# Multi-agent Semantic Web Systems: Data and Metadata

Fiona McNeill

School of Informatics

24th January 2013

# Data and Metadata

## Examples, 1

- pottery fragment: site of discovery
- packet of crisps: average salt content
- person: date of birth

## Examples, 2

- academic paper: date of publication
- map: scale (e.g. 1:25,000)
- audio files: sampling rate
- digital photo: make of camera used
- database entry: who entered the data
- web-page: topic

- Metadata: data about data

# More on metadata

**Hillman http://dublincore.org/documents/usageguide/2003**

A metadata record consists of a set of attributes, or elements, necessary to describe the resource in question.

# Associating Metadata with a Resource, I

Embedding: the metadata is physically contained in the resource.
Mainly relevant for digital resources, e.g., as a file header.

## Embedded metadata (Postscript)

```
%!PS-Adobe-2.0
%%Creator: dvips 5.526 Copyright 1986, 1994 Radical ...
%%Title: Paper.dvi
%%CreationDate: Tue Sep 13 12:38:42 1994
%%Pages: 24
%%BeginProcSet: tex.pro
/TeXDict 250 dict def TeXDict begin /N{def}def ...
```

# Associating Metadata with a Resource, 2

Aboutness: the metadata is a separate resource, and 'points' to the resource it is about.

# Associating Metadata with a Resource, 2

Aboutness: the metadata is a separate resource, and 'points' to the resource it is about.

## Resource Identifiers

What scheme can we use for globally identifying resources?

- digital resources: use URIs (Uniform Resource Identifiers)
  - Similar to URLs, but more general: URIs don't have to be addressable.

# Advantage of Explicit Metadata

- Discovering resources, both by software agents and by humans (searching, browsing).
- Compare web with a structured database:
  - database records can be searched according to the field

**DB Query**

```
SELECT   Author, Title
FROM     Catalogue
WHERE     Author = ”Burns”
```

# Advantage of Explicit Metadata

Google

Web Images Groups News Froogle more »

burns Search

Search: the web pages from the UK

**Web**

Robert **Burns** Country: the official Robert **Burns** site
The Robert **Burns** works archive, with full text indexed and searchable online.
www.robert**burns**.org/ - 24k - 10 Jan 2006 - Cached - Similar pages

MedlinePlus: **Burns**
**Burns**. ... Overviews; **Burns** (Mayo Foundation for Medical Education and Research) ...
Treatment; **Burns**: Taking Care of **Burns** (American Academy of Family ...
www.nlm.nih.gov/medlineplus/**burns**.html - 31k - 10 Jan 2006 - Cached - Similar pages

www.sciencedirect.com/science/journal/03054179
Similar pages

Welcome to **Burns** Guitars - **Burns** Guitars
The **Burns** Shadows Custom Edition comes with its own unique hard case finished ... The new **Burns** Shadows model has been awarded 5 Stars Guitarist Choice for ...
www.**burns**guitars.com/ - 19k - 10 Jan 2006 - Cached - Similar pages

# Formal Metadata Schemes

- Library catalogue cards adopt informal conventions for expressing metadata.

- What about formal conventions for recording computer-based metadata?

- Especially metadata about digital objects ...

- Example: Dublin Core Metadata Initiative

# Dublin Core (DC)

- Initiated by librarians
- Well established and widely used metadata standard
- 15 elements for describing resources
- *a small language for making a particular class of statements about resources*
- The resource is the implicit subject of the statements

**Example of DC Statements**

```
Title   = "A Red, Red Rose"
Creator = "Robert Burns"
Date    = 1794
Type    = poem
```

# Simple DC Elements

## DCMES

| Content | Intellectual Property | Instantiation |
|---|---|---|
| Coverage | Creator | Date |
| Description | Contributor | Format |
| Type | Publisher | Identifier |
| Relation | Rights | Language |
| Title | | |
| Subject | | |
| Source | | |

DCMES = Dublin Core Metadata Element Set

# How Elements are Defined

## Creator

An entity primarily responsible for making the content of the resource.
Examples of Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.

## Format

The physical or digital manifestation of the resource.
Typically, Format may include the media-type or dimensions of the resource.
Format may be used to identify the software, hardware, or other equipment needed to display or operate the resource.

# More on Elements

- Elements are not functions: they can be repeated.

**Repeated Elements**

```
Title = ”In the Heart of the Moon”
Creator = ”Ali Farke Touré”
Creator = ”Toumani Diabaté”
```

- There is no mandatory constraint on element values; but recommended best practice is to use a ‘controlled vocabulary’.

- Some DC Qualifiers (next slide) provide the latter.

# Simple and Qualified Dublin Core

Simple DC: 15 elements listed above

Qualified DC:

- Additional 3 elements: Audience, Provenance and RightsHolder

- Qualifiers - these extend or refine the original 15 elements

# Qualifiers: Refinement

## Element Refinement

Making the meaning of an element more specific.

## Example: Refinements of Date

Used when more than one date is needed.

```
dateSubmitted = 2001-01-31
dateAccepted  = 2001-10-01
```

# Qualifiers: Encoding Scheme

## Encoding Scheme

Provides controlled vocabulary or formatting structure to aid interpretation of an element value.

## Example: Controlled Vocabulary for Language

Value of Language element is selected from list registered by ISO 639-2 (Alpha-3 Code)

Language = eng

## Example: YYYY-MM-DD format for dates (W3CDTF)

dateSubmitted = 2001-01-31

# Generalising the notion of Resource

- In the Semantic Web vision, anything can be a resource.
- The data / metadata distinction is blurred.
- Challenge: representing knowledge about resources on a web-scale.

# Challenges to ‘controlled vocabulary’

## Johann Strauss

```
Title = ”Wiener Waltz”
Creator = ”Johann Strauss”
```

## Wikipedia Entry

- Johann Strauss I (1804-1849), or Johann Strauss Sr., composer, popularizer of the waltz
- Johann Strauss II (1825-1899), or Johann Strauss Jr., composer, known as the ”Waltz King”, son of Johann I
- Johann Strauss III (1866-1939), composer, son of Eduard Strauss and grandson of Johann I

# More on Identifiers

- Problems with ambiguous names.
- Problems with synonymous names.

**Synonyms (Aliases)**

```
J. Strauss I
Johann Strauss Vater
Johan Strauss, Sr.
Johann Strauß sr.
Johann Straus sr.
Johann Strauss Sr
Johann Strauss Snr.
```

# Unique Identifiers, I

- DBPedia (http://dbpedia.org): semi-automatic transformation of Wikipedia into RDF.
- Every resource that is the subject of a page in Wikipedia has a corresponding URI in DBpedia.

## DBPedia URIs

```
Wikipedia: http://en.wikipedia.org/wiki/Johann_Strauss_I
DBpedia: http://DBpedia.org/resource/Johann_Strauss_I
```

# Unique Identifiers, 2

- MusicBrainz (`http://musicbrainz.org`): user-maintained 'metadatabase' for music

- Collects and makes available information such as artist name, release title, and the list of tracks that appear on a release

- Each artist receives an ArtistID `http://musicbrainz.org/artist/UUID`, where UUID is a (128-bit) Universally Unique Identifier in its 36 character ASCII representation.

# RDF Background

- Dublin Core provides a syntax and a vocabulary for talking about resources.

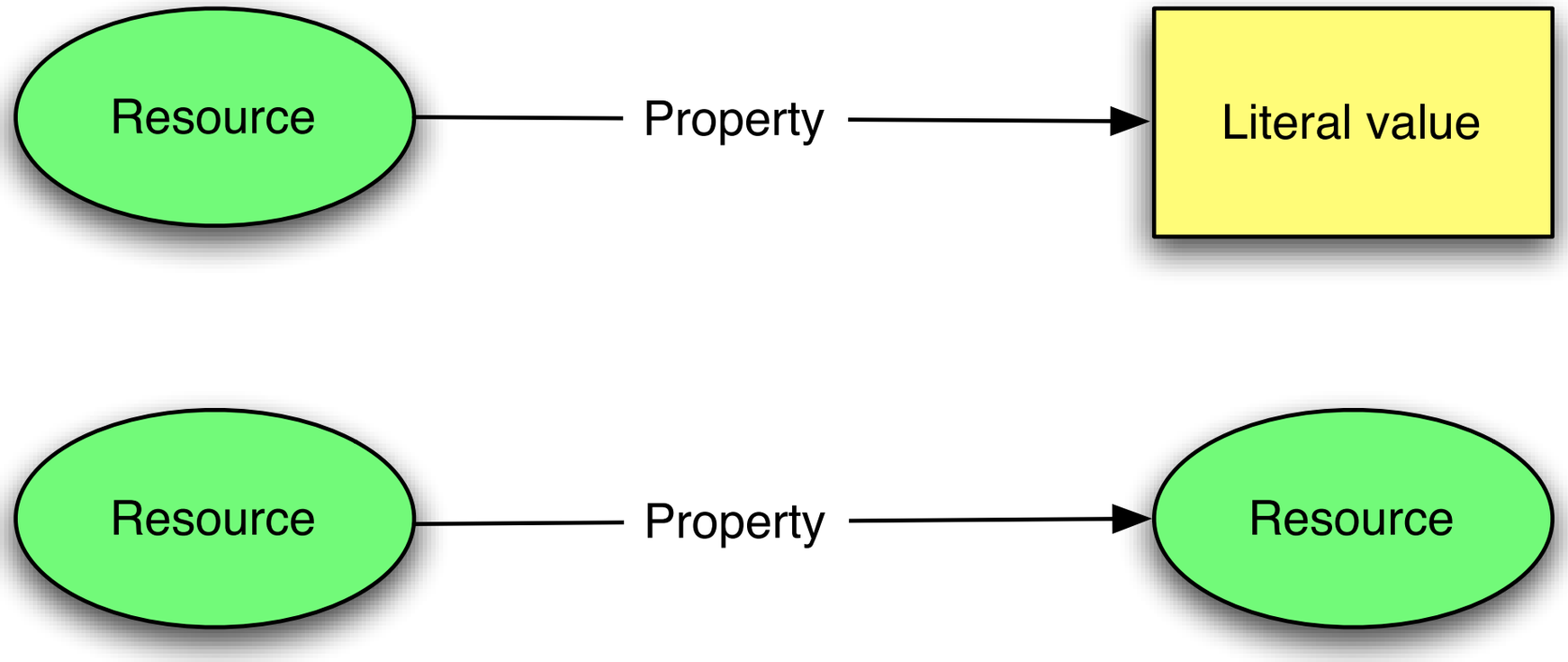
- The vocabulary is given by the elements (Title, Creator, Format, ...)

- Lots of different, specialised vocabularies for talking about different objects / domains.

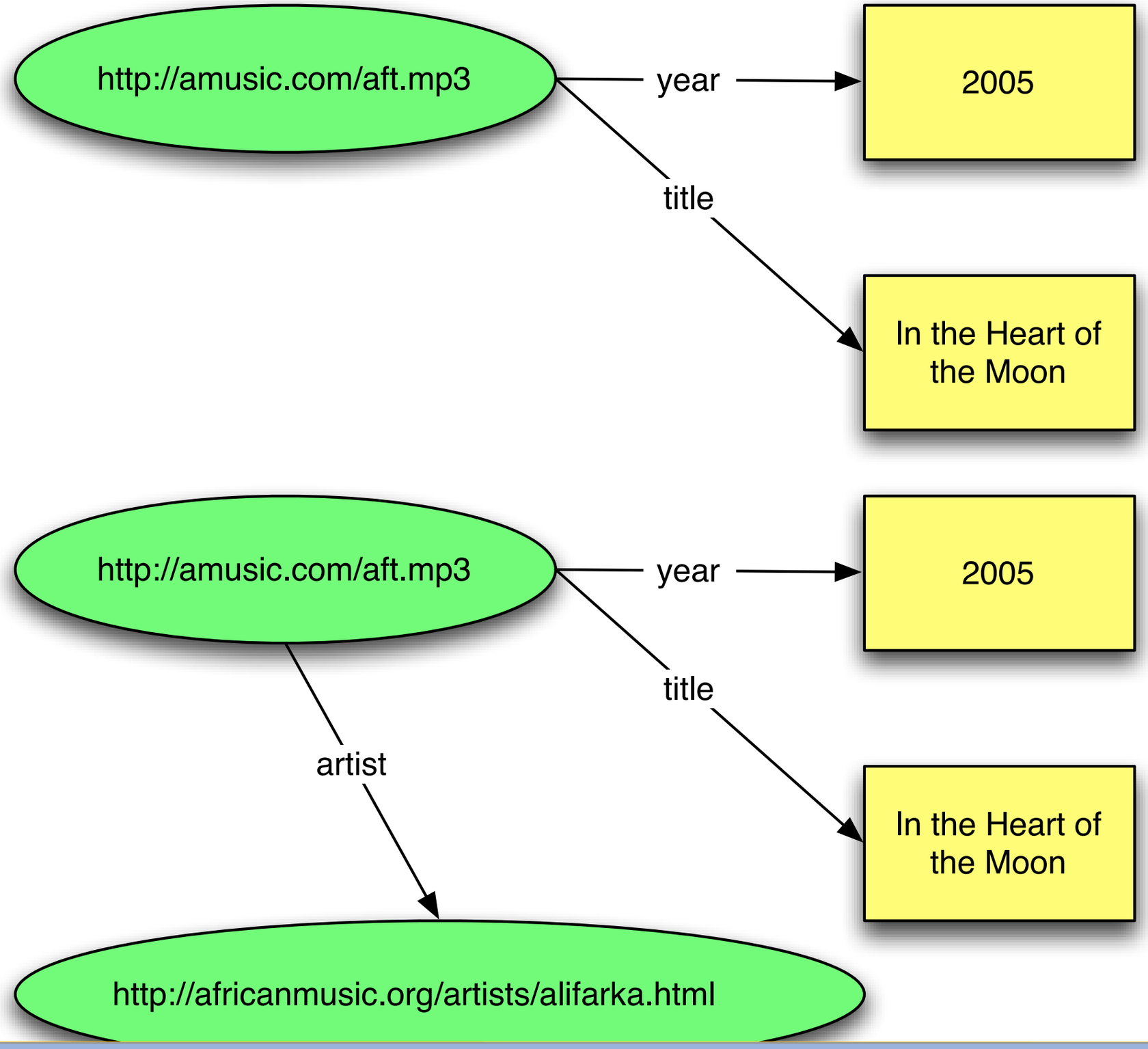
- W3C decided to build infrastructure where users can make assertions using their own vocabularies:
    - Resource Description Framework (RDF)

- RDF Working Group established in 1997

# RDF Data Model



# RDF Example 1



# Syntax: Dublin Core v RDF

## Dublin Core

```
Title = ”In the Heart of the Moon”
Date  = ”2005”
Identifier = dbpedia:In_the_Heart_of_the_Moon
Creator = dbpedia:Ali_Farka_Touré
```

## RDF Style

```
dbpedia:In_the_Heart_of_the_Moon dc:title ”In the Heart of the Moon” .
dbpedia:In_the_Heart_of_the_Moon dc:date ”2005” .
dbpedia:In_the_Heart_of_the_Moon dc:creator dbpedia:Ali_Farka_Touré .
```

# RDF Syntax

- RDF statements identify a resource being described; a specific property; and value of the property.
- Terminology:
  - subject (e.g., dbpedia:In_the_Heart_of_the_Moon)
  - predicate (e.g., dc:date)
  - object (e.g., ”2005”)

**RDF Triples**

$$\overbrace{\texttt{dbpedia:In\_the\_Heart\_of\_the\_Moon}}^{\text{subject}} \quad \overbrace{\texttt{dc:date}}^{\text{predicate}} \quad \overbrace{\texttt{”2005”}}^{\text{object}} \;.$$

- objects can be literals (e.g. strings) or resources.
- subjects can only be resources.
- more usual relational syntax:

date(dbpedia:In_the_Heart_of_the_Moon, ”2005”)

# Processing RDF Statements

- RDF is designed to make machine-processable statements.

- Two things required:

1. a machine-processable syntax for expressing RDF statements ⇒ usually XML

2. a machine-processable system for unambiguously identifying subjects, predicates and objects ⇒ URIs

# URIs

- Uniform Resource Identifier (URI): a simple and extensible means for identifying a resource.

**Examples of Resources**

an electronic document, an image, a source of information with a consistent purpose (e.g., "today's weather report for Los Angeles"), a service (e.g., an HTTP-to-SMS gateway), a collection of other resources

- Uniform Resource Location (URL): a special kind of URI that specifies a network location.
- A URI does not need to identify a network-accessible resource.

# More on URIs

## Example URIs

1. `http://www.ietf.org/rfc/rfc2396.txt`
2. `http://www.example.com/my/fictitious/example`
3. `ftp://ftp.is.co.za/rfc/rfc1808.txt`
4. `mailto:John.Doe@example.com`
5. `news:comp.infosystems.www.servers.unix`

- (1)–(2) are HTTP URIs.
- Originally intended to identify information resources (or documents), i.e., things which
  - carry some semantic content;
  - can be represented digitally.

# Summary

- Dublin Core is a good concrete illustration of a formal metadata scheme.

- Motivation: more effective methods for finding resources on the web.

- Illustrates a protracted standardization effort (started in 1994, DCMES became an ISO standard in 2003).

- Simple language: restricted set of elements, key-value pairs.

- Some extensibility via qualifiers.

# Summary

- Metadata inevitably leads to describing concrete resources (e.g., people)

- But names are often ambiguous, and hard for machines to deal with.
  - China: than 1.1 billion people share just 129 surnames (cf. 'Identity Crisis' paper reference at http://sites.google.com/site/masws09/uris).

- Various approaches for generating unique identifiers for resources.
  - E.g., OpenID for people.

# Task

- Choose 3 things.
- Write down as much metadata about them as you can.
- Consider whether each piece of metadata is functional or not.
- What possible sources of confusion might there be?