

# Announcements

---

- Getting “signed off”

# Metadata (or maybe meta-data)

*Lawrence Snyder*  
*University of Washington, Seattle*

# Metadata In The News

- Most Americans hadn't heard of metadata until recently when Edward Snowden told everyone the NSA is keeping their phone metadata:

- Date
- Time
- No. Phoning
- No. Called
- Duration
- (Location)

This is data about a (mobile) phone call, but it's content is not recorded

# Metadata Is Very Informative

- NSA just collects metadata, who cares, right?
- If someone calls a “suicide hotline” no one is listen to the call except the receiver ... right?
- Maybe, but metadata can be used to make logical inferences
- Fact: Using anonymous metadata alone it's possible to determine if a caller is male / female
- Using cellphone “hand-offs” it's possible to figure out where a caller lives, where they work and those two facts usually can get you their name

# What Does Metadata Tell You?

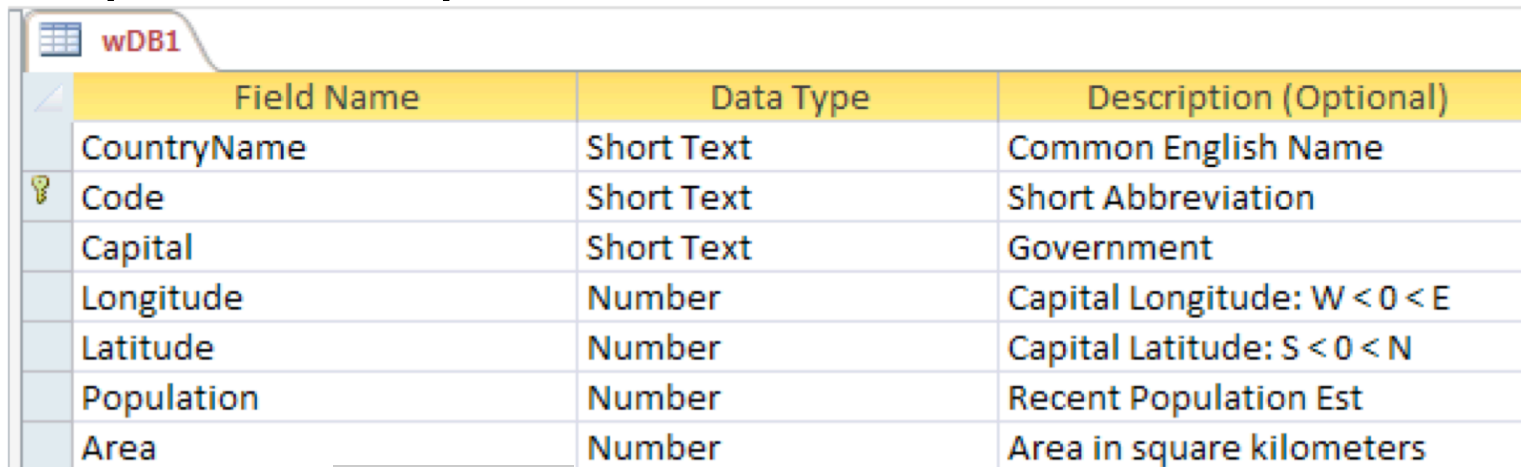
- What does this scenario tell you:
  - Woman calls an obstetrician
  - Minutes later, she calls a woman she calls often, and always calls on holidays like mother's day
  - Immediately calls a man, who she has generally phoned in late evenings
  - Next she calls Planned Parenthood, an abortion provider
- You didn't hear the conversations, so nothing has been revealed (recall privacy definition), right???

# Metadata Is An Important Idea

- We have discussed tags before
  - HTML – describes page layout
  - Oxford English Dictionary – aided in look & look-up
  - XML – Today's topic
    - Extensible Markup Language
    - Easy to learn because YOU make it up
    - Introduce the idea today
- Metadata doesn't REQUIRE tags, there are other ways of giving it, but tags are most common

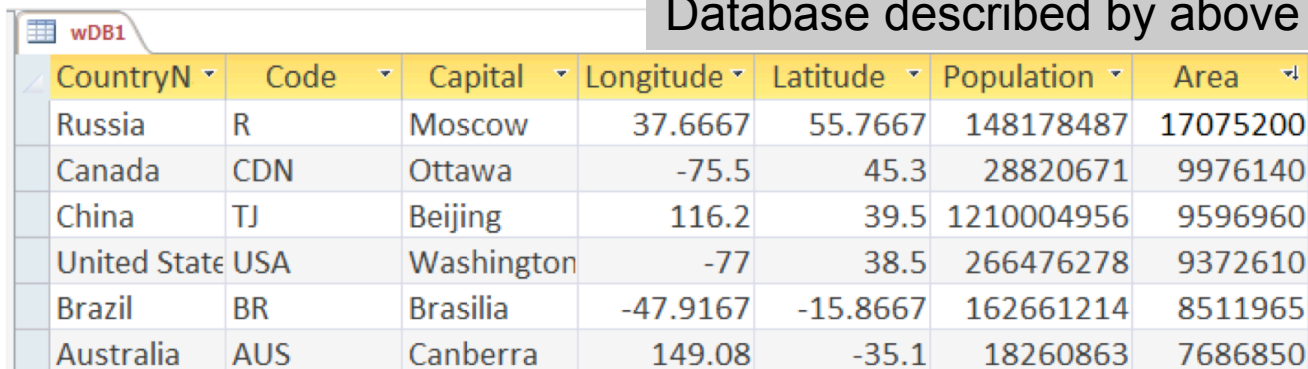
# Metadata In Relational Database

- Access (Microsoft's Relational Database System) captures metadata in a small table



Field Name	Data Type	Description (Optional)
CountryName	Short Text	Common English Name
Code	Short Text	Short Abbreviation
Capital	Short Text	Government
Longitude	Number	Capital Longitude: W < 0 < E
Latitude	Number	Capital Latitude: S < 0 < N
Population	Number	Recent Population Est
Area	Number	Area in square kilometers

Metadata



CountryN	Code	Capital	Longitude	Latitude	Population	Area
Russia	R	Moscow	37.6667	55.7667	148178487	17075200
Canada	CDN	Ottawa	-75.5	45.3	28820671	9976140
China	TJ	Beijing	116.2	39.5	1210004956	9596960
United State	USA	Washington	-77	38.5	266476278	9372610
Brazil	BR	Brasilia	-47.9167	-15.8667	162661214	8511965
Australia	AUS	Canberra	149.08	-35.1	18260863	7686850

Database described by above metadata

# Metadata Separation

- Metadata describes what the data is, but because the tags can be distinguished from the content, it *separates* itself from the content – that's smart

Separate the content and its tags entirely from the processing – produce an annotated data-only file



# XML Tags Invented for the OED

**byte** (balt). *Computers*. [Arbitrary, prob. influenced by bit sb.<sup>4</sup> and bite sb.] A group of eight consecutive bits operated on as a unit in a computer. **1964** *Blaauw & Brooks in IBM Systems Jnl.* III. 122 An 8-bit unit of information is fundamental to most of the formats [of the System/360]. A consecutive group of *n* such units constitutes a field of length *n*. Fixed-length fields of length one, two, four, and eight are termed bytes, halfwords, words, and double words respectively. **1964** *IBM Jnl. Res. & Developm.* VIII. 97/1 When a byte of data appears from an I/O device, the CPU is seized, dumped, used and restored. **1967** *P. A. Stark Digital Computer Programming* xix. 351 The normal operations in fixed point are done on four bytes at a time. **1968** *Dataweek* 24 Jan. 1/1 Tape reading and writing is at from 34,160 to 192,000 bytes per second.

```
<e><hg><hw>byte</hw> <pr><ph>baIt</ph></pr></hg>. <la>Computers</la>. <etym>
Arbitrary, prob. influenced by <xr><x>bit</x></xr> <ps>n.<hm>4</hm> </ps>and
<xr><x>bite</x> <ps>n.</ps> </xr></etym> <s4>A group of eight consecutive bits
operated on as a unit in a computer.</s4> <q><q><q>1964</q><a>Blaauw</a>
& <a>Brooks</a> <bib>in</bib> <w>IBM Systems Jnl.</w> <lc>III. 122</lc>
<qt>An 8-bit unit of information is fundamental to most of the formats <ed>of
the System/360</ed>.&es.A consecutive group of <i>n</i> such units constitutes
a field of length <i>n</i>.&es.Fixed-length fields of length one, two, four,
and eight are termed bytes, halfwords, words, and double words respectively.
</qt></q><q><q>1964</q> <w>IBM Jnl. Res. & Developm.</w> <lc>VIII.
97/1</lc> <qt>When a byte of data appears from an I/O device, the CPU is
seized, dumped, used and restored.</qt></q> <q><q> 1967</q> <a>P. A. Stark
</a> <w>Digital Computer Programming</w> <lc>xix. 351</lc> <qt>The normal
operations in fixed point are done on four bytes at a time.</qt><q><q><q>
1968</q> <w>Dataweek</w> <lc>24 Jan. 1/1</lc> <qt>Tape reading and writing is
at from 34,160 to 192,000 bytes per second.</qt></q></qp></e
```

# Using Metadata

- Metadata is usually a description of what the data is
  - Knowing what the data is, as in the OED, allows us to process it better for users
  - Here's an example: Search OED for def of “binary”
    - Without metadata, get 8,311 hits ... of which one is the definition
    - With metadata, get each definition in order ... how?  
<e><hg><hw>binary</hw> ... </hg> ... <e>

# Metadata Describes Data

- Metadata is data about data ... a description of what the data is
    - Knowing what the data is, as in the OED, allows us to process it better for users
    - Here's an example: Search OED for def of “binary”
      - Without metadata, get 8,311 hits ... which one is the definition?
      - With metadata, get each definition in order ... how?
- `<e><hg><hw>binary</hw> ... </hg> ... <e>`

The Principle: We can program computers to better help us if we say what the content is

# So, What's XML Good For?

Pretty Much Everything!

- Do you recognize
  - .docx
  - .pptx
  - .xlsx
- It's how to annotate data so (general) software can process it
- Consider an example...

# Enter The World of XML

- The Extensible Markup Language (XML) the tool for defining metadata; YOU think up the tags ... it is a self-defining language!
  - The usual rules for tags apply
    - Enclose in < and > and use lowercase ONLY
    - Start tag `<mynewtag>` and End tag `</mynewtag>`
    - Tags must always be matched or self-terminated
    - Tags can have attributes (think those up, too) of form  
`attributename="valueInQuotes"`
    - Use `.xml` as the file extension
    - Always start with “standard text” (shown later)

# Example of XML

- Suppose I want to record information about this class; using XML, I might write:

```
<class dept="cse">  
  <catalog qsr="true" credits="5">  
    <num>120</num>  
    <lec len="50" num="3">M, W, F</lec>  
    <lab len="50" num="2"> Tu,Th </lab>  
    <descrip>  
      Must-know computing knowledge for the  
      21st century</descrip>  
  </catalog>  
  <teach>L. Snyder</teach>  
</class>
```

I invented the tags; they make sense to me, and I could write software to process such descriptions

# Learning XML

- Since we think up the tags ourselves, it's the easiest language in the world to learn, right?
- Right.
- It's trivial?!
- Not quite ... there is a little technique, and we'll do that now
  
- Tags can serve in three roles ...

# Ways To Use Tags

- **Identity** – tag it so you know what it is

```
<name>George Washington</name>
```

```
<gen>Orsinus</gen><spe>orca</spe>
```

- **Affinity** – all properties of a thing should be grouped together

```
<personal>
```

```
  <name>George Washington</name>
```

```
  <height>6' 2"</height>
```

```
  <teeth>Wooden</teeth>
```

```
  <home>Mount Vernon</home>
```

```
</personal>
```



# Ways To Use Tags (continued)

- **Collection** – enclose a group of items of the same type in a collective tag

```
<presidents>  
  <prez num="1"><personal><name>George ...  
  <prez num="2"><personal><name>John ...  
  <prez num="3"><personal><name>Thomas ...  
  ...  
  <prez num="44"><personal><name>Barack ...  
</presidents>
```

- These uses become intuitive quickly

# Collecting Data About My Travels

- XML is a good tool for archiving information and then displaying it as a Web page
- Suppose this is my goal

## Places I've Traveled

### Washington State



The State of Washington is a fun place to visit. We toured Spokane, Grand Coulee Dam, Seattle's Space Needle and Mt. Rainier, which wasn't rainy at all, but beautiful in the sun!

### Oregon



South of Washington is Oregon. It is at the end of the old Oregon Trail. It is an unusual place. First, the University of Oregon's team is called the Ducks. Also, Mt. Bachelor is near the Sisters; with so many women around, why is it still a bachelor?

### California



California seems to be a republic, but not a banana republic. More like an orange republic. We visited San Francisco, San Quentin, the Monterey Bay Aquarium, LA and Hollywood. We didn't see any stars, but we were not there in the dark either.

### Alaska



Alaska is kind of hard to describe -- it is enormous. We visited Anchorage, of course, but we also saw Denali (Mt. McKinley), where we saw grizzly bears. In Fairbanks on the summer solstice, we saw the Midnight Sun Baseball game -- no lights, just sunshine!

# Ex:

Classify  
tag types:  
Identity  
Affinity  
Collection

```
- <travels>
  - <visit>
    <sight>Washington State</sight>
    - <action flag="wash.gif">
      The State of Washington is a fun place to visit. We toured Spokane,
      Grand Coulee Dam, Seattle's Space Needle and Mt. Rainier, which
      wasn't rainy at all, but beautiful in the sun!
    </action>
  </visit>
  - <visit>
    <sight>Oregon</sight>
    - <action flag="oregon.jpg">
      South of Washington is Oregon. It is at the end of the old Oregon
      Trail. It is an unusual place. First, the University of Oregon's team is
      called the Ducks. Also, Mt. Bachelor is near the Sisters; with so
      many women around, why is it still a bachelor?
    </action>
  </visit>
  - <visit>
    <sight>California</sight>
    - <action flag="california.png">
      California seems to be a republic, but not a banana republic. More
      like an orange republic. We visited San Francisco, San Quentin, the
      Monterey Bay Aquarium, LA and Hollywood. We didn't see any
      stars, but we were not there in the dark either.
    </action>
  </visit>
</travels>
```

# Summary

- Metadata is data about data
- Tags are a common form of metadata
- XML is main technology for metadata spec.
- Three roles for tags to fill ... you're building a tree
- By separating data from processing, expertise can be exploited, flexibility, wide usage
- We used metadata to add an image

# Example:

- Plop a standard header on it, develop the “style” for it (*next time*) and it’s ready to display

```
<?xml version = "1.0" encoding="UTF-8" ?>
<?xml-stylesheet type="text/xsl" href="weCookTags.xsl"?>
- <travels>
  - <visit>
    <sight>Washington State</sight>
    - <action flag="wash.gif">
      The State of Washington is a fun place to visit. We toured Spokane,
      Grand Coulee Dam, Seattle's Space Needle and Mt. Rainier, which
      wasn't rainy at all, but beautiful in the sun!
    </action>
  </visit>
  - <visit>
    <sight>Oregon</sight>
    - <action flag="oregon.jpg">
      South of Washington is Oregon. It is at the end of the old Oregon
      Trail. It is an unusual place. First, the University of Oregon's team is
      called the Ducks. Also, Mt. Bachelor is near the Sisters; with so
      many women around, why is it still a bachelor?
    </action>
  </visit>
  - <visit>
```