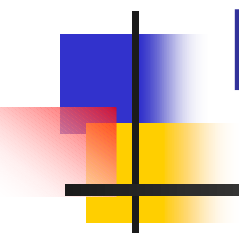
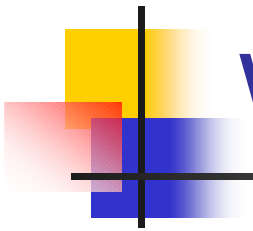


Module 3: All About Metadata

James Powell, Director
Web Application Research & Development
Virginia Tech



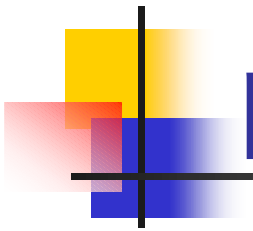
What is Metadata?

- Metadata is data about data – for example:
  - An abstract is metadata about a paper.
  - Herman Melville and “Moby Dick” are metadata about a book.
- Any piece of information that describes an object can be considered metadata.

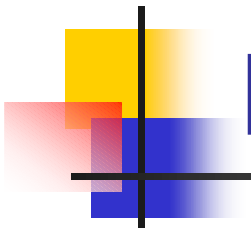
What are some examples?

- Common metadata elements that can be used to describe various types of documents, such as books, articles, or web pages include:
  - Title
  - Author
  - Subject
  - Keywords
  - Description
  - Abstract

Metadata is for software, not people



- Metadata is for software (like search engines), not people.
- But, search engines may display some metadata in result sets.



Metatags

- Metatags are a type of standard HTML tags that can be included in web pages.
- They are data containers with their own distinct format:

```
<META name=“title” content=“Moby Dick”>
<META name=“date” content=“May 2, 2003”>
<META name=“genus” content=“ginko”>
```

Metatags are invisible to humans

- Unless you view the source of a web page, you will not see metatags when you browse the web.
- Metatags are used by web servers and web browsers to perform a variety of tasks, including web page indexing.

Metatags can contain metadata

- Since metatags are flexible data containers, you can use them to contain metadata about a web page.
- On the web, most search engines look for:

```
<META name="keywords" content="some keywords">
```

and

```
<META name="description" content="a description
  of the page's contents">
```

Metatags look like image tags

- HTML metatags don’t wrap around any external text.
- You won’t see <META></META> in an HTML document.
- Metatags are similar to image tags in that all of their content is embedded within the tag itself:

```
<IMG SRC=“balloon.gif” alt=“picture of a
  balloon>
<META name=“title” alt=“Joy of Balloons”>
```

Where do they go?

- HTML Metatags are only valid within the header portion of an HTML document.
  - The header of an HTML document is the content between the <HEAD> start and </HEAD> end tag.
- You can include as many <META> tags within the header as you need. This is important, because metatags have other uses beyond containers for metadata.

Dublin Core Metadata

- Dublin Core is a name for a set of metadata designed to describe web content. It represents an international standard addressing characteristics that many web pages have in common including creator, title, description, and language (we'll review each element in more detail shortly)

Dublin Core Facts

- Originally developed in 1996
- There are 15 dublin core metadata elements
- Each element has a simple, descriptive name
- Each element is optional, and can occur as many times as needed
- Many elements may be associated with a predefined set of terms (a controlled vocabulary).
- Current version is 1.1 (issued July 1999)
- Official site for Dublin Core documentation is http://www.purl.org/metadata/dublin_core_elements

The five elements you should always include

- Title
- Creator (author)
- Subject (keywords)
- Description
- Publisher

Other elements you can include

- Contributor
- Date
- Type
- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights

DC.TITLE

- The title element indicates the name by which a web page is formally known.

Example:

```
<META name="dc.title" content="The Lost World">
```

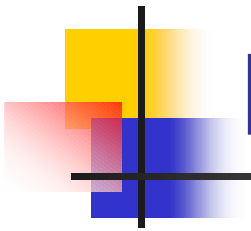
Note: When using Dublin Core metadata in HTML metatags, common practice is to use the dc prefix to indicate the type of metadata. It is perfectly acceptable to repeat the metatag using a different name (for example, title without the dc prefix) to ensure that search engines that are not specifically aware of the Dublin Core standard use your metadata.

DC.CREATOR

- The creator element refers to the person or organization or other type of entity that is primarily responsible for the document or resource.

Example:

```
<META name="dc.creator" content="Arthur Conan
   Doyle">
```



DC.SUBJECT

- The subject element contains the subject or the keywords associated with the document or resource being described.
- Often, subjects are selected from a predefined set of managed terms (a controlled vocabulary). This results in more consistent use of terms.

DC.DESCRIPTION

- The description element can contain a variety of styles of descriptive text about a document or resource, such as:
  - Abstract
  - Table of contents
  - Executive summary

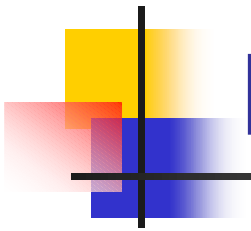
DC.PUBLISHER

- The publisher element refers to the person or organization responsible for making the document or resource available.
- Example:

```
<META name="dc.publisher" content="O'Reilly and
  Associates, Inc.">
```

DC.CONTRIBUTOR

- The contributor element contains information about a person, organization or service (same types of values as creator) that may have had some role in preparing the contents of a document or service.



DC.DATE

- The date can be used to refer to various times within the lifecycle of the document or resource, such as the creation date, last update date, last indexed date, etc.
- Dates should be formatted using a standard (such as ISO 8601), and should follow the YYYY-MM-DD format.

DC.TYPE

- The type element describes the nature or genre of the document or service being described.
- A working group has created a draft document which describes a control vocabulary for the type element:

collection
dataset
event
image
interactive resource
model
party
physical object
place
service
software
sound
text

DC.FORMAT

- Format refers to the type of electronic document being described.
- It is recommended that you use MIME types to indicate document type. Here are some examples:

  Web page: text/html

```
<META name="dc.format" content="text/html">
```

  PDF document: application/pdf

```
<META name="dc.format" content="application/pdf">
```

  MS Word document: application/msword

```
<META name="dc.format" content="application/msword">
```

DC.IDENTIFIER

- An identifier is a unique string that can be used to locate the item.
- The recommendation is that the identifier conform to a standard appropriate for the document type: ISBN for a book, URL for a web page, etc.
- Example:

```
<META name="dc.identifier"
   content="http://www.vt.edu/index.html">
```

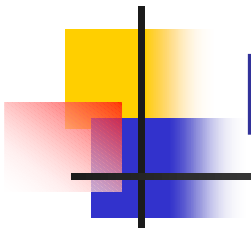
DC.SOURCE

- If the item is derived from a previous work, then the source element should be included as a pointer to each item from which is was derived.
- The source value in the derived work should be an identifier for the item from which it was derived (same as its <DC.identifier> value).

DC.LANGUAGE

- The language element indicates the primary language of the document's content.
- The recommended format for this value is a two digit language code (ISO639), followed by a 2 digit country code (ISO3166).
- Example:

```
<META name="dc.language" content="en-uk">
```



DC.RELATION

- The relation element indicates that there is some type of relationship between this document and other documents.
- The value for this element should be the identifier of the document to which this one is related.

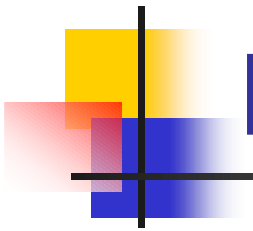
DC.COVERAGE

- Coverage can be used in a number of different ways, depending on the type of object described. According to the DC docs, coverage can be:
  - spatial location (a place name or geographic coordinates)
  - temporal period (a period label, date, or date range)
  - jurisdiction (such as a named administrative entity)

DC.RIGHTS

- The rights element contains textual information about copyright and the owner of the intellectual property contained within the object being described.
- Information may often be expressed in terms of the laws of a specific region, but potential users should make no assumptions about other uses.

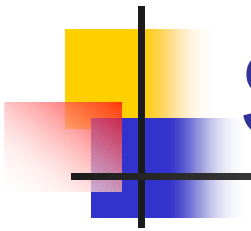
Resource Description Framework



- RDF allows you to embed XML-based metadata in other types of XML documents.
- An XML namespace reference that indicates the type of metadata being embedded supports validation and interoperability.

XML/RDF Dublin Core Example

```
<?xml version="1.0"?>
  <!DOCTYPE rdf:RDF SYSTEM "http://dublincore.org/2000/12/01-dcmes-xml-
   dtd.dtd">
  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
           xmlns:dc ="http://purl.org/dc/elements/1.1/">
    <rdf:Description rdf:about="http://dublincore.org/">
      <dc:title>WARD - Home Page</dc:title>
      <dc:description>Web Application Research and Development web
        site.</dc:description>
      <dc:date>2000-07-31</dc:date>
      <dc:format>text/html</dc:format>
      <dc:identifier>http://www.ward.vt.edu/</dc:identifier>
      <dc:language>en</dc:language>
    </rdf:Description>
  </rdf:RDF>
```



Summary

- There are many types of domain specific metadata, but on the web, basic metatags (keywords, description), Dublin Core metatags, and RDF/XML Dublin Core elements are the best methods for describing content intended for wide audiences.