Metadata

hussein suleman

uct cs honours 2009

Data vs. Metadata

- Data refers to digital objects that contain useful information for information seekers.
- Metadata refers to standardised descriptions of objects, digital or physical.
- Many systems manipulate metadata records, which contain pointers to the actual data.
- The definition is fuzzy as metadata contains useful information as well and in some cases could contain all the data e.g., metadata describing a person.

An Example of Metadata

- Object:

- Metadata
  - name: chalk
  - owner: hussein
  - colour: white
  - size: 2.5
  - description: used to write on board
  - location: honours lecture room
  - source: Waltons Stationers

Another Metadata Example

- Object:

- Metadata
  - colour: white
  - title: RG123
  - owner: UCT
  - lifetime: 2 months
  - size: 1
  - identifier: RG123
  - description: white powdery stick

Metadata Comparisons

- Metadata
  - colour: white
  - title: RG123
  - owner: UCT
  - lifetime: 2 months
  - size: 1
  - identifier: RG123
  - description: white powdery stick

- Metadata
  - name: chalk
  - owner: hussein
  - colour: white
  - size: 2.5
  - description: used to write on board
  - location: honours lecture room
  - source: Waltons Stationers

What problems can occur?

Types of Metadata

- Descriptive
  - title, author, type, format, …
- Structural
  - part, subpart, relation, child, …
- Administrative
  - location, identifier, submitter, …
- Preservation
  - resolution, capture device, watermark, …
- Provenance
  - source archive, previous version, source format, …

Creating Metadata

- Follow metadata guidelines.
- Use terms from controlled vocabularies.
- Avoid duplication of information across fields.
- Use accepted standards for common elements.
  - e.g., ISO 8601 for dates
    - 2005-03-03 instead of 03/03/05
- Use XML-based encoding according to standardised Schema/DTD.

Dublin Core

- Dublin Core is one of the most popular and simplest metadata formats.
- 15 elements with recommended semantics.
- All elements are optional and repeatable.

Title	Creator	Subject
Description	Publisher	Contributor
Date	Type	Format
Identifier	Source	Language
Relation	Coverage	Rights

DC in HTML

- <META NAME=DC.Creator CONTENT="Tony Gill">
- <META NAME=DC.Title CONTENT="ADAM Quick Guide to Metadata">
- <META NAME=DC.Subject CONTENT="ADAM, Dublin Core, internet cataloguing, metadata">
- <META NAME=DC.Description CONTENT="A short ADAM guide to metadata, particularly Dublin Core.">
- <META NAME=DC.Date CONTENT="1997-11-21">

Source: http://adam.ac.uk/adam/metadata.html

DC Metadata in XML

```
<title>02uct1</title>
<creator>Hussein Suleman</creator>
<subject>Visit to UCT </subject>
<description>the view that greets you as you emerge from the tunnel under the freeway - WOW - and, no, the mountain isnt that close - it just looks that way in 2-D</description>
<publisher>Hussein Suleman</publisher>
<date>2002-11-27</date>
<type>image</type>
<format>image/jpeg</format>
```

DC Metadata in Valid Qualified XML

```
<oaidc:dc xmlns="http://purl.org/dc/elements/1.1/"
   xmlns:oaidc="http://www.openarchives.org/OAI/2.0/oai_dc/"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
   http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
   <title>02uct1</title>
   <creator>Hussein Suleman</creator>
   <subject>Visit to UCT </subject>
   <description>the view that greets you as you emerge from the tunnel
   under the freeway - WOW - and, no, the mountain isnt that close - it
   just looks that way in 2-D</description>
   <publisher>Hussein Suleman</publisher>
   <date>2002-11-27</date>
   <type>image</type>
   <format>image/jpeg</format>
   <identifier>http://www.husseinsspace.com/pictures/200230uct/02uct1.jpg
   </identifier>
   <language>en-us</language>
   <relation>http://www.husseinsspace.com</relation>
   <rights>unrestricted</rights>
</oaidc:dc>
```

DC Qualifiers

- Dublin Core has been considered TOO simple for many applications – not enough semantics.
- Some DC terms have had qualifiers added to make the meaning more specific.
  - For example,
    - date.created instead of just date
    - relation.hasPart instead of just relation
- In general, qDC can be dumbed-down (*that's a technical term in interoperability*) to DC by ignoring qualifications.

What Metadata Format?

- Do NOT just use Dublin Core all the time!
- Every project has its own metadata/data requirements, therefore most use an internal format.
- For maximum interoperability,
  - Map metadata to most descriptive format for use by close collaborators.
  - Map metadata to DC for use by all and sundry.
- How do we “map” metadata formats?

Metadata Transformation

- Use XML parser to parse data.
- Use SAX/DOM to extract individual elements and generate new format.
- Example (to convert UCT to DC):

  -
    ```
    my $parser = new DOMParser;
    my $document = $parser->parsefile (‘uct.xml’)->getDocumentElement;
    foreach my $title ($document->getElementsByTagName (‘title’))
    {
       print “<title>”.$title->getFirstChild->getData.”</title>\n”;
    }
    foreach my $author ($document->getElementsByTagName (‘author’))
    {
       print “<creator>”.$author->getFirstChild->getData.”</creator>\n”;
    }
    print “<publisher>UCT</publisher>\n”;
    foreach my $version ($document->getElementsByTagName (‘version’))
    {
       foreach my $number ($version->getElementsByTagName (‘number’))
       {
          print “<identifier>”.
                $number->getFirstChild->getData.”</identifier>\n”;
       }
    }
    ```

- There must be an easier way …

Metadata Transformation (XSLT) 1/2

```
<stylesheet version='1.0'
 xmlns='http://www.w3.org/1999/XSL/Transform'
 xmlns:oaidc='http://www.openarchives.org/OAI/2.0/oai_dc/'
 xmlns:dc='http://purl.org/dc/elements/1.1/'
 xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance'
 xmlns:uct='http://www.uct.ac.za'
>

<!--
   UCT to DC transformation
   Hussein Suleman
   v1.0 : 24 July 2003
-->

   <output method="xml"/>

   <variable name="institution"><text>UCT</text></variable>
```

Metadata Transformation (XSLT) 2/2

```
   <template match="uct:uct">
      <oaidc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
                http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
         <dc:title><value-of select="uct:title"/></dc:title>
         <apply-templates select="uct:author"/>
         <element name="dc:publisher">
            <value-of select="$institution"/>
         </element>
         <apply-templates select="uct:version"/>
      </oaidc:dc>
   </template>

   <template match="uct:author">
      <dc:creator>
         <value-of select="."/>
      </dc:creator>
   </template>

   <template match="uct:version">
      <dc:identifier>
         <value-of select="uct:number"/>
      </dc:identifier>
   </template>

</stylesheet>
```

- Ok, but map to what?

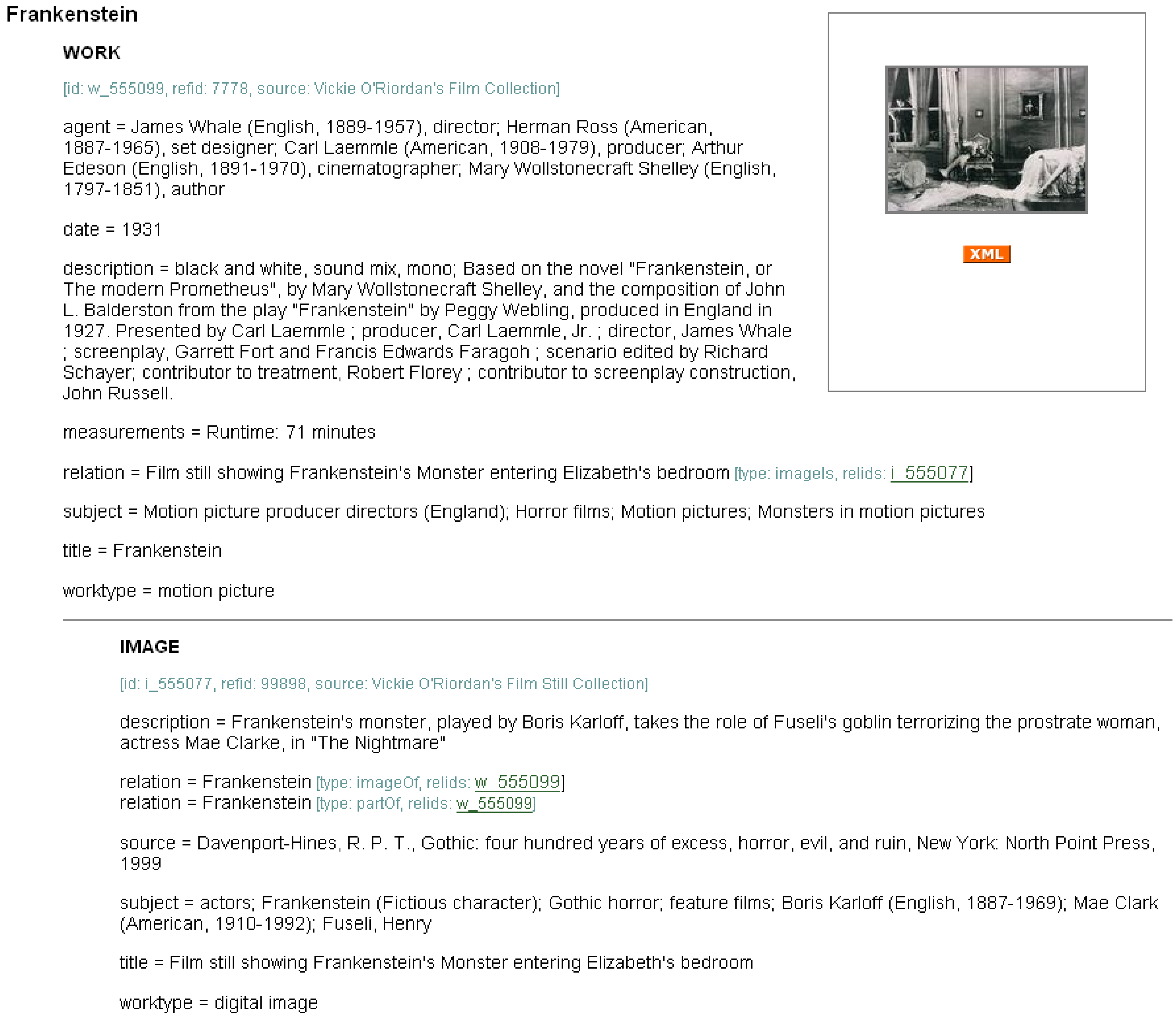
Example: RFC1807

```
<rfc1807 xmlns="http://info.internet.isi.edu:80/in-notes/rfc/files/rfc1807.txt“
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance“
      xsi:schemaLocation="http://info.internet.isi.edu:80/in-notes/rfc/files/rfc1807.txt
      http://www.openarchives.org/OAI/1.1/rfc1807.xsd">
        <bib-version >1</bib-version>
        <id >395</id>
        <entry >2007-01-01</entry>
        <organization >University of Cape Town - Department of Computer Science</organization>
        <title >Using Payment Gateways to Maintain Privacy in Secure Electronic Transactions</title>
        <type >Conference Paper</type>
        <author >Arnab, Alapan</author>
        <author >Hutchison, Andrew</author>
        <other_access >url:http://pubs.cs.uct.ac.za/archive/00000395/</other_access>
        <language >en</language>
        <abstract >Because many current payment systems are poorly implemented, or of incompetence,
             private data of consumers such as payment details, addresses and their purchase history can
             be compromised. Furthermore, current payment systems do not offer any non-repudiable
             verification to a completed transaction, which poses risks to all the parties of the transaction --
             the consumer, the merchant and the financial institution. One solution to this problem was
             SET, but it was never really a success because of its complexity and poor reception from
             consumers. In this paper, we introduce a third party payment system that aims to preserve
             privacy by severing the link between their purchase and payment records, while providing a
             traceable transaction that maintains its integrity and is non-repudiable. Our system also
             removes much of the responsibilities placed on the merchant with regards to securing
             sensitive data related to customer payment, thus increasing the potential of small businesses
             to take part in e-commerce without significant investments in computer security.</abstract>
</rfc1807>
```

Example: MARC

- <marc:record xmlns:marc="http://www.loc.gov/MARC21/slim" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.loc.gov/MARC21/slim http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd" type="Bibliographic">
- <marc:leader>00000coc 2200000uu 4500</marc:leader>
- <marc:controlfield tag="001">444652</marc:controlfield>
- <marc:controlfield tag="003">SzGeCERN</marc:controlfield>
- <marc:datafield tag="035" ind1="" ind2="">
  - <marc:subfield code="9">CERCER</marc:subfield>
  - <marc:subfield code="a">2195309</marc:subfield>
- </marc:datafield>
- <marc:datafield tag="037" ind1="" ind2="">
  - <marc:subfield code="a">nlin.SI/0006040</marc:subfield>
- </marc:datafield>
- <marc:datafield tag="041" ind1="" ind2="">
  - <marc:subfield code="a">eng</marc:subfield>
- </marc:datafield>
- <marc:datafield tag="088" ind1="" ind2="">
  - <marc:subfield code="9">nlin/0006040</marc:subfield>
- </marc:datafield>
- <marc:datafield tag="100" ind1="" ind2="">
  - <marc:subfield code="a">Bibikov, P N</marc:subfield>
- </marc:datafield>
- <marc:datafield tag="245" ind1="" ind2="">
  - <marc:subfield code="a">Derivation of R-matrix from local Hamiltonian density</marc:subfield>
- </marc:datafield>
- <marc:datafield tag="260" ind1="" ind2="">
  - <marc:subfield code="c">2000</marc:subfield>
- </marc:datafield>
- <marc:datafield tag="269" ind1="" ind2="">
  - <marc:subfield code="a">Geneva</marc:subfield>
  - <marc:subfield code="b">CERN</marc:subfield>
  - <marc:subfield code="c">23 Jun 2000</marc:subfield>
- </marc:datafield>
- <marc:datafield tag="300" ind1="" ind2=""> <marc:subfield code="a">18 p</marc:subfield> </marc:datafield> <marc:datafield tag="520" ind1="" ind2=""> <marc:subfield code="a">A computer algebra algoritm for solving the quantum Yang-Baxter equation is presented. It is based on the Taylor expansion of R-matrix which is developed up to the order \lambda^6. As an example the classification of 4x4 R-matrices is given.</marc:subfield> </marc:datafield> <marc:datafield tag="595" ind1="" ind2=""> <marc:subfield code="a">LANL EDS</marc:subfield> </marc:datafield> <marc:datafield tag="650" ind1="1" ind2="7"> <marc:subfield code="2">SzGeCERN</marc:subfield> <marc:subfield code="a">Nonlinear Systems</marc:subfield> </marc:datafield> <marc:datafield tag="690" ind1="C" ind2=""> <marc:subfield code="a">PREPRINT</marc:subfield> </marc:datafield> <marc:datafield tag="690" ind1="C" ind2=""> <marc:subfield code="a">CERN</marc:subfield> </marc:datafield> <marc:datafield tag="695" ind1="" ind2=""> <marc:subfield code="9">LANL EDS</marc:subfield> <marc:subfield code="a">Exactly Solvable and Integrable Systems</marc:subfield> </marc:datafield> <marc:datafield tag="695" ind1="" ind2=""> <marc:subfield code="9">LANL EDS</marc:subfield> <marc:subfield code="a">Nonlinear Sciences</marc:subfield> </marc:datafield> <marc:datafield tag="710" ind1="" ind2=""> <marc:subfield code="5">CERN?</marc:subfield> </marc:datafield> <marc:datafield tag="856" ind1="4" ind2=""> <marc:subfield code="u">http://documents.cern.ch/cgi-bin/setlink?base=preprint&amp;categ=nlin.SI&amp;id=0006040</marc:subfield> <marc:subfield code="y">Access to fulltext document</marc:subfield> </marc:datafield> <marc:datafield tag="859" ind1="" ind2=""> <marc:subfield code="f">sergey.paston@pobox.spbu.ru>David Dallman &lt;David.Dallman@cern.ch</marc:subfield> </marc:datafield> <marc:datafield tag="916" ind1="" ind2=""> <marc:subfield code="s">n</marc:subfield> <marc:subfield code="w">200026</marc:subfield> </marc:datafield> <marc:datafield tag="960" ind1="" ind2=""> <marc:subfield code="a">11</marc:subfield> </marc:datafield> <marc:datafield tag="961" ind1="" ind2=""> <marc:subfield code="c">20060112</marc:subfield> <marc:subfield code="h">1202</marc:subfield> <marc:subfield code="l">CER01</marc:subfield> <marc:subfield code="x">20000626</marc:subfield> </marc:datafield> <marc:datafield tag="963" ind1="" ind2=""> <marc:subfield code="a">PUBLIC</marc:subfield> </marc:datafield> <marc:datafield tag="970" ind1="" ind2=""> <marc:subfield code="a">002195309CER</marc:subfield> </marc:datafield> <marc:datafield tag="980" ind1="" ind2=""> <marc:subfield code="a">PREPRINT</marc:subfield> </marc:datafield>
- </marc:record>

Example: VRA-Core



Frankenstein

WORK

[id: w_555099, refid: 7778, source: Vickie O'Riordan's Film Collection]

agent = James Whale (English, 1889-1957), director; Herman Ross (American, 1887-1965), set designer; Carl Laemmle (American, 1908-1979), producer; Arthur Edeson (English, 1891-1970), cinematographer; Mary Wollstonecraft Shelley (English, 1797-1851), author

date = 1931

description = black and white, sound mix, mono; Based on the novel "Frankenstein, or The modern Prometheus", by Mary Wollstonecraft Shelley, and the composition of John L. Balderston from the play "Frankenstein" by Peggy Webling, produced in England in 1927. Presented by Carl Laemmle ; producer, Carl Laemmle, Jr. ; director, James Whale ; screenplay, Garrett Fort and Francis Edwards Faragoh ; scenario edited by Richard Schayer; contributor to treatment, Robert Florey ; contributor to screenplay construction, John Russell.

measurements = Runtime: 71 minutes

relation = Film still showing Frankenstein's Monster entering Elizabeth's bedroom [type: imageIs, relids: i_555077]

subject = Motion picture producer directors (England); Horror films; Motion pictures; Monsters in motion pictures

title = Frankenstein

worktype = motion picture

IMAGE

[id: i_555077, refid: 99898, source: Vickie O'Riordan's Film Still Collection]

description = Frankenstein's monster, played by Boris Karloff, takes the role of Fuseli's goblin terrorizing the prostrate woman, actress Mae Clarke, in "The Nightmare"

relation = Frankenstein [type: imageOf, relids: w_555099]
relation = Frankenstein [type: partOf, relids: w_555099]

source = Davenport-Hines, R. P. T., Gothic: four hundred years of excess, horror, evil, and ruin, New York: North Point Press, 1999

subject = actors; Frankenstein (Fictious character); Gothic horror; feature films; Boris Karloff (English, 1887-1969); Mae Clark (American, 1910-1992); Fuseli, Henry

title = Film still showing Frankenstein's Monster entering Elizabeth's bedroom

worktype = digital image

Other Metadata Standards

- IMS Metadata Specification
  - Courseware object description.
- EAD
  - Library finding aids to locate archived items.
- METS
  - Descriptive, administrative and structural encoding for metadata of digital objects
- MODS
  - Richer than DC, subset of MARC21
- MPEG21-DIDL
  - Structural descriptions of complex multimedia objects

Automatic Metadata Extraction

- Create metadata automatically from a digital object.
- Embedded Metadata
  - e.g., MP3 tags
- Heuristic Techniques
  - e.g., The first string that looks like a date is the date of publication
- Machine Learning
  - e.g., Neural networks
- Dictionary Techniques
  - e.g., If it looks like a name, it could be an author

References

- Dublin Core Metadata Initiative (2005). DCMI Metadata Terms. Available http://dublincore.org/documents/dcmi-terms/
- Dublin Core Metadata Initiative (2004). Dublin Core Metadata Element Set, Version 1.1: Reference Description. Available http://dublincore.org/documents/dces/
- IMS Global Learning Consortium, Inc. (2001). IMS Learning Resource Meta-Data Information Model, Version 1.2.1 Final Specification. Available http://www.imsglobal.org/metadata/imsmdv1p2p1/imsmd_infov1p2p1.html
- Lasher, R. and D. Cohen (1995). A Format for Bibliographic Records. Network Working Group, RFC1807. Available http://www.ietf.org/rfc/rfc1807.txt
- Library of Congress (2002). Encoded Archival Description (EAD), Official EAD Version 2002 Web Site. Website http://www.loc.gov/ead/
- Library of Congress (2005). MARC Standards. Website http://www.loc.gov/marc/
- Library of Congress (2005). Metadata Encoding and Transmission Standard. Website http://www.loc.gov/standards/mets/
- Library of Congress (2005). Metadata Object Description Schema. Website http://www.loc.gov/standards/mods/
- Visual Resources Association Data Standards Committee. (2007). VRA Core 4.0. Available http://www.vraweb.org/projects/vracore4/index.html
- XML Cover Pages (2005). MPEG-21 Part 2: Digital Item Declaration Language (DIDL). Website http://xml.coverpages.org/mpeg21-didl.html