Dottorato di Ricerca

Le Digital Humanities: aspetti metodologici e pratici

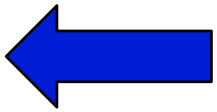
Enrica Salvatori (enrica.salvatori@unipi.it)
Vittore Casarosa (casarosa@isti.cnr.it)

Pisa, 28 Marzo 2019

Dottorato di Ricerca

Refresher on Computer Fundamentals and Data Representation

- Brief History of computers
- Architecture of a computer
- Data representation within a computer
- Metadata



Description of “content”

- In the libraries: bibliographic records
  - Classification and cataloguing
  - MARC standard

- In the Web: metadata
  - Resources and metadata
  - The Dublin Core metadata schema

Emerging new requirements (early 90's)

- Increase in the amount of information available on-line (data bases, repositories, the Web, etc)
- Increase in the variety of information available on-line (text, sound, images, video, 3D, etc)
- Scholarly publishing (open access and non-open access)
  - Self-publishing in **Institutional Repositories**
- Need to describe (in some way) the "content" of the Web
  - Description of information not always done by "specialists"
- Description of the content of the Web done through **metadata**

Some definitions of Metadata

- Machine-understandable information about Web resources or other things (Tim Berners-Lee 1997)
- Data associated with objects which relieves their potential users of having to have full advance knowledge of their existence or characteristics; a user might be a program or a person (Lorcan Dempsey 1998)
- Structured data about resources that can be used to help support a wide range of operations (Michael Day, 2001)
- Structured data about data (DCMI 2003)
- Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource (NISO 2004)

Resources in the Web

- The “objects” of interest in the Web are generically called **resources**
- A resource is anything that has identity. For example, a resource may be an electronic document, an image, a service (e.g., "today's weather report for Pisa"), a collection of other resources
- Not all resources are network “downloadable"; e.g., human beings, corporations, and bound books in a library can also be considered resources

The Internet generation

"No, you weren't downloaded.
Your were born."

Resources in the Web

- The “objects” of interest in the Web are generically called **resources**
- A resource is anything that has identity. For example, a resource may be an electronic document, an image, a service (e.g., "today's weather report for Pisa"), a collection of other resources
- Not all resources are network “downloadable"; e.g., human beings, corporations, and bound books in a library can also be considered resources
- All resources are identified by a URI (Uniform Resource Identifier)

Several types of URIs
URI Syntax

<scheme name> : <hierarchical part> [? <query>] [# <fragment>]

any://example.com:8042/over/there?name=ferret#nose
___/ _____________/ _____/ _____/ __/
scheme authority path query fragment

- ftp://ftp.is.co.za/rfc/rfc1808.txt
- **http://www.ietf.org/rfc/rfc2396.txt**
- ldap://[2001:db8::7]/c=GB?objectClass?one
- mailto:John.Doe@example.com
- news:comp.infosystems.www.servers.unix
- tel:+1-816-555-1212
- telnet://192.0.2.16:80/
- urn:oasis:names:specification:docbook:dtd:xml:4.1.2bb

Resources and metadata

Metadata can be associated with any resource: physical, digital, abstract resource, etc.

- HTML documents
- digital images
- databases
- books
- museum objects
- archival records
- metadata records
- Web sites
- collections
- services
- physical places
- people
- institutions
- abstract “works”
- concepts
- events

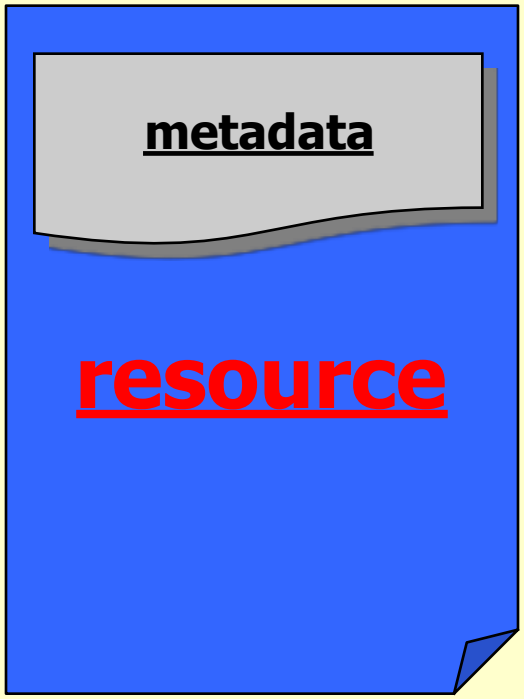
Not simply a cataloguing record

- An important reason for creating descriptive metadata is to facilitate discovery of relevant information, as it serves the same functions in resource discovery as good cataloging does by:
  - allowing resources to be found by relevant criteria
  - identifying resources
  - bringing similar resources together
  - distinguishing dissimilar resources
  - giving location information

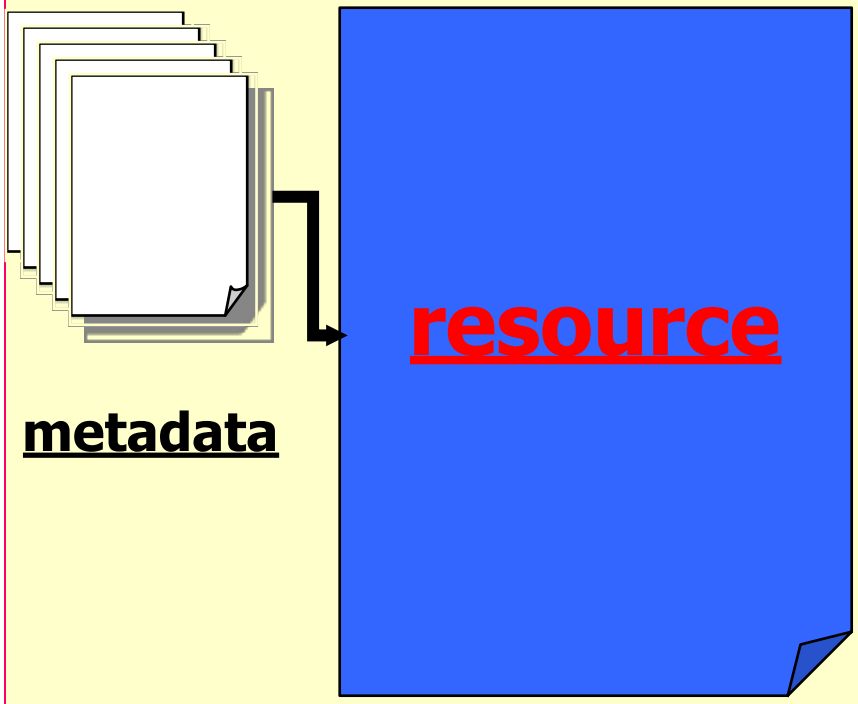
- In addition to resource discovery, metadata can
  - help organize electronic resources
  - facilitate interoperability and legacy resource integration
  - provide digital identification
  - support archiving and preservation

Storing Metadata

embedded metadata



stand-alone metadata



Storing metadata

- Metadata can be embedded in a digital object or it can be stored separately. Metadata is often embedded in HTML documents and in the headers of image files
- Storing metadata with the object it describes ensures the metadata will not be lost, obviates problems of linking between data and metadata, and helps ensure that the metadata and object will be updated together
- However, it is impossible to embed metadata in some types of objects (for example, artifacts). Also, storing metadata separately can simplify the management of the metadata itself and facilitate search and retrieval. Therefore, metadata is commonly stored in a database system and linked to the objects described

Possible types of metadata

- *Descriptive metadata* describe a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords
- *Structural metadata* indicate how compound objects are put together, for example, how pages are ordered to form chapters
- *Technical metadata* indicate hardware or software used in converting an item/object to a digital format, or in storing or displaying it
- *Administrative metadata* provide information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. Two subsets of administrative data are:
  - *Rights management metadata,* which deal with intellectual property rights
  - *Preservation metadata,* which contain information needed to archive and preserve a resource

Dublin Core Metadata Initiative

- Dublin Core
  - Dublin: Dublin, Ohio, 1995
  - Core: minimal set of broad and generic elements
- Dublin Core was originally developed with an eye to describing document-like objects
  - Descriptions easy to create (unlike MARC)
- Despite initial focus, has proved to be general enough to describe “any” type of objects
  - unlike catalog records, often tied to specific application fields
- It is now a widely used international standard
  - ISO Standard 15836-2003
  - NISO Standard Z39.85-2007
  - IETF RFC 5013

IETF, RFC and W3C

- IETF: Internet Engineering Task Force
  - Groups of experts proposing and defining new technologies and applications in the web, that might become “Internet standards”
  - “We believe in running code and rough consensus”
- RFC: Request For Comments
  - Publications numbered sequentially starting with 1 (RFC 1 in 1969, RFC 8005 in Oct. 2016)
  - The way to publish and define the standards in the web
- W3C: World Wide Web Consortium
  - Consortium of industries and universities to ensure compatibility in the adoption of new standards
  - Issues recommendations (W3C standards) and certifications

Fifteen Core Elements (1996)

Definition of **elements** (or **terms**) to describe **resources**

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

Characteristics of the Dublin Core

- All elements optional
- All elements repeatable
- Elements may be displayed in any order
- International in scope
- Extensible (Qualified Dublin Core)
- Dublin Core Principles
  - Dumb-Down
  - One-to-One
  - Appropriate Values

The “Dumb-Down” principle

- The fifteen core elements are usable with or without qualifiers
  - Qualifiers make elements more specific:
  - Element Refinements narrow meanings, never extend
  - Encoding Schemes give context to element values
- If your software encounters an unfamiliar qualifier, look it up –or **just ignore it!**

The “One-to-One” principle

- Describe one manifestation of a resource with one record
  - Example: a digital image of the Mona Lisa is not described as if it were the same as the original painting
- Separate descriptions of resources from descriptions of the agents responsible for those resources
  - Example: email addresses and affiliations of creators are attributes of the creator, not the resource
- If needed, group related descriptions into a “description set” (record)

The “Appropriate Values” priciple

- Use elements, element refinements and qualifiers to meet the needs of your local context, but. . .
- Remember that your metadata may be interpreted by machines and people, so. . .
- Consider whether the values you use will aid discovery outside your local context and. . .
- Make decisions about your local practices accordingly

Simple and Qualified Dublin Core

- Simple Dublin Core is limited to the original 15 elements
- Qualified Dublin Core includes, in addition:
  - New Elements
  - Qualifiers
    - Element Refinements
    - Encoding Schemes
      - Syntax Encoding Scheme
      - Vocabulary Encoding Scheme

That’s the end, folks

Many thanks for your attention (and your endurance)

Vittore Casarosa
casarosa@isti.cnr.it