# Metadata

## Considerations for digital libraries

Tefko Saracevic, Ph.D.

# ToC

- Definitions
- Context & problems: digital resources
    - libraries & the Web
- Precursors
- Types of metadata
- Examples
- Issues in applications

# Definitions

- "Data about data"
  - glib, content-free catchphrase, but very popular
- Structured data about information resources describing their elements and functions
- Value added information which enables information objects to be:
  - identified
  - represented
  - managed
  - accessed & searched
  - preserved

# What?

- Machine understandable information for digital resources (particularly on the Web) - emphasis on **machine**
  - description of what a resource* or part is all about
    - e.g. labeling title, author, source, subjects …
    - a simple description – what is metadata?

*subsumes: textual documents, pictures, illustrations, movies, simulations, art objects, artifacts, software … – “anything that has an identity”

# Force for metadata developments: the Web

- Fastest growing technology in history
- Explosive growth of WWW provided
  - ubiquity of information and access
  - but also information chaos & anarchy
    - growing difficulty in identifying, searching & retrieving
    - ‘lost in an ocean’ metaphors

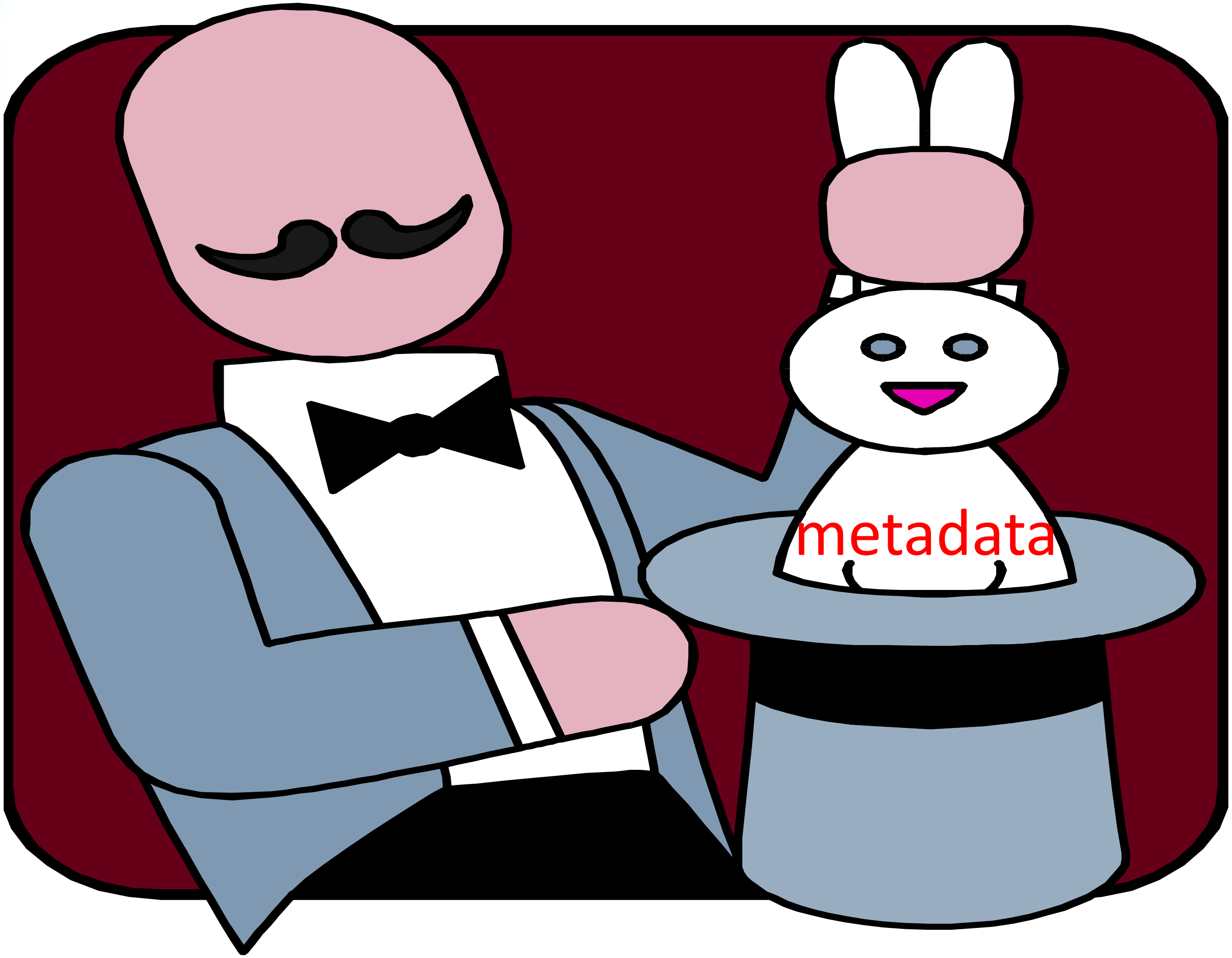
# Problem

- To organize & search the Web needed: knowledge about the structure of data
  - but Web data & databases fuzzy
  - structures vary widely; no consistency
  - constantly evolve over time
  - lack of agreement about meaning of even simple terms & concepts in structure

# Solution

- Some standardized description or language to increase functionality
  - a mechanism for a more precise description of things on the Web
- Future: Going from machine-readable to machine-understandable
  - semantic Web
  - missing in original Web architecture

→ METADATA !


metadata

# Where?

- In volatile digital environments
  - metadata describe electronic resources, texts & multimedia
  - metadata exist or have meaning only in relation to the referenced document or object
    - provide information about the object

# Why?

- To standardize description of electronic resources in collection(s) in order to:
  - aid in **identification, organization, & location** of objects (documents)
  - enable effective **search** of variety of objects distributed all over
  - sometimes also to provide **controls** (e.g. validation, rights, provenance, ratings ...)

# Importance

- Standard metadata descriptions are a prerequisite to
  - common use
  - effective searching
  - ‘intelligent’ roaming by agents
  - validation, ratings,

# Precursor: markup languages

- SGML – Standard General Markup Language
  - – granddaddy (standard in 1986)
  - – marks elements within documents
    - ▪ derived from old markups for typesetting
    - ▪ adapted by communities producing electronic documents
    - ▪ machine independent – main reason for success
      - – transportable from one hardware & software to another; substitutes strings
    - ▪ many extensions & specific applications

# Principles

- ALL markup language must specify
  - what markup means
  - what markup is allowed
  - what markup is required
  - how markup is distinguished from text
- All markup languages & applications follow these principles
  - underlying concepts are fairly simple but they get very confusing real fast.

# Followed by extensions

- HTML (technical specification)
- HTML (Wikipedia)
- Most famous & successful
  - allows for metatags in the Head
    - but these are not used much, even discouraged by some
    - in the body could be indirect

- XML (technical)
- XML – (Wikipedia)
- Extensible Markup Language
  - data format for structured document interchange & interoperability on WWW
    - increases functionality of SGML & combines with ease of use of HTML

# Standards for metadata

- Many standards developed or proposed
- Depends on need of a domain & purpose in application

- Conflicts between need for
  - specialized standards domain or community specific, and
  - generic standards enabling resource sharing/use/discovery across domains

# Who specifies metadata standards?

- Formal groups
  - national & international standards organizations - ISO, ANSI, NISO
- Informal groups
  - WWW Consortium (W3C)
  - Dublin Core Metadata Initiative
  - Standards at the Library of Congress

# Proliferation

- Currently: proliferation of metadata standards activities -many domains
  - a lot of confusion & incompatibility
  - in document description & libraries
    - coordination through liaisons & a number of projects in the U.S & internationally
      - strength: domain experts involvement
      - weakness: limited perspective; re-invention

# Sample of metadata projects

- Encoded Archival Description (EAD)
- Text Encoding Initiative (TEI) - international consortium for standards for digital texts
- Geospacial data - Federal Geographic Data Committee
- Z39.50 standards
  - information retrieval

- Understanding metadata
  - NISO publication exactly aimed at what the title says
  - includes simple description of various standards with examples
  - also listing of metadata sites

# Libraries

- In libraries metadata have a rich tradition long preceding the Web (but not called metadata)
  - – cataloging rules, standards widely applied
- MARC (Machine Readable Cataloging)
  - a computer-readable format that is used for bibliographic records
  - enabled worldwide exchange of cataloging records
  - but long standing problems with searching

# Dublin* Core Metadata Initiative

“making it easier to find information”

- international initiative to define a core set of metadata for description of digital resources
  - Web oriented
- wide interest & a lot of work, but not widely applied on the Web

* named for a 1995 conference in Dublin, Ohio, seat of OCLC, not Molly Malone’s fair city

- set of 15 elements:
  - Title
  - Creator
  - Subject
  - Description
  - Publisher
  - Contributor
  - Date
  - Type
  - Format
  - Identifier
  - Source
  - Language
  - Relation
  - Coverage
  - Rights

# How does it look like?

A Dublin Core record for a short poem, encoded as part of a Web e page using the <META> tag in HTML (from: Univ of Queensland Introduction to Metadata)

Dublin Core metadata

```
<HTML> !4.0!
<HEAD>
<TITLE>Song of the Open Road</TITLE>
<META NAME="DC.Title" CONTENT="Song of the Open
   Road">
<META NAME="DC.Creator" CONTENT="Nash, Ogden">
<META NAME="DC.Type" CONTENT="text">
<META NAME="DC.Date" CONTENT="1939">
<META NAME="DC.Format" CONTENT="text/html">
<META NAME="DC.Identifier"
   CONTENT="http://www.poetry.com/nash/open.html">
</HEAD>
<BODY><PRE>
I think that I shall never see
A billboard lovely as a tree.
Indeed, unless the billboards fall
I'll never see a tree at all.
</PRE></BODY>
</HTML>
```

# Comparing schemes

- Crosswalks: tables showing similarities & differences between metadata schemes
  - coping with different metadata standards
- Examples:
  - MARC to Dublin Core Crosswalk by LoC
  - Metadata Standards Crosswalk by Getty
  - A Repository of Metadata Crosswalks – D-Lib Magazine

# Semantic Web

## a hope for a future Web

- Effort by W3C (World Wide Web Consortium) led by Tim Berners-Lee, developer of the Web
  - "The **Semantic Web** provides a common framework that allows **data** to be shared and reused across application, enterprise, and community boundaries."
- Based on Resource Description Framework (RDF)
- So far more a vision of extension of the Web & creation of various tools than real viable application
  - and a bit nebulous as well
- Not clear how may be applied widely, even if viable

# Library interoperability

- Library catalogs mostly bound by proprietary software
- Middleware needed
  - e.g. protocols (based on Z39.50) provide for interaction of clients with many servers (catalogs)
- Problems remain with semantic interoperability

# Metadata & digitization

- Metadata assignment (cataloging) a key component in digitization of resources and electronic publishing
- Choices: a spectrum of possibilities to select & apply metadata
- Search for automation in assigning metatags
  - to speed up the process and make it economical
  - as yet progress incremental
- connection with cataloging, indexing

# Decisions, decision

- How & what to plan for metadata creation in conjunction with digital libraries?
- Target audience?
- Scope and depth?
- What to adopt? plug-in a scheme?
- How to integrate metadata projects?
- Needed skills? training? staffing?

# Issue: $$$$

- Costs of metadata: HUGE
  - operations, making decisions are complex & involved
  - large effort - time, personnel
  - learning many new things included
- Cooperative activities essential
- Libraries pushed out of libraries

RUTGERS

# Criticisms of metadata

- Too complicated
- Subjective & depends on context
- There is no end to metadata
- Other methods e.g. automatic by search engines, accomplish search & discovery effectively & efficiently
  - so who needs metadata?

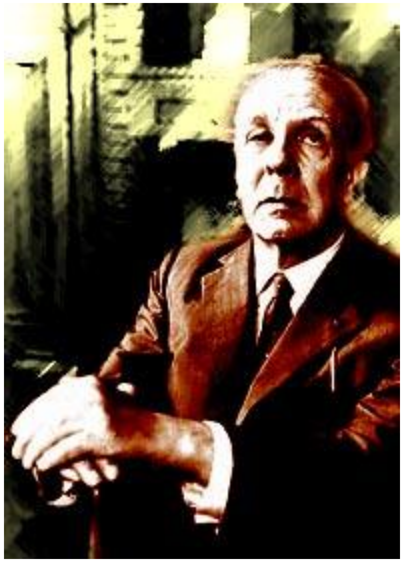
# In conclusion

- Effective access to digital resources depends on metadata
- Today, there are efforts to derive metadata automatically, using Natural Language Processing (NLP) methods
- Maybe automation of assigning metadata is the future?

# Dedicated to:

## Jorge Luis Borges

1899-1986



The Library of Babel – explore the idea

# One of delightful Jorge Louis Borges quotes

"These ambiguities, redundancies, and deficiencies recall those attributed by Dr. Franz Kuhn to a certain Chinese encyclopedia entitled *Celestial Emporium of Benevolent Knowledge*. On those remote pages it is written that animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance."

-- Essay: "The Analytical Language of John Wilkins"

metadata
Web
information
description
standards
resources
digital
markup
data
documents
libraries
META
CONTENT
NAME
HTML
search
searching
many
cataloging
electronic
projects
application
interoperability
machine
Language
Metadata
Dublin
Core
effective
Library
Natural
e.g
Initiative
HEAD
BODY
language
growing
SGML
WWW
creation
text
long
proliferation
objects
access
allows
texts
document
extensions
structure
languages
also
automation
see
lot
even
real
W3C
Consortium
time
validation
Processing
principles
decisions
domain
things
XML
Song
needed
technical
methods
digitization
international
never
functionality
activities
making
widely
hardware
Semantic
TITLE
problems
structured
Common
resource
elements
using
context
tree
variety
specific
meaning
standard
definitions
provide
applied
groups
NISO
domains
metatags
simple
Description
library
part
future
assigning
object
Road
applications
Open
concepts
community
PRE
software
examples
ratings
need
Wikipedia
followed
across
depends
use
viable
Z39.50
catalogs
set