

# Embedded Semantic Markup, schema.org, the Common Crawl, and Web Data Commons: Big Web Data

Jason Ronallo

Associate Head, Digital Library Initiatives  
North Carolina State University Libraries

@ronallo

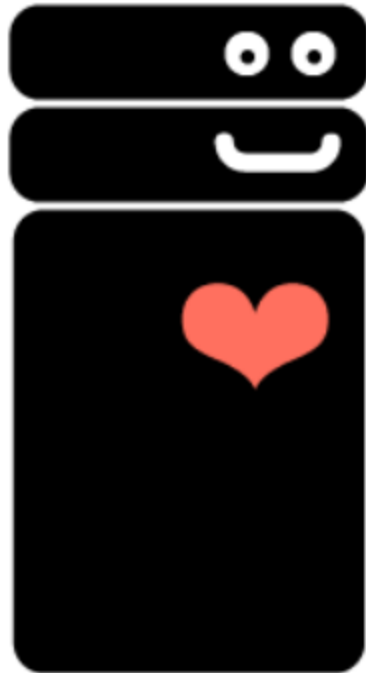
jason\_ronallo@ncsu.edu

# Outline

- Embedded Semantic Markup
- Schema.org
- Examples
- Common Crawl
- Web Data Commons
- Preliminary Research

# How Search Engines Work

1. Robots crawl the Web
2. Process and index crawl data
3. Try to answer search queries with the most relevant results



# Embedded Semantic Markup

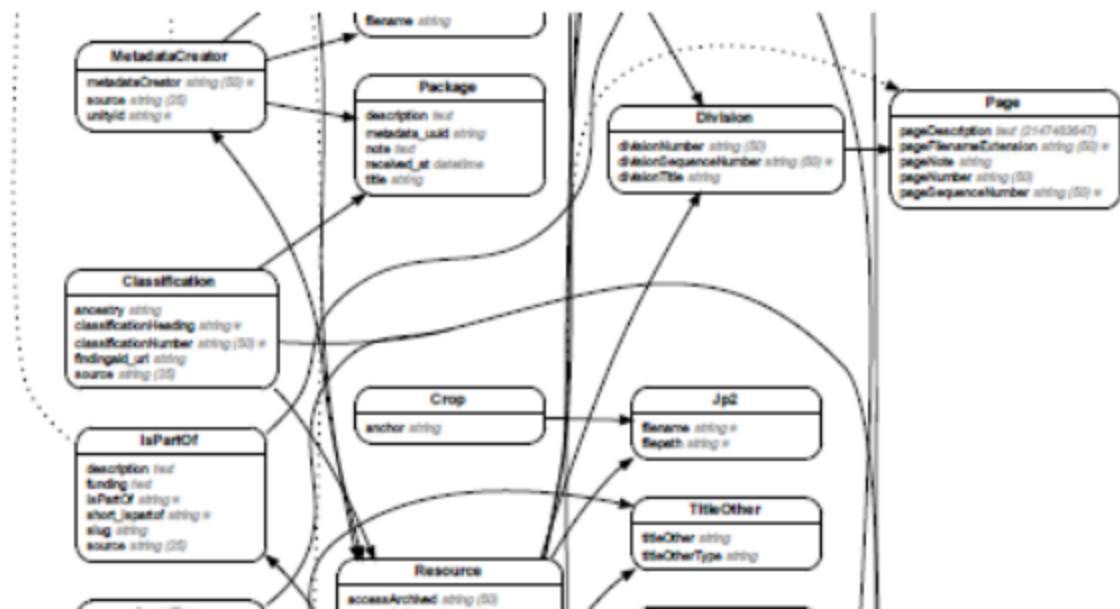
[Jason Ronallo](#) is the Associate Head of Digital Library Initiatives at [NCSU Libraries](#).

# Embedded Semantic Markup

```
<span itemscope
  itemtype="http://schema.org/Person">
  <a itemprop="url"
    href="http://twitter.com/ronallo">
    <span itemprop="name">Jason Ronallo</span>
  </a> is the <span itemprop="jobTitle">
  Associate Head of Digital Library
  Initiatives</span> at
  <span itemprop="affiliation" itemscope
    itemtype="http://schema.org/Library">
    <span itemprop="name"> <a itemprop="url"
      href="http://lib.ncsu.edu">NCSU Libraries</a></span>
  </span>.
</span>
```

# Why use embedded semantic markup?

- A way to *structure* data in HTML
- It is meant for machines--that's why it is hidden!
- Your eyes are on the Web site (Maintain this data in one place and keep it in sync)
- Rich Metadata  $\Rightarrow$  Rich Embedded Data



# Schema.org

- Shared, Web-scale, single-stop vocabulary for describing the content of Web pages.
- Maintained by the major search engines (Bing, Google, Yahoo, Yandex)
- Everything is a Thing
- 407 Types of Things
- 545 Properties of Things
- Everything from Airport to Library to Volcano
- Single site for documentation makes it easy to use (no fragmentation)
- Expanding and open to proposals to update the schema (see SchemaBibEx W3C Community Group)

\* Numbers from last time I checked early in 2013.

# Examples



Limit your search

Topic ▾

- Architecture 17543
- Campus and Town 7860
- Agriculture 6114
- Sports 5511
- Community and Extension 4248

[more »](#)

Decade ▾

- Undated 24892
- 1960s 8081
- 1950s 7291
- 1970s 5209
- 1940s 4784

[more »](#)

Buildings >

Location >

Subject >

Names >

Collection >

Format >

# NCSU Libraries' Rare and Unique Digital Collections

Documenting NC State and North Carolina through historic photographs, architectural drawings, yearbooks, archival materials, and more.



Architecture



Agriculture



Campus and Town



Community and Extension



Sports



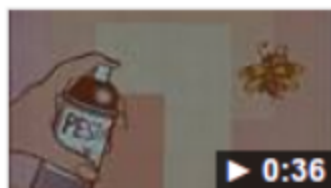
Student Life

# Schema.org types on the Rare & Unique Materials site

- Photograph
- CreativeWork
- LandmarksOrHistoricalBuildings
- Person
- Organization
- Event
- PostalAddress
- GeoCoordinates
- VideoObject
- AudioObject
- and more...

# Rich Snippets: Rare and Unique Materials Video

[USDA Public Service Film, "Bug Sprays and Pets" a P...](#)



[d.lib.ncsu.edu/collections/.../ua024-002-bx0149-066-0...](http://d.lib.ncsu.edu/collections/.../ua024-002-bx0149-066-0...)

May 5, 2013

Using a **bug spray** in your home? That's fine.  
But make sure you remove **pets** and their food  
and water first ...

Third result in Google video search for "bug sprays and pets."

# Library Home Page

## Search

**All** | [Articles](#) | [Books & Media](#) | [Our Website](#)

Search books, articles, journals, & library website

Search

**More Research Tools:** [Databases](#) | [Journal Titles](#) | [Citation Builder](#)

## Technology

[Technology Lending](#)

[Create Digital Media](#)

[Makerspace](#)

## Studying

[Reserve a Room](#)

[GroupFinder](#)

## Courses

[Course Tools](#)

[Course Reserves](#)

[LOBO](#)

 **NCSU LIBRARIES**

**POSTER DESIGN  
CONTEST  
WIN \$500**

**DEADLINE  
JULY 31**

**SUBMIT YOUR POSTER DESIGN FOR THE UPCOMING  
NORTH CAROLINA LITERARY FESTIVAL.**

## The James B. Hunt Jr. Library



**Open now!**

- [Hunt Library on Storify](#)
- [In the News](#)
- [bookBot](#)
- [Help support Hunt](#)
- [Think and Do](#)

## My #HuntLibrary



Your recent and popular photos of the Hunt Library

## Technology Available Now at D. H. Hill

- Laptops** 117 of 135
- Tablets** 8 of 46

## News



## Exhibits



## Events



# Library Home Page Footer

**NCSU Libraries** 2 Broughton Drive, Raleigh, NC 27695-7111 (919) 515-3364 [Contact Us](#)

[Copyright](#) | [Disability Services](#) | [Privacy Statement](#) | [Staff Only](#)

[D. H. Hill Library](#) | [Hunt Library](#) | [Design Library](#) | [Natural Resources Library](#) | [Veterinary Medicine Libra](#)

Embedded semantic markup includes the Libraries name, URL, logo (hidden), address, and telephone number.

# Answers Instead of Search Results

[Alan Alda - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Alan\\_Alda](https://en.wikipedia.org/wiki/Alan_Alda)

Alphonso Joseph D'Abruzzo (born January 28, 1936), better known as **Alan Alda**, is an American actor, director, screenwriter, and author. A six-time Emmy ...

[Robert Alda](#) - [Elizabeth Alda](#) - [Arlene Alda](#) - [Same Time, Next Year](#) (film)

[Alan Alda - IMDb](#)

[www.imdb.com/name/nm0000257/](http://www.imdb.com/name/nm0000257/)

Includes filmography, biography, and upcoming television appearances.

[Biography](#) - [By type](#) - [135 photos](#) - [Awards](#)

[Alan Alda](#)

[www.alanalda.com/](http://www.alanalda.com/)

A description for this result is not available because of this site's robots.txt – learn more.

[Alan Alda Challenges Scientists to Explain: What Is Ti...](#)

[news.sciencemag.org](#) › [News](#) › [ScienceInsider](#) › [December](#)



## Alan Alda

Alphonso Joseph D'Abruzzo, better known as Alan Alda, is an American actor, director, screenwriter, and author. [Wikipedia](#)

**Born:** January 28, 1936 (age 76), [The Bronx](#)

**Spouse:** [Arlene Alda](#) (m. 1957)

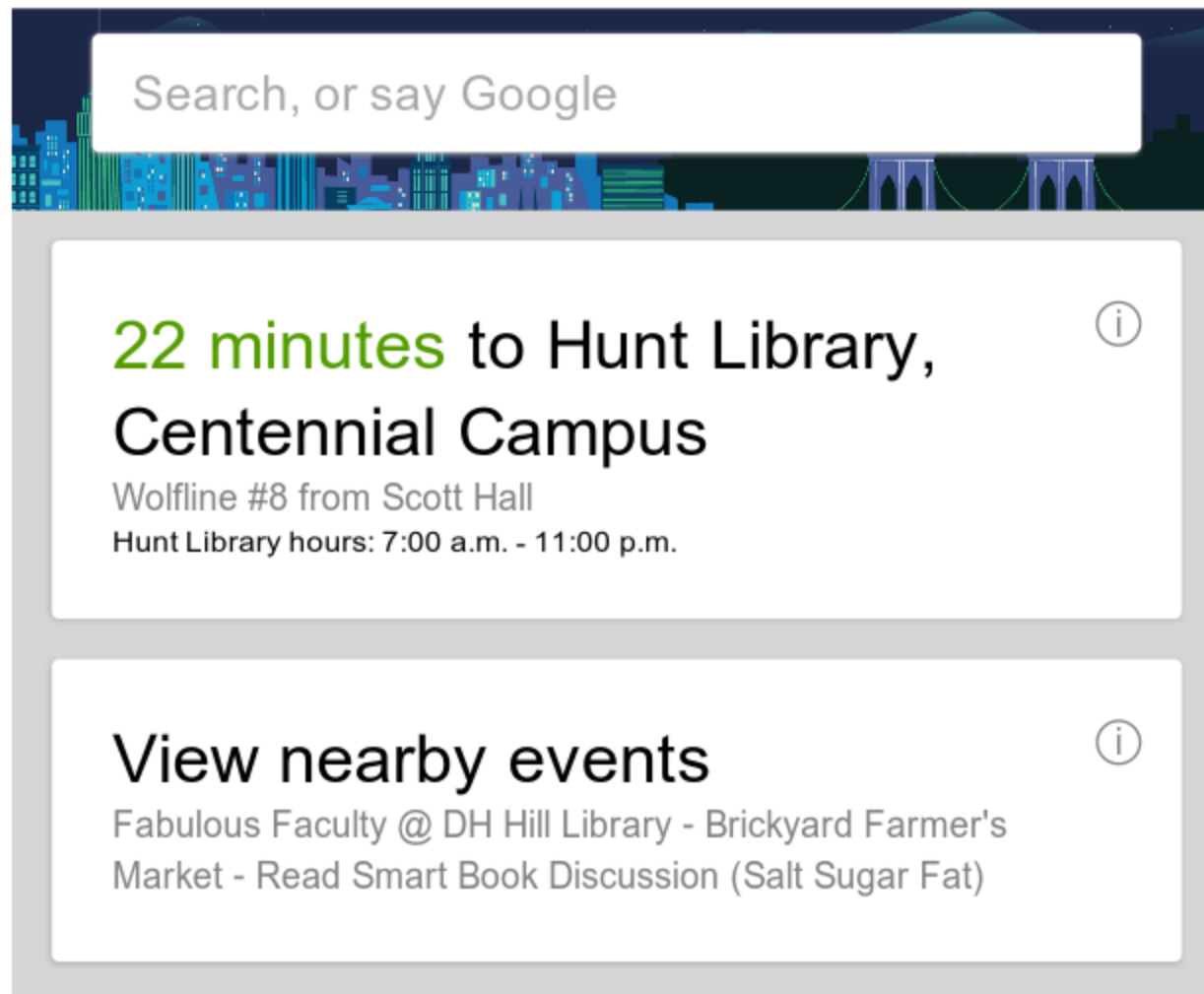
**Parents:** [Robert Alda](#), [Joan Browne](#)

**Books:** [Never Have Your Dog Stuffed And Other Things I've Learned](#), [More](#)

**Children:** [Elizabeth Alda](#), [Beatrice Alda](#)

Movies and TV shows

# Library Services and Google Now



Search, or say Google

**22 minutes** to Hunt Library,  
Centennial Campus (i)

Wofline #8 from Scott Hall  
Hunt Library hours: 7:00 a.m. - 11:00 p.m.

**View nearby events** (i)

Fabulous Faculty @ DH Hill Library - Brickyard Farmer's Market - Read Smart Book Discussion (Salt Sugar Fat)

# **Research Question: Are Academic Libraries Publishing Embedded Structured Data?**

## **How about digital collections?**

Get some rough idea of the landscape of use of embedded semantic markup and schema.org among academic institutions and academic libraries.

---



# Common Crawl

**Common Crawl is a non-profit foundation dedicated to providing an open repository of web crawl data that can be accessed and analyzed by everyone.**



- Over 5 billion Web pages (3,005,629,093 for the set I looked at)
- 40,600,000 domains
- ~81TB total
- Other sets of crawl data are being added (Blekko)
- Uses PageRank so is a snapshot of the current most popular part of the Web

<http://commoncrawl.org/>

# Web Data Commons

## Extracting Structured Data from the Common Web Crawl

<http://webdatacommons.org/>

Domains with Triples	2,286,277
Typed Entities	1,811,471,956
Triples/Statements	7,350,953,995

Cost to extract the data from the Common Crawl: \$398

# What's an N-Quad?

`_:node6eecc231551a72e90e7efb3dc3fc26`  
<http://schema.org/Photograph/name> "Mary Travers singing live  
on stage" <http://d.lib.ncsu.edu/collections/catalog/0228376> .

## **Subject Predicate Object *Context***

An N-Quad is an RDF statement that also includes a context piece at the end. Context is the URL of the HTML page from which the data was extracted.

Line-based format makes it easier to do some rough parsing.

# Methodology

1. Grab all of the Web Data Commons extracted N-Quads (7,350,953,995 of them) from the August 2012 Common Crawl Corpus
2. Use commandline tools (cat & grep) to boil things down to just N-Quads that contain ".edu" somewhere, anywhere
3. Further reduce by university (duke.edu, nccu.edu, ncsu.edu, unc.edu)
4. Even further reduce to just libraries (library.duke.edu, lib.ncsu.edu, lib.ncsu.edu)
5. Run some scripts over these smaller batches to get some results

**All very much a crude pass at this!**

# Total Statements (N-Quads)

All triples	7,350,953,995
All .edu	8,178,985
duke.edu	58,867
nccu.edu	79
ncsu.edu	9,339
unc.edu	52,751

These are all the statements that contain the text (.edu, duke.edu, nccu.edu, ncsu.edu, unc.edu) anywhere in the N-Quad.

# Unique Contexts/Pages

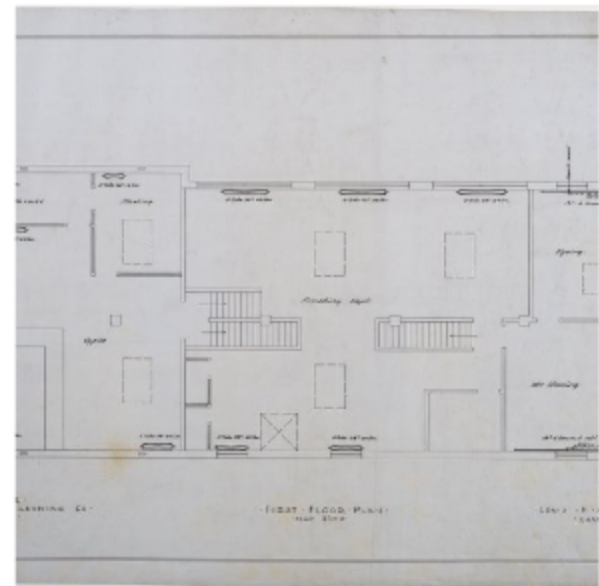
duke.edu	55,344	library.duke.edu	1,123
nccu.edu	2	n/a*	
ncsu.edu	664	lib.ncsu.edu	155
unc.edu	2,837	lib.unc.edu	503

These are the number of unique contexts (HTML pages) that are included in the Common Crawl and that have included some embedded semantic markup that Web Data Commons has extracted.

\* Uncertain how to target just NCCU Libraries.

# Digital Collections at NCSU: Rare & Unique Materials

1. <http://d.lib.ncsu.edu/collections/>
2. <http://d.lib.ncsu.edu/collections/catalog/0228376>
3. <http://d.lib.ncsu.edu/collections/catalog/bh2127pnc001>
4. [http://d.lib.ncsu.edu/collections/catalog/unccmc00145-002-ff0003-002-004\\_0002](http://d.lib.ncsu.edu/collections/catalog/unccmc00145-002-ff0003-002-004_0002)



# Use of Schema.org

145,351 N-Quads from all 8,178,985 .edu N-Quads use schema.org.

duke.edu	1901	library.duke.edu	1660
nccu.edu	3		
ncsu.edu	326	lib.ncsu.edu	102
unc.edu	301	lib.unc.edu	25

\* These numbers look at the whole quad and not just the context. So these universities and libraries might not actually be using schema.org (or may have been using schema.org but the documents that have schema.org have not been crawled by the Common Crawl).



# Most Popular Types Used by TRLN Parent Institutions

- <http://www.w3.org/2006/vcard/ns> (Electronic business cards)
- <http://www.w3.org/2002/12/cal/icaltzd> (Calendar events)
- <http://vocab.sindice.com/xfn#mePage> (XHTML Friends Network)
- <http://schema.org/BlogPosting>

# Most Popular Schema.org Types Used by TRLN Parent Institutions

- <http://schema.org/CollectionPage> (Duke)
- <http://schema.org/VideoObject> (All TRLN Libraries)
- <http://schema.org/BlogPosting> (All TRLN Libraries)
- <http://schema.org/Event> (Duke)
- <http://schema.org/Person> (UNC)
- <http://schema.org/ScholarlyArticle> (UNC)

# OK, What Does This Mean?

## Preliminary Thoughts.

- Common Crawl is crawling .edu domains (Good.)
- Common Crawl is crawling TRLN libraries (Great!)
- Common Crawl is not extensively crawling, especially digital special collections :-)
- We could probably improve what gets crawled by increasing our PageRank.
- Academic institutions are using some embedded markup and starting to use schema.org.
- Our use as of this is a just start. We could be doing more.

**What questions about  
Big Web Data would  
you be interested in?**

---

# Bonus Slide: What Could Libraries and Archives Do With These Web Technologies?

Libraries and archives can be both producers and consumers of this data.

- Improve discoverability of our services and collections
- Domain-specific vertical search engines
- The public interoperability API to our data (replace library-specific APIs)
- Publish data sets and harvest them from others
- Archiving the Web with improved metadata
- What are your ideas?

# Links

- [http://en.wikipedia.org/wiki/Microdata\\_\(HTML\)](http://en.wikipedia.org/wiki/Microdata_(HTML))
- <http://en.wikipedia.org/wiki/RDFa>
- <http://schema.org/>
- <http://www.w3.org/community/schemabibex/>
- <http://commoncrawl.org/> and [URL search tool](#)
- <http://webdatacommons.org/>

NCSU sites that use embedded semantic markup (Microdata) and Schema.org:

- [Student Leadership Initiative](#)
- [NCSU Libraries Rare & Unique Materials](#)

# Credits

- Friendly Robot by Sean Hannan
- Google Now HTML and CSS derived from Bennett Feely  
<http://codepen.io/bennettfeely/details/Ftczh>

# Jason Ronallo

@ronallo

<http://jronallo.github.io/>

jason\_ronallo@ncsu.edu