Minitex

A Publication of the Minitex Digitization, Cataloging & Metadata Education Unit

# DIGITIZATION, CATALOGING & METADATA MAILING

## March

## Schema.org Pilot Project

*Sara Ring, Minitex/DCME*

Have you heard of Schema.org? Perhaps only in reference to OCLC’s summer 2012 announcement that stated all the records in WorldCat.org have been “marked up” with schema.org? The original announcement was printed in the July/August *Mailing*:

www.minitex.umn.edu/Communications/Mailing/2012/07JulyAugust.pdf

Since then, I have been thinking about the significance of this announcement and often would wonder what I could do to learn more about schema.org. What value does it hold for libraries? For end users? I decided to try a small pilot at Minitex to learn more about schema.org and to find out how feasible it is to “mark up” and expose our library data on the web.

### What Is Schema.org?

The Schema.org initiative—launched in 2011 by Google, Bing and Yahoo! and later joined by Yandex (Russian Search Engine) provides a core vocabulary for markup that helps search engines and other web crawlers better understand the data on your website. For example, you might have a library web page with a list of new dvds you have added to your collection. HTML tags only tell browsers how to display/format the content on your page. Using a simplified example, you might have the movie title, “Argo,” in a list on your webpage and have it marked up like so:

```
<ul>
<li>Argo</li>
</ul>
```

However, the HTML doesn’t provide information about what is meant by “Argo.” Is it a place, a movie, or a name? It doesn’t tell a search engine anything other than “this is a string of text (Argo) that appears in a list. By using schema.org vocabulary along with the microdata format (tags introduced with HTML 5), you can add meaning to your HTML content that is more understandable to search engines. This concept of structured data isn’t new to library staff who work with MARC records. We code the descriptive text found in our bibliographic databases in the MARC format so our library systems can make sense of the data and display it to end users in various ways.

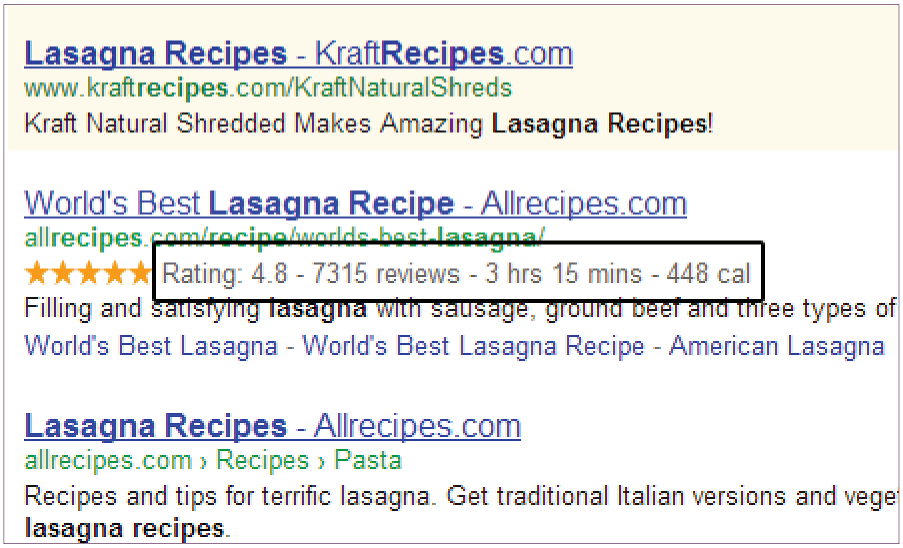
Search engines are using the schema.org marked up content on the web in very interesting ways! Let’s take a look at an example. If you look at Figure 1 on the the next page, you will see a list of recipes resulting from a search I performed in Google using the keywords: *lasagna recipe*.

Notice that though it is not the top result, the recipe appearing second in the list, “World’s Best Lasagna Recipe,” gives me more information up front to help me decide whether to click through to the website or not. I know it has a 4.8 star rating from 7,315 reviews, it will take approximately 3 hours and 15 min. to make, and has 448 calories per serving.

What you are seeing in Figure 1 is Google's Rich Snippet information, and it is only possible because the website, allrecipes.com, marked up their content using schema.org and microdata. Tip: An easy way to tell if a site is using structured data is to copy the website url into Google's structured data testing tool:

www.google.com/webmasters/tools/richsnippets

When I copy the allrecipes.com lasagna recipe url (allrecipes.com/recipe/worlds-best-lasagna) into this tool, I'm presented with a labeled view of the schema.org/microdata. It shows the schema.org "type" used and "property" information. The schema.org type "recipe" was used on the lasagna recipe webpage. The schema.org properties used are listed underneath the type information. Here are just some of the properties used for our lasagna recipe webpage:



**Lasagna Recipes** - Kraft**Recipes**.com
www.kraft**recipes**.com/KraftNaturalShreds
Kraft Natural Shredded Makes Amazing **Lasagna Recipes**!

World's Best **Lasagna Recipe** - Allrecipes.com
allrecipes.com/**recipe**/worlds-best-**lasagna**/
★★★★★ Rating: 4.8 - 7315 reviews - 3 hrs 15 mins - 448 cal
Filling and satisfying **lasagna** with sausage, ground beef and three types of
World's Best Lasagna - World's Best Lasagna Recipe - American Lasagna

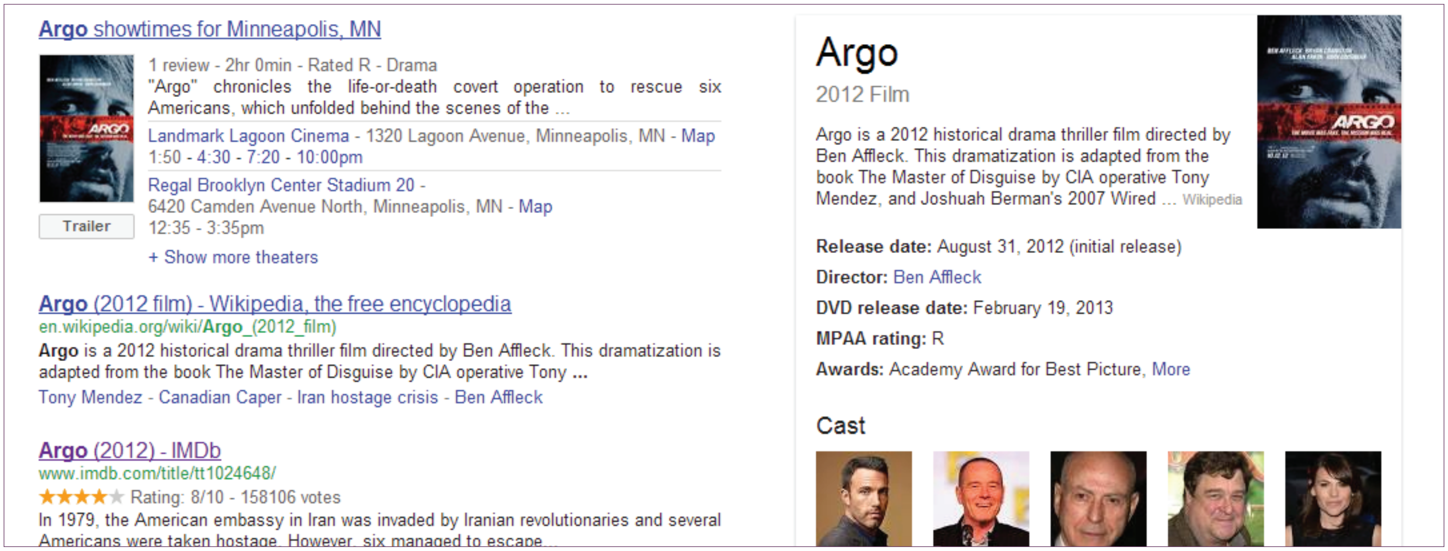
**Lasagna Recipes** - Allrecipes.com
allrecipes.com › Recipes › Pasta
Recipes and tips for terrific lasagna. Get traditional Italian versions and vege
**lasagna recipes**.

*Figure 1: Recipe Search Results in Google*

- image
- name
- aggregaterating
- description
- author
- recipeyield
- ingredients
- preptime
- cooktime
- totaltime
- nutrition

Google is able to pull out some of this information in their Rich Snippet view search results (see Figure 1 above), in particular, *aggregaterating*, *totaltime*, and *nutrition*.

Let's look at one other way that Google is making use of structured data. Here I've searched Google for the movie "Argo." The information that appears on the right side of the page is being pulled from the Internet Movie Database, imdb.com.



**Argo** showtimes for Minneapolis, MN
1 review - 2hr 0min - Rated R - Drama
"Argo" chronicles the life-or-death covert operation to rescue six Americans, which unfolded behind the scenes of the ...
Landmark Lagoon Cinema - 1320 Lagoon Avenue, Minneapolis, MN - Map
1:50 - 4:30 - 7:20 - 10:00pm
Regal Brooklyn Center Stadium 20 -
6420 Camden Avenue North, Minneapolis, MN - Map
12:35 - 3:35pm
Trailer
+ Show more theaters

**Argo** (2012 film) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Argo**_(2012_film)
**Argo** is a 2012 historical drama thriller film directed by Ben Affleck. This dramatization is adapted from the book The Master of Disguise by CIA operative Tony **...**
Tony Mendez - Canadian Caper - Iran hostage crisis - Ben Affleck

**Argo** (2012) - IMDb
www.imdb.com/title/tt1024648/
★★★★★ Rating: 8/10 - 158106 votes
In 1979, the American embassy in Iran was invaded by Iranian revolutionaries and several Americans were taken hostage. However, six managed to escape ...

Argo
2012 Film
Argo is a 2012 historical drama thriller film directed by Ben Affleck. This dramatization is adapted from the book The Master of Disguise by CIA operative Tony Mendez, and Joshuah Berman's 2007 Wired ... Wikipedia

**Release date:** August 31, 2012 (initial release)
**Director:** Ben Affleck
**DVD release date:** February 19, 2013
**MPAA rating:** R
**Awards:** Academy Award for Best Picture, More

Cast

*Figure 2: Google Knowledge Graph*

And, guess what? If you go to the imdb.com page for Argo and look at their HTML, they are also using schema.org and microdata to mark up their content. The Argo movie information that you are seeing in Figure 2 is part of Google's Knowledge Graph project:
www.google.com/insidesearch/features/search/knowledge.html

Marking up your webpages using schema.org/microdata will not necessarily improve your search engine rankings. What it will do is provide more information up front to end users and probably increase click-throughs to your website from the major search engines. And, it will allow for search engines to make sense of your information/data in new and interesting ways, like the Rich Snippet and Google Knowledge Graph projects mentioned earlier in this article.

## How It Works

As I mentioned earlier, I wanted to experiment with the schema.org vocabulary to find out how steep the learning curve was and, hopefully to glean some of the benefits for using it. I have noticed that there hasn't been a lot in the literature on using and implementing schema.org/microdata within a library context.

I first met with our IT staff at Minitex, and they were very interested in trying out the process. We have a collection of about 17 oral history interview recordings with library staff from our region on our website, and I thought this type of content might be ideal to start with. I worked with our web developer to mark up just one Minitex oral history webpage:
www.minitex.umn.edu/40th/Baldwin.aspx

We first took a look at the commonly used Schema.org types. They are:

- Creative works (CreativeWork, Book, Movie, MusicRecording, Recipe)
- Embedded non-text objects: AudioObject, ImageObject, VideoObject
- Event
- Organization
- Person
- Place
- Product
- Review

Next, we looked at the existing content on the Minitex Oral History interview page for Jerry Baldwin. We decided that the most logical Schema.org type to use for the Jerry Baldwin Interview page was *CreativeWork*.

We were then able to look at all the allowable schema.org properties that could be used under the *CreativeWork* type. This included properties like:

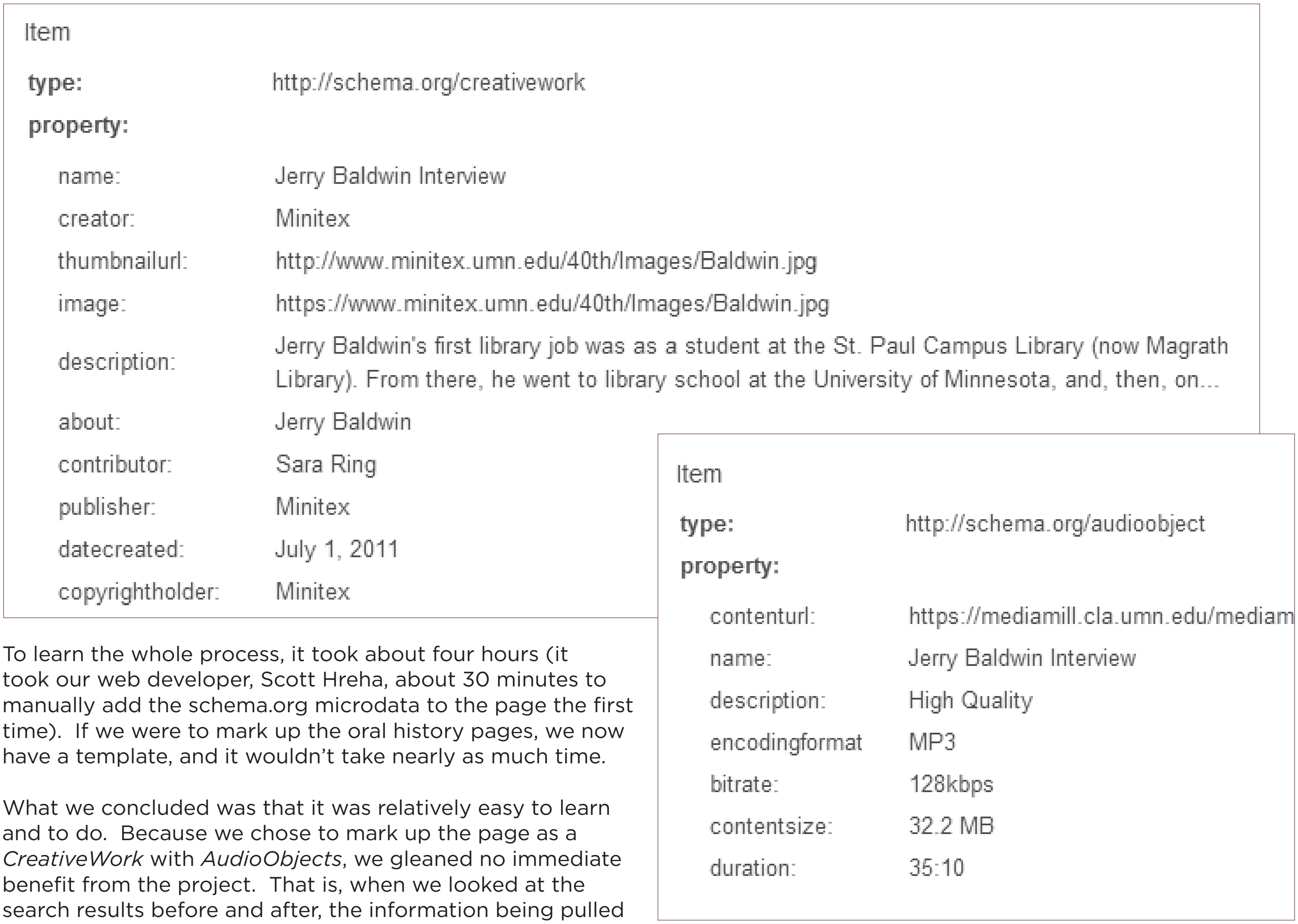
- description (description of the item)
- about (subject matter of content)
- image (url of an image of the item)
- dateCreated (the date on which the *CreativeWork* was created)

We also had audio files linked to and described on this page, so we were able to use properties like this for the .mp3 and streaming files:

- bitrate (bitrate of media object)
- contentSize (file size in mega/kilo bytes)
- duration (duration of audio recording)

After deciding on what schema.org type and properties to use for our webpage content, our web developer went ahead and edited the HTML to embed the microdata and schema.org terms. If you wish to view the mark up, look at the source view of this page:
www.minitex.umn.edu/40th/Baldwin.aspx

Using the Google Structured Data Testing Tool, here's a quick snapshot that will give you a sense of the properties we used for the page:



Item

| | |
|---|---|
| **type:** | http://schema.org/creativework |
| **property:** | |
| name: | Jerry Baldwin Interview |
| creator: | Minitex |
| thumbnailurl: | http://www.minitex.umn.edu/40th/Images/Baldwin.jpg |
| image: | https://www.minitex.umn.edu/40th/Images/Baldwin.jpg |
| description: | Jerry Baldwin's first library job was as a student at the St. Paul Campus Library (now Magrath Library). From there, he went to library school at the University of Minnesota, and, then, on... |
| about: | Jerry Baldwin |
| contributor: | Sara Ring |
| publisher: | Minitex |
| datecreated: | July 1, 2011 |
| copyrightholder: | Minitex |

Item

| | |
|---|---|
| **type:** | http://schema.org/audioobject |
| **property:** | |
| contenturl: | https://mediamill.cla.umn.edu/mediam |
| name: | Jerry Baldwin Interview |
| description: | High Quality |
| encodingformat | MP3 |
| bitrate: | 128kbps |
| contentsize: | 32.2 MB |
| duration: | 35:10 |

To learn the whole process, it took about four hours (it took our web developer, Scott Hreha, about 30 minutes to manually add the schema.org microdata to the page the first time). If we were to mark up the oral history pages, we now have a template, and it wouldn't take nearly as much time.

What we concluded was that it was relatively easy to learn and to do. Because we chose to mark up the page as a *CreativeWork* with *AudioObjects*, we gleaned no immediate benefit from the project. That is, when we looked at the search results before and after, the information being pulled out and displayed within the search results was the same.
It is my impression that the schema.org item types that are currently being used by Google in Rich Snippets and the Google Knowledge Graph tend to be video (movies), recipes, and events. That is not to say that this information will not be used by the major search engines in the future, and this shouldn't be the only deciding factor when considering using schema.org/microdata.

## OCLC And Schema.org

As mentioned at the start of this article, OCLC has marked up the records in WorldCat.org using schema.org terms. This means that the bibliographic data in WorldCat.org is now available for use by web crawlers that can make use of the metadata in their search indexes and other applications (wouldn't it be great to see WorldCat bibliographic data incorporated into the Google Knowledge Graph?!). Our library catalogs and other bibliographic resources are mostly hidden to search engines as they are now. The point of marking up your library resources in this way is to expose your data beyond the library community and on the web.

Schema.org does have a *Book* type and properties associated with it. But, there are some properties that are not included in the schema.org vocabulary that are needed for the WorldCat.org data, such as holdings count and carrier type. OCLC is working with the Schema.org community to develop and add a set of vocabulary extensions to WorldCat data. Schema.org is working with a number of other industries to provide similar sets of extensions for other specific use cases. The most recently announced community was Good Relations (an e-commerce schema):

blog.schema.org/2012/11/good-relations-and-schemaorg.html

Richard Wallis, technology evangelist at OCLC, states on his blog (http://dataliberate.com/2012/11/the-correct-end-of-your-telescope-viewing-schema-org-adoption):

> *My personal hope being that the resulting proposals, if and when adopted by Schema.org, will enable libraries, publishers, interest groups, universities, retailers, OCLC, and others to not only publish data about their resources in a way that the search engines can understand, but also have a light weight way to interconnect them to each other and authoritative identifiers for place, name, subject, etc., that will help us begin to form a distributed web of bibliographic data*

I would like to see more libaries marking up their content for the web using schema.org/microdata. It is another way to put our information in the path where the users already are—on the web. In addition to marking up library catalog records, I could see schema.org markup being used on library webpages that post events, book reviews, new materials added to the collection, and subject lists. Schema.org isn't a large enough vocabulary to fully describe our rich library resources, but it is what is supported by all the major search engines, and that says a lot!

I'll also note that there are a few applications and tools starting to emerge that make marking up your content with schema.org terms even easier. I've listed a few below.

### Drupal module

drupal.org/project/schemaorg

### Wordpress plugins

wordpress.org/extend/plugins/tags/schemaorg

### Microdata generator

www.microdatagenerator.com ■

Cataloging & Metadata

## RDA Notes

*Mark K. Ehlert, Minitex/DCME*

### RDA Implementation

Before digging into the details, I want to point to a message MARCIVE's Mary Mastraccio posted to the AUTOCAT electronic listserv in response to a question on RDA preparations. Her comments later appeared on the RDA Toolkit Blog. I recommend all catalogers give it a read:

www.rdatoolkit.org/blog/511

### RDA Implementation: LC

By the time you receive this issue of the *Mailing*, the Library of Congress RDA implementation date of March 31 will be just around the corner. What does this mean for catalogers?

- LC's implementation date applies only to the Library of Congress. Other libraries can set their own start-up date.
- LC's implementation refers to their original cataloging— i.e., all new cataloging, whether for recent materials or old materials,—will be done using RDA. Other libraries' definitions of *RDA implementation* appear to follow this same lead.
- When LC copy catalogers encounter AACR2 records, they, more often than not, will check it over and move the record into their catalog. Other institutions will mirror this workflow, though the occasional flipping of an older record to RDA form is certainly an option.

When LC makes the move, the number of RDA bibliographic records with the "DLC" moniker in the 040 field will certainly increase. However, there are plenty of AACR2 records from LC still in the pipeline, so the change will not be sharply apparent on Monday morning, April 1. Bear in mind too that LC's output concentrates on print format materials; those who handle, say, audiovisual works may or may not see an increase in RDA records for those materials, as it's unknown how many other libraries have opted to match LC's implementation date.

Nor does LC or any other library that I'm aware of have plans to institute large-scale flipping projects to convert older bibliographic records already in the catalog to fully-compliant RDA forms. Just as many catalogs still have AACR1 records or those built following the even older ALA cataloging rules, LC's catalog like so many others will continue to incorporate AACR2 and RDA records into the mix.

### RDA Implementation: OCLC

Speaking of flipping, OCLC members should be aware of a new RDA cataloging policy that is set to take effect on March 31. It lays out the expectations for both original and copy cataloging under RDA in the WorldCat environment:

www.oclc.org/en-US/rda/new-policy.html

The opening paragraph makes plain that OCLC is not forcing its membership to shift over to RDA exclusively—new AACR2 records will continue to be a part of the WorldCat database along with their RDA