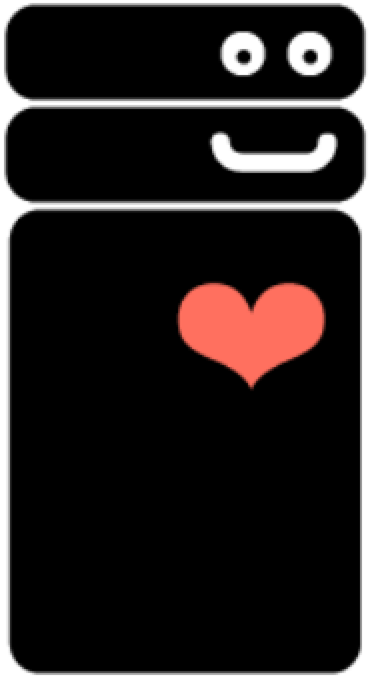
Embedded Semantic Markup, schema.org, the Common Crawl, and Web Data Commons: What Big Web Data Means for Libraries and Archives

Jason Ronallo
NCSU Libraries
@ronallo

Slides: https://ronallo.com/presentations/2013-dlf

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



Semantics

Semantics in HTML

```
<ol>
  <li>First item</li>
  <li>Second item</li>
  <li>Third item</li>
</ol>
```

HTML5 Semantics

```
<nav></nav>
<header></header>
<article>
  <section></section>
  <section></section>
</article>
<footer></footer>
```

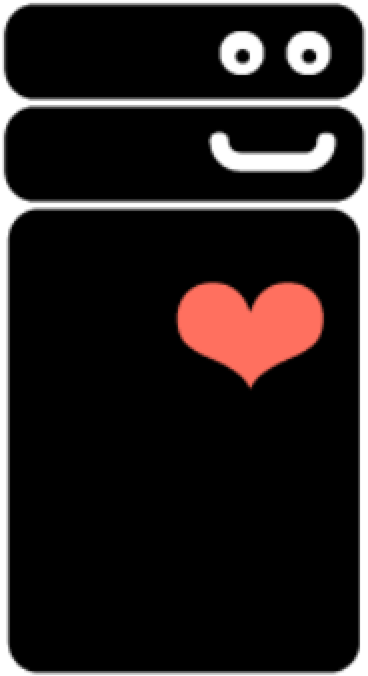
Trapped Knowledge

Embedded Semantic Markup

Embedded Semantic Markup Example

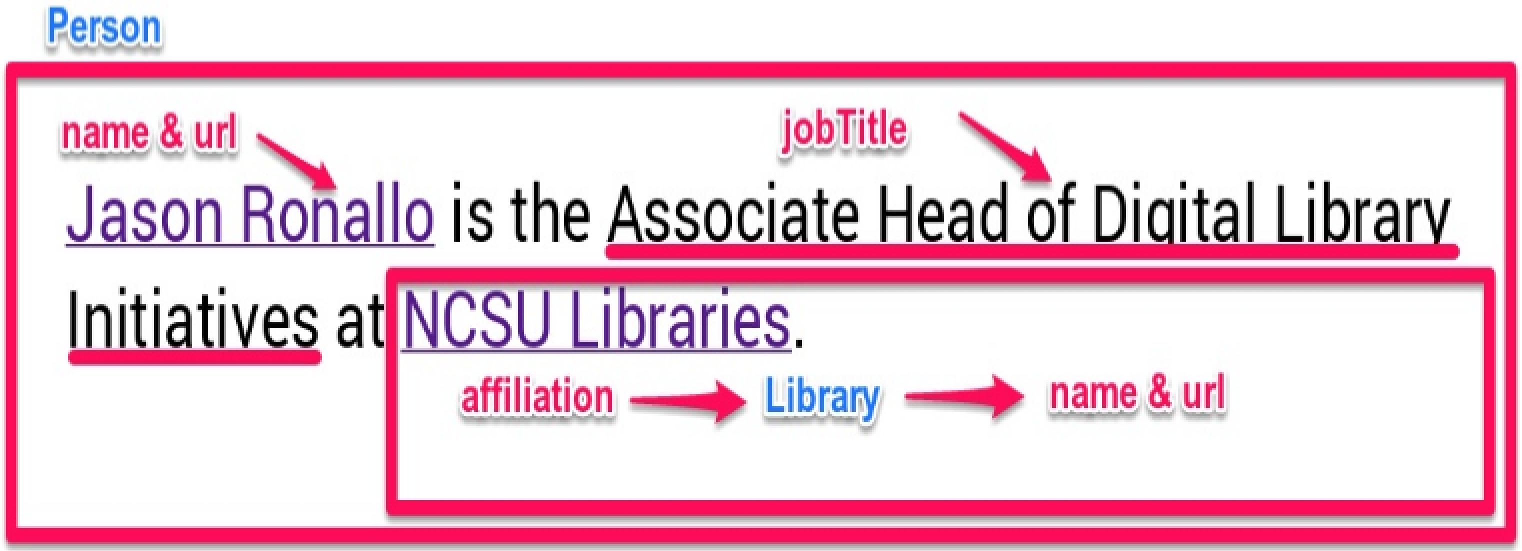
Jason Ronallo is the Associate Head of Digital Library Initiatives at NCSU Libraries.

Embedded Semantic Markup Is Hidden Annotations Meant for Machines



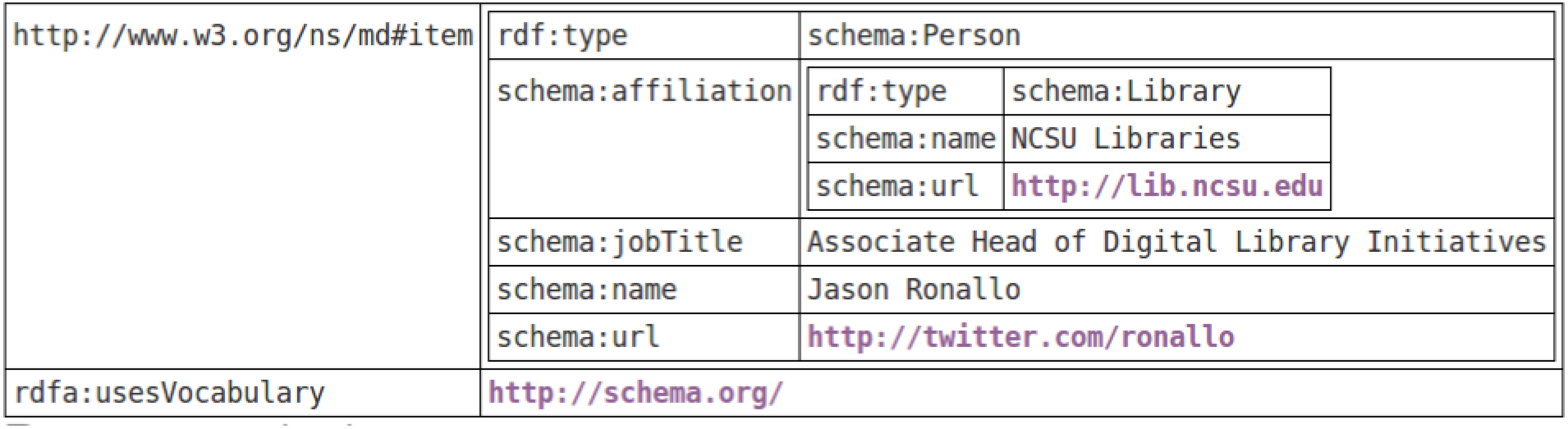
Embedded Semantic Markup Exposed

Person has the properties name, url, jobTitle, and affiliation. The affiliation is with a Library that has a name and url.



Jason Ronallo is the Associate Head of Digital Library Initiatives at NCSU Libraries.

Embedded Semantic Markup Structure



http://www.w3.org/ns/md#item	rdf:type	schema:Person	
	schema:affiliation	rdf:type	schema:Library
		schema:name	NCSU Libraries
		schema:url	http://lib.ncsu.edu
	schema:jobTitle	Associate Head of Digital Library Initiatives	
	schema:name	Jason Ronallo	
	schema:url	http://twitter.com/ronallo	
rdfa:usesVocabulary	http://schema.org/		

Embedded Semantic Markup HTML

```
<span itemscope itemtype="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="Library">
    <span itemprop="name">
      <a itemprop="url" href="http://lib.ncsu.edu">
       NCSU Libraries</a>
     </span>
  </span>.
</span>
```

JSON Serialization

```
{"items": [
  { "type": [ "http://schema.org/Person" ],
   "properties": {
    "url": [ "http://twitter.com/ronallo" ],
    "name": [ "Jason Ronallo" ],
    "jobTitle": [ "Associate Head of Digital Library Initiatives" ],
    "affiliation": [
     { "type": [ "http://schema.org/Library" ],
      "properties": {
       "name": [ "NCSU Libraries" ],
       "url": [ "http://lib.ncsu.edu/" ]
      }
     }
    ]
   }
  }
 ]
}
```

RDF (Turtle)

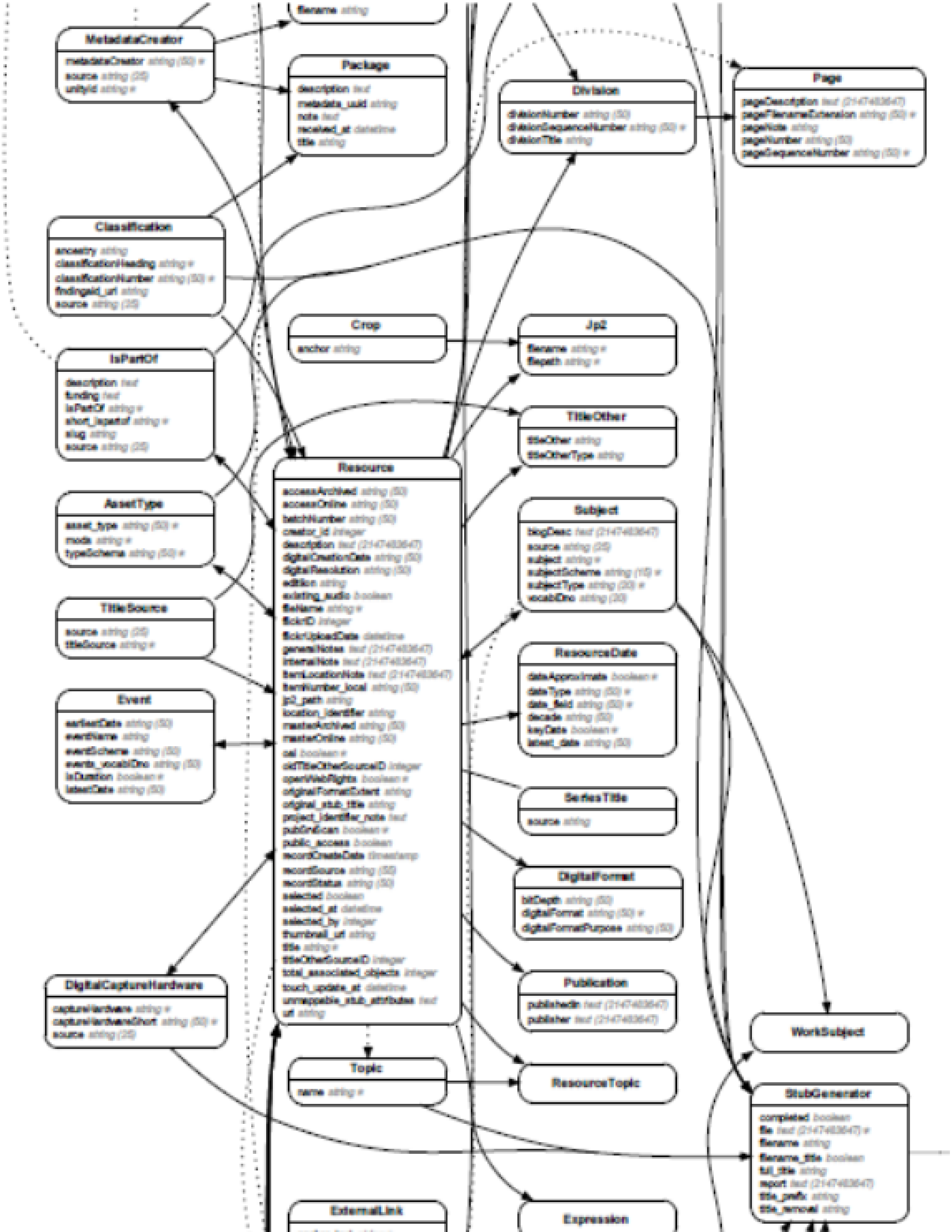
```
@prefix md: <http://www.w3.org/ns/md#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfa: <http://www.w3.org/ns/rdfa#> .
@prefix schema: <http://schema.org/> .

<> md:item ( [ a schema:Person;
          schema:affiliation [ a schema:Library;
               schema:name "NCSU Libraries";
               schema:url <http://lib.ncsu.edu> ];
          schema:jobTitle "Associate Head of Digital Library Initiatives";
          schema:name "Jason Ronallo";
          schema:url <http://twitter.com/ronallo> ] );
  rdfa:usesVocabulary schema: .
```

Types of
Embedded Semantic Markup

- Microformats
- RDFa (Lite)
- Microdata

Why use embedded semantic markup?

- A way to *structure* data in HTML
- To communicate with machines
- Your eyes are on the Web site (Maintain this data in one place and keep it in sync)
- Rich Metadata to Rich Embedded Data



filename string
MetadataCreator
metadataCreator string (50) *
source string (25)
unityid string *
Package
description text
metadata_uuid string
note text
received_at datetime
title string
Division
divisionNumber string (50)
divisionSequenceNumber string (50) *
divisionTitle string
Page
pageDescription text (2147483647)
pageFilenameExtension string (50) *
pageNote string
pageNumber string (50)
pageSequenceNumber string (50) *
Classification
ancestry string
classificationHeading string *
classificationNumber string (50) *
findingaid_url string
source string (25)
Crop
anchor string
Jp2
filename string *
filepath string *
IsPartOf
description text
funding text
isPartOf string *
short_ispartof string *
slug string
source string (25)
TitleOther
titleOther string
titleOtherType string
Resource
accessArchived string (50)
accessOnline string (50)
batchNumber string (50)
creator_id integer
description text (2147483647)
digitalCreationDate string (50)
digitalResolution string (50)
edition string
existing_audio boolean
fileName string *
flickrID integer
flickrUploadDate datetime
generalNotes text (2147483647)
internalNote text (2147483647)
itemLocationNote text (2147483647)
itemNumber_local string (50)
jp2_path string
location_identifier string
masterArchived string (50)
masterOnline string (50)
oai boolean *
oldTitleOtherSourceID integer
openWebRights boolean *
originalFormatExtent string
original_stub_title string
project_identifier_note text
pubOnEScan boolean *
public_access boolean
recordCreateDate timestamp
recordSource string (50)
recordStatus string (50)
selected boolean
selected_at datetime
selected_by integer
thumbnail_url string
title string *
titleOtherSourceID integer
total_associated_objects integer
touch_update_at datetime
unmappable_stub_attributes text
url string
AssetType
asset_type string (50) *
mods string *
typeSchema string (50) *
Subject
blogDesc text (2147483647)
source string (25)
subject string *
subjectScheme string (15) *
subjectType string (20) *
vocabIDno string (20)
TitleSource
source string (25)
titleSource string *
ResourceDate
dateApproximate boolean *
dateType string (50) *
date_field string (50) *
decade string (50)
keyDate boolean *
latest_date string (50)
Event
earliestDate string (50)
eventName string
eventScheme string (50)
events_vocabIDno string (50)
isDuration boolean *
latestDate string (50)
SeriesTitle
source string
DigitalFormat
bitDepth string (50)
digitalFormat string (50) *
digitalFormatPurpose string (50)
DigitalCaptureHardware
captureHardware string *
captureHardwareShort string (50) *
source string (25)
Publication
publishedIn text (2147483647)
publisher text (2147483647)
WorkSubject
Topic
name string *
ResourceTopic
StubGenerator
completed boolean
file text (2147483647) *
filename string
filename_title boolean
full_title string
report text (2147483647)
title_prefix string
title_removal string
ExternalLink
Expression

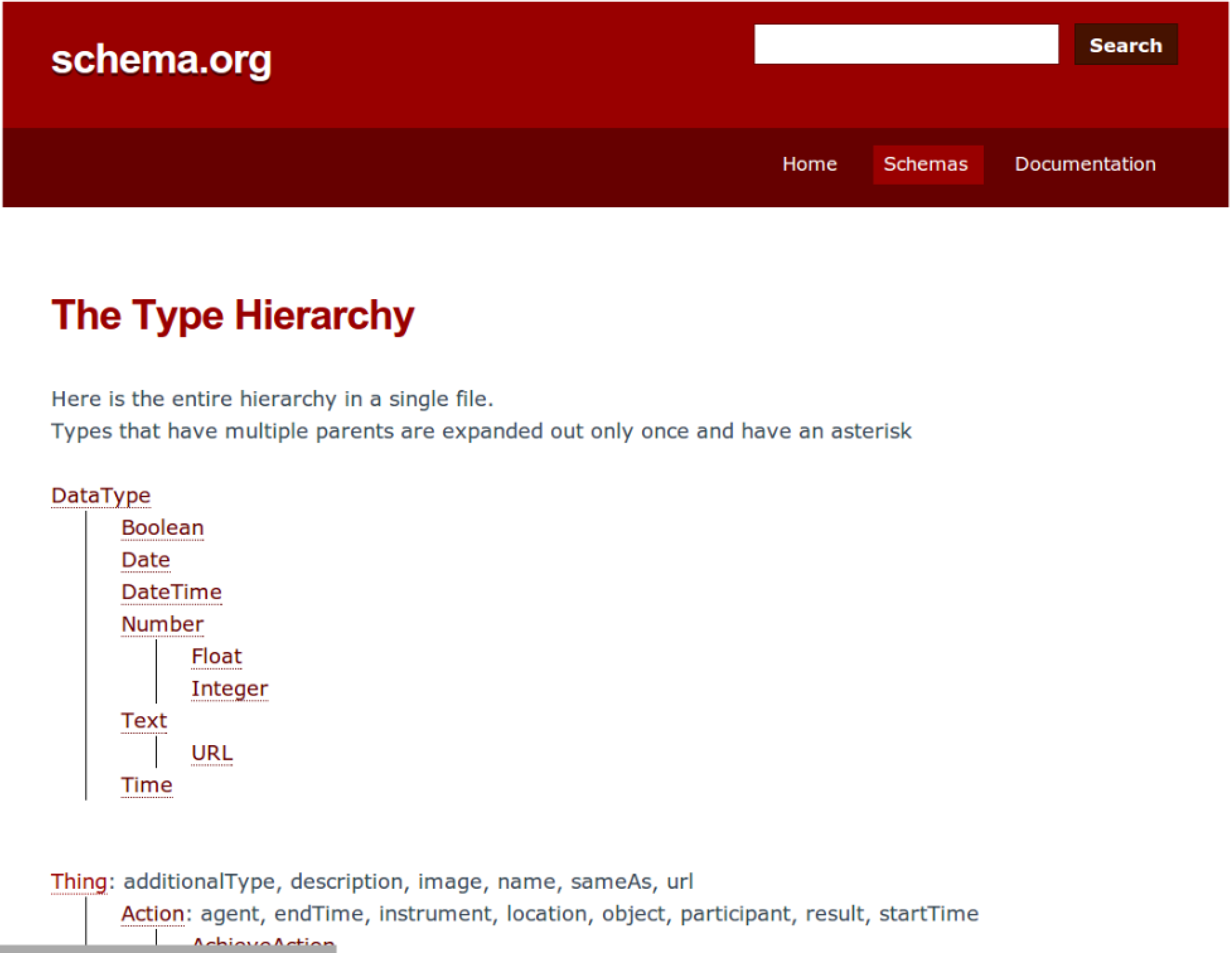
The End of Dumbed Down Metadata

Vocabularies for Understanding

Schema.org

Schema.org

- Shared, Web-scale, single-stop vocabulary for describing the content of Web pages.
- Released 2011
- Maintained by the major search eninges (Bing, Google, Yahoo, Yandex)
- Everything is a Thing
- 407+ Types of Things (Numbers from early 2013)
- 545+ Properties of Things
- Everything from Airport to Library to Volcano
- Expanding and open to proposals to update the schema (see SchemaBibEx W3C Community Group)
- Single site for documenation. Easy to use. No fragmentation.

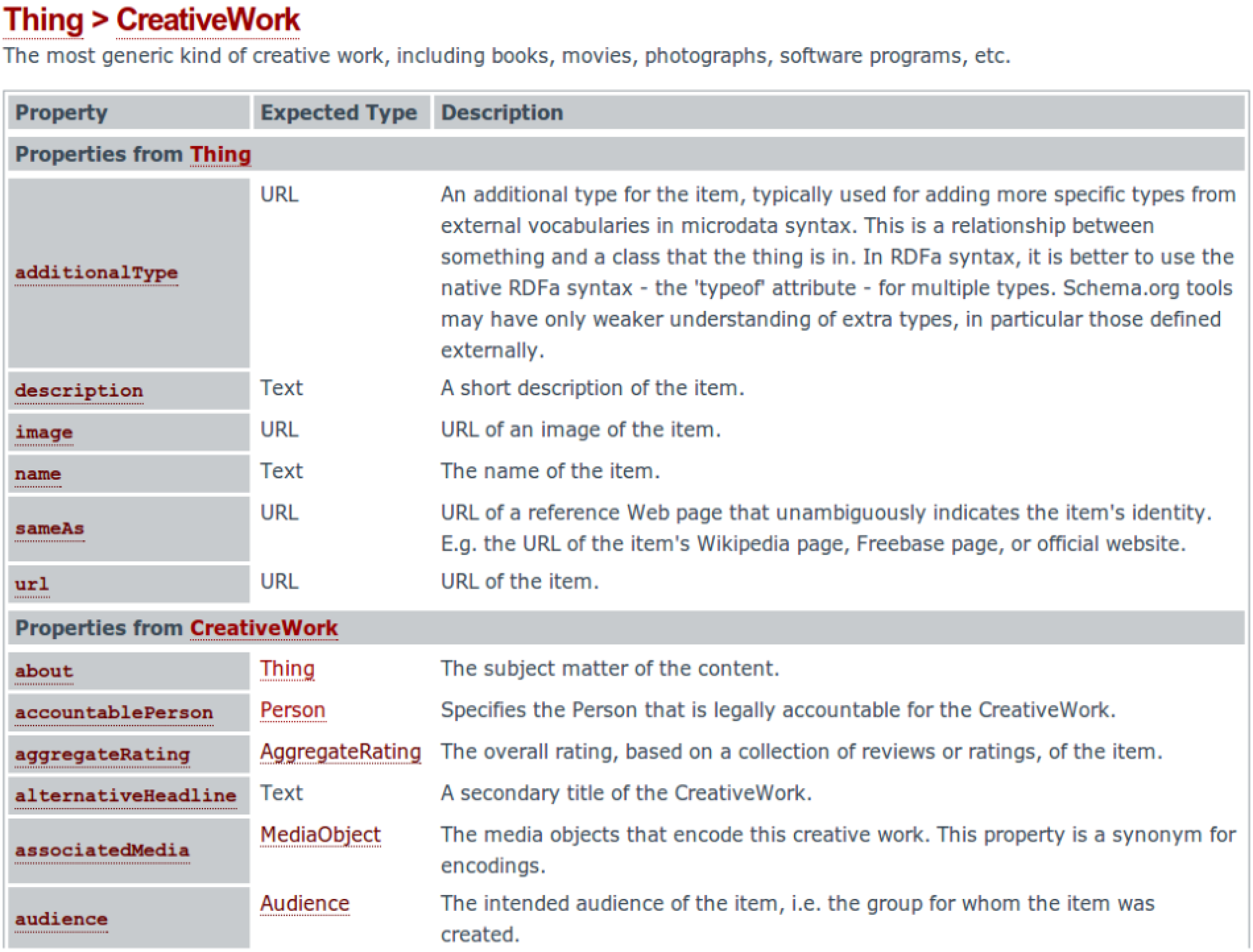


The Type Hierarchy

Here is the entire hierarchy in a single file.
Types that have multiple parents are expanded out only once and have an asterisk

- DataType
  - Boolean
  - Date
  - DateTime
  - Number
    - Float
    - Integer
  - Text
    - URL
  - Time

- Thing: additionalType, description, image, name, sameAs, url
  - Action: agent, endTime, instrument, location, object, participant, result, startTime
    - AchieveAction



Thing > CreativeWork

The most generic kind of creative work, including books, movies, photographs, software programs, etc.

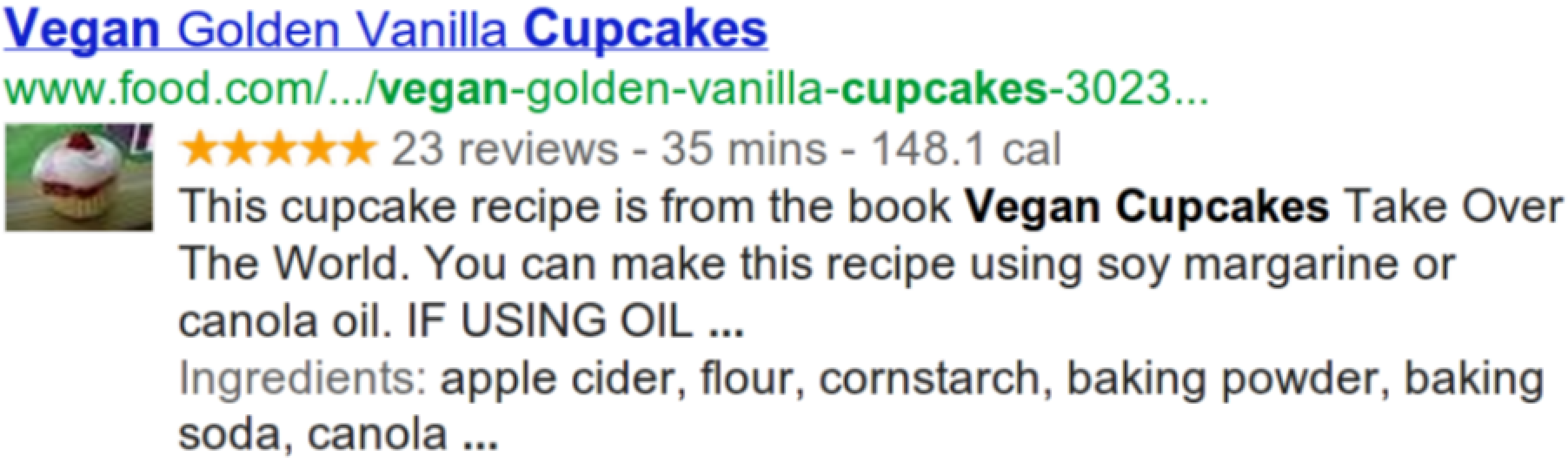
Property	Expected Type	Description
Properties from Thing		
additionalType	URL	An additional type for the item, typically used for adding more specific types from external vocabularies in microdata syntax. This is a relationship between something and a class that the thing is in. In RDFa syntax, it is better to use the native RDFa syntax - the 'typeof' attribute - for multiple types. Schema.org tools may have only weaker understanding of extra types, in particular those defined externally.
description	Text	A short description of the item.
image	URL	URL of an image of the item.
name	Text	The name of the item.
sameAs	URL	URL of a reference Web page that unambiguously indicates the item's identity. E.g. the URL of the item's Wikipedia page, Freebase page, or official website.
url	URL	URL of the item.
Properties from CreativeWork		
about	Thing	The subject matter of the content.
accountablePerson	Person	Specifies the Person that is legally accountable for the CreativeWork.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
alternativeHeadline	Text	A secondary title of the CreativeWork.
associatedMedia	MediaObject	The media objects that encode this creative work. This property is a synonym for encodings.
audience	Audience	The intended audience of the item, i.e. the group for whom the item was created.

Why use Schema.org?

- Growing implementation base.
- Software implementations (CMS).
- With implementations and known consumers, other consumers will follow.

Improve Discoverability on the Open Web

Rich Snippets



Vegan Golden Vanilla **Cupcakes**

www.food.com/.../**vegan**-golden-vanilla-**cupcakes**-3023...

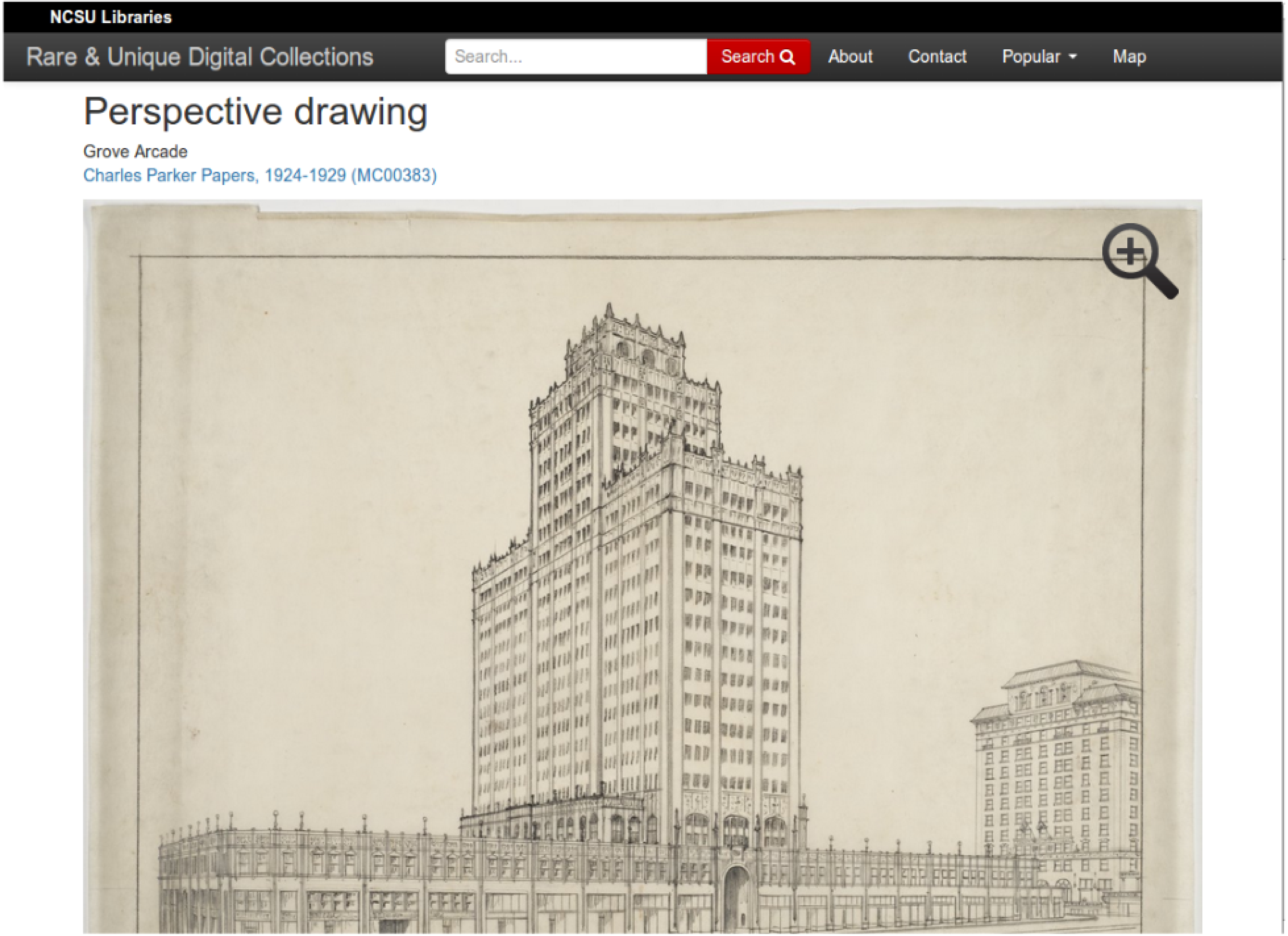
★★★★★ 23 reviews - 35 mins - 148.1 cal

This cupcake recipe is from the book **Vegan Cupcakes** Take Over The World. You can make this recipe using soy margarine or canola oil. IF USING OIL ...

Ingredients: apple cider, flour, cornstarch, baking powder, baking soda, canola ...

Library Examples

- NCSU Libraries
- Future Possibilities

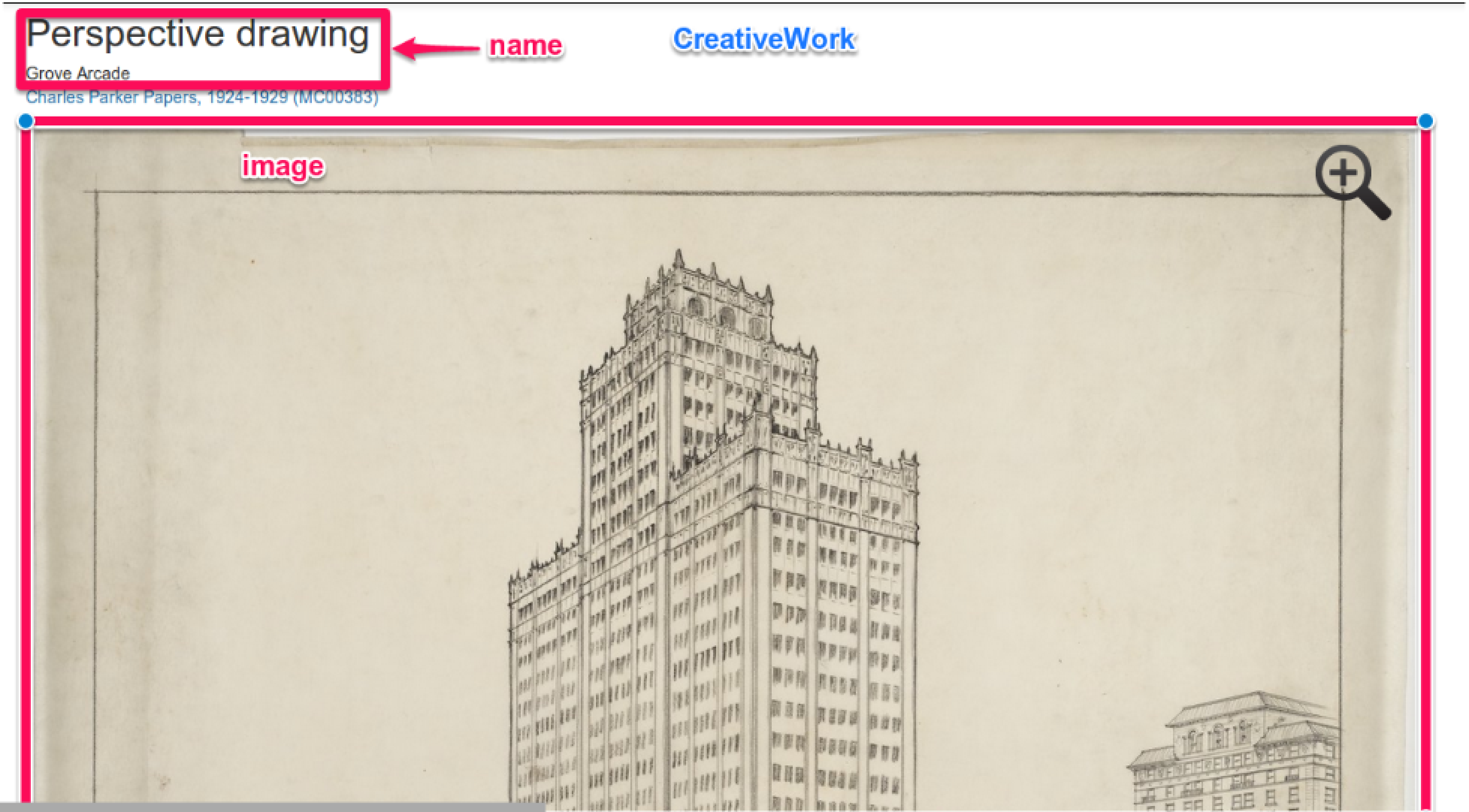


Perspective drawing

Grove Arcade

Charles Parker Papers, 1924-1929 (MC00383)

A Grove Arcade Drawing is a http://schema.org/CreativeWork

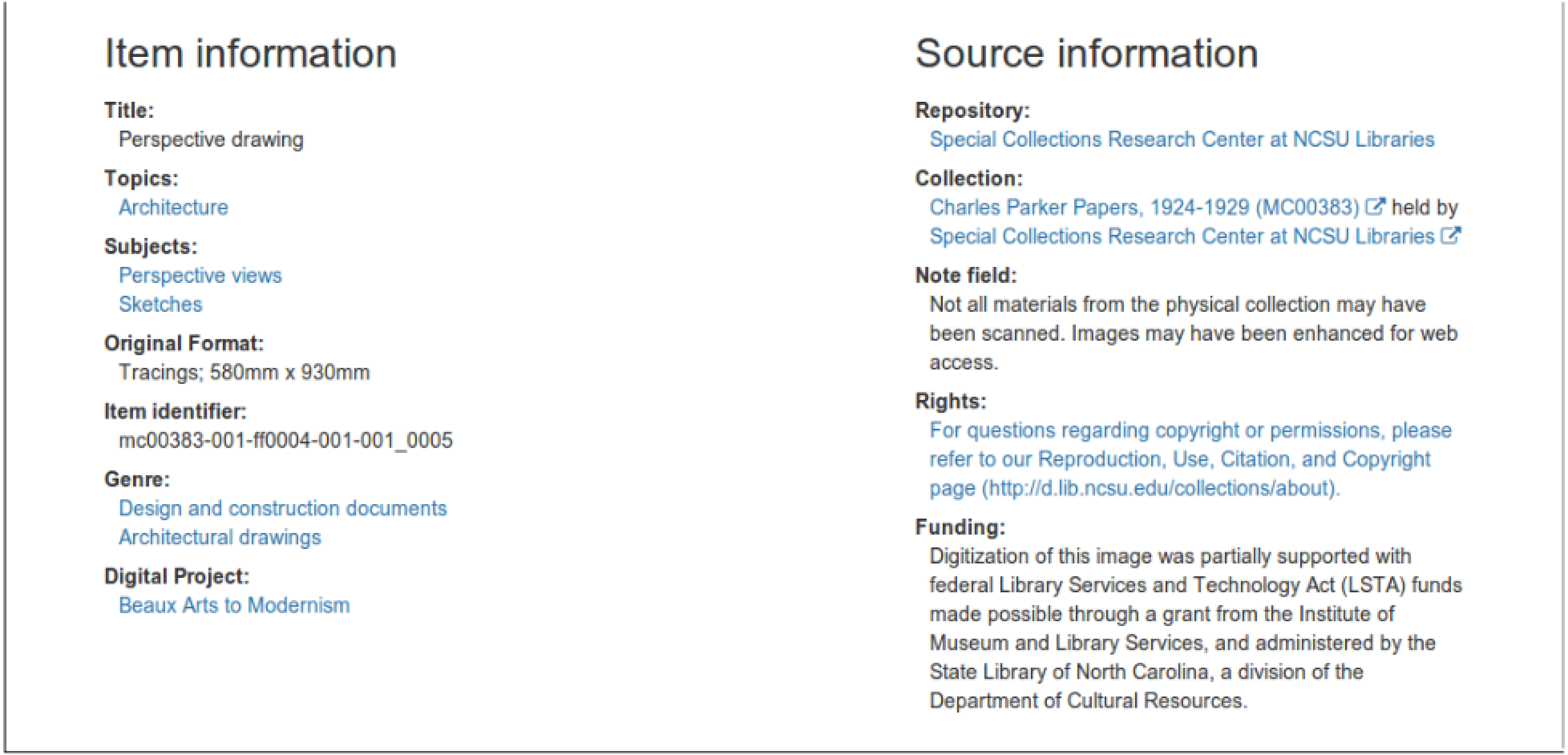


Perspective drawing

Grove Arcade

Charles Parker Papers, 1924-1929 (MC00383)

Item Information



Item information

Title:
Perspective drawing

Topics:
Architecture

Subjects:
Perspective views
Sketches

Original Format:
Tracings; 580mm x 930mm

Item identifier:
mc00383-001-ff0004-001-001_0005

Genre:
Design and construction documents
Architectural drawings

Digital Project:
Beaux Arts to Modernism

Source information

Repository:
Special Collections Research Center at NCSU Libraries

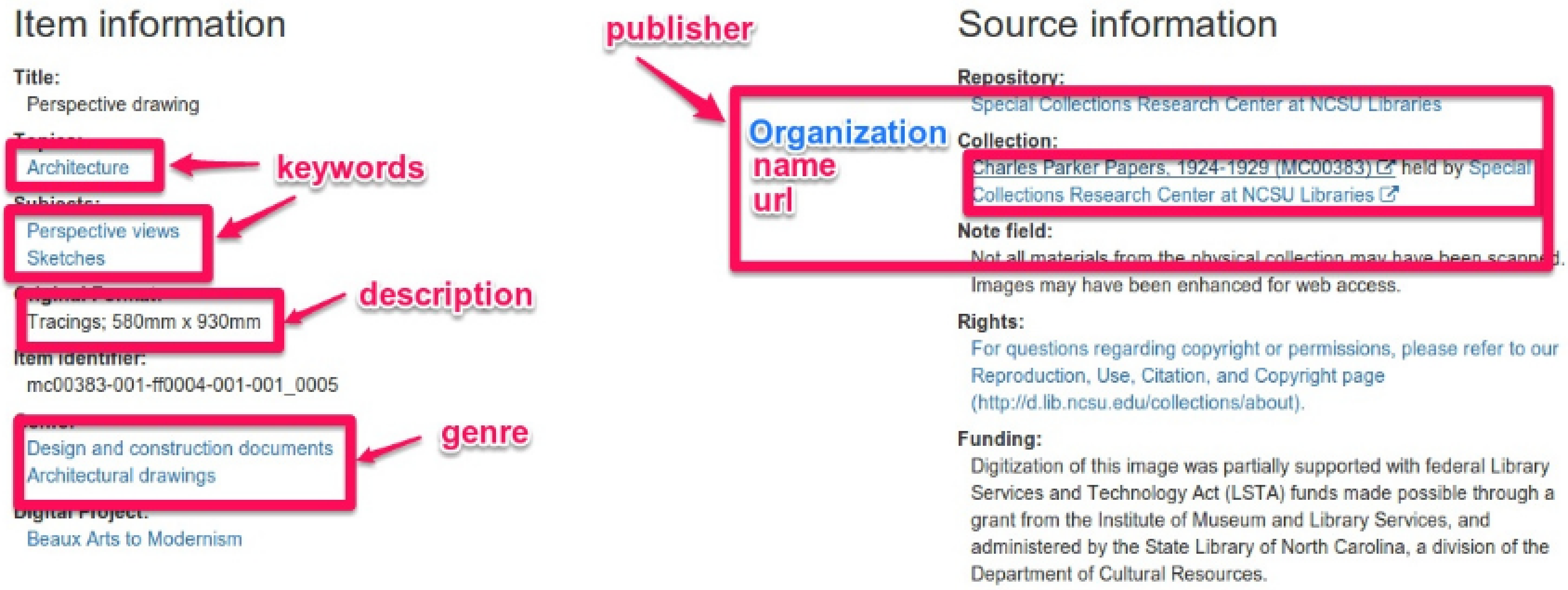
Collection:
Charles Parker Papers, 1924-1929 (MC00383) held by Special Collections Research Center at NCSU Libraries

Note field:
Not all materials from the physical collection may have been scanned. Images may have been enhanced for web access.

Rights:
For questions regarding copyright or permissions, please refer to our Reproduction, Use, Citation, and Copyright page (http://d.lib.ncsu.edu/collections/about).

Funding:
Digitization of this image was partially supported with federal Library Services and Technology Act (LSTA) funds made possible through a grant from the Institute of Museum and Library Services, and administered by the State Library of North Carolina, a division of the Department of Cultural Resources.

Embedded Item Information



Item information

Title:
Perspective drawing

Topics:
Architecture

keywords

Subjects:
Perspective views
Sketches

Original Format:
Tracings; 580mm x 930mm

description

Item identifier:
mc00383-001-ff0004-001-001_0005

Genre:
Design and construction documents
Architectural drawings

genre

Digital Project:
Beaux Arts to Modernism

publisher

Organization
name
url

Source information

Repository:
Special Collections Research Center at NCSU Libraries

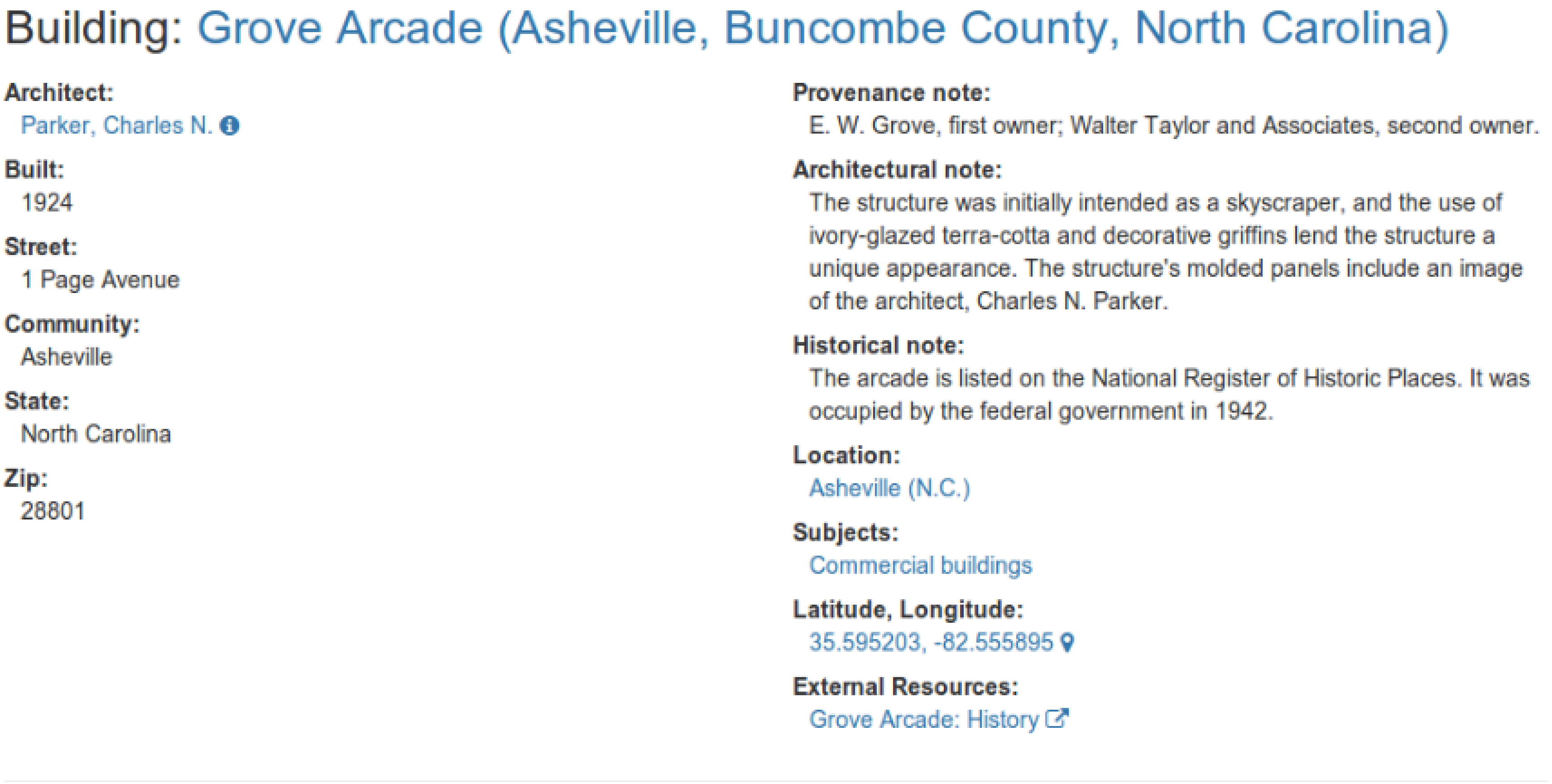
Collection:
Charles Parker Papers, 1924-1929 (MC00383) held by Special Collections Research Center at NCSU Libraries

Note field:
Not all materials from the physical collection may have been scanned. Images may have been enhanced for web access.

Rights:
For questions regarding copyright or permissions, please refer to our Reproduction, Use, Citation, and Copyright page (http://d.lib.ncsu.edu/collections/about).

Funding:
Digitization of this image was partially supported with federal Library Services and Technology Act (LSTA) funds made possible through a grant from the Institute of Museum and Library Services, and administered by the State Library of North Carolina, a division of the Department of Cultural Resources.

Building Information



Building: Grove Arcade (Asheville, Buncombe County, North Carolina)

Architect:
Parker, Charles N.

Built:
1924

Street:
1 Page Avenue

Community:
Asheville

State:
North Carolina

Zip:
28801

Provenance note:
E. W. Grove, first owner; Walter Taylor and Associates, second owner.

Architectural note:
The structure was initially intended as a skyscraper, and the use of ivory-glazed terra-cotta and decorative griffins lend the structure a unique appearance. The structure's molded panels include an image of the architect, Charles N. Parker.

Historical note:
The arcade is listed on the National Register of Historic Places. It was occupied by the federal government in 1942.

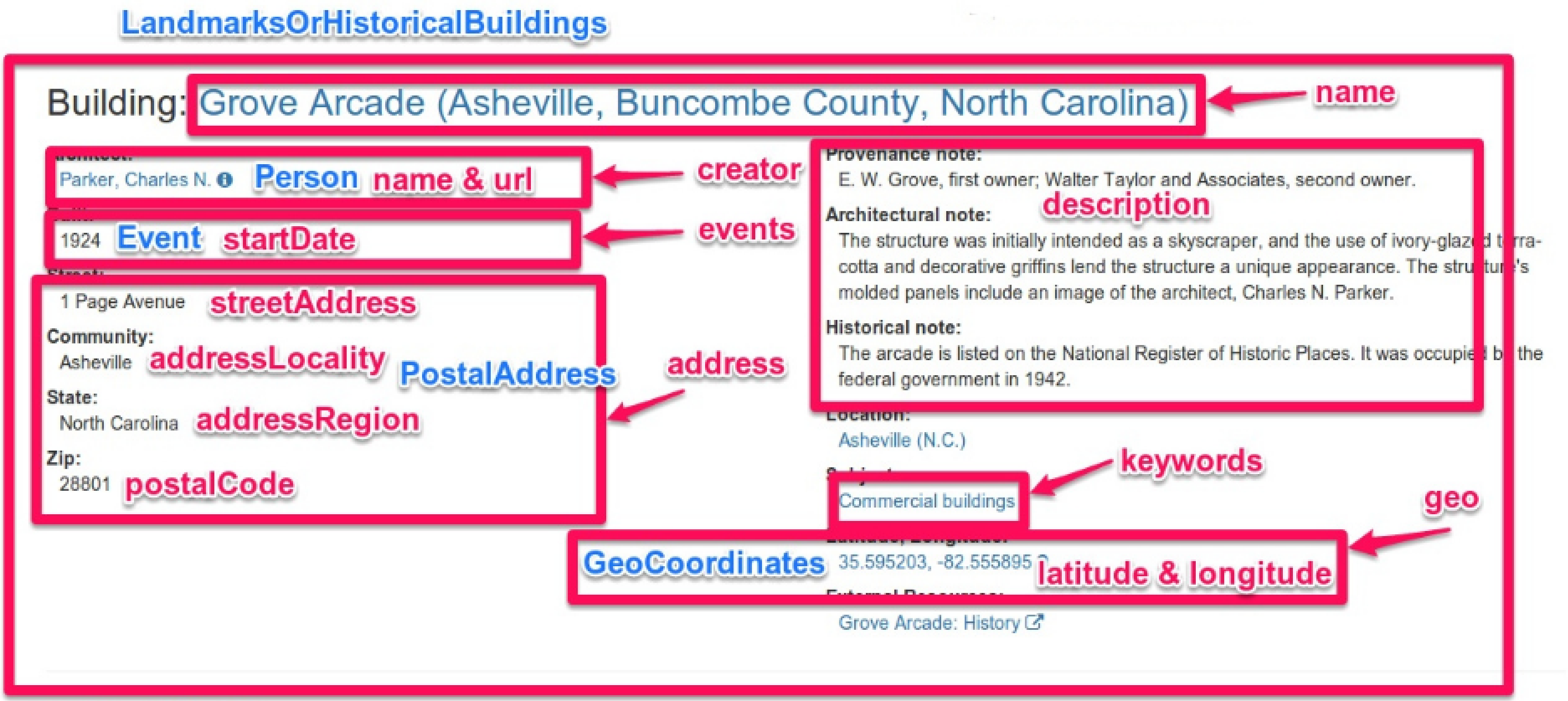
Location:
Asheville (N.C.)

Subjects:
Commercial buildings

Latitude, Longitude:
35.595203, -82.555895

External Resources:
Grove Arcade: History

Embedded Building Information



Building: Grove Arcade (Asheville, Buncombe County, North Carolina)

Architect:
Parker, Charles N.

Built:
1924

Street:
1 Page Avenue

Community:
Asheville

State:
North Carolina

Zip:
28801

Provenance note:
E. W. Grove, first owner; Walter Taylor and Associates, second owner.

Architectural note:
The structure was initially intended as a skyscraper, and the use of ivory-glazed terra-cotta and decorative griffins lend the structure a unique appearance. The structure's molded panels include an image of the architect, Charles N. Parker.

Historical note:
The arcade is listed on the National Register of Historic Places. It was occupied by the federal government in 1942.

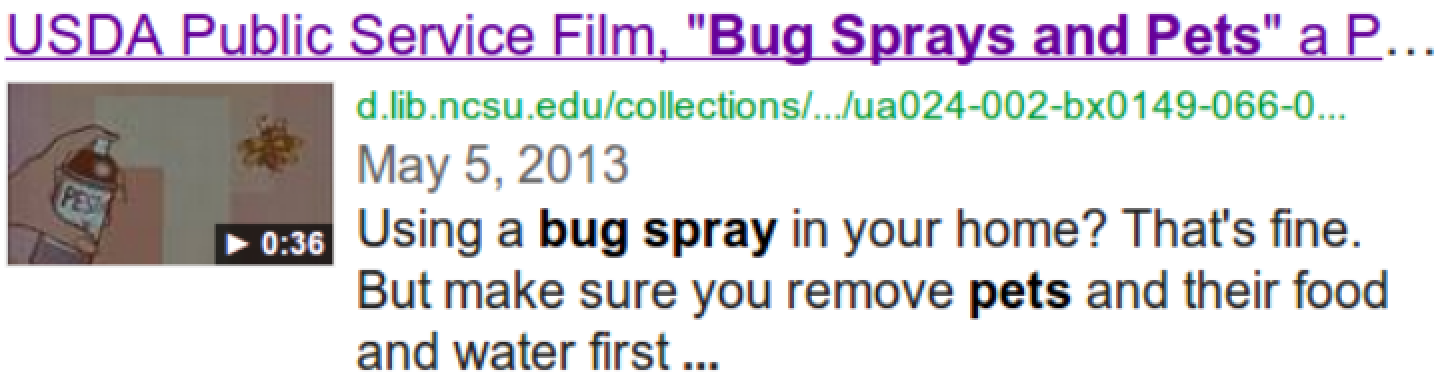
Location:
Asheville (N.C.)

Subject:
Commercial buildings

Latitude, Longitude:
35.595203, -82.555895

External Resources:
Grove Arcade: History

Rich Snippets: Video



USDA Public Service Film, "**Bug Sprays and Pets**" a P…

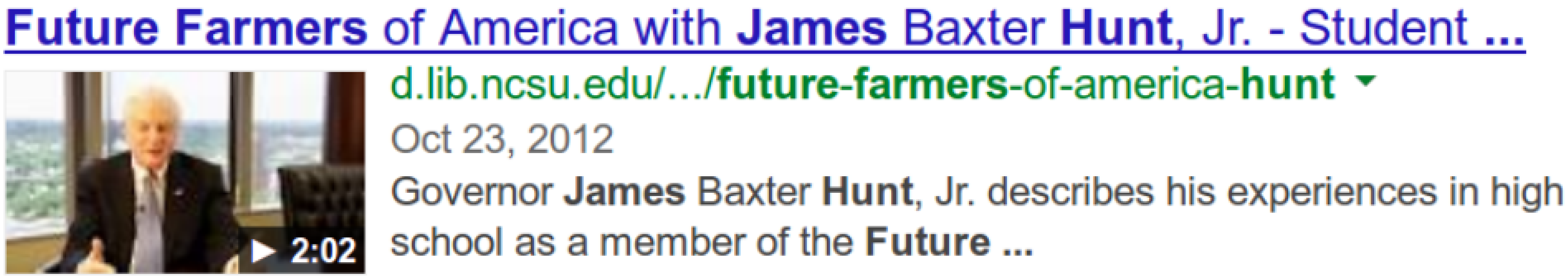
▶ 0:36

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."



Future Farmers of America with **James** Baxter **Hunt**, Jr. - Student **...**

▶ 2:02

d.lib.ncsu.edu/.../**future-farmers**-of-america-**hunt**

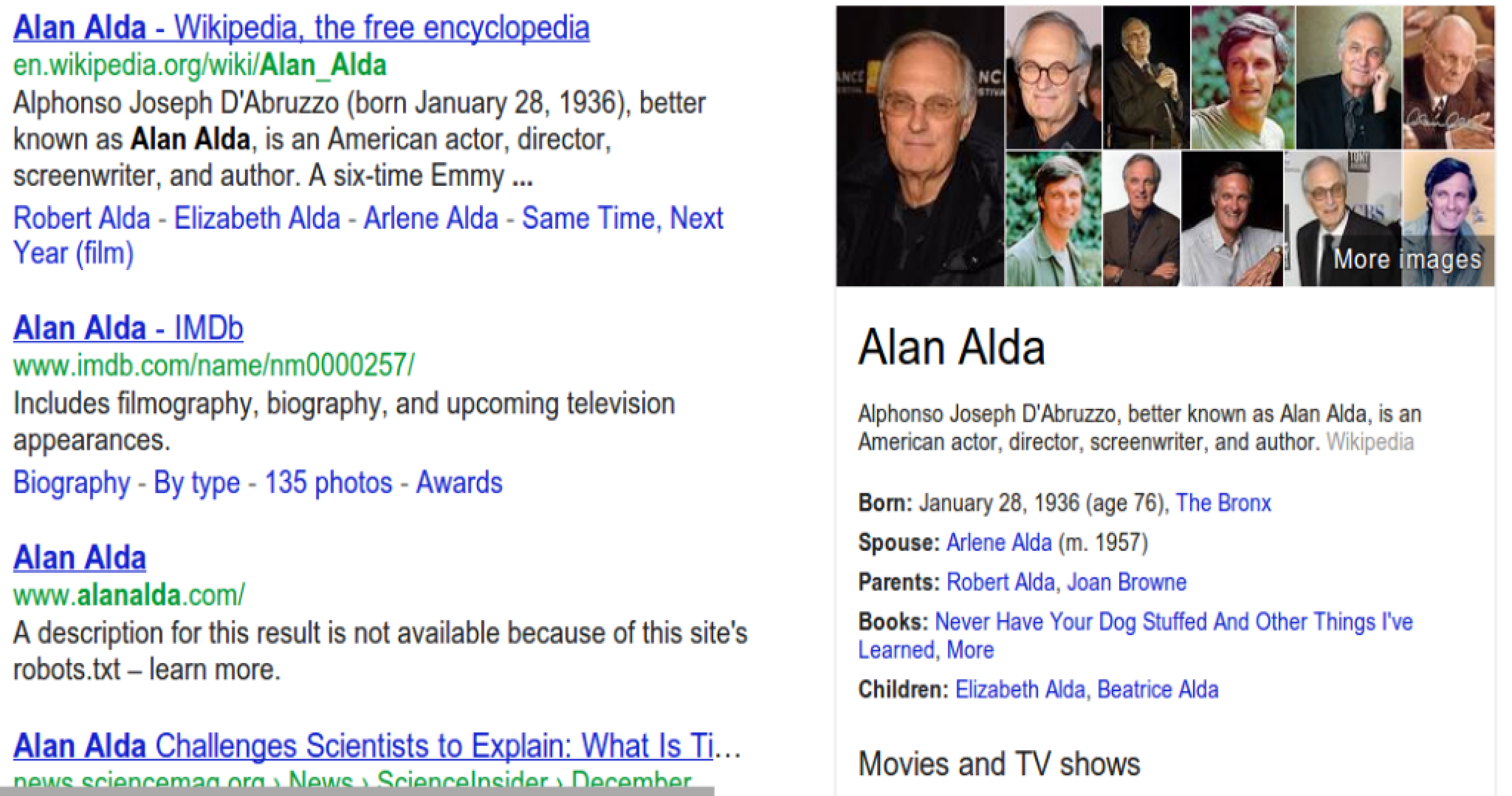
Oct 23, 2012

Governor **James** Baxter **Hunt**, Jr. describes his experiences in high school as a member of the **Future ...**

Second result in Google for "jim hunt future farmers".

Future Possibilities

Answers instead of Search Results



Alan Alda - Wikipedia, the free encyclopedia
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/
Includes filmography, biography, and upcoming television appearances.
Biography - By type - 135 photos - Awards

Alan Alda
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

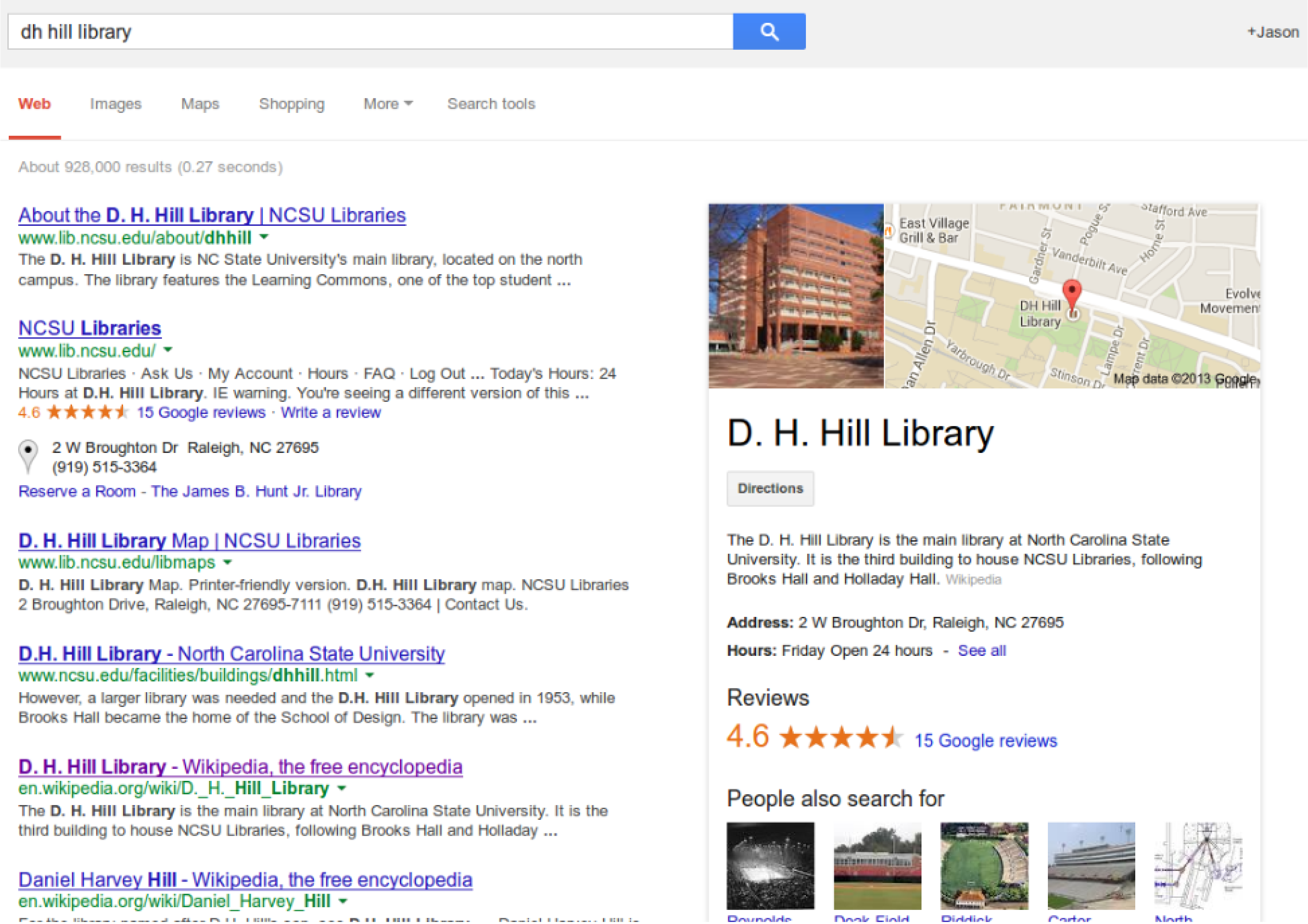
Spouse: Arlene Alda (m. 1957)

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



About 928,000 results (0.27 seconds)

About the **D. H. Hill Library** | NCSU Libraries
www.lib.ncsu.edu/about/**dhhill** ▾
The **D. H. Hill Library** is NC State University's main library, located on the north campus. The library features the Learning Commons, one of the top student ...

NCSU Libraries
www.lib.ncsu.edu/ ▾
NCSU Libraries · Ask Us · My Account · Hours · FAQ · Log Out ... Today's Hours: 24 Hours at **D.H. Hill Library**. IE warning. You're seeing a different version of this ...
4.6 ★★★★★ 15 Google reviews · Write a review

2 W Broughton Dr Raleigh, NC 27695
(919) 515-3364
Reserve a Room - The James B. Hunt Jr. Library

D. H. Hill Library Map | NCSU Libraries
www.lib.ncsu.edu/libmaps ▾
D. H. Hill Library Map. Printer-friendly version. **D.H. Hill Library** map. NCSU Libraries 2 Broughton Drive, Raleigh, NC 27695-7111 (919) 515-3364 | Contact Us.

D.H. Hill Library - North Carolina State University
www.ncsu.edu/facilities/buildings/**dhhill**.html ▾
However, a larger library was needed and the **D.H. Hill Library** opened in 1953, while Brooks Hall became the home of the School of Design. The library was ...

D. H. Hill Library - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/D._H._**Hill_Library** ▾
The **D. H. Hill Library** is the main library at North Carolina State University. It is the third building to house NCSU Libraries, following Brooks Hall and Holladay ...

Daniel Harvey **Hill** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Daniel_Harvey_**Hill** ▾
For the library named after D.H. Hill's son, see **D.H. Hill Library**. ... Daniel Harvey Hill is

D. H. Hill Library

Directions

The D. H. Hill Library is the main library at North Carolina State University. It is the third building to house NCSU Libraries, following Brooks Hall and Holladay Hall. Wikipedia

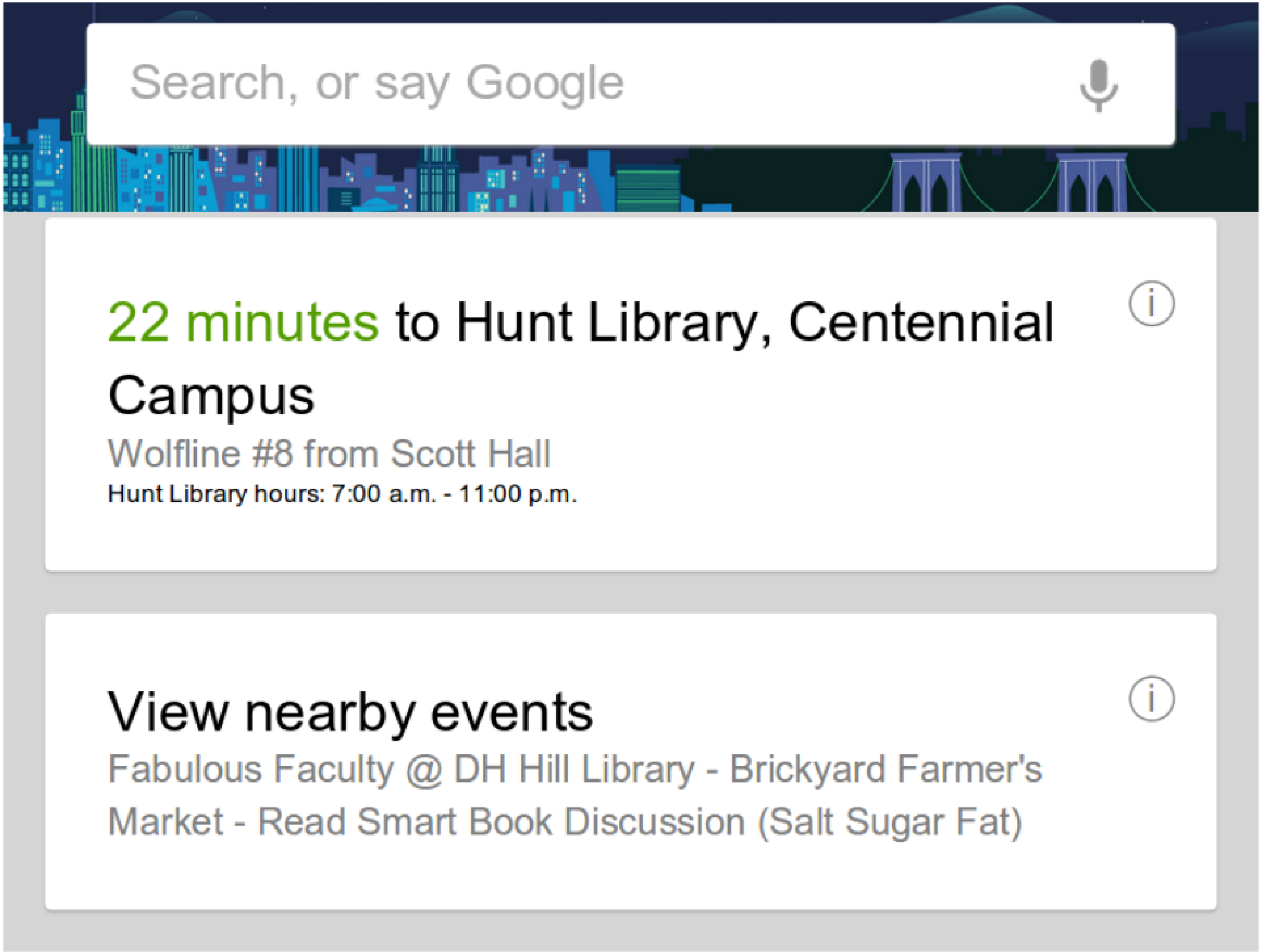
Address: 2 W Broughton Dr, Raleigh, NC 27695

Hours: Friday Open 24 hours - See all

Reviews

4.6 ★★★★★ 15 Google reviews

People also search for



Search, or say Google

22 minutes to Hunt Library, Centennial Campus

Wolfline #8 from Scott Hall

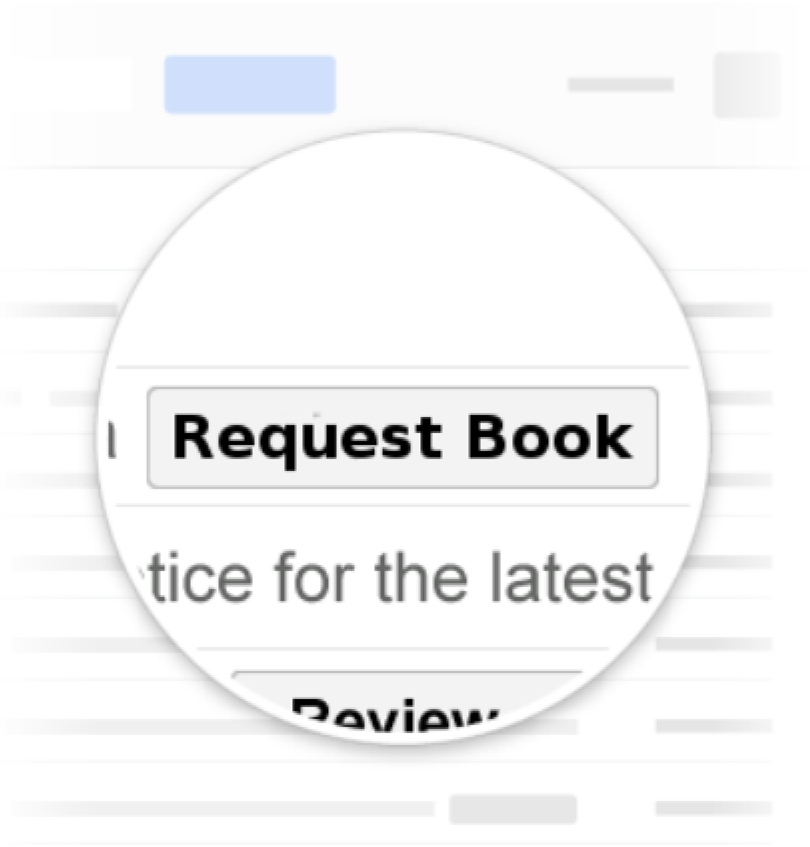
Hunt Library hours: 7:00 a.m. - 11:00 p.m.

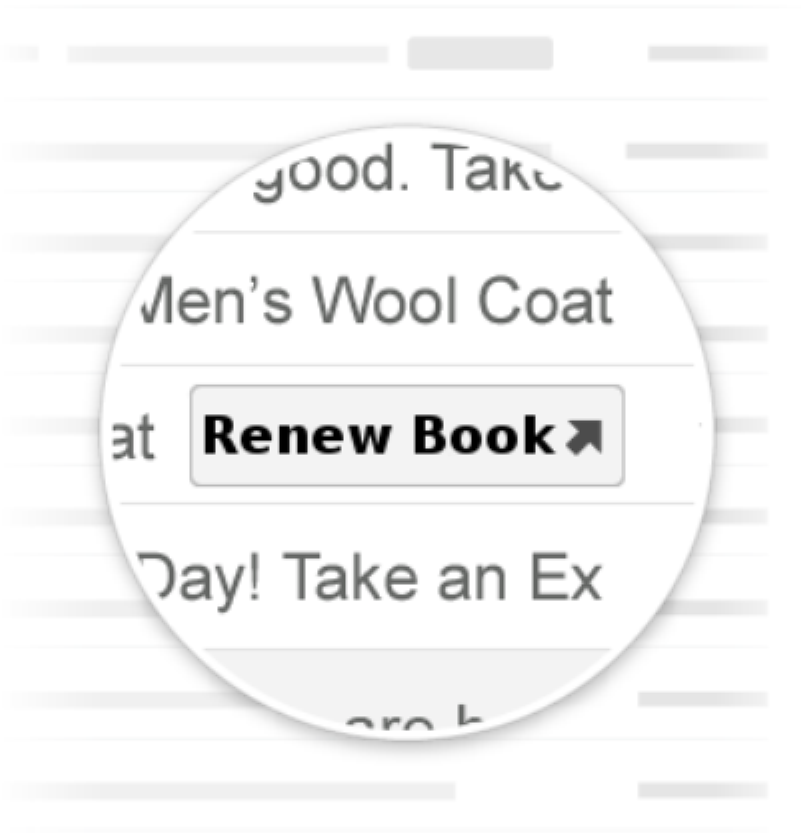
View nearby events

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

Save the Time of the Reader

Schema.org Activities in Gmail





Embedded Semantic Markup + a Web-scale Vocabulary = The Semantic Web?

Yes

Research

Research Questions

Are academic institutions publishing embedded semantic markup?

Are academic libraries?

What kind of data are they publishing?

What syntaxes are they using?

What vocabularies?

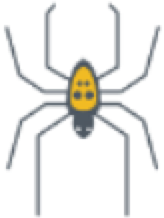
How can you even answer these questions then?

- Ask them
- Crawl the Web
- Beg a search engine
- Give up

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.

- Uses PageRank so is a snapshot of the current most popular part of the Web
- Over 5 billion Web pages (3,005,629,093 for the most recent (2012) set)
- 40,600,000 domains
- ~81TB total
- Big Web Data



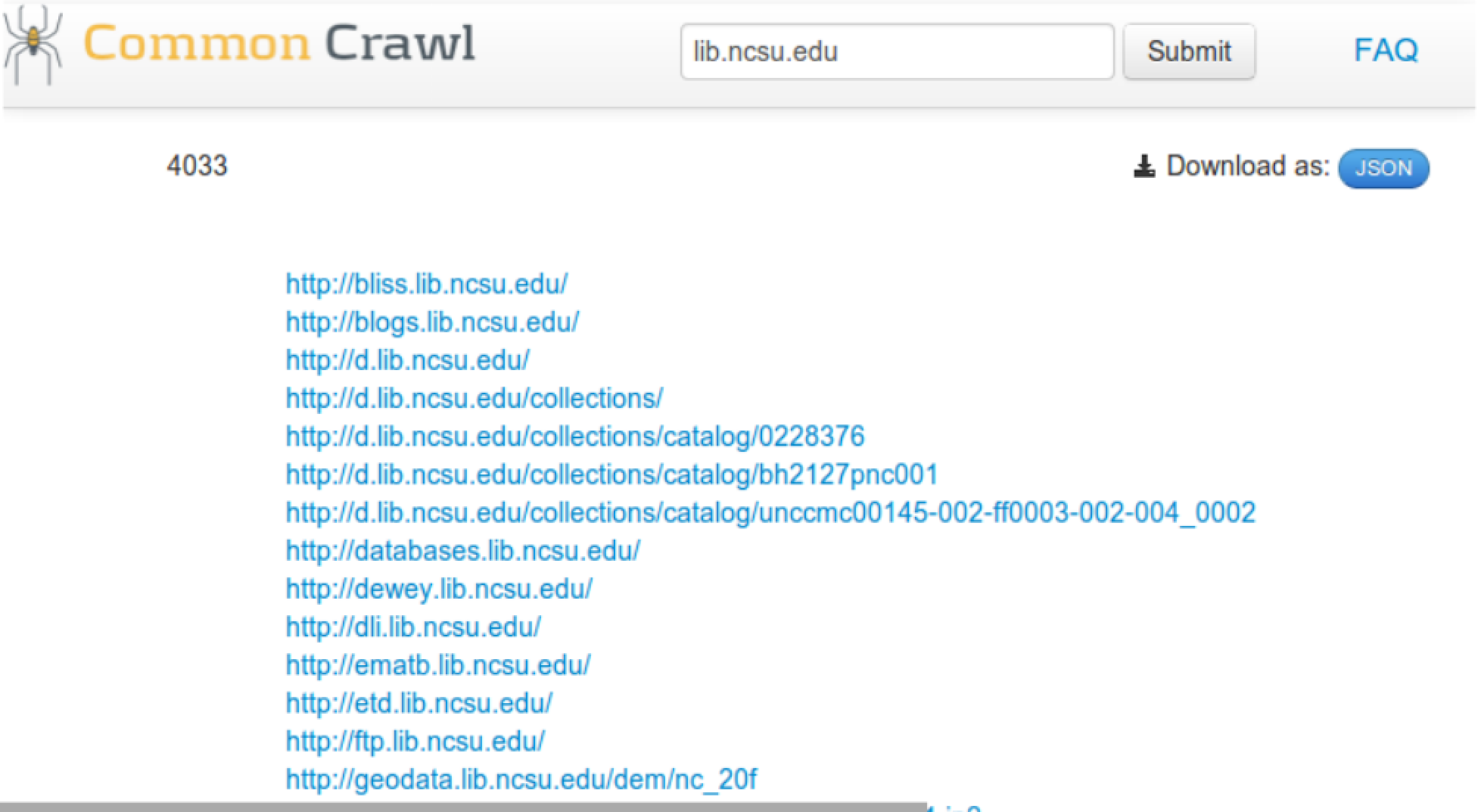
http://commoncrawl.org/

Uses of the Common Crawl

- Free to access, cheap to use (AWS)
- Startups trying out business ideas (Swiftkey)
- Norvig Web Data Science Award
    - Associating concepts on the Web
    - Reading-level analysis and search engine
- Teaching Big Data skills in colleges

Common Crawl URL Search

http://urlsearch.commoncrawl.org/?q=lib.ncsu.edu



4033

Download as: JSON

http://bliss.lib.ncsu.edu/
http://blogs.lib.ncsu.edu/
http://d.lib.ncsu.edu/
http://d.lib.ncsu.edu/collections/
http://d.lib.ncsu.edu/collections/catalog/0228376
http://d.lib.ncsu.edu/collections/catalog/bh2127pnc001
http://d.lib.ncsu.edu/collections/catalog/unccmc00145-002-ff0003-002-004_0002
http://databases.lib.ncsu.edu/
http://dewey.lib.ncsu.edu/
http://dli.lib.ncsu.edu/
http://ematb.lib.ncsu.edu/
http://etd.lib.ncsu.edu/
http://ftp.lib.ncsu.edu/
http://geodata.lib.ncsu.edu/dem/nc_20f

How to get to the Embedded Semantic Markup?

Web Data Commons

Extracting Structured Data from the Common Crawl

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

Percentage of the Common Crawl corpus with embedded structured data? 12.3%

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

http://webdatacommons.org/

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.

Summary 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. Analyze all of these RDF statements and output CSV.
4. Index in Solr and view with Blacklight.

Extraction code and documentation links:
https://ronallo.com/presentations/2013-dlf

Caveats

I have not done any calculations to determine what proportion of each site was crawled. So the comparisons here are just raw numbers. A very large university site might have lots of crawled pages, but it could be a smaller percentage of their whole website than a smaller number here from a university with a smaller website.

For the libraries they may have lots of materials that are not under the library subdomain which would skew those numbers significantly. I was also only looking for subdomains that included "lib.", "library.", and "libraries.". It could be that a lot of academic libraries have their site under a subdirectory or some other subdomain.

Total Resulting Statements (N-Quads)

All statements	7,350,953,995
All .edu	12,182,975
.edu context	8,178,985

Research Questions

Are academic institutions publishing embedded semantic markup?

Are academic libraries?

What kind of data are they publishing?

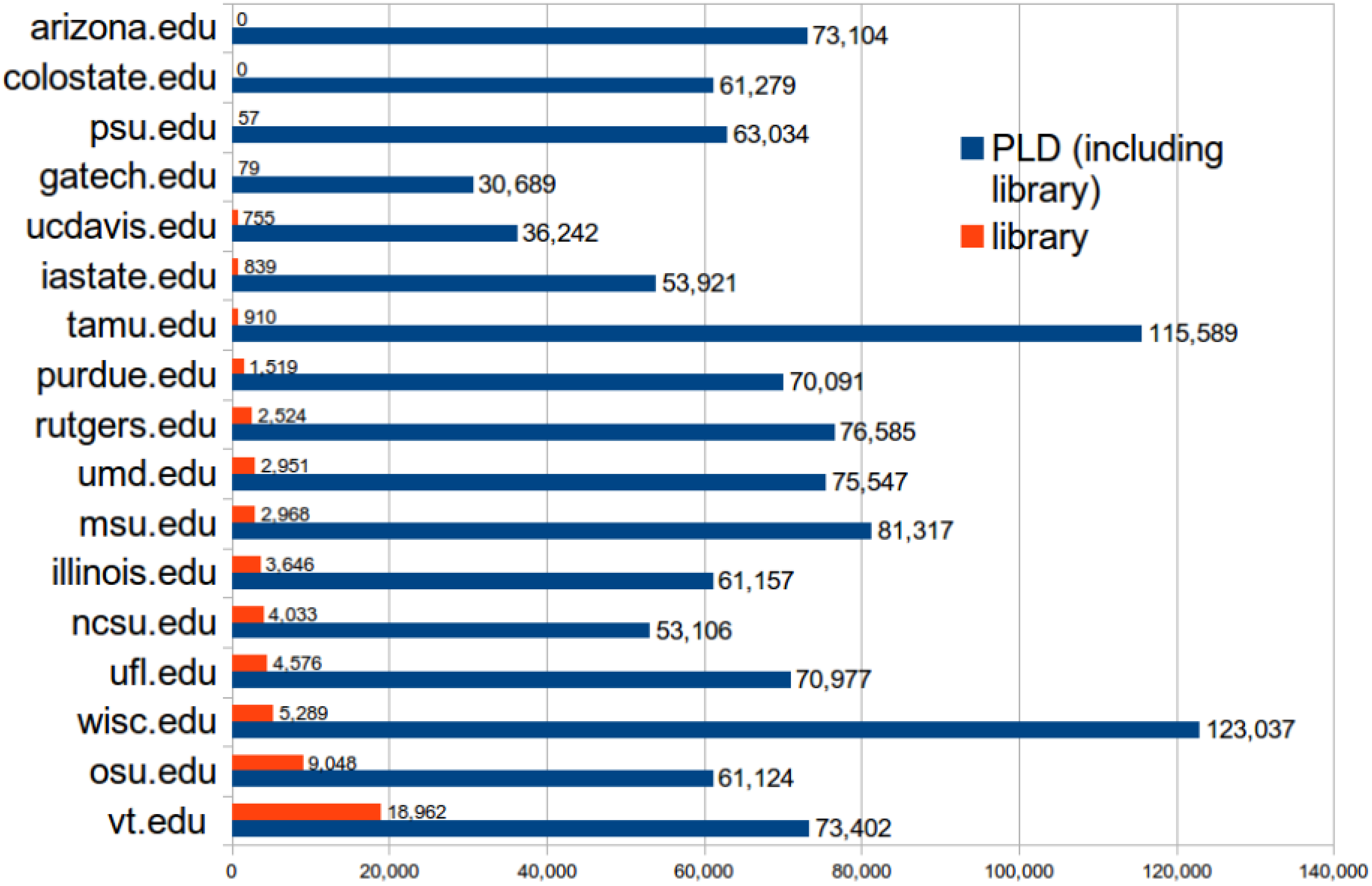
What syntaxes are they using?

What vocabularies?

NC State Univ. Peer Institutions

- Colorado State Univ.
- Georgia Institute of Technology
- Iowa State Univ.
- Michigan State Univ.
- Ohio State Univ.
- Pennsylvania State Univ.
- Purdue Univ.
- Rutgers Univ.-New Brunswick
- Texas A & M Univ.
- Univ. of Arizona
- Univ. of California-Davis
- Univ. of Florida
- Univ. of Illinois at Urbana-Champaign
- Univ. of Maryland-College Park
- Univ. of Wisconsin-Madison
- Virginia Polytechnic and State Univ.

Common Crawl URLs



% of Common Crawl URLs w/ Embedded Semantic Markup

psu.edu	18.35%
illinois.edu	8.87%
ucdavis.edu	5.58%
osu.edu	3.87%
rutgers.edu	3.45%
arizona.edu	1.97%
msu.edu	1.67%
tamu.edu	1.58%
ufl.edu	1.51%
gatech.edu	1.20%
ncsu.edu	1.19%
iastate.edu	0.71%
wisc.edu	0.69%
umd.edu	0.35%
colostate.edu	0.26%
purdue.edu	0.20%
vt.edu	0.16%

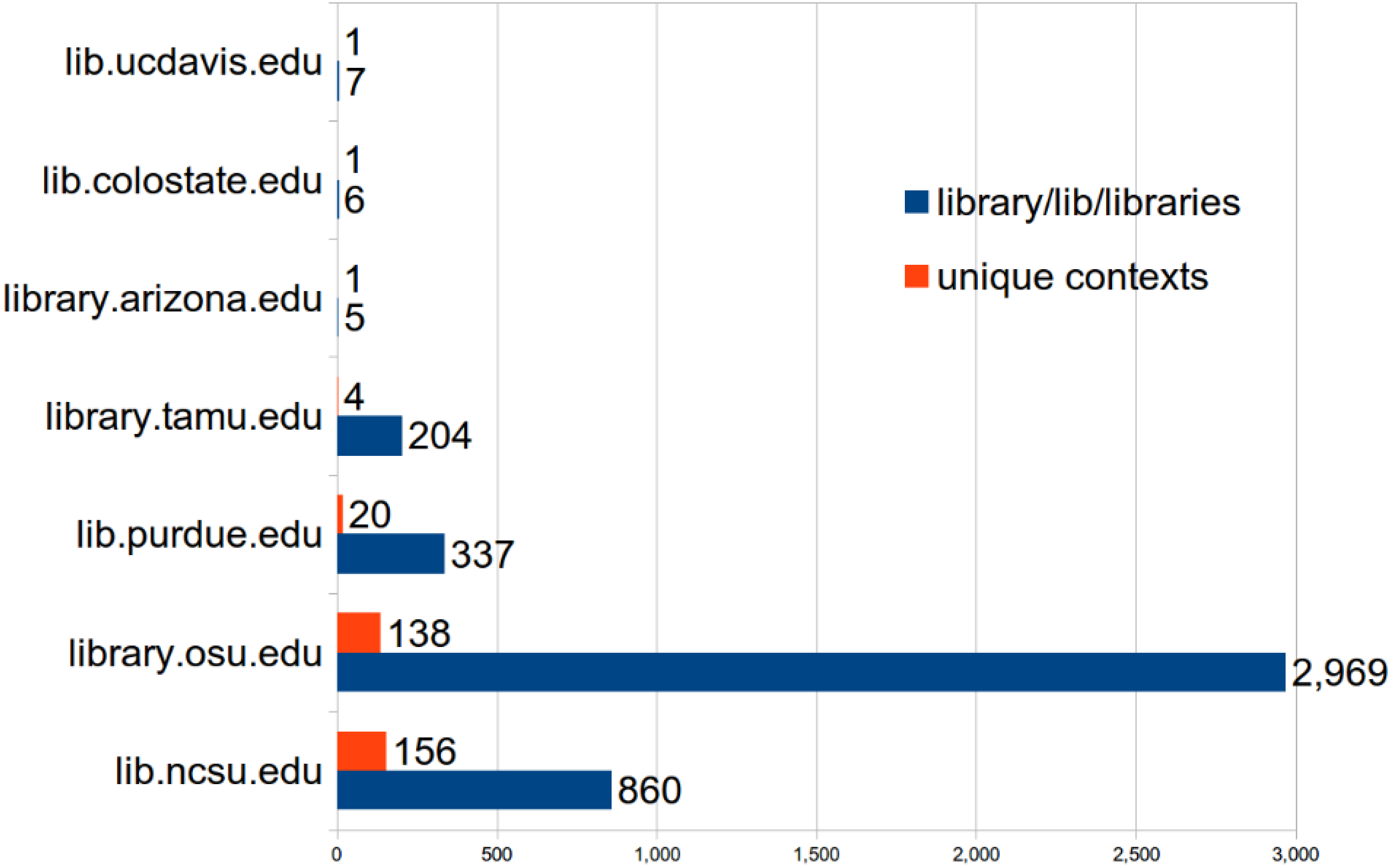
What kind of content is psu.edu describing?

- hcard (Microformat)
- hcalendar (Microformat)
- open graph protocol (RDFa)

Web Data Commons Library Contexts

- lib.*.edu
- library.*.edu
- libraries.*.edu

Web Data Commons Library Triples and Unique Contexts



General Academic Institution Stats

Syntaxes Used by Academic Institutions

mf-hcard	5,854,493
rdfa	1,337,528
mf-hcalendar	770,228
mf-xfn	456,184
microdata	285,296
mf-geo	51,565
mf-hresume	5,363
mf-hreview	2,908
mf-hlisting	48

Schema.org Types Used by Academic Institutions

LocalBusiness	20,565	**CollectionPage**	1,275
PostalAddress	17,267	Blog	991
CollegeOrUniversity	11,554	University	550
Organization	10,172	CollegeorUniversity	420
WebPage	8,846	Review	372
Article	8,351	NewsArticle	316
BlogPosting	3,511	Place	313
Person	2,071	**ScholarlyArticle**	298
Event	1,539	SportsEvent	289
EducationalOrganization	1,508	Thing	168

Some library and archives-related schema.org types

CollectionPage	1275	duke.edu ncsu.edu
ScholarlyArticle	298	santafe.edu
ContactPoint	139	lynn.edu duke.edu
Photograph	20	ncsu.edu
Library	18	berkeley.edu
CreativeWork	12	ncsu.edu
LandmarksOrHistoricalBuildings	11	ncsu.edu
Book	6	ohiolink.edu
GeoCoordinates	4	marshall.edu

What about digital collections?

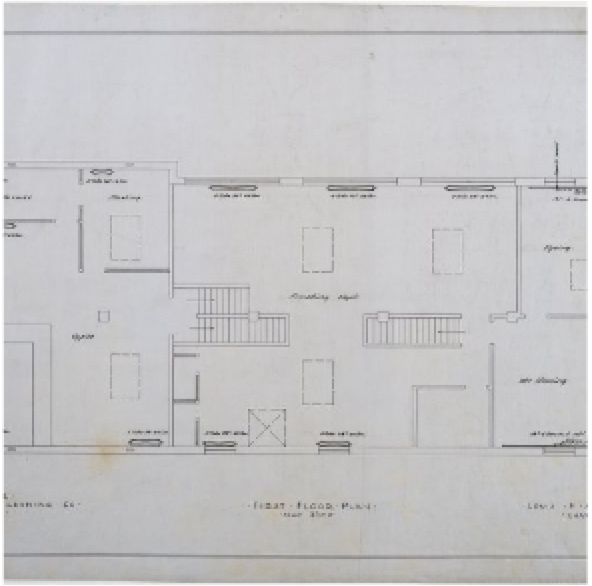
- Are digital collections represented in the Common Crawl?
- Are they publishing embedded semantic markup?

Have a digital collection with a sitemap? Please, go to this URL right now:

http://go.ncsu.edu/sitemap

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



2013 Crawl

Libraries as Producers and Consumers of Big Web Data

Libraries as Producers

- Improve discoverability of our services and collections.
- The public interoperability API to our data (replace library-specific APIs).
- Enable new services we haven't thought of.

Libraries as Consumers

- Learn from what others are successfully doing.
- Help develop best practices for publishing data in HTML.
- Create (or seed) domain-specific vertical search engines.
- Web preservation.
- Become more familiar with handling big data.

Open Research

Code, documenation, data sets, and slides with speaker notes: http://ronallo.com/presentations/2013-dlf

Please submit your digital collection sitemaps: http://go.ncsu.edu/sitemap

Links

- 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/
- Norvig Web Data Science Award http://norvigaward.github.io/

NCSU Links

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

- Student Leadership Initiative http://d.lib.ncsu.edu/student-leaders
- NCSU Libraries Rare & Unique Materials http://d.lib.ncsu.edu/collections

Credits

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

Jason Ronallo

@ronallo

jronallo@gmail.com

http://ronallo.com

Please submit your digital collection sitemaps:
http://go.ncsu.edu/sitemap

Code, documenation, data sets, and slides with speaker notes: http://ronallo.com/presentations/2013-dlf