EMORY
LIBRARIES &
INFORMATION
TECHNOLOGY

Metadata: Key Concepts for Digital Collections

Summer Institute for Digital Scholarship: 6/16/14

Emily Porter, Metadata Analyst
Woodruff Library, Emory University

Session Outline

Overview of important concepts for metadata in digital collections & systems

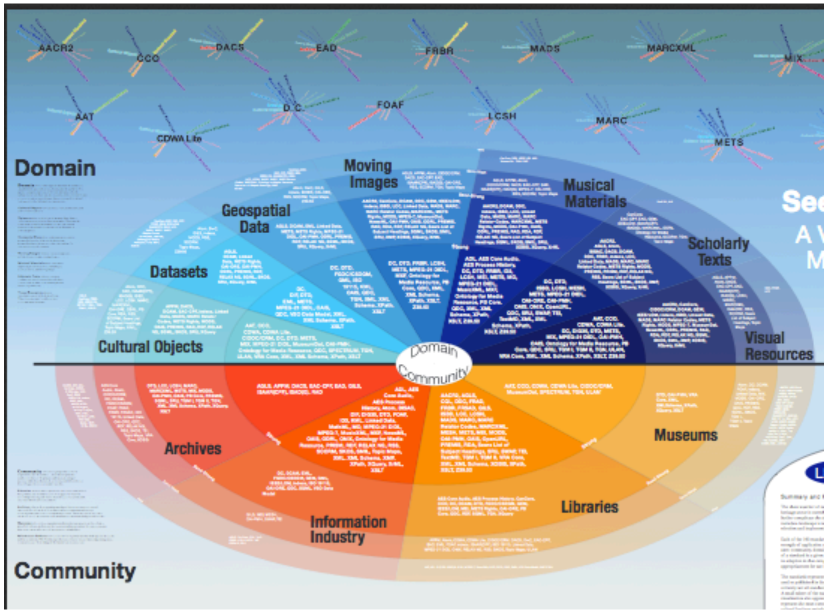
- Metadata Types, Communities, Contexts
- Schemas: Concepts and Terms
- Implementing: Platforms and Systems
- Dublin Core
- Web Findability/Search Engine Optimization
- Trends and Changing Practices
- References and Resources

METADATA

Types, Communities, Contexts

Communities and Domains

- Specific schemas and standards across domains
  - **Cultural Heritage**:
    - Galleries/Libraries/Archives/ Museums (GLAM)
  - Scientific
  - Geospatial
  - Media/Broadcasting/ Audiovisual
  - Educational/Learning Management
  - Web/Social Media



Seeing Standards: A Visualization of the Metadata Universe (Jenn Riley, Devin Becker)

Digital Collections vs. MARC Metadata

Cataloging/MARC

- "Library cataloging, as a form of metadata, has traditionally had well-defined goals…" (Karen Coyle)
  - → The catalog
- Benefits from mature encoding standards, well-defined content standards and system contexts
- Bibliographic, published content
- Established infrastructure for large-scale sharing of records

Digital Objects' Metadata

- Parallel practices emerged in non-Library domains
- Highly contextual and often highly customized:
  - Specific collection/content focused
  - End user/interface needs
  - Wide variety of destination systems/environments
- Granularity/hierarchy of objects in a collection
  - Group photo on a page within a scrapbook
  - Unpublished content / provenance

Types of Metadata: Cultural Heritage

Type	Notes	Schemas/Standards
Descriptive	Primary focus for today’s overview: search and discovery metadata	Dublin Core; MODS; VRA, PBCore; EAD; MARC/RDA
Administrative/Technical/Rights (Repository/digitization)	Information about the digital object in a system How it was generated How it can be used/accessed/modified	Administrative: METS; Repository-generated Technical: MIX; PBCore Rights: PREMIS; METS
Structural (Repository/digitization)	How to reconstruct a complex digital object	METS/ALTO; Repository-generated
Preservation (Repository/digitization)	Tracks changes to a digital object over time	PREMIS; Repository-generated (audit trail)

Considerations: Planning for a Project

- **Which schema should I use?**
  - Depends on system constraints, local practices, content type needs
- **Who will create the metadata?**
  - Does metadata already exist? Can it be re-used?
  - Data entry resources; training and support; usability of interface
  - Consistency and standards
  - Uniqueness of resource/preservation needs – is this the only record?
- **What do the users need?**
  - Display labels (vs. schema element names)
  - Search result screens
  - Search, browse, filter, sort
  - Special navigation/presentation needs
  - Knowledge domain/vocabulary needs
  - Basic web discovery, or library catalog integration?
  - Individual item descriptions, or project/collection level?
- **Documentation**

METADATA

Schemas:
Concepts and Terms

What’s a “Schema”?

- **Element**
  - Single, distinct metadata “field” or property in a resource description
  - Generally implemented as name/value pairs (sometimes with labels)
    - Creator: Doe, John
- **Element Set**
  - Set of metadata elements which together compose a larger content standard or schema
- **Schema**
  - Formally structured, machine-processable Element Set (often XML)
  - Collection of elements with documented usage requirements, data types, and data entry conventions
  - Typically require technical validation against XML, DTD/XSD
- **Application Profile**
  - Organizational or consortial interpretation of a schema
  - Local practices and business rules for how to utilize a schema
    - Dublin Core Library Application Profile
- **Data Dictionary**
  - Technical documentation for a schema or element set

Metadata Encoding

- **Element Data Types**
  - Specify what type of information can be stored in a field
    - String (text – but can be alphanumeric)
    - Date/Time (conforms to a date encoding scheme)
    - Number (can be manipulated as a number)
    - URI (URL)
- **Element Cardinality**
  - Required, optional
  - Repeatable (and minimum/maximum entries)
- **Data Formats**
  - The storage and encoding format for metadata
    - Database
    - Flat file (tab-delimited, CSV)
    - XML
    - HTML
- **Validation**
  - Technical error-checking of the markup or values
- **Crosswalks/Transformations**
  - Mapping elements from one standard to another (documentation)
  - Technical transformation of data elements from one standard to another (e.g. XSLT)

Populating Elements in a Schema

- **Content standards**
  - Set of guidelines for populating fields
  - Recommended values, element usage (which may be more specific than the schema indicates), punctuation/style guide
  - Often evolve around particular content types
    - Examples: RDA, DACS, CCO, AACR2, CSDGM
- **Value standards** (see DC cheat sheet)
  - Rules/constraints for how the data can be entered
  - Re-usable across schemas/standards
  - May correlate to data types
    - Date formats (e.g. W3CDTF: YYYY-MM-DD)
  - **Controlled vocabularies** and authorities
    - Restrict entries to a controlled list of terms
    - Enhance consistency and data entry entry (e.g. consistent names)
    - Maintained by formal bodies with subject matter expertise
    - Enable batch updates of values over time
    - Examples: Library of Congress Subject Headings, Name Authority File

METADATA

Implementing: Platforms and Systems

Metadata in System Environments

Environment	Product Examples	Metadata Support
ILS/Cataloging	ExLibris Voyager ExLibris Alma	• Mostly MARC focused • Some DC/MODS • Controlled vocabulary integration
Content/Asset Management (CMS/DAMS)	Drupal, Cascade Server, Extensis Portfolio Server, ContentDM, WordPress, Omeka	• HTML metadata • Some Dublin Core • Some custom metadata
Discovery	ExLibris Primo* EBSCO Discovery Service	• OAI ingest • Map incoming data (misc. XML*; Dublin Core*; MARC/MARCXML)
Presentation	LUNA, ArtStor, ContentDM, Omeka	• Dublin Core • Some custom schemas • Some VRA • Some OAI export
Repository	DSPACE, ContentDM, Rosetta, Fedora, ePrints	• Package/create desc. schemas • Custom pres. schemas (PREMIS, METS, other) • OAI ingest/export

ECDS Institute Tools

- **Tour app builder**
  - HTML titles and descriptions: stop pages, homepage, splash page
  - Basic image/video metadata
  - Twitter and Facebook metadata for "stops"
- **WordPress**
  - Limited metadata entries without plugins (.com account)
  - Add Meta Tags Plugin enables qualified DC, Facebook, Twitter, Schema.org
  - Standard WordPress themes are optimized for general SEO best practices
- **Omeka**
  - Support for Dublin Core, other schemas (if self-hosting)
  - Metadata-oriented plugins (.net account):
    - CSV import
    - Google Analytics
    - Library of Congress Suggest
    - COinS
    - OAI harvester
    - Shared Shelf
    - (additional plugins for self-hosting)
  - Tags as categorization/ discovery options

METADATA

Dublin Core

Dublin Core 101

- Conceived in 1995 in as a “core set of semantics for web-based resources” to enhance search and retrieval in response to rapidly growing web
- Benefits:
  - Universal translator: understood by many systems; common mapping point for metadata across schemas
  - Relatively small element set (easy to populate)
  - Flexible and easy to encode as XML, embed in HTML, or store in a spreadsheet
  - Elements utilized by other schema (PB Core, Darwin Core)

Dublin Core: Challenges

- 1:1 principle - what are you describing?
  - Digital copy vs. analog original vs. born digital original
  - Strict meaning: only describe the *digital* instance
    - Digital Library Federation best practice: use 1:1 “when practical”
    - Consider context for end users; local practices; consistency across collections
    - Is this the canonical metadata record?
- Lack of specificity; ambiguity of some elements
- Looseness of requirements (all optional, repeatable)
- Loss of data when mapping from richer schemas

Dublin Core: Simple vs. Qualified

- Simple Dublin Core
  - 15 “core” elements (no attributes or qualifiers)
  - Issues of interpretation: date, relation, coverage
  - No way to specify attributes for controlled vocabulary names/encoding schemes
- Qualified Dublin Core
  - Refinements that add context to core elements
    - date.created
    - relation.isPartOf
    - coverage.spatial
    - format.extent
  - Qualifed Dublin Core is “deprecated”; evolved into DCMI Terms Vocabulary

Dublin Core: Elements (aka “Simple”)

1. Title
2. Description
3. Type
4. Subject
5. Coverage
6. Relation
7. Source
8. Creator
9. Contributor
10. Publisher
11. Rights
12. Date
13. Format
14. Identifier
15. Language

See Dublin Core Cheat Sheet for more detail

Dublin Core: Elements + Qualifiers

- abstract
- accessRights
- accrualMethod
- accrualPeriodicity
- accrualPolicy
- alternative
- audience
- available
- bibliographicCitation
- conformsTo
- *contributor*
- *coverage*
- created
- *creator*
- *date*
- dateAccepted
- dateCopyrighted
- dateSubmitted
- *description*
- educationLevel
- extent
- *format*
- hasFormat
- hasPart
- hasVersion
- *identifier*
- instructionalMethod
- isFormatOf
- isPartOf
- isReferencedBy
- isReplacedBy
- isRequiredBy
- issued
- isVersionOf
- *language*
- license
- mediator
- medium
- modified
- provenance
- *publisher*
- references
- *relation*
- replaces
- requires
- *rights*
- rightsHolder
- *source*
- spatial
- *subject*
- tableOfContents
- temporal
- *title*
- *type*
- valid

Dublin Core: Points of Confusion

- Type vs. Format
  - Type = original content
  - Format = physical or digital file characteristics
  - Digitized page of a letter
    - Type = text
    - Format = image/tiff
- Source vs. Relation
  - Source = original source (generally analog) from which digital copy/excerpt was made
  - Relation = multiple types of relationships
- Creator vs. Contributor
  - Primary vs. secondary role in creation
- Date vs. Coverage
  - Date related to content lifecycle vs. date within content

Dublin Core: Missing Concepts

Elements/concepts that are hard to express in DC, but available in other schemas:

- Roles for creators/contributors: editor, photographer, cinematographer, advisor, donor
- Audience
- Edition or version
- Place of publication
- Location (museum/repository/library name)
- Style/genre/material technique
- Notes
- Metadata record information

Sample Record: Digitized Map

Identifier: http://digitalgallery.emory.edu/luna/servlet/detail/EMORYUL~3~3~2187~100303

Relation: Atlas of Atlanta and Vicinity, 1928

Coverage: Atlanta

Coverage: United States

Coverage: Fulton County; Cobb County

Coverage: Georgia

Creator: U.S. Coast and Geodetic Survey

Date: 1930

Description: Color map showing Pleasant Hill Church and Chattahoochee River.

Format: image/jpeg

Title: City of Atlanta: Sheet 51. Construction Department, William A. Hansell, Chief; S.P. Floore, Topographic Engineeer in charge. Topography by E.J. Essick and J.B. Leachman. Control by U.S. Coast and Geodetic Survey and City of Atlanta Mapping Division. Surveyed in 1927. Williams & Heintz Co., Wash, D.C.

Publisher: U.S. Coast and Geodetic Survey and City of Atlanta Mapping Division

Rights: The City of Atlanta has granted Emory University, Woodruff Library, permission to digitize, distribute, display and geo-reference maps produced by the U.S. Coast Guard and Geodetic Survey and the City of Altanta Mapping Division in a 1928 survey published as the Atlas of Atlanta and VIcinity. Emory may digitize, display, and georeference the maps in electronic formats, including free public access to maps on the web. The City of Atlanta does not attest to the accuracy of the image. The Maps Content, including all images and text, are for personal, educational, and non-commercial use only.

Type: Atlas Map

Type: Image

Sample Record: Digitized Photo

Identifier: MSS1218_B001_I117_P0002

Title: Surveying crew staking out a grove, 1936

Description: Recto: Surveying crew staking out a grove 1936, [Tung Grove Development Co., Florida]

Date: 1936

Subject: African American men.

Subject: Employees.

Subject: Surveying.

Type: Image

Type: Black-and-white photographs

Format: 05.61 x 07.47 inches

Related: Robert Langmuir African American Photograph Collection, MSS1218, Manuscript, Archives, and Rare Book Library, Emory University

Rights: Emory University does not control copyright for this image. This image is made available for individual viewing and reference for educational purposes only such as personal study, preparation for teaching, and research. Your reproduction, distribution, public display or other re-use of any content beyond a fair use as codified in section 107 of US Copyright Law is at your own risk. We are always interested in learning more about our collections. If you have information regarding this photograph, please contact marbl@emory.edu.

Sample Record: Web Application

Title: Battle of Atlanta

Creator: Pollock, Daniel

Subject: American Civil War (1861-1865)

Subject: Battlefields

Subject: Battle of Atlanta

Description: This website is a resource for getting in touch with the Battle of Atlanta in history and in memory. It is a means of engaging with the recorded past and the remembered past of a particularly fierce fight between North and South on July 22, 1864. This site combines a narrative of battlefield events with images from the Cyclorama and other visual and textual artifacts into a digital tour guide with contextual links.

Publisher: Emory Center for Digital Scholarship, Emory University, Atlanta, GA

Contributor: Tullos, Allen

Contributor: Varner, Jay

Contributor: Croxall, Brian

Date: 2013

Type: InteractiveResource

Format: text/html

Identifier: http://dev.emorydisc.org/battleofatlanta/tour/battle-of-atlanta/

Language: English

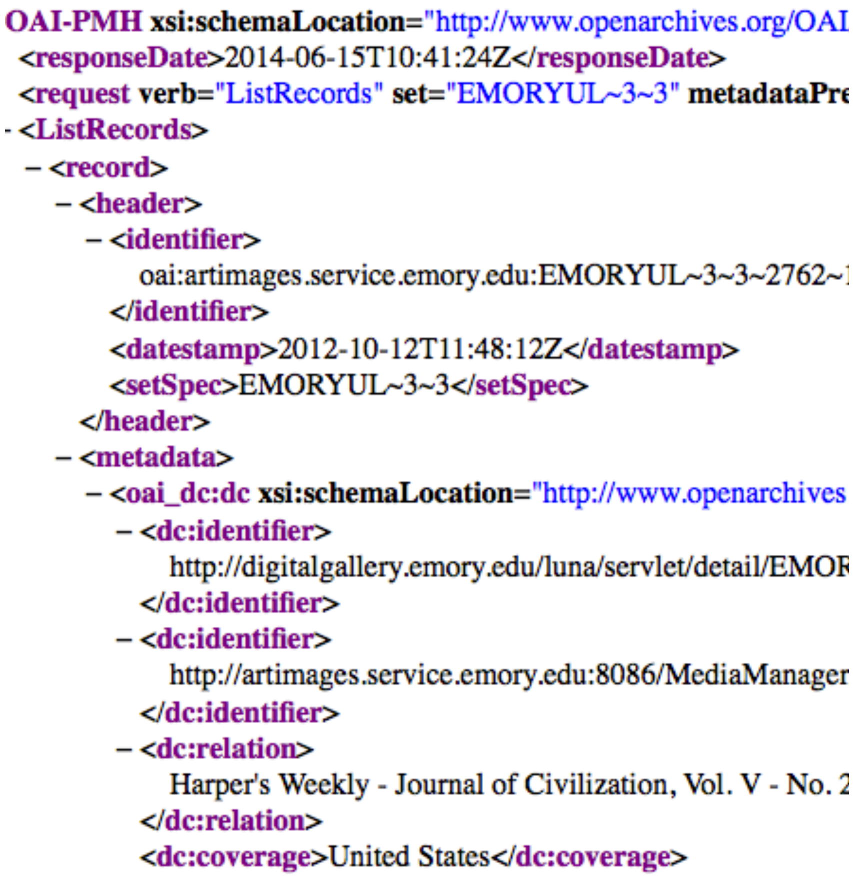
Coverage: 1864

Populating Dublin Core – General Tips

- Use structured entries over free-text when possible
  - Controlled vocabularies, URLs, date formats etc.
- If an element’s value is unknown, don’t create it (avoid “Unknown” or blank element entries)
  - ~~Publisher: unknown~~
- Create separate element instances for multiple entries
  - <subject>Term 1</subject>
  - <subject>Term 2</subject>
  - Vs. <subject>Term 1; Term 2</subject>
- (See the Dublin Core cheat sheet for more guidelines)

Harvesting and Sharing with OAI-PMH

- Open Archives Initiative Protocol for Metadata Harvesting
- Standard data mechanism for importing/exporting records across systems or organizations (XML)
- Many systems automate OAI
- Simple Dublin Core is *always* required within OAI
  - You can also transmit your records in additional formats, too (e.g. supply both MODS and DC versions of records)



```
OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI
 <responseDate>2014-06-15T10:41:24Z</responseDate>
 <request verb="ListRecords" set="EMORYUL~3~3" metadataPre
- <ListRecords>
  – <record>
    – <header>
      – <identifier>
          oai:artimages.service.emory.edu:EMORYUL~3~3~2762~
        </identifier>
        <datestamp>2012-10-12T11:48:12Z</datestamp>
        <setSpec>EMORYUL~3~3</setSpec>
      </header>
    – <metadata>
      – <oai_dc:dc xsi:schemaLocation="http://www.openarchives
        – <dc:identifier>
            http://digitalgallery.emory.edu/luna/servlet/detail/EMOR
          </dc:identifier>
        – <dc:identifier>
            http://artimages.service.emory.edu:8086/MediaManager
          </dc:identifier>
        – <dc:relation>
            Harper's Weekly - Journal of Civilization, Vol. V - No. 2
          </dc:relation>
          <dc:coverage>United States</dc:coverage>
```

METADATA

Web Find-ability:
Search Engine Optimization,
Social Media Metadata

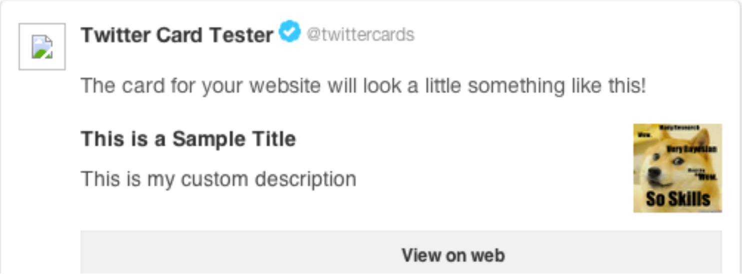
EMORY
LIBRARIES &
INFORMATION
TECHNOLOGY

Guidelines for SEO

- **Filenames**
  - Short and sweet: human and machine readable
  - Google prefers hyphens to separate words
- **Use good document structure**
  - Use heading tags (<h1>, <h2> etc.)
  - Make sure your content level title is formatted as a Heading 1/<h1>
  - Google looks at the document as a whole
- **Title <title>**
  - Shown as link text in search results (~70 chars)
  - Provide unique titles for all pages
  - Best practice – include web site name appended to end of individual resource title
- **Description <meta name="description">**
  - Google uses this as the excerpt in search results (~160 chars)
- **HTML 5**
  - More restrictive on <meta> tags' name attributes (new registry)
- **Google webmaster account**
  - Google sitemap – XML outline of your site (helps Google discover your content)
  - Track and refresh Google crawls of your site
  - Test structured data for rich snippets
- **Google Analytics account**
  - Track visitor demographics, content usage
  - Configure site search for your account to see search terms
- **Maturity of site; incoming links**
  - The longer and more established your site is, the better
  - When re-designing a site, make sure to provide redirects to your most visited pages

Social Media Metadata

- Emerging, proprietary standards offer additional context for sharing/ discovery in social media platforms
  - Override and enhance default values
  - Some re-usability across platforms
  - Can be automated
- **Twitter Cards**: customize thumbnail, description, title, creator information, type of content
  - Tour app integration
- **Facebook OpenGraph**: customize thumbnail, description, title, language/ locale, type of content
  - Tour app integration
- **Pinterest – Rich Pins**: product information, related links



Twitter Card Tester @twittercards

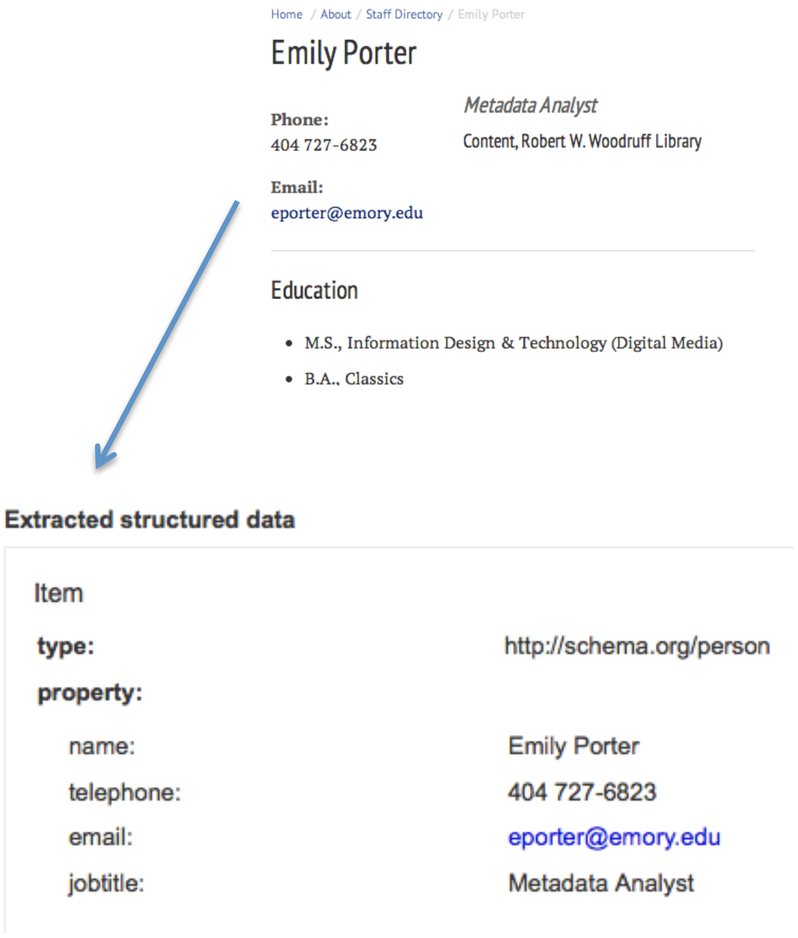
The card for your website will look a little something like this!

This is a Sample Title

This is my custom description

Search Engine Initiatives

- Shift from <meta> tags, keywords, Dublin Core in <head>
- Schema.org
  - Google, Bing, Yahoo
  - Creative Work definition
  - Person definition
- Microdata, microformats, RDFa
  - Embedded structured metadata as attributes within your HTML content
- Rich snippets
  - More contextual display in search results
- Google Structured Data tool (for testing)



Home / About / Staff Directory / Emily Porter

Emily Porter

Metadata Analyst

Content, Robert W. Woodruff Library

Phone:
404 727-6823

Email:
eporter@emory.edu

Education

- M.S., Information Design & Technology (Digital Media)
- B.A., Classics

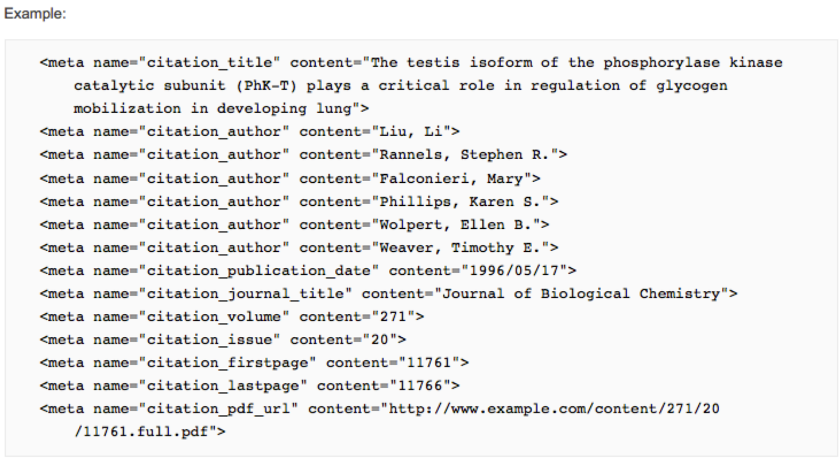
Extracted structured data

Item

type:	http://schema.org/person
property:	
name:	Emily Porter
telephone:	404 727-6823
email:	eporter@emory.edu
jobtitle:	Metadata Analyst

Google Scholar – Tactics for Inclusion

- Very specific markup requirements
- Prefers <meta> tags in one of the following formats:
  - Highwire Press
  - Eprints
  - BE Press
  - PRISM
- Recommends *against* use of Dublin Core (not granular enough)
- Crawling/indexing challenges when mixing scholarly articles with other types of content



Example:

```
<meta name="citation_title" content="The testis isoform of the phosphorylase kinase
    catalytic subunit (PhK-T) plays a critical role in regulation of glycogen
    mobilization in developing lung">
<meta name="citation_author" content="Liu, Li">
<meta name="citation_author" content="Rannels, Stephen R.">
<meta name="citation_author" content="Falconieri, Mary">
<meta name="citation_author" content="Phillips, Karen S.">
<meta name="citation_author" content="Wolpert, Ellen B.">
<meta name="citation_author" content="Weaver, Timothy E.">
<meta name="citation_publication_date" content="1996/05/17">
<meta name="citation_journal_title" content="Journal of Biological Chemistry">
<meta name="citation_volume" content="271">
<meta name="citation_issue" content="20">
<meta name="citation_firstpage" content="11761">
<meta name="citation_lastpage" content="11766">
<meta name="citation_pdf_url" content="http://www.example.com/content/271/20
    /11761.full.pdf">
```

METADATA

Trends and Evolving Practices

EMORY
LIBRARIES &
INFORMATION
TECHNOLOGY

Trends

- **RDF/Linked Data**
  - Web model: making connections across distributed data vs. creating one-off records
  - Opening up closed systems
  - Shifting from unstructured text to ids/URIs: “use URIs as names for things”
  - Mixing metadata elements and vocabularies
  - Enabling new types of search engines/queries, machine learning

- **GIS/geocoding**
  - Extracting coordinates and mapping from place-names
  - Tate Museum: Art Maps

- **Crowdsourcing**
  - Metadata Games
  - Ancient Lives/Zooniverse
  - Langmuir African American Photographs

- **Automation**
  - Extraction of technical metadata
  - Embedding metadata
  - Text analysis
  - Image analysis

General shift toward distributed, role-based creation of metadata

METADATA

Resources and References

Metadata – General Overviews

- *Understanding Metadata (NISO)*
  - http://www.niso.org/publications/press/UnderstandingMetadata.pdf
- *Metadata for digital collections: a how-to-do-it manual*
  - Miller, Steven J. *Metadata for digital collections : a how-to-do-it manual.* New York : Neal-Schuman Publishers, 2011.
  - http://www.neal-schuman.com/nealschuman/companionwebsite/metadata-digital-collections
- *Digital Library Federation: Best Practices for Shareable Metadata*
  - Wiki site: http://webservices.itcs.umich.edu/mediawiki/oaibp/index.php/ShareableMetadataPublic
  - Also in print as *Best Practices for OAI PMH Data Provider Implementations and Shareable Metadata (2007)*
- *Glossary of Metadata Standards*
  - *http://www.dlib.indiana.edu/~jenlrile/metadatamap/seeingstandards_glossary_pamphlet.pdf*
- *An Introduction to Metadata (JISC)*
  - *http://www.jiscdigitalmedia.ac.uk/guide/an-introduction-to-metadata*
- *Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files (*Federal Agencies Digitization Initiative (FADGI) - Still Image Working Group)
  - http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf (includes helpful overviews of different types of metadata)

Search Engine/Social Media Resources

- Twitter Cards metadata tags reference:
  - https://dev.twitter.com/docs/cards/markup-reference
- Facebook OpenGraph metadata tags reference:
  - https://developers.facebook.com/docs/opengraph/howtos/maximizing-distribution-media-content
- WordPress - Add Meta Tags plugin:
  - http://wordpress.org/plugins/add-meta-tags/
- Google Webmaster:
  - http://www.google.com/webmasters/
- Google Analytics:
  - http://www.google.com/analytics/ (account setup)
  - https://support.google.com/analytics/answer/1012264?hl=en (site search)
- Google Scholar inclusion reference:
  - http://scholar.google.com/intl/en-US/scholar/inclusion.html#indexing
- Google: Rich Snippets
  - https://support.google.com/webmasters/answer/99170?hl=en