

Linked Data Enlightenment: Lessons Learned from LUX

<https://lux.collections.yale.edu/>

Robert Sanderson

Senior Director for Digital Cultural Heritage
Yale University

*robert.
sanderson
@yale.edu*

Yale



Overview

- Introduction to LUX (~10 minutes)
- Demo (~10 minutes)
- Lessons Learned:
 - Usability (~15-20 minutes each)
 - Technology
 - Data Modeling
- Discuss!

What is LUX?

A ground-breaking discovery
and research platform,
providing unified digital
access to the collections of
our museums, libraries and
archives

Launch Date: June 1st 2023

robert.
sanderson
@yale.edu



17,012,680

Objects



4,910,154

Concepts



5,677,302

People & Organizations



38,208

Events



586,814

Places



13,122,257

Works

Cultural & Natural History

Linked Data
Enlightenment



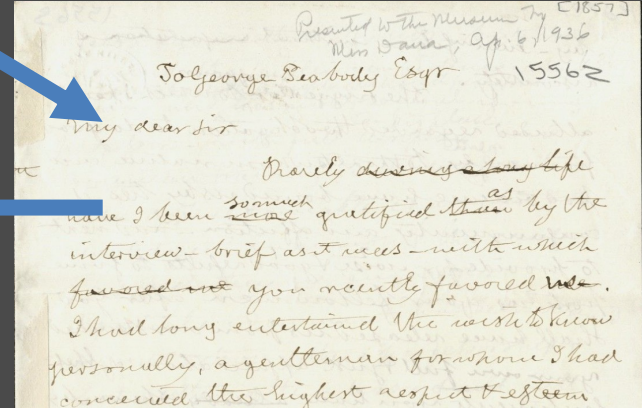
Yale Peabody Museum



Yale Center for British Art



Yale University Art Gallery



Yale University Library

robert.
sanderson
@yale.edu

Yale



Connecting Yale's Heritage

- Reconciling Across Collections
 - Need one knowledge base, not multiple
- Enriching with External Knowledge
 - From 20 other knowledge bases
- 41+ Million Records
 - Automation needed at scale
 - Standards for knowledge management

robert.
sanderson
@yale.edu

Yale



Name

James Dwight Dana

詹姆斯·德怀特·丹纳

ジェームズ・デーナ

جیمس دوايت دانا

जेम्स द्वाइट डेना



Yale Contributing Records

<https://images.peabody.yale.edu/data/agent/d/1d/d1ded813-9a24-48a1-895f-98d508564636.json>

<https://linked-art.library.yale.edu/node/3747e6c8-5dc1-4e9f-b965-c5b2e90f392a>

<https://linked-art.library.yale.edu/node/d7294d60-0229-4d56-9240-a99cda64251f>

<https://media.art.yale.edu/content/lux/agt/44763.json>

External Contributing Records

<https://d-nb.info/gnd/116020849>

<https://data.bnf.fr/ark:/12148/cb122841716>

<http://id.loc.gov/authorities/names/n50055685>

<http://viaf.org/viaf/27128721>

<http://vocab.getty.edu/ulan/500117085>

<http://www.wikidata.org/entity/Q315366>

LUX Data: Linked Art

A metadata profile and API, collaboratively designed to work across cultural heritage organizations, that is easy to publish and enables a variety of consuming applications.

Linked Art provides a **Standards** based metadata profile,
... which **Consistently** solves problems from real data,
... is designed for **Usability** and ease of implementation,
... which are prerequisites for **Sustainability**

robert.
sanderson
@yale.edu

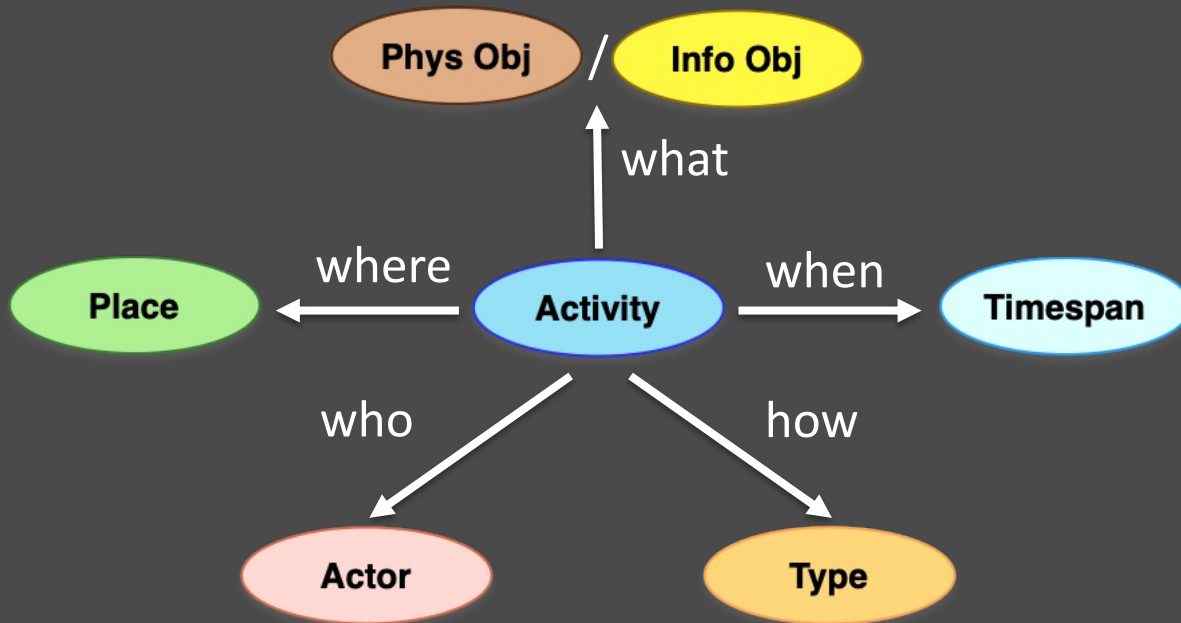
Yale



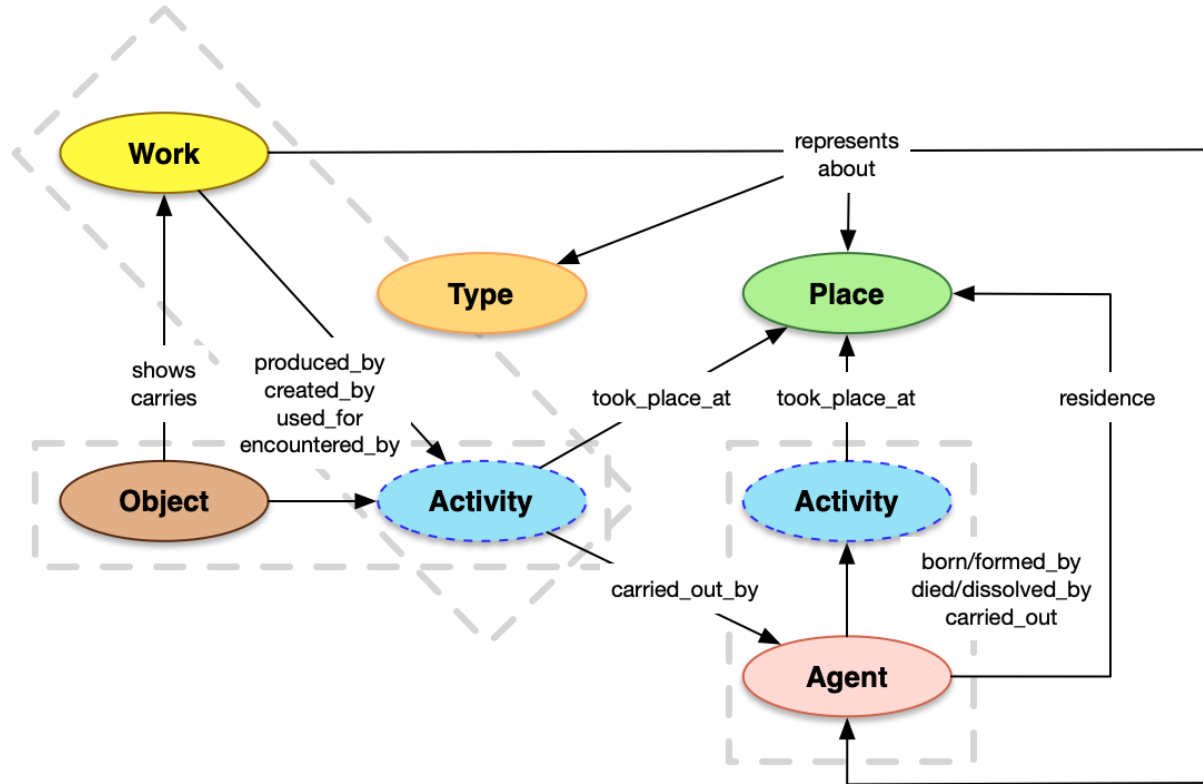
<https://linked.art/>



Conceptual Model Baseline



Linked Art Model in LUX



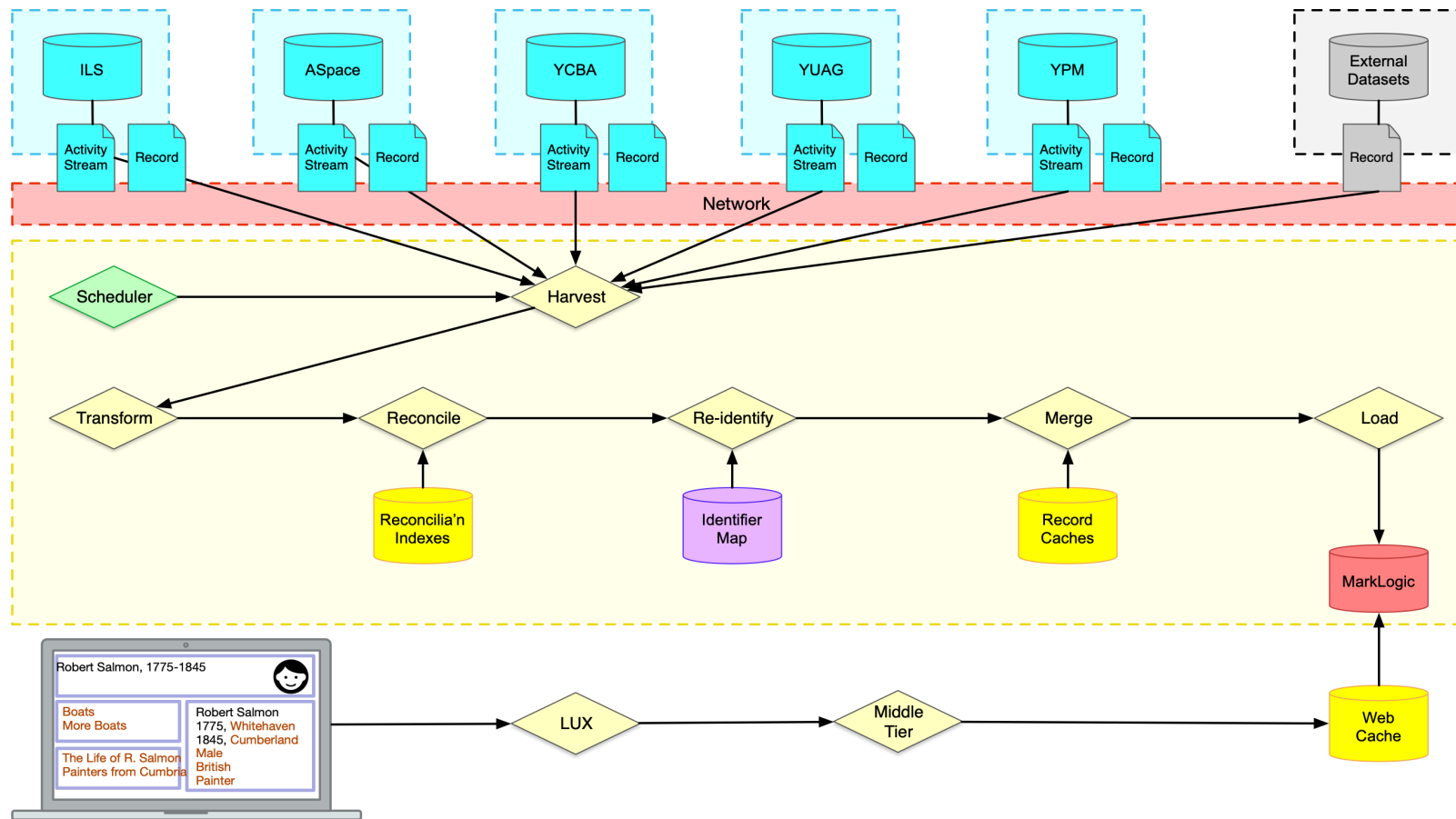
How Did We Do This?

- Collaboration!
 - Trust and cultural transformation has led to generosity, accountability and excellence
- Hard Work
 - By many people over the past five years
- Sophisticated Technology
 - LUX is a knowledge graph, not a database



LUX Technical Architecture

Linked Data
Enlightenment



robert.
sanderson
@yale.edu

Yale

Impact

- Truly Game Changing for our Sector
 - Unique in the Cultural Heritage sector
 - Deep and widespread interest from 50+ peers
- Ease of Access, Teaching & Learning, and Research
 - LUX directly furthers the mission of the University
 - Responsibility to preserve and make accessible

robert.
sanderson
@yale.edu

Yale

Demo!

<https://lux.collections.yale.edu/>

*robert.
sanderson
@yale.edu*

Categories of Lessons Learned

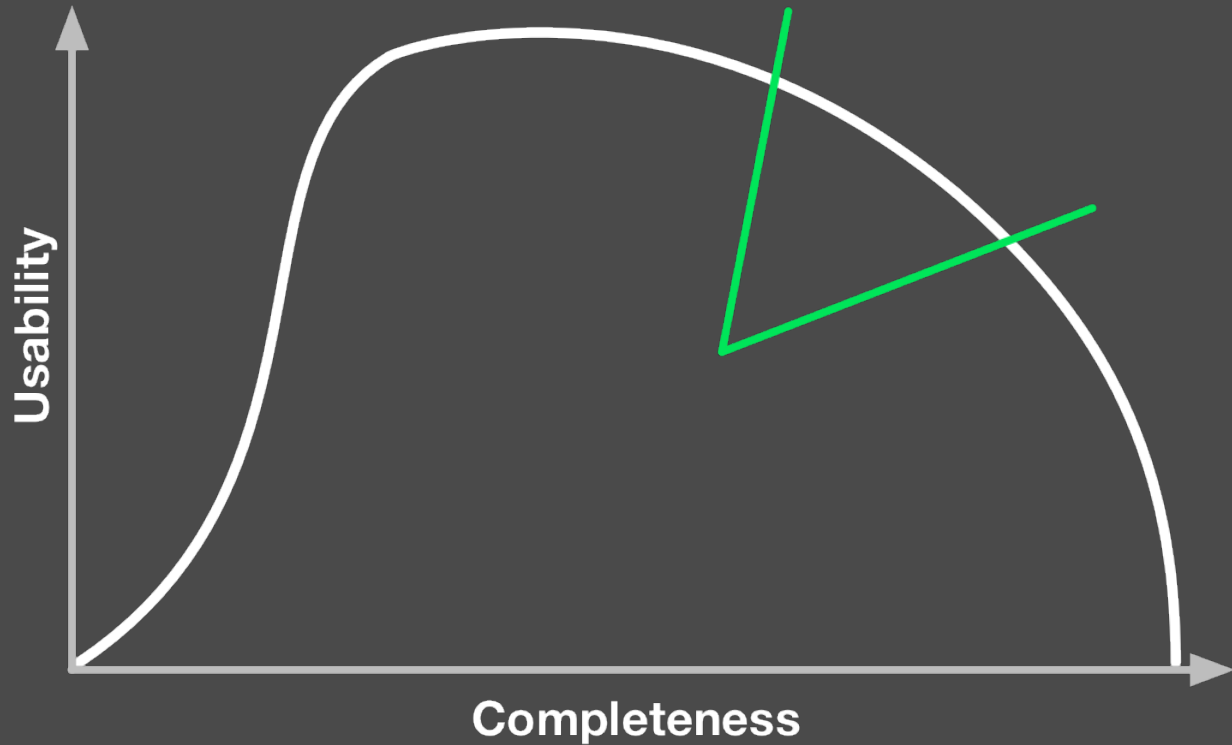
- Usability
 - Ensuring developers are effective and happy
- Technology
 - Finding the right intersection of data and functionality
- Data Modeling
 - Some linked data ideals are useful, ... some are not

4 lessons in each category ... “You won’t believe number 12!”

Usability

1. LOUD: Linked Open Usable Data
2. Usable data requires developer infrastructure
3. Hypermedia is more usable than queries
4. Records are necessary for usability

1. Usability vs Completeness



Linked Art + IIF Design Principles

1. Scope design through shared **use cases**
2. Design for **international** use
3. Make **easy things easy**, complex things possible
4. **Avoid dependency** on specific technologies
5. Use **REST** / Don't break the web / Don't fear the network
6. **Separate concerns**, keep APIs & systems loosely coupled
7. Design for **JSON-LD**, using LOD principles
8. Follow existing **standards** & best practices, when possible
9. **Define success**, not failure
10. Solve issues at the **right level** (in modeling section)

Data Usability is Necessary for Everyone

For LUX we needed data usability for:

- Collection Managers ... translating domain knowledge
- Software Engineers ... transforming data from CMS
- Data Engineers ... reconciling and merging aggregated data
- Software Engineers ... building backend discovery system
- UX Analysts ... designing interactions and interfaces
- Software Engineers ... building front end user interfaces

Application quality is dependent on data usability

How I See It

**I'm learning
through osmosis.**



How the Devs See It



The Fires of Conceptual Modeling


2. Development Infrastructure


- **Documentation**
 - With annotated, working examples to cut and paste
 - <https://linked.art/>
- **Validators**
 - Syntax: JSON-Schema ; Semantics: SHACL / ShEX
 - [https://github.com/linked-art/\(json|shacl\)-validator](https://github.com/linked-art/(json|shacl)-validator)
- **Code Libraries**
 - With built in validation and convenience functions
 - LinkedArt.js (javascript), Cromulent (python)

3. Hypermedia is more Usable than Search






Linked Data
Enlightenment

Search Needed?


LUX: Yale Collections Discovery About LUX Open Access Help 

University of California, Berkeley, 3/23/1868- 

Works Created or Published Works About

-  **University of California publications in geological sciences**
Creator University of California, Berkeley
Work Types Texts
Imprint Berkeley, University of California Press [etc.] 1893-
Languages English
Identifiers ils:yut:478672...
-  **Publications in geological sciences ;**
Creator University of California, Berkeley
-  **Publications in philology ;**
Creator University of California, Berkeley
-  **Publications in economics ;**
Creator University of California, Berkeley
-  **University of California publications in mathematics**
Creator University of California, Berkeley
Work Types Texts
Imprint Berkeley and Los Angeles, Calif. : University of California Press, 1943-1960.
Languages English
Identifiers ils:yut:4739664...

[Show all 258 results](#)



Name
University of California, Berkeley (French)
Universiteit van Californië - Berkeley (Dutch)
加州大學柏克萊分校 (Chinese transliterated Pinyin without tones)
جامعة كاليفورنيا، بركلي (Arabic)
केलिफोर्निया विश्वविद्यालय, बर्कले (Hindi)
[Show All](#)

Nonpreferred Terms
Universidad de California en Berkeley (Spanish)
Université de Californie à Berkeley (French)
University of California, Berkeley, U.S.A. (French)
カリフォルニア大学バークレー校 (Japanese)
Universidade da Califórnia em Berkeley (Portuguese)
[Show All](#)

Additional Names
Berkeley University of California
University of California Berkeley
University of California, Berkeley.
University of California, Berkeley (Estados Unidos)

Type
Group

Formation Date
3/23/1868

Type

Record Data

robert.
sanderson
@yale.edu

Yale



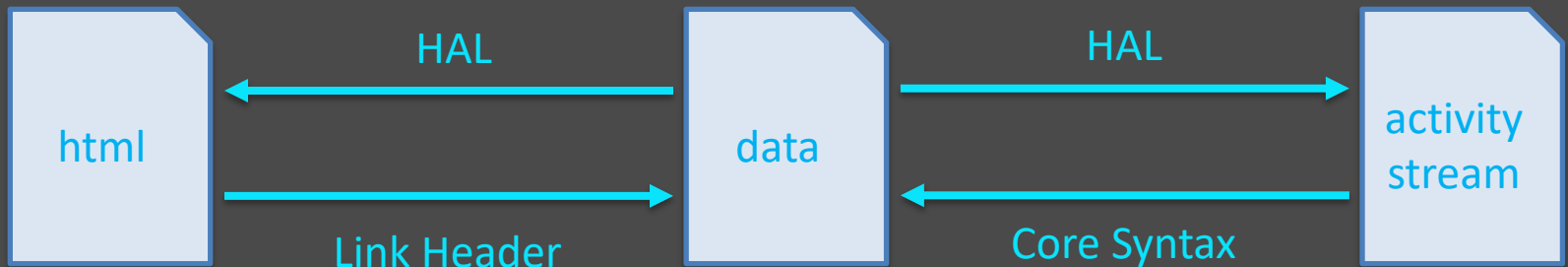
HAL: Hypertext Application Language

```
::  
"lux:agentAgentMemberOf": {  
  "href": "https://lux.collections.yale.edu/api/search?...",  
"lux:agentCreatedPublishedWork": {  
  "href": "https://lux.collections.yale.edu/api/search?...",  
"lux:agentRelatedAgents": {  
  "href": "https://lux.collections.yale.edu/api/related-list?...",  
"lux:agentRelatedConcepts": {  
  "href": "https://lux.collections.yale.edu/api/related-list?...",  
"lux:agentRelatedSubjects": {  
  "href": "https://lux.collections.yale.edu/api/facets?...",  
"lux:agentWorkAbout": {  
  "href": "https://lux.collections.yale.edu/api/search?...",  
...  
...
```

robert.
sanderson
@yale.edu

Hypermedia for “SEO” / Discovery

- HAL link from the data to the web page
- HAL link from the data to the activity stream
- Link header from web page to the data
- Activity stream links to the data (IIF Change Discovery)



4. The Record is a Necessary Construct

“The Graph” is not a comprehensible unit.
Constructing records from a graph at run-time is hard.

We need records for usability:

- **Retrieval** Anchored and sufficiently complete
- **Indexing** What is the unit of search?
- **Faceting** Otherwise arbitrary triple counting
- **Metadata** Rights, administrative, technical
- **Deletion** Is the triple still needed?

Technology

5. Multi-modal functionality is necessary
6. Graphs should be optimized for the application
7. Inference is impractical for Cultural Heritage data
8. Cross-institutional, real-time data is still impossible

5. Multi-Modal Functionality

To benefit from Linked Data, a user interface must have access to both record-based and graph-based functionality at the same time

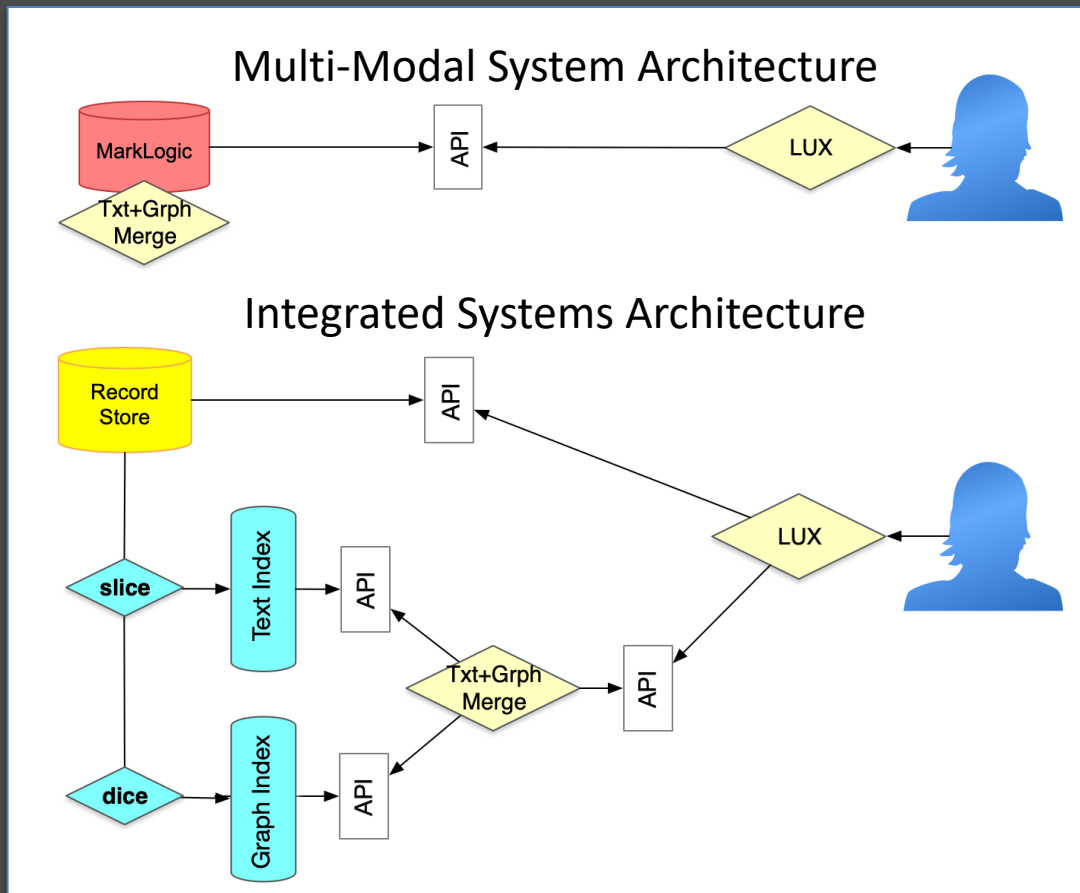
Semantic queries across collections requires the graph

Note: the graph allows full data normalization (e.g. names)

Facets, “anywhere” search, and data permissions require records

Having both together is what gives LUX its superpowers

Architecture is an Easy Choice



6. Graph Optimization

Records are JSON-LD, but not every triple is needed for search

We **materialize only necessary relationships** between records

We **reduce the number of joins** by creating artificial triples from triple paths within the record:

```
Object encountered_by/part/carried_out_by Person
```

becomes

```
Object lux:agentOfEncounter Person
```

without losing semantic precision in the JSON-LD record

7. The Myth of Inference

To enable semantic inferencing, data must be:

- ~~• Objective~~
- ~~• Precise~~
- ~~• Certain~~
- ~~• Complete~~

Cultural Heritage is none of these

The Myth of Inference

Lesson 7a:

Don't sell LOD based on the ability to “infer new facts”

Lesson 7b:

Inference exacerbates existing errors, which leads to hand-wringing about bias, authority, reputation damage. Hard to overcome perfectionism, without adding extra challenges.

**Waiting for Perfect Data means Never Starting
And Inference needs “Perfect” Data**

8. Metasearch is Still Impossible

Currently 5 internal, 20 external sources, but adding to both.

Challenge 1: System availability

- Network latency is a killer, even internally
- Systems aren't always available

Challenge 2: Computational expense

- Would need to do reconciliation, enrichment, merging and query at run time, following links between records – impossible & wasteful!
- Conversely, system of record data changes slowly, no use case requires sub-hour timeframe for propagation

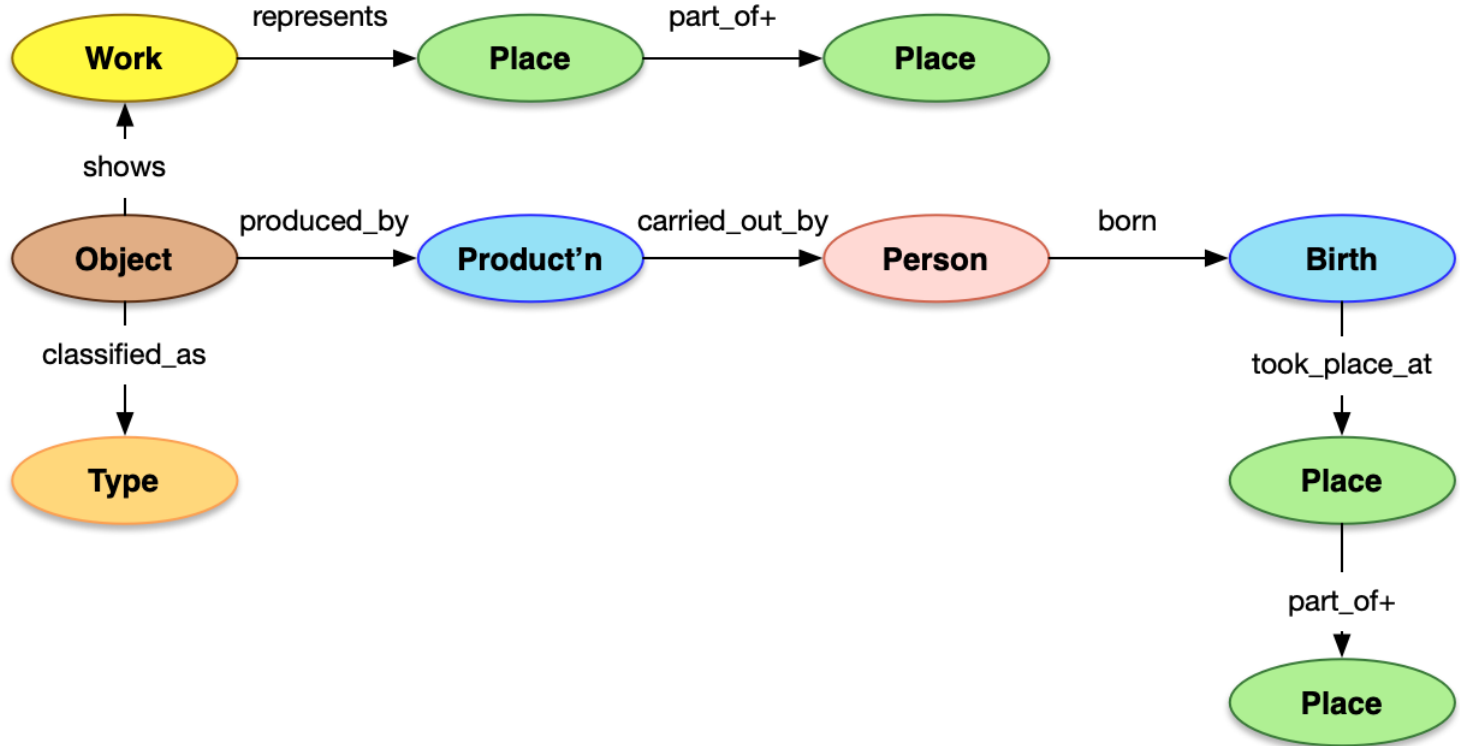
Harvesting and preprocessing is still the right approach

Yosemite Valley, Glacier Point Trail, Albert Bierstadt, 1873
<https://artgallery.yale.edu/collections/objects/4964>



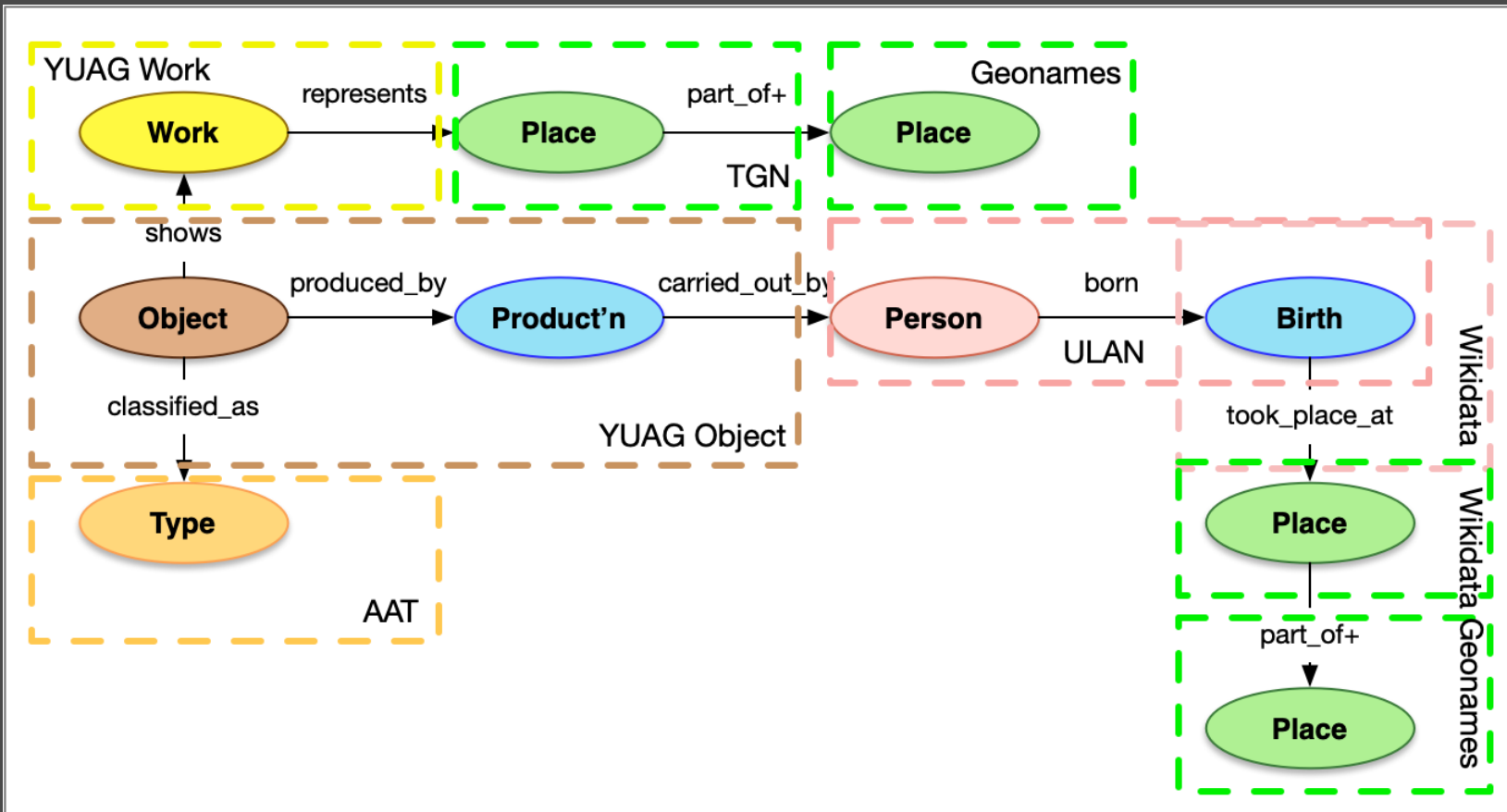
Paintings
of the US South West
by Europeans?

Graph Model



Graph Metasearch Across Systems?!

Linked Data
Enlightenment



robert.
sanderson
@yale.edu

Yale

8b. Search is Application Functionality

- If everyone is harvesting and processing data, then search is application-specific functionality
- So ... No need for a SPARQL endpoint (heresy?!)
- Other applications will have very different interactions and requirements over the same data
 - YPM want to use LUX infrastructure, but need new query patterns

Make it easy to discover, harvest and use data:

Open Licenses, JSON-LD, Activity Streams, HAL links

Data Modeling

9. Unique identities doesn't mean unique identifiers
10. Conceptual models should not be domain-specific
11. Don't confuse Class with Classification
12. Predicate reuse is illogical and impractical (!)

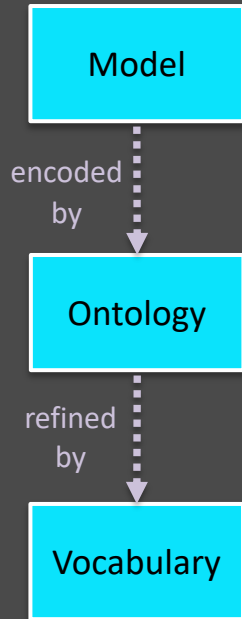
9. Unique Identity, Multiple Identifiers

Should every entity have exactly one URI, theoretically?

- No one has all the knowledge; everyone can contribute
- No single central platform will work for all
- Inclusion of viewpoints and information is a choice
- Every new dataset will mint a new URI ...
- And should publish other systems' URIs:
 - 5* Linked Data is **link** to others, not use others in situ

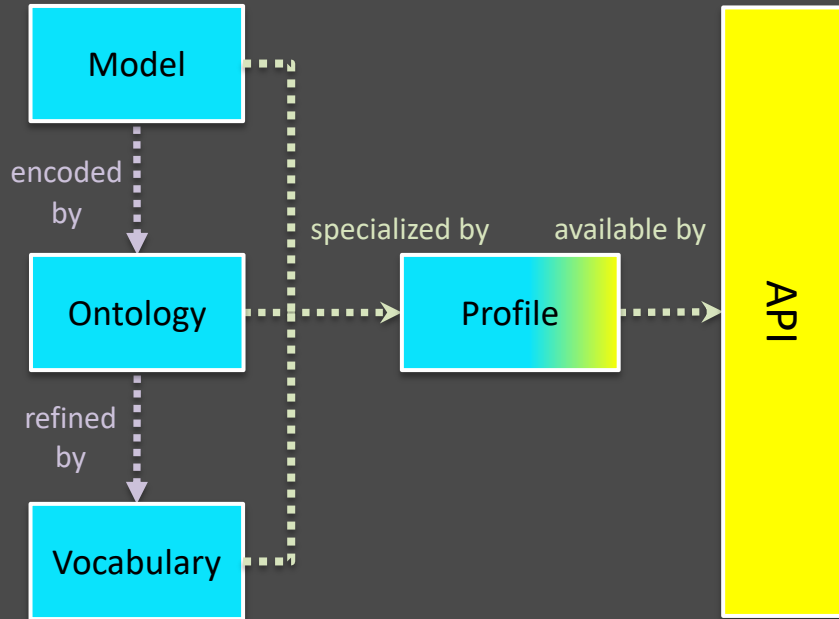
Multiple URIs for the same entity is a feature not a bug

Sidebar: Solve at the Right Level



- **Conceptual Model**
 - Abstract way to think about the world, holistically, consistently and coherently
- **Ontology**
 - Shared set of terms to encode that thinking in a logical, machine-actionable way
- **Vocabulary**
 - Curated set of sub-domain specific terms, to make the ontology more concrete

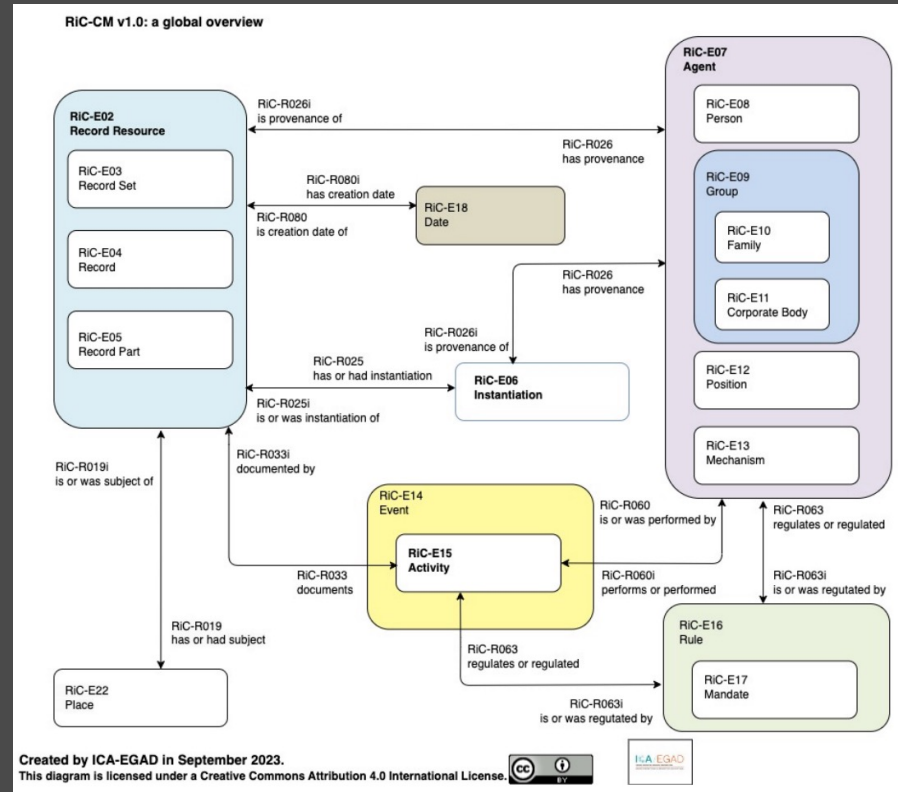
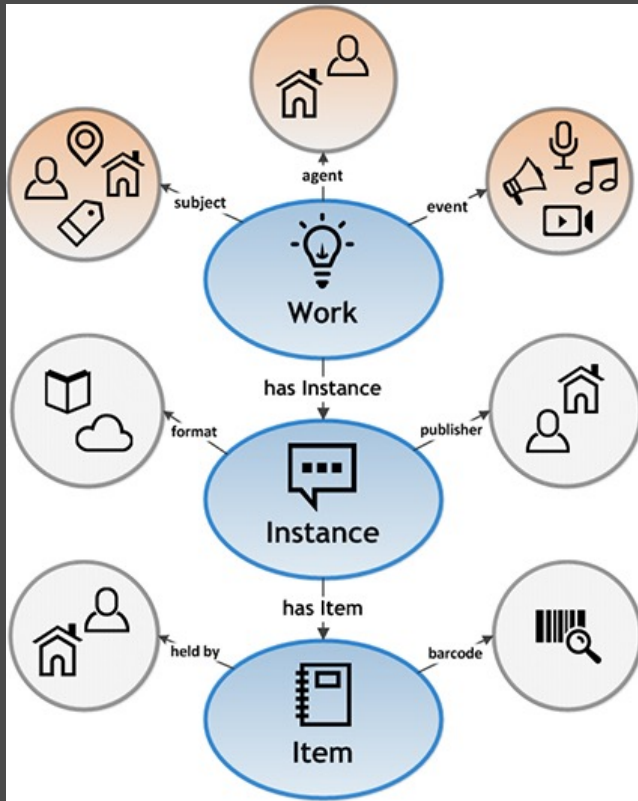
Sidebar: Solve at the Right Level



A **Profile** is a selection of appropriate **abstractions**, to encode the **scope** of what can be described.

An **API** is a selection of appropriate **technologies**, to give **access** to the data managed using the profile.

10. Conceptual Models should be General

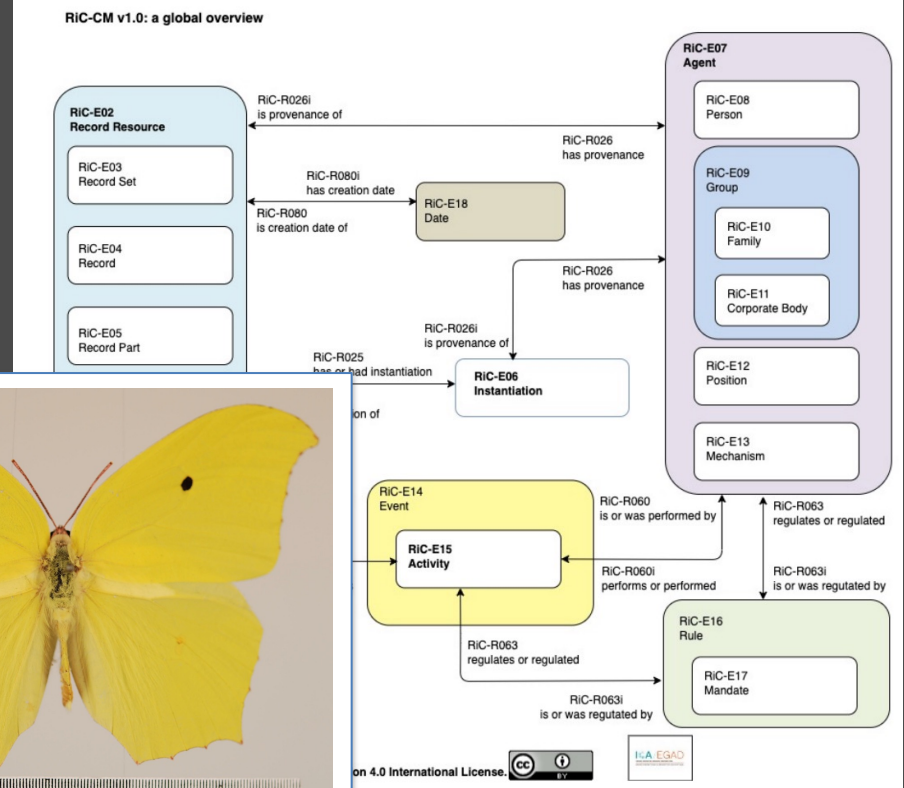
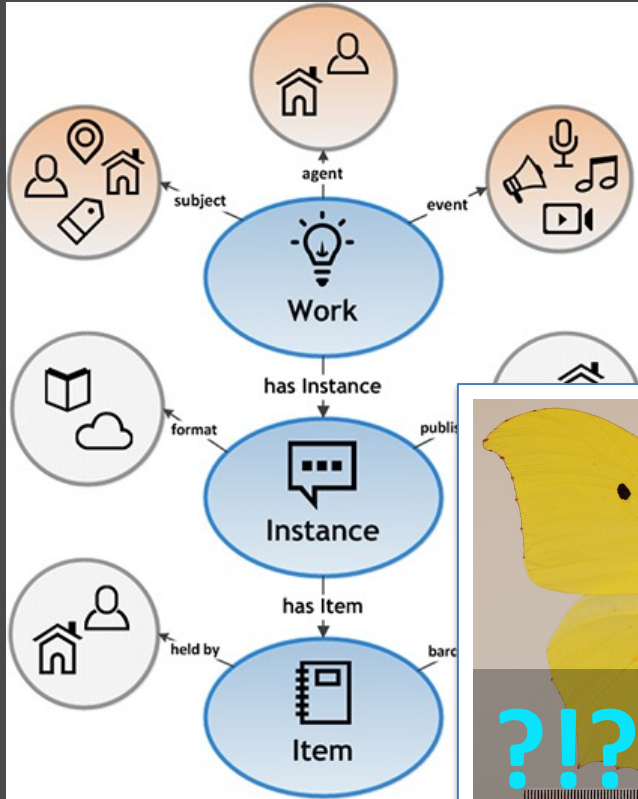


Created by ICA-EGAD in September 2023.

This diagram is licensed under a Creative Commons Attribution 4.0 International License.



10. Conceptual Models should be General



robert.
sanderson
@yale.edu

Domain Specific Models Limit Usage

Linked Data
Enlightenment



imgflip.com

robert.
sanderson
@yale.edu

Yale



Domain Specific Models Limit Usage

Linked Data
Enlightenment



imgflip.com

robert.
sanderson
@yale.edu

Yale



Domain Specific Models Limit Usage

- General, easy to understand, conceptual model has broader applicability, better usability, and can be profiled
- Highly specialized models are confusing even to experts
 - Object/Works split in LUX is largest source of confusion
- Interoperability is constrained across related domains
- Using multiple conceptual models is more confusing

Use Profiles to restrict the scope of a more general model

11. Separate Class and Classification

- Class: A modeling construct for the abstract set of entities that all have the associated features/properties (model)
 - Person, Group, Physical Object, Event,...
- Classification: An assignment of a terminological instance to another instance (vocabulary)
 - Painting, Fossil, Corporation, Exhibition, ...
- No need for a class that does not define new properties

Few general classes, with core relationships between them

12. Predicate Reuse is Illogical

- Each ontology* expresses a single conceptual model.
- Each model is expressed by a single set of ontological terms.
- Those terms are classes and predicates.
- We should use exactly one conceptual model.
- Therefore, we should use exactly one set of ontological terms.
- Therefore, **we should not use classes or predicates from more than one ontology.**

* The technical representation of the ontology may be split between files and namespaces, c.f. CRM + Extensions

Predicate Reuse is Impractical

Knowledge integration is mapping between conceptual models, which is data engineering and requires humans to interpret the subjective, contextual domain

Ontologies are not lego bricks, they are personalities:
more than one and you get chaos

Summary

- Data must be usable to be used
- Usable data means better, cheaper and faster products
- Records must exist, with hypermedia and relationships
- Systems must deal with graphs and records, simultaneously and efficiently
- Forget inference and SPARQL, federated or otherwise
- Use a general conceptual model/ontology for interoperability
- Specialize via domain-specific vocabulary

Thank You!
Discuss!

<https://lux.collections.yale.edu/>