

# So You Think You Want to **MIGRATE TO RDF?**

Steven Anderson  
Eben English  
Boston Public Library

---

Slides: [goo.gl/csBcd9](https://goo.gl/csBcd9)

# ■ RDF: NO FURTHER KITTENS





# ■ RDF 101: GRAPH

A data model specifying “statements about resources in the form of subject–predicate–object expressions.”

`<http://example.org/item/123>`

`<http://purl.org/dc/terms/type>`

`<http://id.loc.gov/vocabulary/resourceTypes/img>` .



# VOCABULARIES

---

Choose wisely.



## ■ VOCABULARIES: REUSE++

“Vocabularies get their value from reuse: the more vocabulary IRIs are reused by others, the more valuable it becomes to use the IRIs (the so-called network effect).”

“This means you should prefer re-using someone else's IRI instead of inventing a new one.”

# VOCABULARIES: FIND YOUR BLISS

## Linked Open Vocabularies (LOV)

`<http://lov.okfn.org/dataset/lov/>`

`<sameAs>`

`<http://sameas.org/>`



# VOCABULARIES: COMBINATIONS

You're not limited to a single vocabulary.

Mix and match at will!

```
@prefix schema: <http://schema.org> .
```

```
@prefix dc: <http://purl.org/dc/elements/1.1/> .
```

```
<http://example.org/item/123>
```

```
    dc:title      "Do you still want to migrate to RDF?"@en ;
```

```
    schema:genre <http://vocab.getty.edu/aat/300258677>
```

```
.
```

# ■ VOCABULARIES: USAGE

So... I just pick a predicate and use it?

Not exactly. There are rules:

- domain
- range
- not all URIs can be used as predicates

## ■ RDF 101: RANGE

"the class or datatype of the **object** in a triple"

`<http://example.org/item/123>`

`<http://purl.org/dc/terms/type>`

`<http://id.loc.gov/vocabulary/resourceTypes/img> .`

# ■ VOCABULARIES: RANGES

Let's say I want to represent this in RDF:

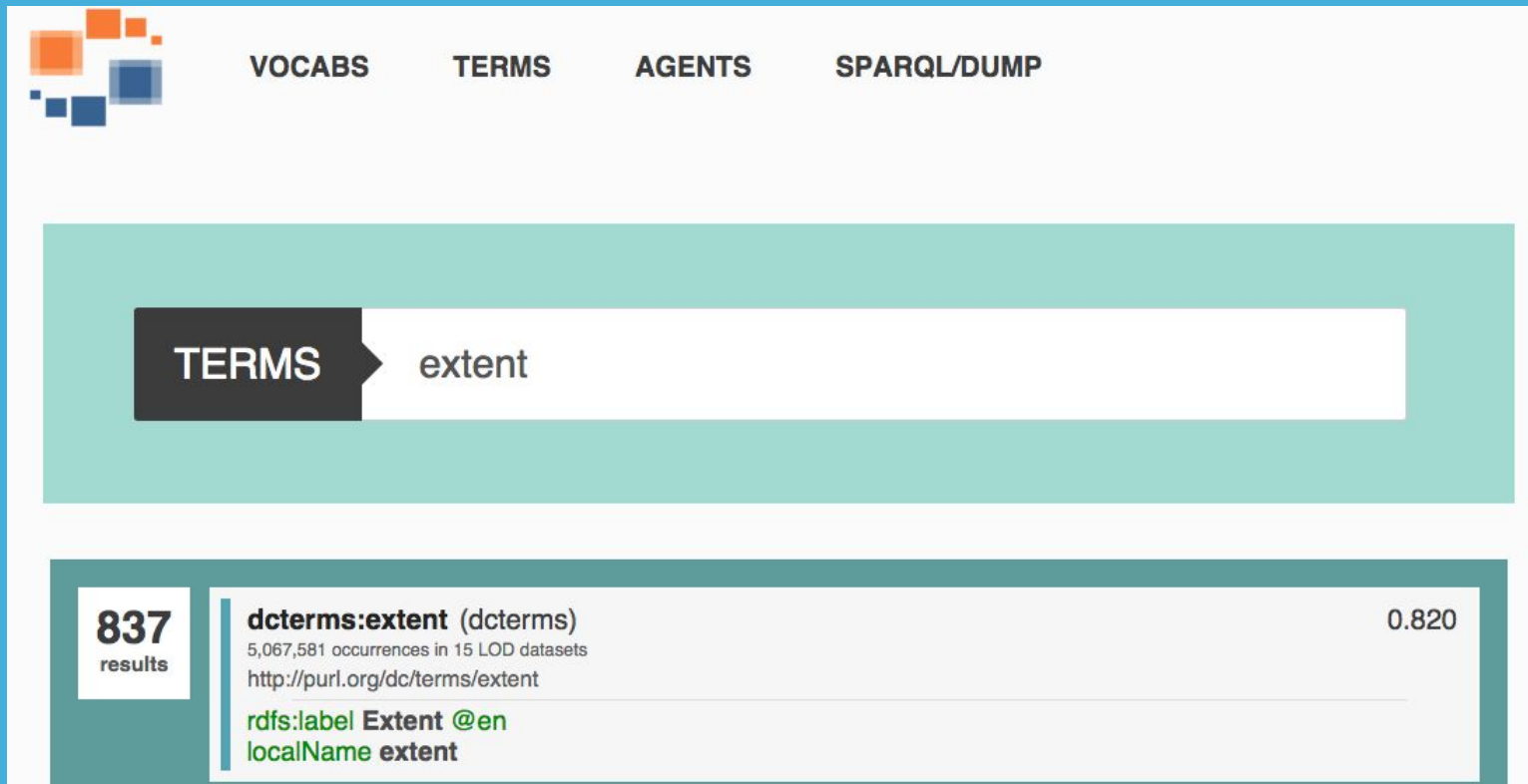
```
<mods:extent>
```

```
  1 photographic print : gelatin silver ; 5 x 7 in.
```

```
</mods:extent>
```

# VOCABULARIES: RANGES

We find a highly-used predicate “dcterms:extent” via LOV:



The screenshot shows the LOV interface with a search for the term 'extent'. The search results are displayed in a table format.

VOCABS	TERMS	AGENTS	SPARQL/DUMP
	TERMS		
	extent		

<b>837</b> results	<b>dcterms:extent</b> (dcterms) 5,067,581 occurrences in 15 LOD datasets <a href="http://purl.org/dc/terms/extent">http://purl.org/dc/terms/extent</a>	0.820
	<b>rdfs:label</b> Extent @en <b>localName</b> extent	

(<http://lov.okfn.org/dataset/lov/terms?q=extent>)

# VOCABULARIES: RANGES

What are the expected values for this predicate?:

## **dcterms:extent**

The range of `dcterms:extent` is the class `dcterms:SizeOrDuration`. All values used with `dcterms:extent` have to be instances of this class. Therefore the property may only be used with non-literal values.

```
ex:myVideo dcterms:extent [ rdf:value "21 minutes" ] .
```

```
ex:myVideo dcterms:extent [ rdf:value "PT21M"^^xsd:duration ] .
```

([http://wiki.dublincore.org/index.php/User\\_Guide/Publishing\\_Metadata#dcterms:extent](http://wiki.dublincore.org/index.php/User_Guide/Publishing_Metadata#dcterms:extent))

# VOCABULARIES: RANGES

But lots of institutions are using `dcterms:extent` with literal values!

- DPLA, Europeana

Isn't this a problem?

- We'd never do this in a DB or XML doc
- Validation is lacking in RDF
- "there are no Semantic Web police"

# VOCABULARIES: RANGES

Have to make a choice:

- Conform to “accepted” usage; ignore official range definition.

OR

- Use a less popular predicate (or mint your own).
  - Fewer harvesters will have out of the box code to understand it...
  - ...but it conforms to the standards, so parsing should be OK



# VOCABULARIES: RANGES

bf:extent does have a range of literal

- but, less adoption than dcterms:extent

## Extent - Property

Number and type of units and/or subunits making up a resource.

### Extent

Number and type of units and/or subunits making up a resource.

**Property:** Extent

**Used With:** Instance

**Expected value(s):** Literal

**URI:** <http://bibframe.org/vocab/extent>

(<http://bibframe.org/vocab/extent.html>)

# ■ RDF 101: DOMAIN

"the class of the **subject** in a triple"

`<http://example.org/item/123>`

`<http://purl.org/dc/terms/type>`

`<http://id.loc.gov/vocabulary/resourceTypes/img> .`

# ■ VOCABULARIES: DOMAINS

The latest thinking is that these mean very little.

- bf:extent has a domain of bf:Instance
- While your object may not explicitly declare this class, this is OK as long as it could also be a “bf:Instance”.
- Beware domain class requirements!
  - required predicates, etc.

# ■ VOCABULARIES: EXTINCTION

A URI is useless if it can't be resolved.

- But URI's have the library community behind them!
- Surely they'll be around forever...

# VOCABULARIES: EXTINCTION

Don't be so sure . . .

## **dcterms:format**

The range of `dcterms:format` is the class `dcterms:MediaTypeOrExtent`. All values used with `dcterms:format` have to be instances of this class. Therefore the property may only be used with non-literal values.

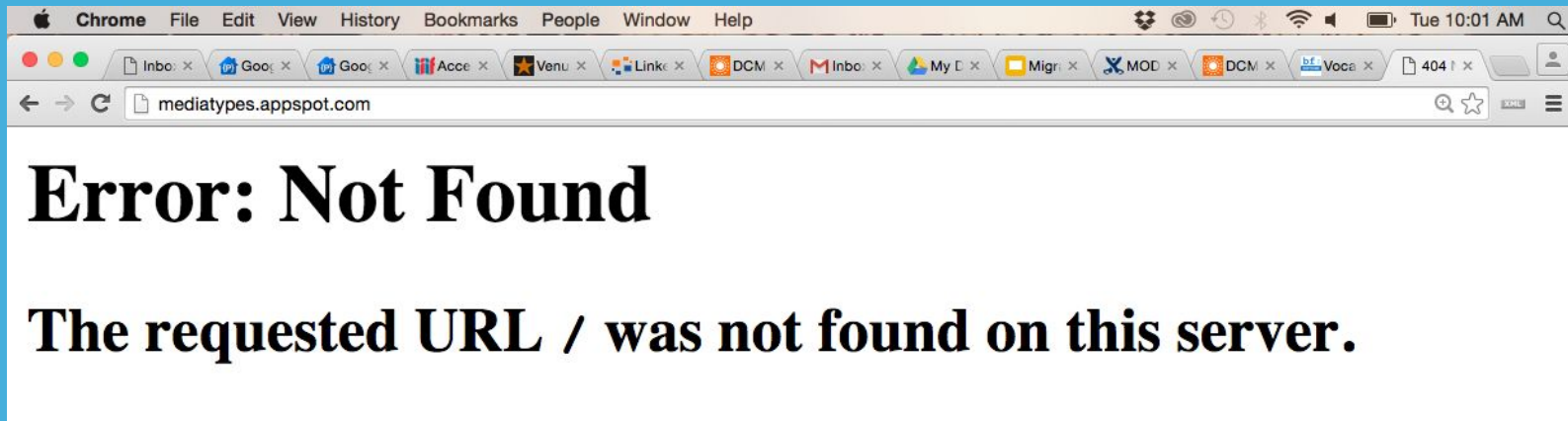
```
ex:myPicture dcterms:format mime:jpeg .
```

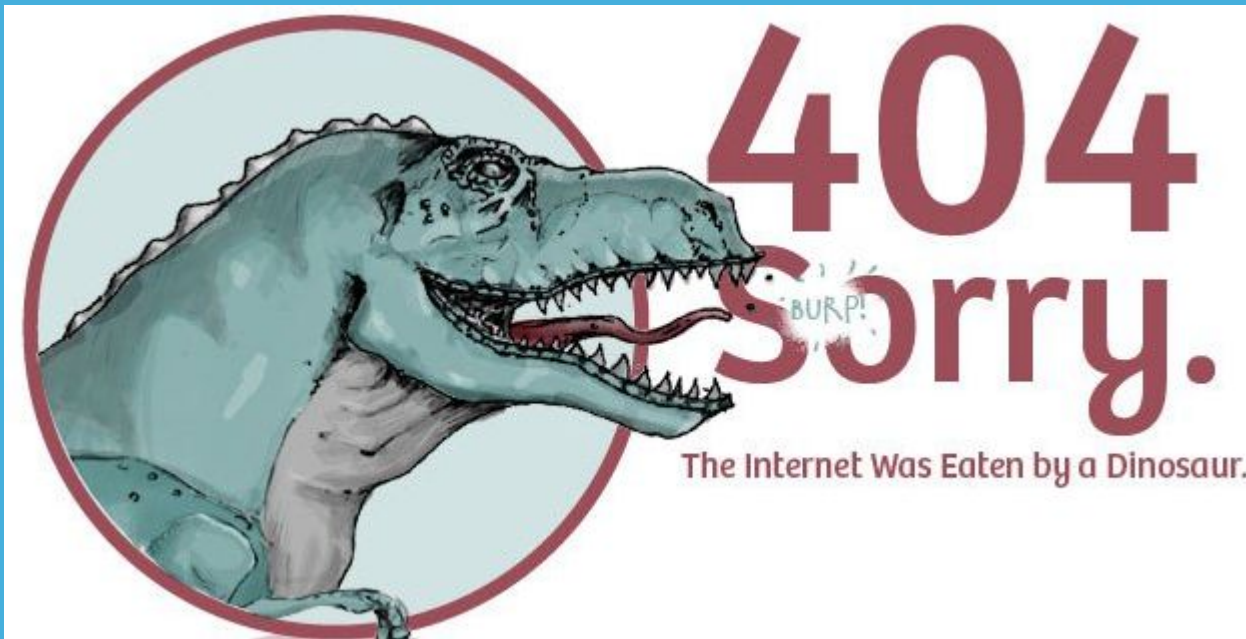
```
@prefix mime: <http://purl.org/NET/mediatypes/> .
```

(<http://dublincore.org/documents/dcmi-terms/#terms-format>)

# VOCABULARIES: EXTINCTION

Try and act surprised...





- Several proposed ideas on handling this but not much practical work has been completed.
- About the best you can currently do is store values locally in some fashion.

# MODELING

---

Get the Tylenol ready...



# MODELING: MINTING PREDICATES

What if no predicate currently exists for my data?

- You can mint your own predicate and/or vocabulary.
- Use a community namespace ([opaquenamespace.org](http://opaquenamespace.org)).
- Get community investment in your predicate.

Don't dumb down your data just to fit a predicate.

- Use your judgement but the fidelity of data is important.
- Standards and systems change... it is your data that lives on.

# MODELING: XML TO RDF

Attributes:

```
<mods:note type="ownership">  
  This pipe belonged to Albert Einstein.  
</mods:note>
```

Unlikely that we're going to find a "hasOwnershipNote" predicate in any namespace.

# MODELING: XML TO RDF

Hierarchies:

```
<mods:originInfo eventType="manufacture">  
  <mods:place>  
    <mods:placeTerm type="text">Cambridge</mods:placeTerm>  
  </mods:place>  
  <mods:publisher>Kinsey Printing Company</mods:publisher>  
</mods:originInfo>
```

We need to associate place and publisher data with “manufacture” event.

# MODELING: **BLANK NODES**

```
@prefix dcterms: <http://purl.org/dc/terms/> .  
@prefix rdag1: <http://rdvocab.info/Elements/> .  
@prefix loc: <http://id.loc.gov/vocabulary/relators/> .
```

```
<http://example.org/item/123>  
  rdag1:manufactureStatement :_1  
  .  
  :_1 loc:pup "Cambridge" ;  
      dcterms:publisher "Kinsey Printing Company"  
  .
```

# MODELING: **BLANK NODES**

AKA “anonymous resource” AKA “bnode”

- Add complexity
- Make data processing more difficult
- Aren't well-supported in some major platforms (Fedora 4)

# MODELING: MINTING OBJECTS

```
@prefix dcterms: <http://purl.org/dc/terms/> .
```

```
@prefix bf: <http://bibframe.org/vocab/> .
```

```
@prefix loc: <http://id.loc.gov/vocabulary/relators/> .
```

```
<http://example.org/item/123>
```

```
    bf:manufacture <http://example.org/provider/123>
```

```
.
```

```
<http://example.org/provider/123>
```

```
    a bf:Provider ;
```

```
    loc:pup "Cambridge" ;
```

```
    dcterms:publisher "Kinsey Printing Company"
```

```
.
```

# MODELING: UN-ORDERED-NESS

## Linked Data Is Merely More Data

Prateek Jain\*, Pascal Hitzler\*, Peter Z. Yeh†, Kunal Verma† and Amit P. Sheth\*

\*Kno.e.sis Center, Wright State University, Dayton, OH

†Accenture Technology Labs, San Jose, CA

### Abstract

In this position paper, we argue that the Linked Open Data (LoD) Cloud, in its current form, is only of limited value for furthering the Semantic Web vision. Being merely a weakly linked “triple collection,” it will only be of very limited benefit for the AI or Semantic Web communities. We describe the corresponding problems with the LoD Cloud and give directions for research to remedy the situation.

### Where We Are

The recent emergence of the “Linked Data” approach for publishing data represents a major step forward in realizing Berners-Lee, Handler and Lassila’s original vision of a web that can “understand and satisfy the requests of people and machines to use the web content”<sup>1</sup> – i.e. the Semantic Web (Berners-Lee et al. 2001). This new approach has resulted in the Linked Open Data (LoD) Cloud (Bizer et al. 2007), which includes more than 70 large datasets con-

explore the different kinds of information related to the locations where it can be found (*Wisconsin*), the locations where it cannot be found (*Iowa, Minnesota*), and the topography of these regions. Thus, in this scenario, the interlinks might help in identifying and analyzing the topographical patterns related to Iowa and Minnesota which make it difficult for this spider to survive in those regions.

However, the current interlinks between datasets in the LoD Cloud – as we will illustrate – are too shallow to realize much of the benefits promised. If this limitation is left unaddressed, then the LoD Cloud will merely be more data that suffers from the same kinds of problems which plague the Web of Documents, and hence the vision of the Semantic Web will fall short.

### What Is Needed

The growing number of datasets available on the LoD Cloud presents a challenge with regards to its usage, since on the one hand datasets such as DBpedia and Freebase offer mas-

Need to preserve order of authors.

(<http://daselab.cs.wright.edu/resources/publications/jain-hitzler-et-al-AAAISS2010.pdf>)

# MODELING: UN-ORDERED-NESS

@prefix dcterms: <http://purl.org/dc/terms/> .

@prefix foaf: <http://xmlns.org/foaf/0.1/> .

@prefix opaque: <http://opaquenamespace.org/ns/foo> .

<http://example.org/item/123>

dcterms:creator <http://example.org/creator/123> ;

opaque:nameOrder "(http://example.org/names/123, http://example.org/names/456)"

.

<http://example.org/creator/123>

a foaf:Person

foaf:firstName "Jane" ;

foaf:lastName "Doe"

.



# USING LINKED DATA

---

Like, IRL

# ■ USING: REAL-WORLD PROBLEMS

## Performance

- real-time lookup is a bottleneck
- data providers aren't always available

## Rate limiting

- id.loc.gov
  - can only hit their endpoint every 3 seconds (slow for multiple URIs).
  - You'll get blocked if you try to use them for any non-trivial and limited Linked Data use case.



- See [scande3.com](http://scande3.com) for how to do this using Rails Linked Data Fragments.
  - Support Blazegraph, Marmotta, and In-Memory thus far (acts as a communication layer to your cache).
- Caveat: cached linked data won't be as up-to-date.
  - LoC's download of LCSH last updated March 2014.

# USING: METADATA ENRICHMENT INTERFACE (MEI)

Add LCSH SUBJECT Term x

Search Term:

Results:

Broader Terms	Actual Search Term	Narrower Terms
Medicine Physiology	<b>Health (0)</b> <i>(Personal health, Wellness)</i>	Health attitudes Exercise Diet Rural health Physical fitness Astrology and health Presidents--United States--Health Rest Mental health Stress management Longevity Vitality Nutrition Sexual health Reproductive health Self-care, Health Animal health Cardiovascular fitness Sleep Health status indicators Relaxation Detoxification (Health) Plant health Alexander technique Public health
	<b>Health (0)</b> <i>(Biography--Health)</i>	Mental health
	<b>Health (0)</b> <i>(Hygiene)</i>	Cleanliness

<https://github.com/boston-library/mei>

# USING: METADATA ENRICHMENT INTERFACE (MEI)

A screenshot of the Metadata Enrichment Interface (MEI) showing a search for 'henry'. The interface includes several input fields:

- A search box containing the text 'henry'.
- A dropdown menu showing 'http://schema.org/experienc'.
- A long, empty text input field.
- A shorter, empty text input field.

(Coming soon courtesy of Villanova University)

# CUSTOM: OREGON DIGITAL CONTROLLED VOCAB MANAGER

- <https://github.com/OregonDigital/ControlledVocabularyManager>
  - <http://opaquenamespace.org>
- Stores in Marmotta
  - If you backup the Marmotta DB, then you have backed up Marmotta (and subsequently your linked data vocabulary).
- Supports:
  - RDFS.label
  - RDFS.comment
  - DC.issued
  - DC.modified

# CUSTOM: DTA VOCAB MANAGER

- Used to power homosaurus.org terms. Based on Oregon Digital Vocab Manager.
  - (Code gemification TBA)
- Stores in Fedora 4 Commons
- Supports:
  - SKOS.prefLabel
  - SKOS.altLabel
  - RDFS.comment
  - DC.issued
  - DC.modified
  - SKOS.broader
  - SKOS.narrower
  - SKOS.related

# CUSTOM: DTA VOCAB MANAGER

## Edit

**\* Identifier ?**

http://homosaurus.org/terms/ ableism

**\* Preferred Term (USE) ?**

ableism

**Use For (UF) ?**

discrimination against people with disabilities + -

**Scope Note (SN) ?**

**Broader Terms (BT) ?**

http://homosaurus.org/terms/ discrimination ▼ + -

**Narrower Terms (NT) ?**

http://homosaurus.org/terms/ ▼ + -

**Related Terms (RT) ?**

http://homosaurus.org/terms/ peopleWithDisabilities ▼ + -

Update Homosaurus



# CONCLUSIONS

---

# CONCLUSIONS: IS IT WORTH IT?

- Migration is never painless.
- What are the real benefits?
  - Public UI users can't tell the difference.
  - Just because your data is in RDF doesn't make it instantly aggregatable or harvestable.
- Local practices still a barrier to sharing.



# CONCLUSIONS: **MAYBE**

But the cupcakes  
are real!



- When tightly-defined data structures exist, and standards are followed, sharing can be successful.
- Don't let today's limitations ruin tomorrow's potential.
- It's where things are going. Deal with it.

# THANKS!

(Not the end - more slides with further links and reading beyond)

---

Steven Anderson  
@scande3  
sanderson[at]bpl.org

Eben English  
@ebenenglish  
eenglish[at]bpl.org

Slides: [goo.gl/csBcd9](https://goo.gl/csBcd9)

# CURRENT COMMUNITY EFFORTS

---

At a glance

# MAPPING: HYDRA DESCRIPTIVE METADATA WORKING GROUP

- As of the last meeting (03/02/2016), will be talking through a combination mapping of two University of California institutions (San Diego and Santa Barbara).
  - <https://wiki.duraspace.org/display/hydra/Descriptive+Metadata+Working+Group>

# ■ MAPPING: MODS IN RDF

- MODS RDF Ontology
  - “Official” representation
    - <https://www.loc.gov/standards/mods/modsrdf/>
- MODS and RDF Descriptive Metadata Subgroup
  - Independent group of institutions working collaboratively
    - Not just doing “MODS in RDF”
    - Use of widely-used vocabularies
    - No blank nodes
    - <https://wiki.duraspace.org/display/hydra/MODS+and+RDF+Descriptive+Metadata+Subgroup>

# ■ MAPPING: MARC IN RDF

Bibliographic Framework Initiative (BIBFRAME.ORG)

- BIBFRAME 2.0 Draft Specifications currently under review



# USING: HYDRA APPLIED LINKED DATA

- Discussions on how to use and implement Linked Data in Hydra
  - <https://wiki.duraspace.org/display/hydra/Applied+Linked+Data+Working+Group>

# MORE READING

---

Fun, yes?

# MORE READING: LINKS

- A guide on using Dublin Core in RDF?
  - [http://wiki.dublincore.org/index.php/User\\_Guide/Publishing\\_Metadata](http://wiki.dublincore.org/index.php/User_Guide/Publishing_Metadata)
- More on Linked Data and RDF?
  - Semantic Web for the Working Ontologist
- A list of datasets available as Linked Data:
  - <https://datahub.io/dataset>
- An explanation of how Bibframe works
  - <http://infomotions.com/blog/2016/03/bibframe/>