret the information I need to read the page. Linked Data challenges us to use the principles of the web--unique identification or resources i.e. URLs, and links between those resources--to the data itself.

URIs, RDF, and Ontologies
URI – “uniform resource identifier”. URL’s are “uniform resource locators”--a type of URI used to uniquely identify pages and locate them. You “point” your browser to them to locate the unique page. In the semantic web, elements of data are also given URIs--using the same syntax. For example The Artful Dodger might have a webpage with a URL. But in the semantic web The Artful Dodger itself--as a thing, not a page, would have a URI.
RDF – the most basic statement about data--subject/predicate/object--a “triple”.
[motion with hands]
For example The Artful Dodger (subject), Is a (predicate), Bar (object); The Artful Dodger (subject), Has a (predicate), Happy Hour (predicate).
Ontologies – collections of information about classes of things and their relationships. For example, in this context - a Bar is a type of Restaurant. If I know that the Artful Dodger (s), Is a (p), Bar (o), then I can infer (an important word) that the Artful Dodger (s) is a (p), restaurant (o).

That’s all I’m going to say about the fundamentals – let’s quickly see how they apply to our vocabularies, and specifically to our core capability of mapping two vocabularies together.



Vocabularies as linked data
Some basics.

- Concepts are a type of **thing**.
- Concepts have unique identifiers (**URIs**).
- A **vocabulary** is about concepts, their properties, and their relationships.
- ***SKOS*** – an RDF-based ontology for controlled vocabularies.

has <uri>
O
p
S
p
O
S
p
O
p
O

28 Jim Morris | 26 Jan 2014

AstraZeneca | R&D

Concepts are Things
Just like The Artful Dodger is a thing. The concept of a “Bar” is also a thing.

Concepts have URIs
Things are identified on the web with URIs

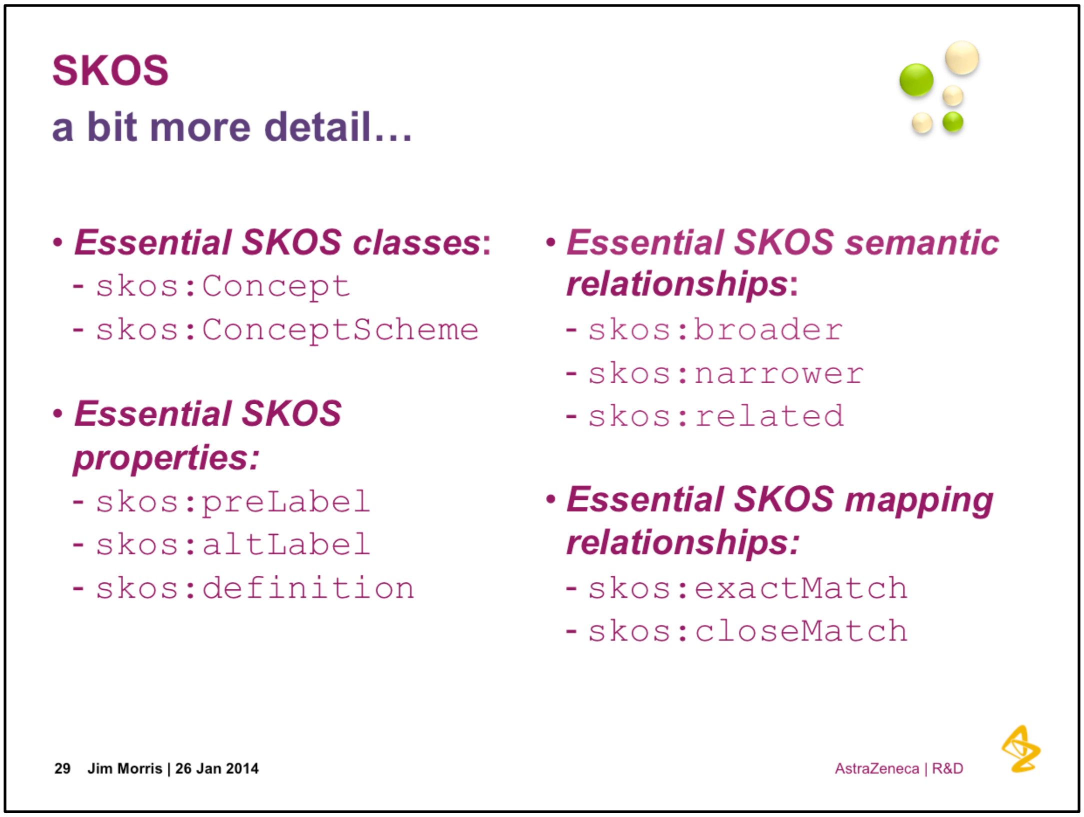
A vocabulary is about concepts, properties, and relationships
Obviously. In fact, it’s interesting to think of vocabularies themselves as little webs of knowledge. Well, they need not be little webs of knowledge unto themselves. They can join the larger web of data.

SKOS –
“Simple Knowledge Organisation System”. SKOS is a standardized set of classes and properties for representing “knowledge organisation systems”, aka vocabularies. Unlike other thesaurus, etc. standards, the fact that SKOS is Linked Data, means it allows for the management of vocabularies in a distributed, linkable, interoperable, way. In other words, as part of the open-ended, web of data.

Lets’ walk through this diagram with a biomedial example using SKOS …
Bladder Cancer (S) is a preferred term of (P) Bladder Cancer (O)
Bladder Cancer (S) has an alternate label – synonym of (P) Bladder Neoplasm (O)
Bladder Cancer (S) has a definition of (P) this… (O)

Bladder Cancer (S) has a broader term (P) of Cancer (O) (which has it’s own properties)

But it’s when we start mapping concepts between vocabularies that we really get the power we need.



SKOS

a bit more detail…

- ***Essential SKOS classes***:
  - `skos:Concept`
  - `skos:ConceptScheme`
- ***Essential SKOS properties:***
  - `skos:preLabel`
  - `skos:altLabel`
  - `skos:definition`
- ***Essential SKOS semantic relationships***:
  - `skos:broader`
  - `skos:narrower`
  - `skos:related`
- ***Essential SKOS mapping relationships:***
  - `skos:exactMatch`
  - `skos:closeMatch`

29 **Jim Morris | 26 Jan 2014**

AstraZeneca | R&D

I want to just outline some very common SKOS terminology as it will help with the following examples.

skos:Concept
skos:ConceptScheme
As mentioned every “term” in a vocabulary is a concept represented by a unique URI.
A ConceptScheme is generally used for a particular thesaurus, or part of a thesaurus. Concepts belong to one or more ConceptSchemes.

skos:prefLabel
skos:altLabel
skos:definition
These are classic thesaurus properties –
prefLabel: preferred name for the term,
altLabel: entry term, “see from”, “used for”, synonym, etc.
definition – self explanatory!

skos:broader
skos:narrower
skos:related
These are the essential pieces to build a hierarchical or ontological framework, within a single vocabulary.

skos:exactMatch
skos:closeMatch
Everything above is within a single vocabulary. These, instead, are used to relate Concepts across vocabularies. e.g. Bladder Cancer in this vocabulary is an exactMatch to Bladder Neoplasm in this vocabulary.