THE UNIVERSITY of EDINBURGH
informatics

Semantic Web Systems

Metadata

Jacques Fleuriot

School of Informatics

In the previous lecture

Possible components of ontologies contain:

- individuals
- classes
- attributes
- relations
- functions
- axioms
- planning rules

Representation considerations:

- trade-off between expressivity and efficiency.
- decidability, soundness, completeness.

In this lecture

- Metadata
  - What, how and why.
- Dublin Core
  - A formal metadata scheme.
- Unique Identifiers
  - Address ambiguous and synonymous names.
- RDF
  - A gentle intro.

Metadata

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
- 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

http://dublincore.org/documents/usageguide (Hillmann, 2005)
http://wiki.dublincore.org/index.php/User_Guide (Rühle et al, 2011-)

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

Associating metadata with a resource

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

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**

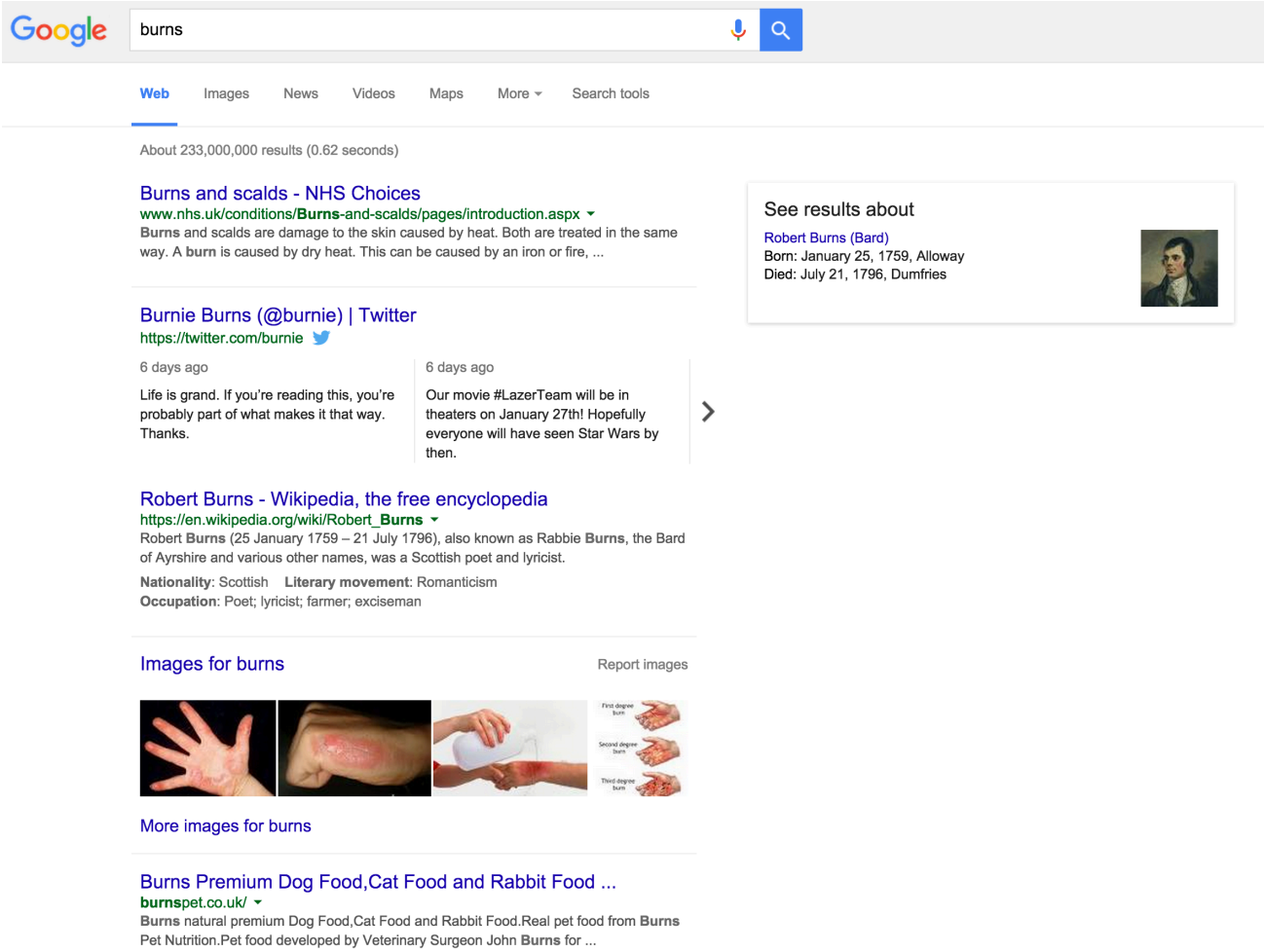
Advantage of explicit metadata

- 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



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

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

Dublin Core Metadata Element Set (DCMES)

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

http://dublincore.org/documents/dces/

How elements are defined

- Creator: An entity primarily responsible for making the content of the resource.
  - Examples of a Creator include a person, an organization, or a service.
  - Typically, the name of a Creator should be used to indicate the entity.
- Format: The file format, physical medium, or dimensions of the resource.
  - Examples of dimensions include size and duration.
  - Recommended best practice is to use a **controlled vocabulary** such as the list of Internet Media Types [MIME].

More on elements

- Elements are not functions: they can be repeated.

Repeated Elements
Title = “In the Heart of the Moon”
Creator = “Ali Farka 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 provide the latter.

Simple and Qualified Dublin Core

- Simple DC: 15 elements listed earlier.

- Qualified DC:
  - Additional 3 elements: Audience, Provenance and RightsHolder.
  - Qualifiers 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

Unique Identifiers

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
Johann Strauss, Sr.
Johann Strauß sr.
Johann Straus sr.
Johann Strauss Sr
Johann Strauss Snr.

Unique Identifiers

- 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

- 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 of the form:

http://musicbrainz.org/artist/UUID

where UUID is a (128-bit) Universally Unique Identifier in its 36 character ASCII representation.

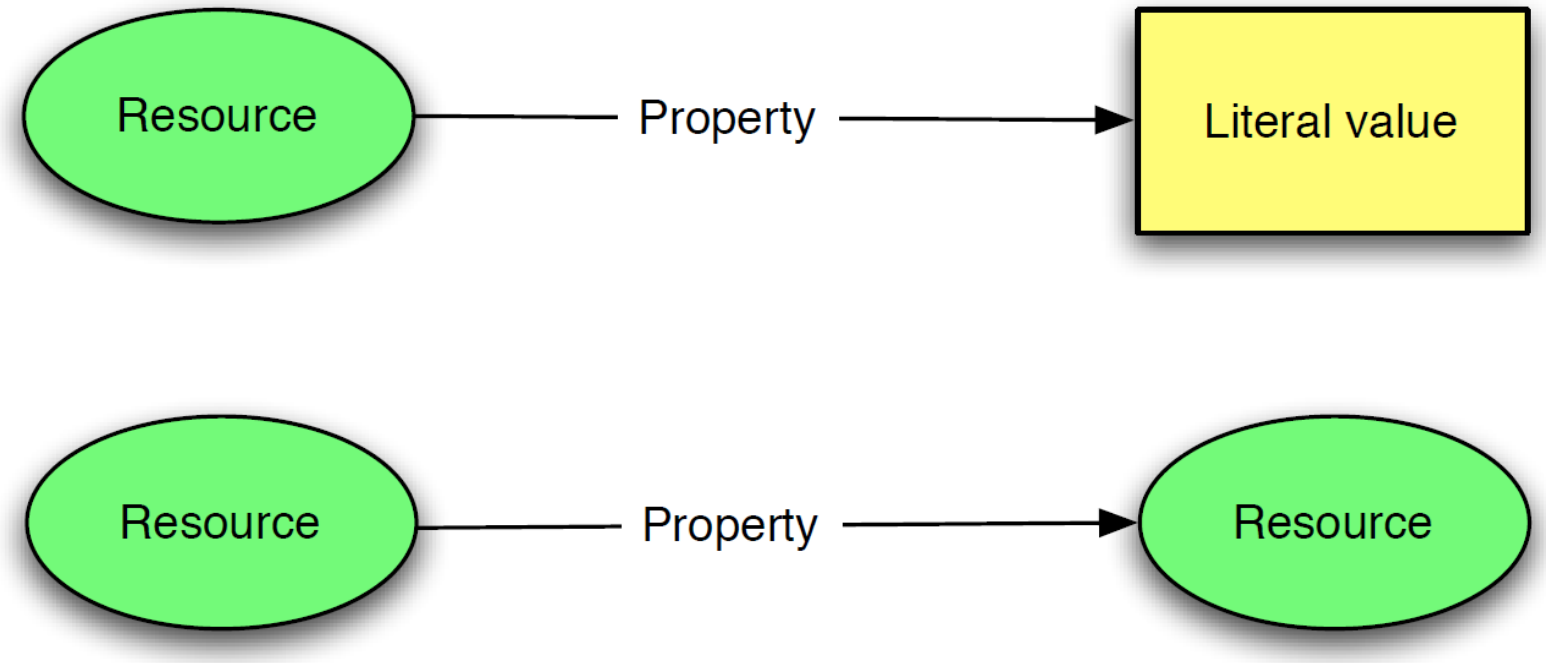
Example: http://musicbrainz.org/artist/9fff2f8a-21e6-47de-a2b8-7f449929d43f

RDF

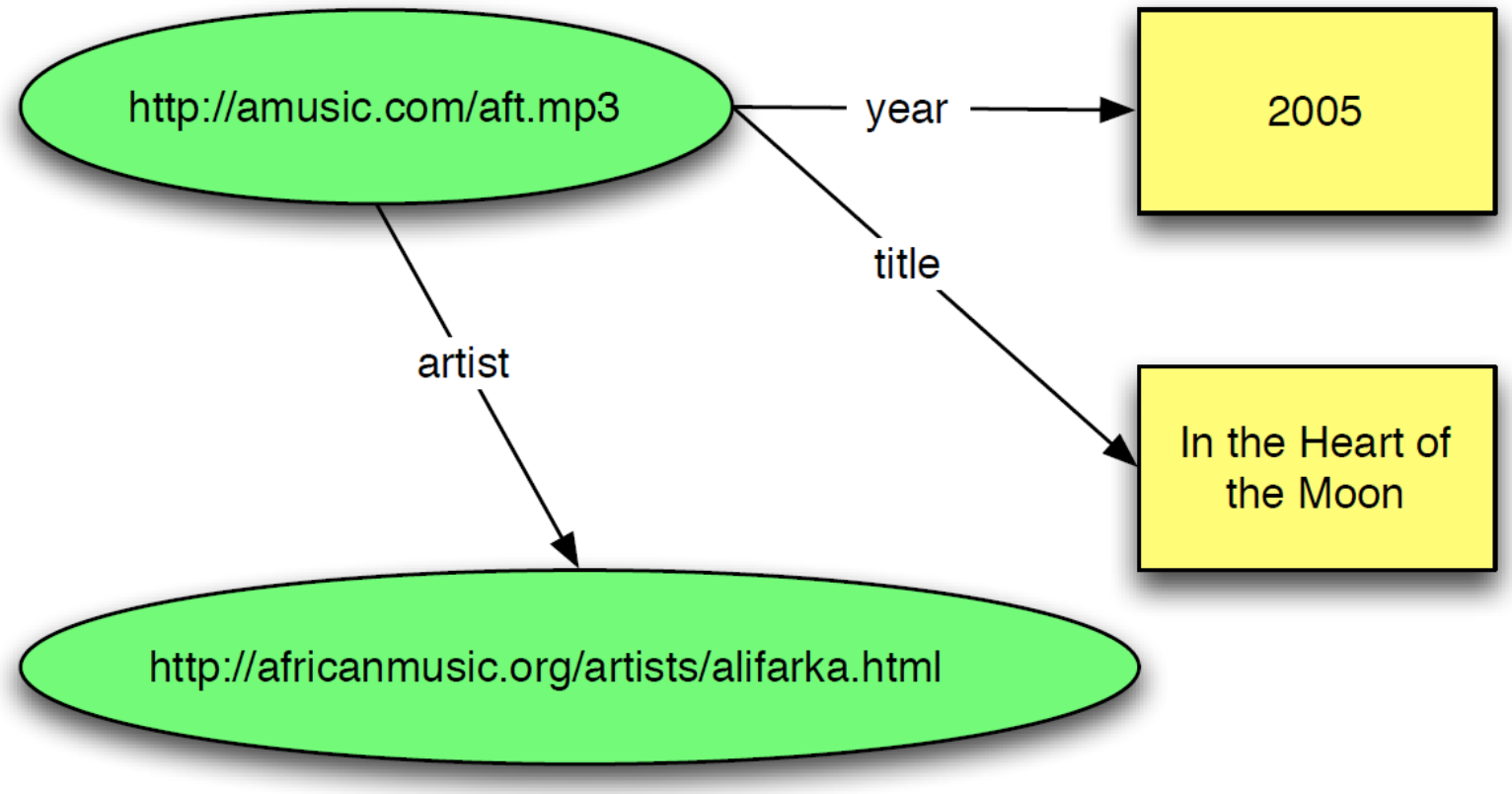
RDF Background

- 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


http://amusic.com/aft.mp3
year
2005
title
artist
In the Heart of the Moon
http://africanmusic.org/artists/alifarka.html

Syntax: Dublin Core vs 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

subject	predicate	object
dbpedia:In_The_Heart_of_the_Moon	dc:date	“2005” .

- Subjects can only be resources.
- Objects can be literals (e.g. strings) or 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://example.com/my/fictitious/example
3. ftp://ftp.is.co.za/rfc/rfc1808.txt
4. mailto:JohnDoe@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.

See also: http://www.ltg.ed.ac.uk/~ht/WhatAreURIs

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 standardisation effort (started in 1994, DC Metadata Element Set, 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: more than 1.1 billion people share just 129 surnames (cf. 'Identity Crisis' paper, referenced 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?