Resource Description: Cataloging & Metadata

CS 431 – February 21, 2007

Carl Lagoze – Cornell University

Acknowledgments

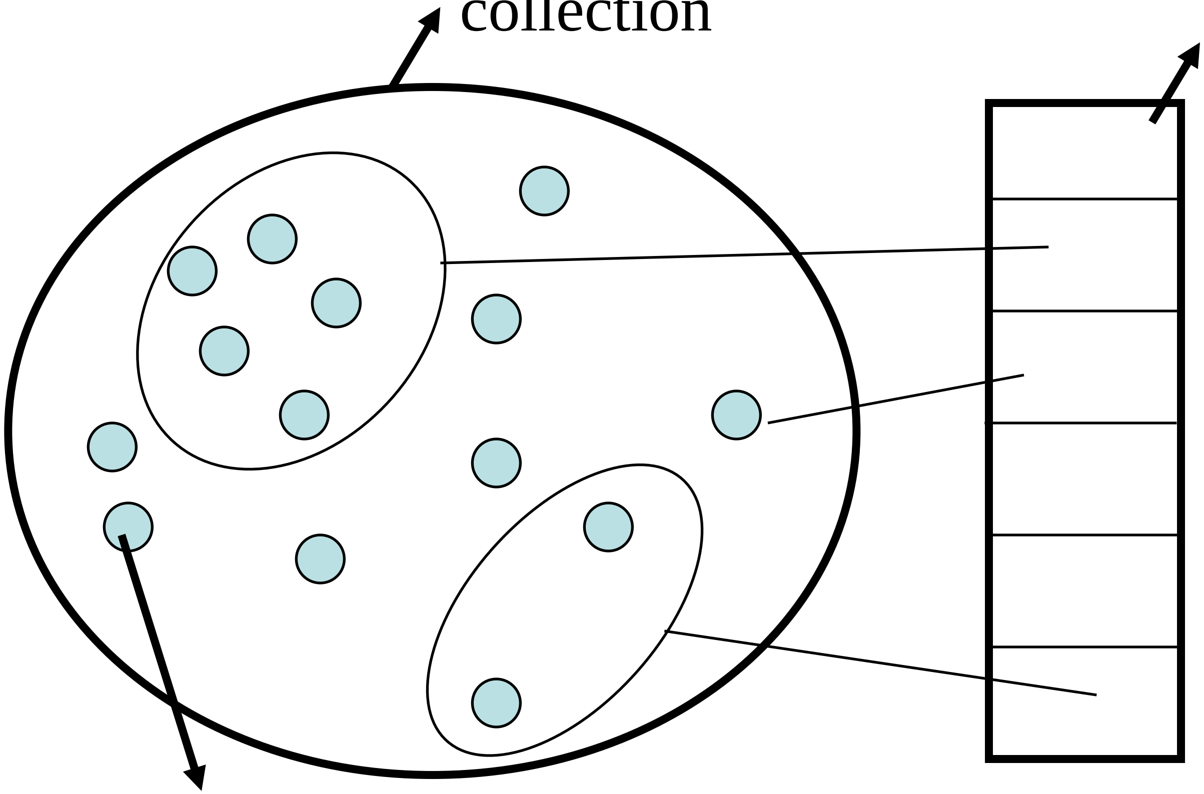
- Andy Powell, Head of Development, Eduserv Foundation, UK
- Tom Baker, Dublin Core Metadata Initiative
- Diane Hillmann, Cornell University
- Erik Wilde, UC Berkeley School of Information

A few points to contextualize this talk

- In parts of it you should forget that the web and Google exist
- In other parts you should be very skeptical about the need for resource descriptions
- In the end you should believe that resource description makes a lot of sense in some contexts

Bibliographic model

establishes equivalence classes to organize information objects for human understanding and management



collection

object = piece of content

Metadata surrogates:
- Identity
- Description
- Structure

NOTE: we are not just talking about describing single web resources (URI)

Objects are Related

IFLA Entity Model

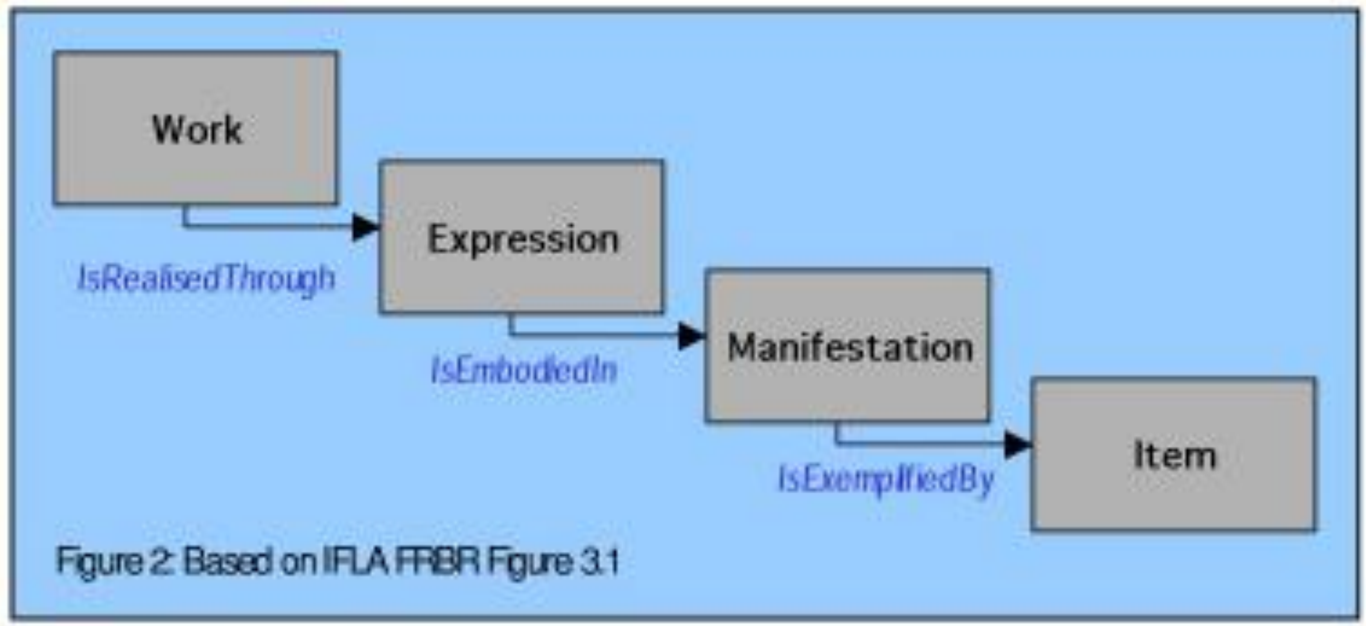


Figure 2: Based on IFLA FRBR Figure 3.1

Some attributes change over time while some change

Cataloging, Metadata, and Resource Description as Order Making

David Levy

Cataloging in the Digital Order

Traditional Library Cataloging

In the beginning…..

A Highly Standardized (interoperable) Process

- LC card distribution begins in 1890s
- AACR (Anglo-American Cataloging Rules) 1960’s – 1970’s, standardized rules for description
- MARC developed (by Henriette Avram) at LC in the 1960s
- OCLC (first bibliographic utility using MARC) in the early 1970s

Controlled Vocabularies

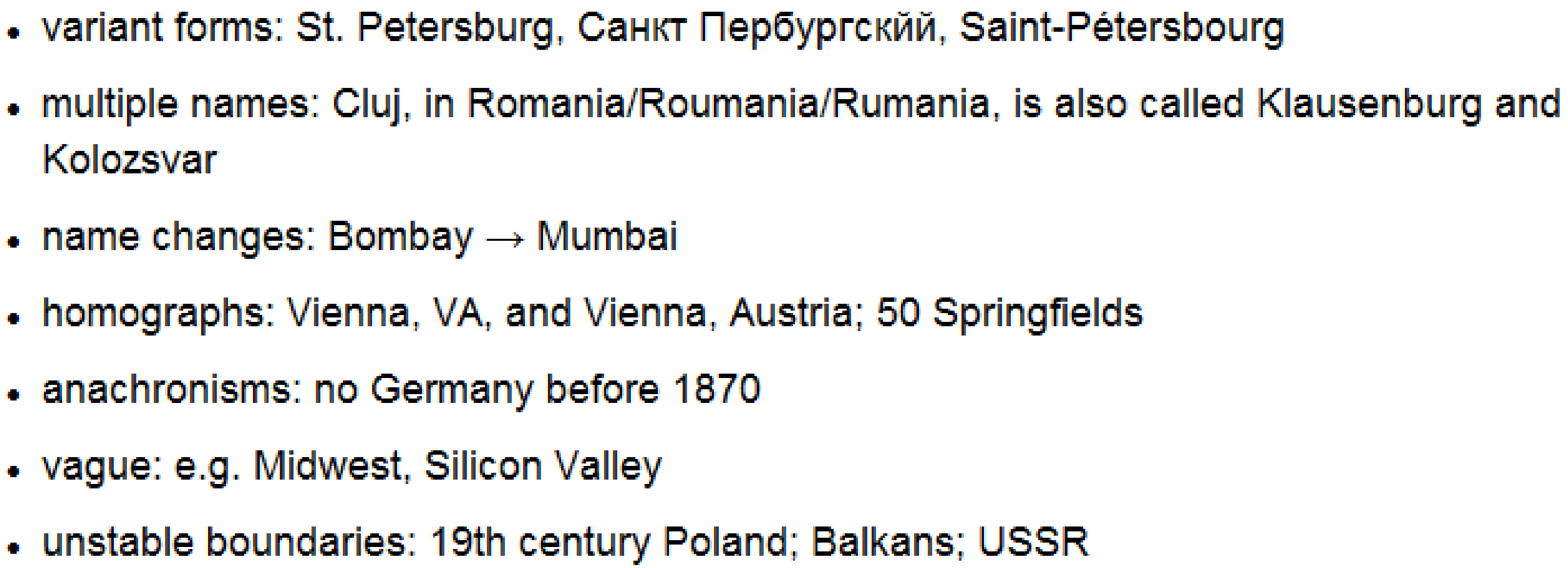
- A standardized set of terms assigned by organizers of information
- Goal is to impose some order in description within a domain
- Can be thought of as a fixed dictionary, artificial language, or vocabulary (cf. namespaces)
  - Names
  - Subject classifications

Problems with names and controlled vocabularies

- We want a label for some thing or category that is used to distinguish one from another
- A thing or category can have multiple names; there are synonyms or aliases
- Different things can sometimes have the same names
  - Homonyms have same syntax or pronunciation
  - Polysemes are words that have many meanings

Problems of Vocabulary Stability

- Places: One particularly troublesome area



- variant forms: St. Petersburg, Санкт Пербургскйй, Saint-Pétersbourg
- multiple names: Cluj, in Romania/Roumania/Rumania, is also called Klausenburg and Kolozsvar
- name changes: Bombay → Mumbai
- homographs: Vienna, VA, and Vienna, Austria; 50 Springfields
- anachronisms: no Germany before 1870
- vague: e.g. Midwest, Silicon Valley
- unstable boundaries: 19th century Poland; Balkans; USSR

From Erik Wilde

“Solution”: Authority Files

- Controlled vocabularies for names (author, corporate), titles, subjects
- Library of Congress
  - http://authorities.loc.gov/webvoy.htm
- OCLC Web Service
  - http://www.oclc.org/research/researchworks/authority/

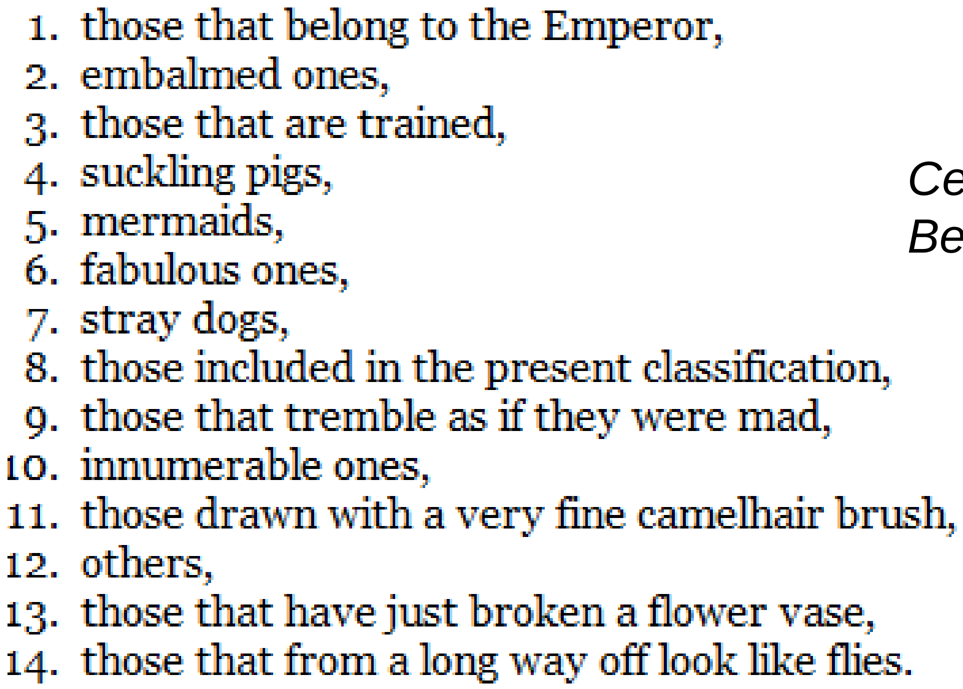
NOTE: Automatic name disambiguation is a VERY INTERESTING computer/information science problem

Dealing with Subjects: Classification

- Categories are equivalence classes
- Classifying is the process of assigning entities to the categories in a classification system
- Claassification performs a series of functions
  - Access points, relationships, browsing, retrieval
- Classification is arbitrary
  - Criteria for categorization reflects a perspective on reality.
  - Remember what Bates said about information

The *fiction* of classification

> …there is no classification of the universe that is not fictional and conjectural.
>
> Jorge Luis Borges



1. those that belong to the Emperor,
2. embalmed ones,
3. those that are trained,
4. suckling pigs,
5. mermaids,
6. fabulous ones,
7. stray dogs,
8. those included in the present classification,
9. those that tremble as if they were mad,
10. innumerable ones,
11. those drawn with a very fine camelhair brush,
12. others,
13. those that have just broken a flower vase,
14. those that from a long way off look like flies.

Celestial Emporium of Benevolent Knowledge

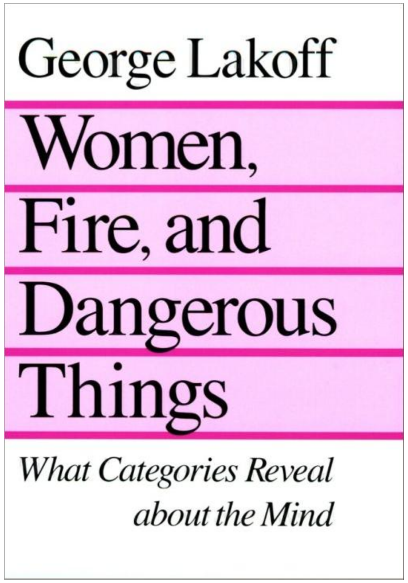
Classification is Problematic

- Historically loaded
  - Race names
  - Ordering
- The world changes
  - AIDS
- Ethno-centric

SORTING THINGS OUT

CLASSIFICATION AND ITS CONSEQUENCES

GEOFFREY C. BOWKER AND SUSAN LEIGH STAR

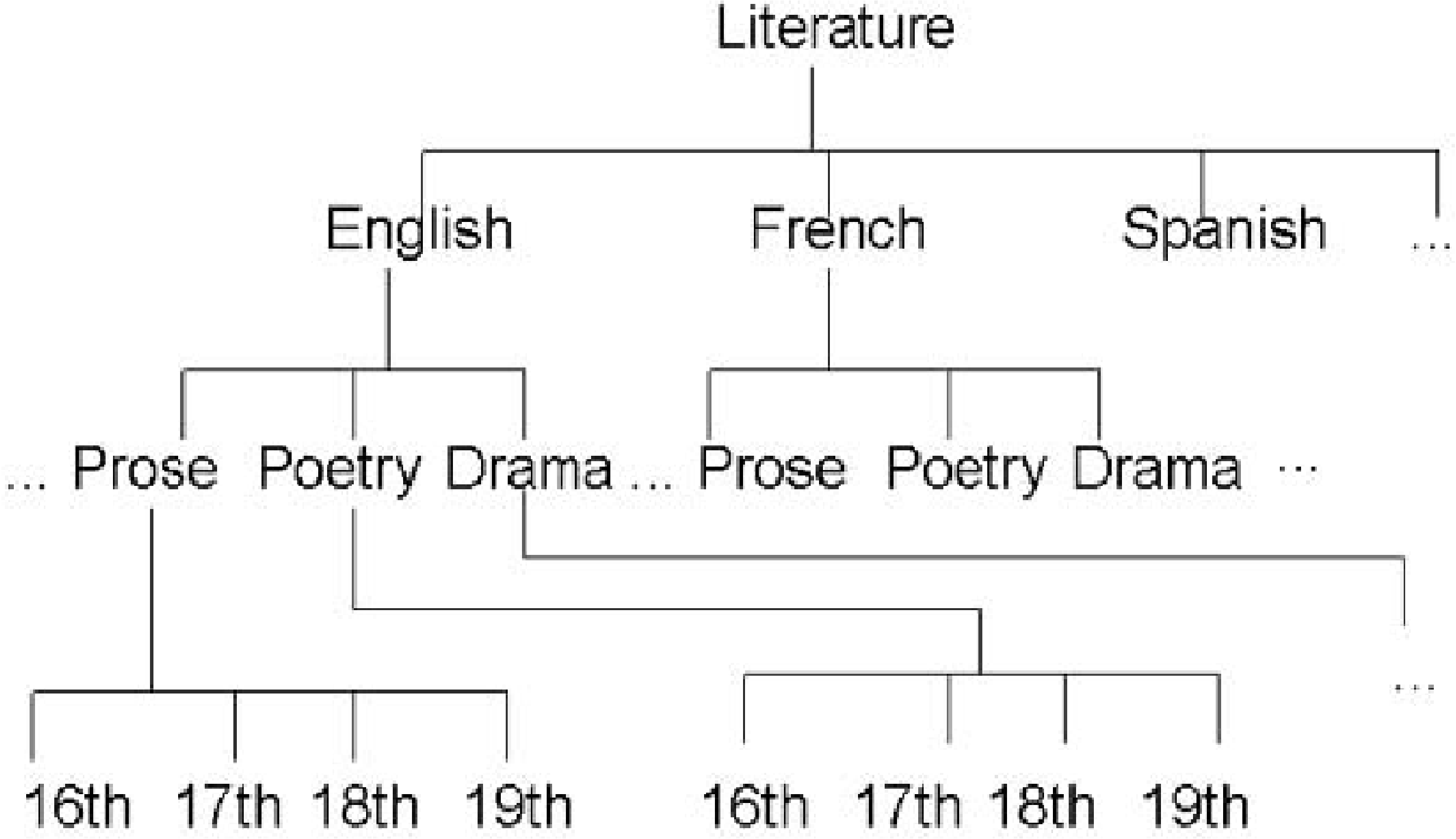


George Lakoff

Women, Fire, and Dangerous Things

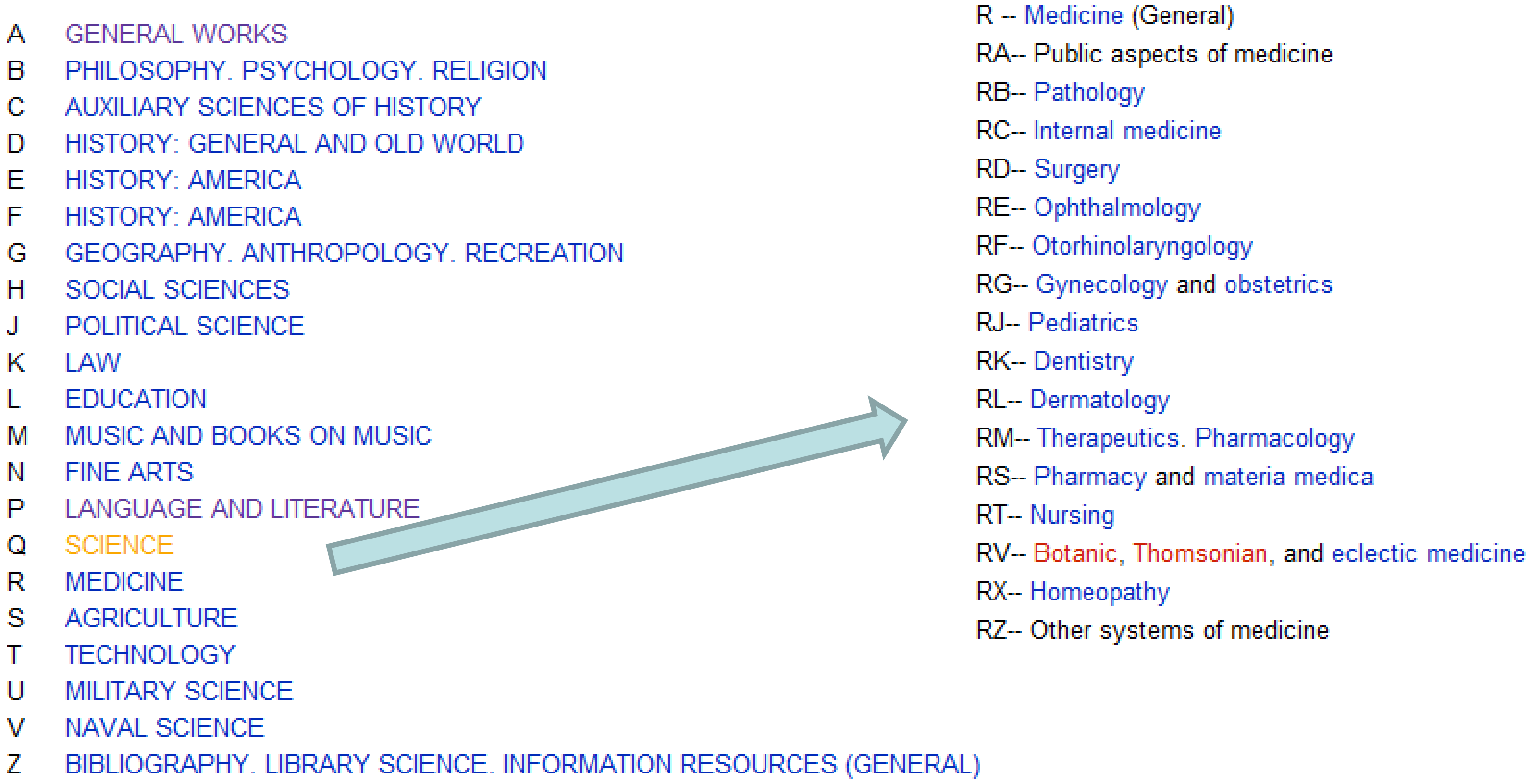
What Categories Reveal about the Mind

Hierarchical Classification



From Erik Wilde

Library of Congress Classification



A	GENERAL WORKS
B	PHILOSOPHY. PSYCHOLOGY. RELIGION
C	AUXILIARY SCIENCES OF HISTORY
D	HISTORY: GENERAL AND OLD WORLD
E	HISTORY: AMERICA
F	HISTORY: AMERICA
G	GEOGRAPHY. ANTHROPOLOGY. RECREATION
H	SOCIAL SCIENCES
J	POLITICAL SCIENCE
K	LAW
L	EDUCATION
M	MUSIC AND BOOKS ON MUSIC
N	FINE ARTS
P	LANGUAGE AND LITERATURE
Q	SCIENCE
R	MEDICINE
S	AGRICULTURE
T	TECHNOLOGY
U	MILITARY SCIENCE
V	NAVAL SCIENCE
Z	BIBLIOGRAPHY. LIBRARY SCIENCE. INFORMATION RESOURCES (GENERAL)

R -- Medicine (General)

RA-- Public aspects of medicine

RB-- Pathology

RC-- Internal medicine

RD-- Surgery

RE-- Ophthalmology

RF-- Otorhinolaryngology

RG-- Gynecology and obstetrics

RJ-- Pediatrics

RK-- Dentistry

RL-- Dermatology

RM-- Therapeutics. Pharmacology

RS-- Pharmacy and materia medica

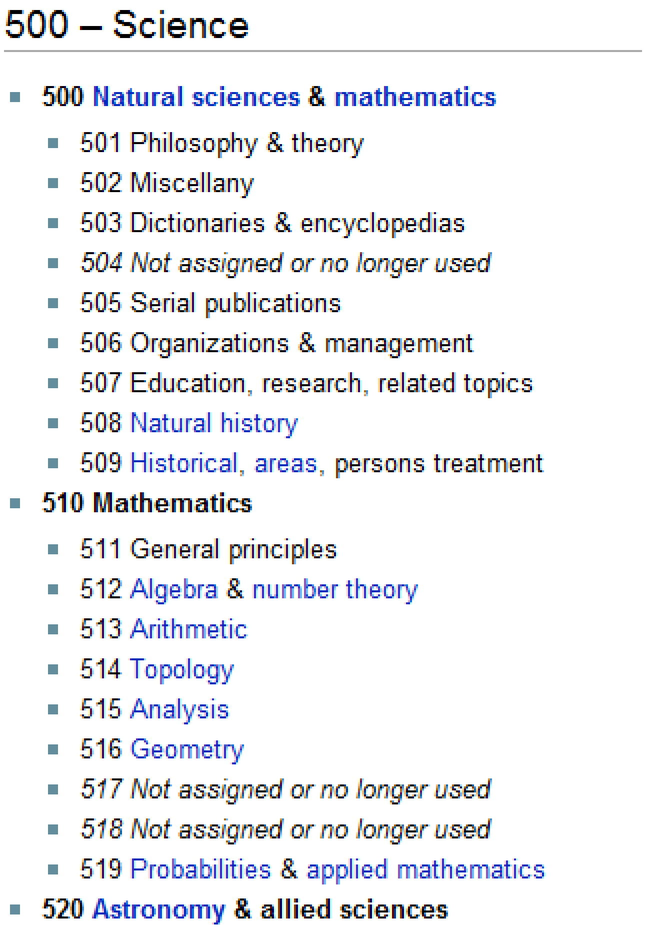
RT-- Nursing

RV-- Botanic, Thomsonian, and eclectic medicine

RX-- Homeopathy

RZ-- Other systems of medicine

Dewey Classification

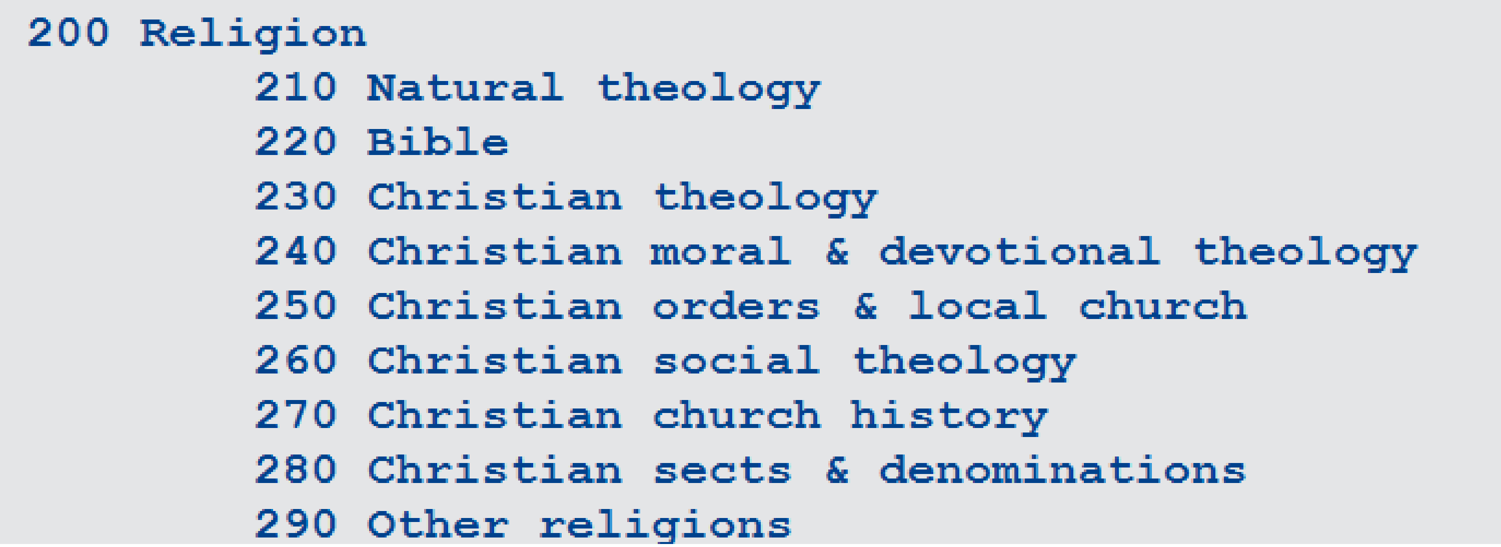


500 – Science

- **500 Natural sciences & mathematics**
  - 501 Philosophy & theory
  - 502 Miscellany
  - 503 Dictionaries & encyclopedias
  - *504 Not assigned or no longer used*
  - 505 Serial publications
  - 506 Organizations & management
  - 507 Education, research, related topics
  - 508 Natural history
  - 509 Historical, areas, persons treatment
- **510 Mathematics**
  - 511 General principles
  - 512 Algebra & number theory
  - 513 Arithmetic
  - 514 Topology
  - 515 Analysis
  - 516 Geometry
  - *517 Not assigned or no longer used*
  - *518 Not assigned or no longer used*
  - 519 Probabilities & applied mathematics
- **520 Astronomy & allied sciences**

From
Wikipedia

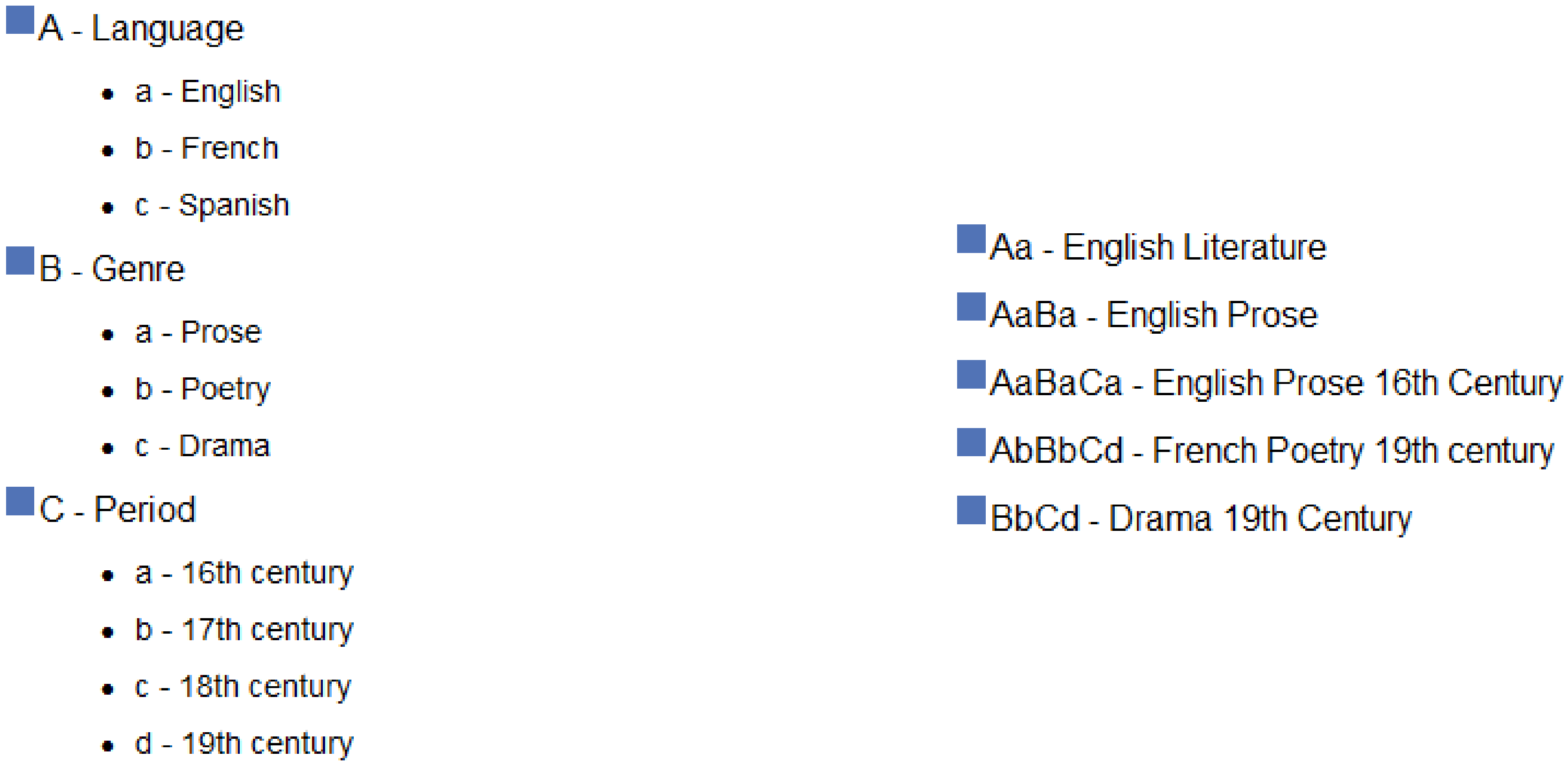
Bias in Dewey



```
200 Religion
        210 Natural theology
        220 Bible
        230 Christian theology
        240 Christian moral & devotional theology
        250 Christian orders & local church
        260 Christian social theology
        270 Christian church history
        280 Christian sects & denominations
        290 Other religions
```

From Erik Wilde

Faceted Classification



- A - Language
  - a - English
  - b - French
  - c - Spanish
- B - Genre
  - a - Prose
  - b - Poetry
  - c - Drama
- C - Period
  - a - 16th century
  - b - 17th century
  - c - 18th century
  - d - 19th century

- Aa - English Literature
- AaBa - English Prose
- AaBaCa - English Prose 16th Century
- AbBbCd - French Poetry 19th century
- BbCd - Drama 19th Century

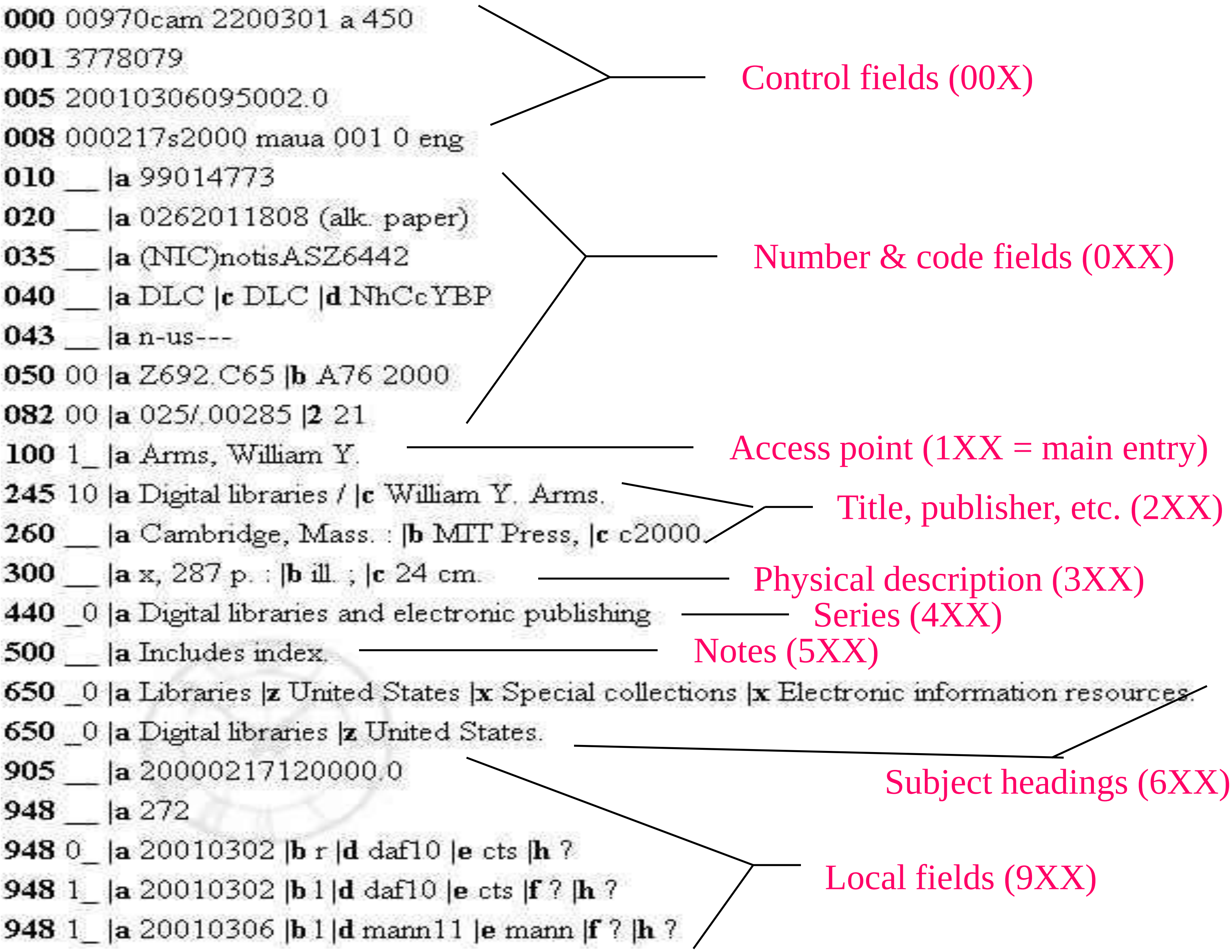
From Erik Wilde

Faceted Browsing

http://browse.guardian.co.uk/search

MARC

- Machine Readable Cataloging
- Bibliographic Types
  - Books
  - Serials
  - Maps
  - Visual materials
  - Sound recordings
  - Computer files
  - Archives and manuscripts
- Authority Records
- Holdings Records



```
000 00970cam 2200301 a 450
001 3778079
005 20010306095002.0
008 000217s2000 maua 001 0 eng
010 __ |a 99014773
020 __ |a 0262011808 (alk. paper)
035 __ |a (NIC)notisASZ6442
040 __ |a DLC |c DLC |d NhCcYBP
043 __ |a n-us---
050 00 |a Z692.C65 |b A76 2000
082 00 |a 025/.00285 |2 21
100 1_ |a Arms, William Y.
245 10 |a Digital libraries / |c William Y. Arms.
260 __ |a Cambridge, Mass. : |b MIT Press, |c c2000.
300 __ |a x, 287 p. : |b ill. ; |c 24 cm.
440 _0 |a Digital libraries and electronic publishing
500 __ |a Includes index.
650 _0 |a Libraries |z United States |x Special collections |x Electronic information resources.
650 _0 |a Digital libraries |z United States.
905 __ |a 20000217120000.0
948 __ |a 272
948 0_ |a 20010302 |b r |d daf10 |e cts |h ?
948 1_ |a 20010302 |b 1 |d daf10 |e cts |f ? |h ?
948 1_ |a 20010306 |b 1 |d mann11 |e mann |f ? |h ?
```

Control fields (00X)

Number & code fields (0XX)

Access point (1XX = main entry)

Title, publisher, etc. (2XX)

Physical description (3XX)

Series (4XX)

Notes (5XX)

Subject headings (6XX)

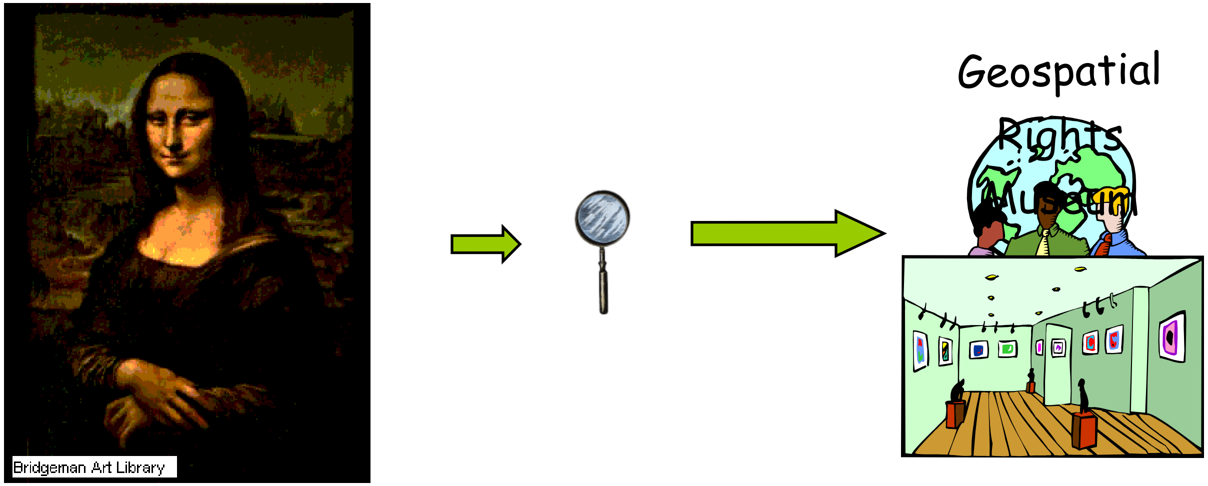
Local fields (9XX)

What’s wrong with this model?

- Expensive
  - Complex (even for its original goal?)
  - Professional intervention (assumes single community of expertise)
- Monolithic
  - One size fits all approach
  - Reflects its centralized system origins
- Bias towards physical artifacts
  - Fixed resources
  - Incomplete handling of resource evolution and other resource relationships

Lenses and Views

- All classification does and should provide a biased *lens* or *view* of reality
- Each view emphasizes certain characteristics and hides others

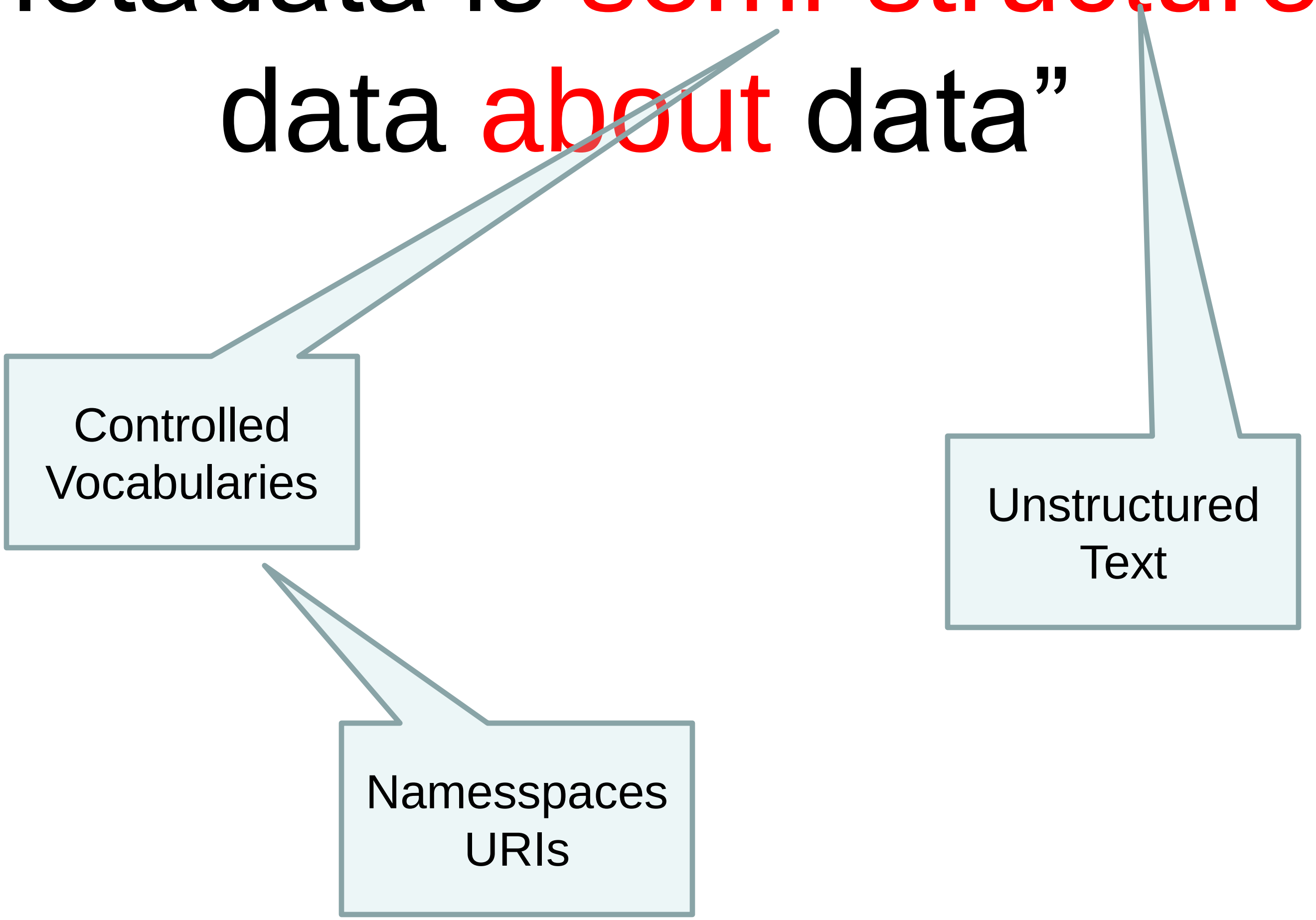


Moving Towards Metadata

- Providing a more “simple” solution
- Accepting that multi-lens view of reality
- Accepting the multiple functions of description
- Adapting to the changing resource context

“Metadata is data about data”

“Metadata is semi-structured data about data”



Controlled Vocabularies

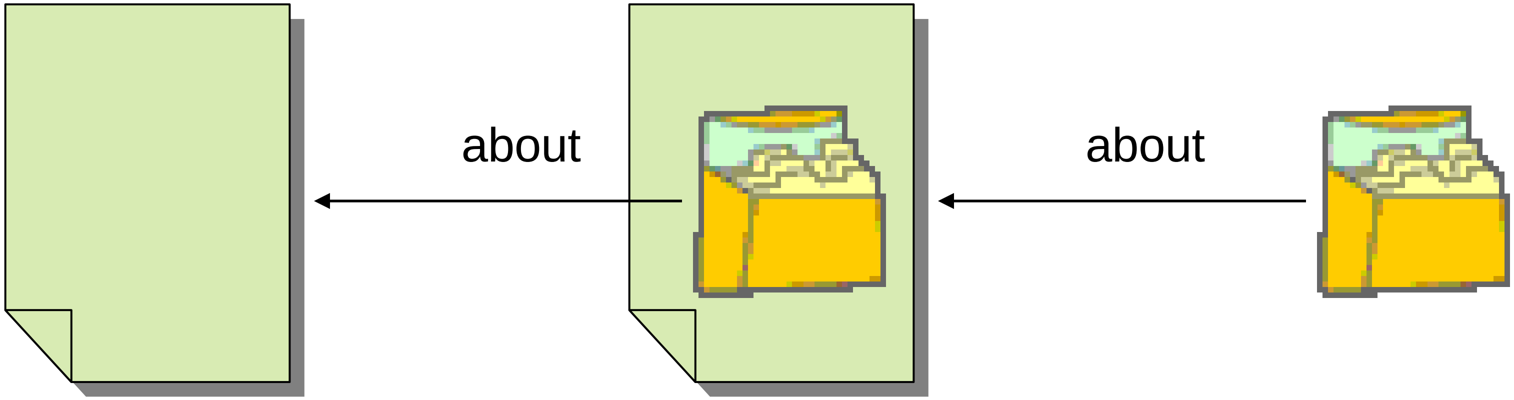
Unstructured Text

Namesspaces URIs

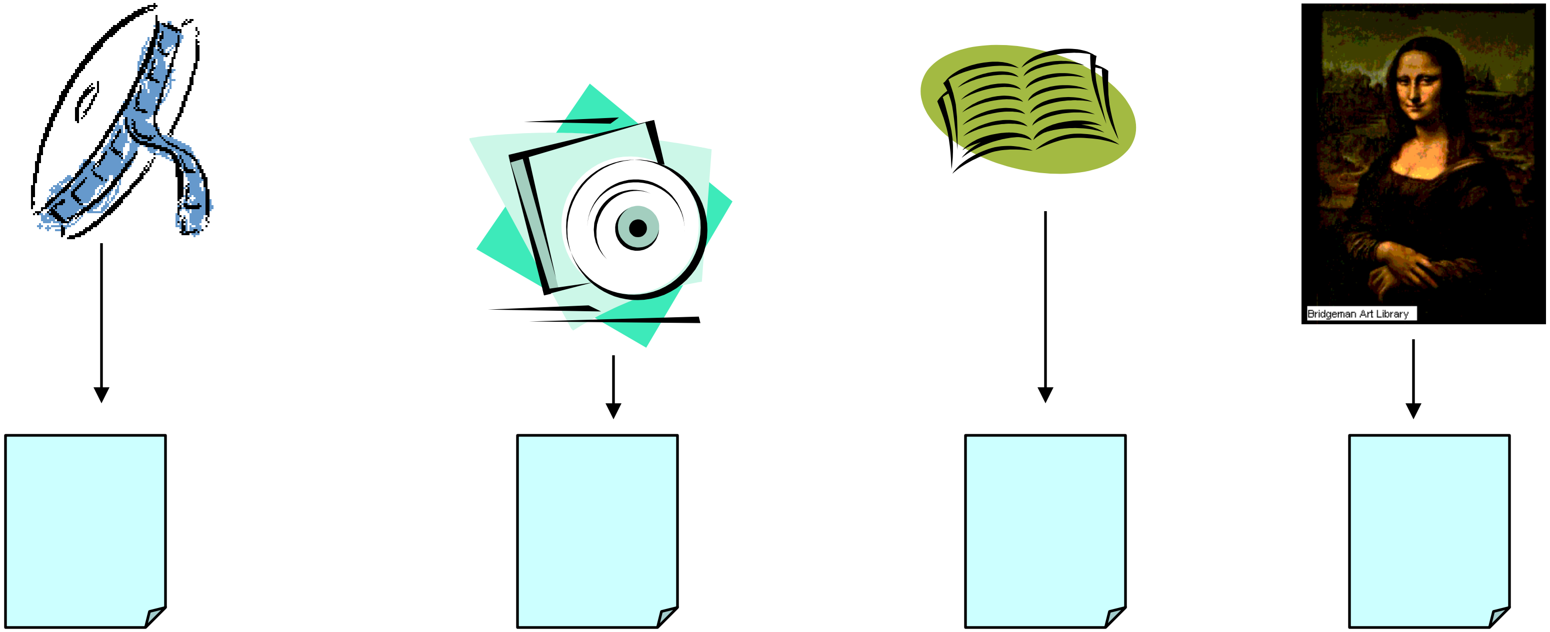
Are metadata and data distinguishable?

- Objectivity?
- Intellectual property?
- Structure?
- Aboutness?

Data/Metadata Polymorphism



Metadata is semi-structured data conforming to commonly agreed upon models, providing operational interoperability in a heterogeneous environment

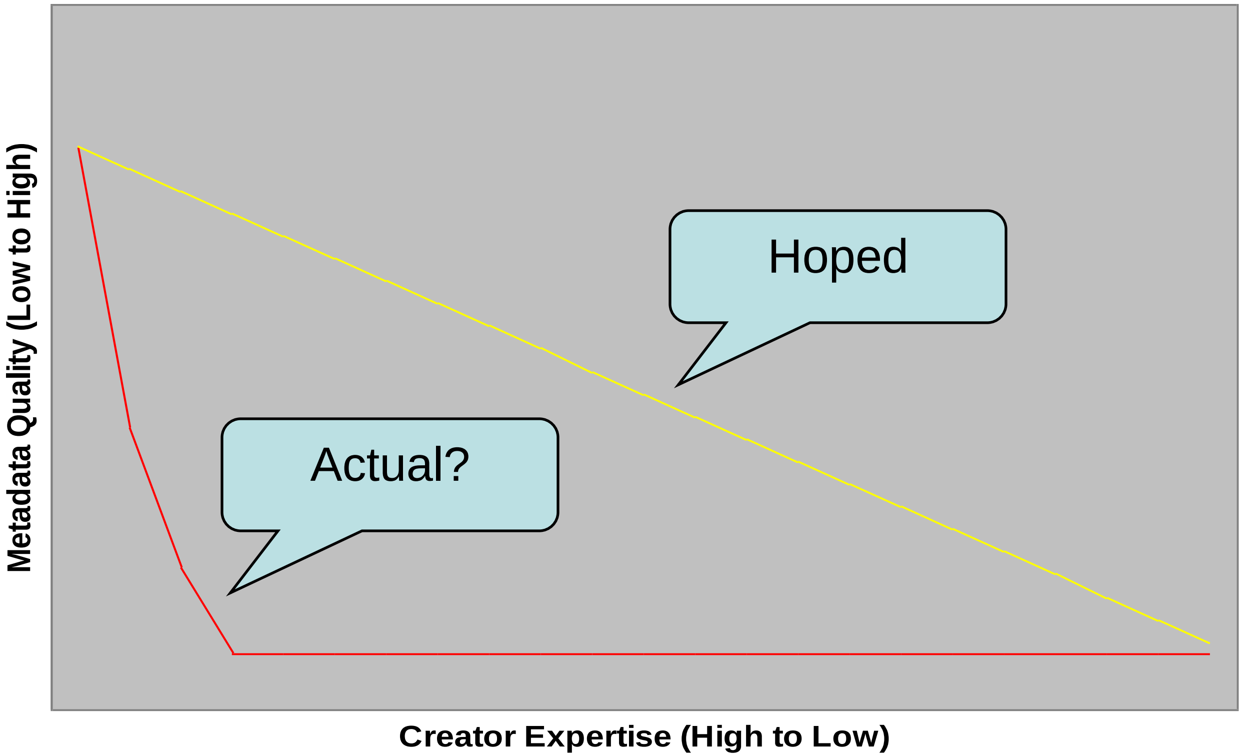


Why hasn’t metadata worked as a general solution for web search?

- No perceived benefit – Search engines keep getting better
- Its all about trust
- People are lazy
- Metadata is hard
- No agreement on one way to describe things

- “Metacrap” - http://www.well.com/~doctorow/metacrap.htm

Metadata Quality as function of Creator Expertise



Contexts for utility of metadata

- non-machine process-able information
  - complex objects
  - services
  - data
- information hiding – intellectual property
- restricted domains
- Establishing relationships among objects (citation matching)
- beyond description and discovery

Dublin Core Metadata Initiative®

Making it easier to find information.

Dublin Core

- Origins at 1994 Web Conference
  - Metadata was necessary for finding things on the web
  - Simple cross-domain vocabulary (15 elements) describing “document-like” objects
- 2004 ISO standard elements
  - http://dublincore.org/documents/dces/

The fifteen Dublin Core Elements

Creator

Contributor

Publisher

Coverage

Source

Title

Date

Type

Rights

Language

Subject

Description

Format

Relation

Identifier

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

Dublin Core Qualifiers

- From loose semantics to more specific description
- Model of “graceful degradation”
  - Support both simplicity and specificity
  - Intra-domain and inter-domain semantics
- Informally three class of qualification
  - Element refinement – from “date” to “date published”, from “contributor” to “illustrator”
  - Value encoding schemes – from “subject” to “LCSH subject”
  - Language

The Dublin Core Vocabulary
http://dublincore.org/documents/dcmi-terms/

Elements

1. Identifier
2. Title
3. Creator
4. Contributor
5. Publisher
6. Subject
7. Description
8. Coverage
9. Format
10. Type
11. Date
12. Relation
13. Source
14. Rights
15. Language

Refinements

Abstract
Access rights
Alternative
Audience
Available
Bibliographic citation
Conforms to
Created
Date accepted
Date copyrighted
Date submitted
Education level
Extent
Has format
Has part
Has version
Is format of
Is part of
Is referenced by
Is replaced by
Is required by
Issued
Is version of
License
Mediator
Medium
Modified
Provenance
References
Replaces
Requires
Rights holder
Spatial
Table of contents
Temporal
Valid

Schemes

Box
DCMIType
DDC
IMT
ISO3166
ISO639-2
LCC
LCSH
MESH
Period
Point
RFC1766
RFC3066
TGN
UDC
URI
W3CTDF

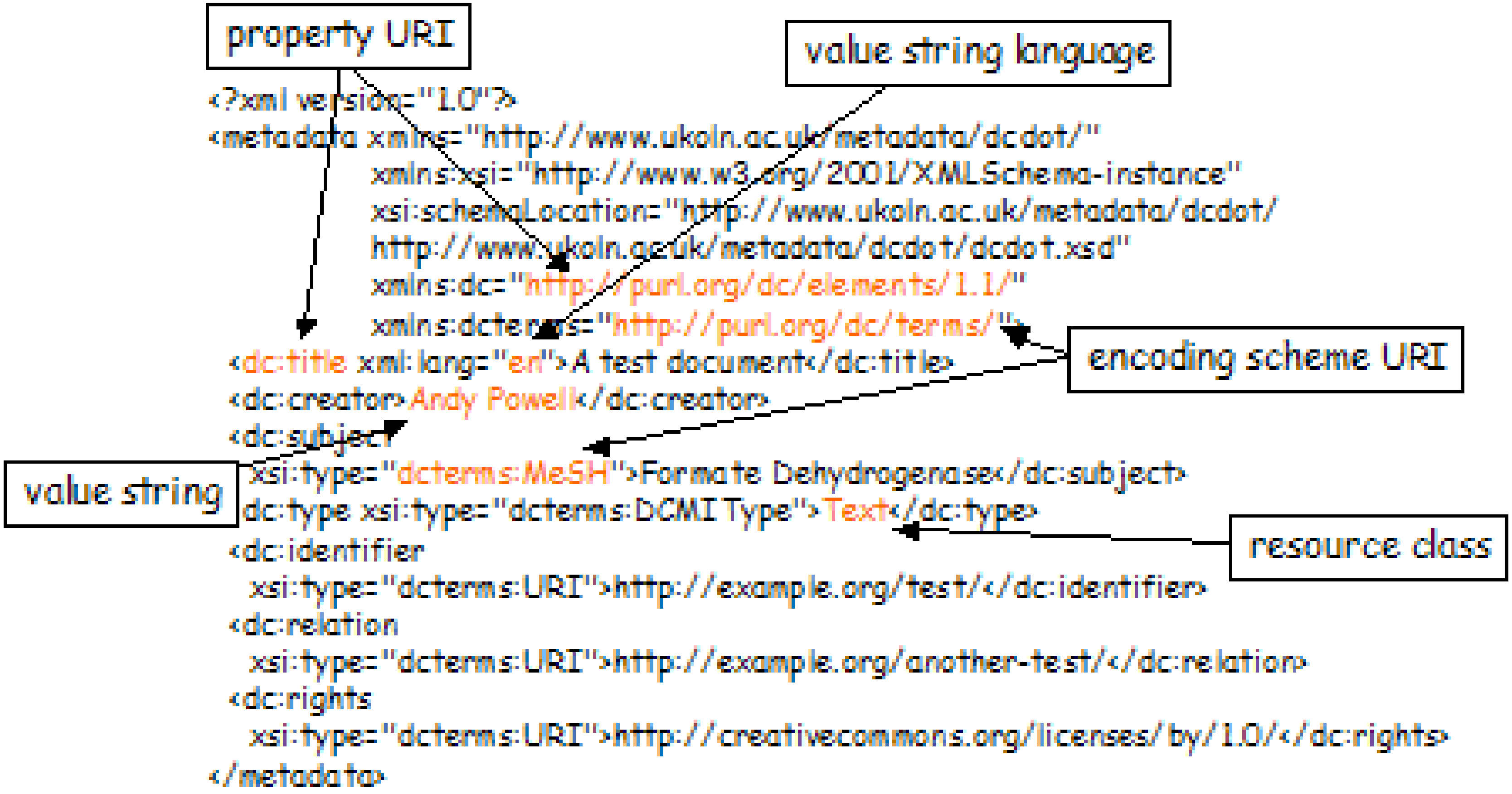
Types

Collection
Dataset
Event
Image
Interactive Resource
Moving Image
Physical Object
Service
Software
Sound
Still Image
Text

Dumb-down

- the process of translating a qualified DC metadata record into a simple DC metadata record is normally referred to as ‘dumbing-down’
- can be separated into two parts:
  - – Property – from refinement to core element
  - – Value – from encoding to basic string

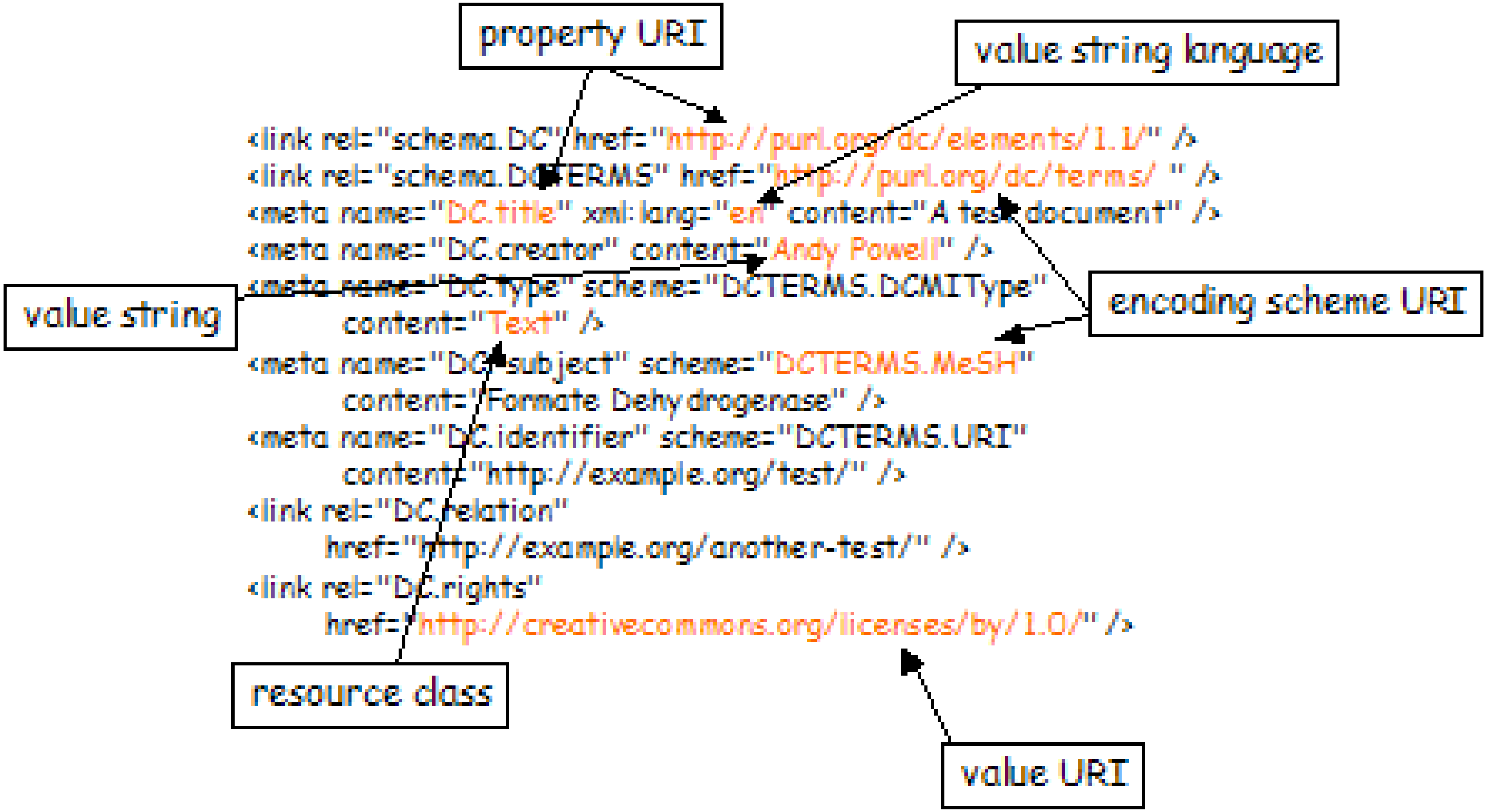
Encoding DC - XML



```
<?xml version="1.0"?>
<metadata xmlns="http://www.ukoln.ac.uk/metadata/dcdot/"
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xsi:schemaLocation="http://www.ukoln.ac.uk/metadata/dcdot/
          http://www.ukoln.ac.uk/metadata/dcdot/dcdot.xsd"
          xmlns:dc="http://purl.org/dc/elements/1.1/"
          xmlns:dcterms="http://purl.org/dc/terms/">
  <dc:title xml:lang="en">A test document</dc:title>
  <dc:creator>Andy Powell</dc:creator>
  <dc:subject
   xsi:type="dcterms:MeSH">Formate Dehydrogenase</dc:subject>
  dc:type xsi:type="dcterms:DCMIType">Text</dc:type>
  <dc:identifier
   xsi:type="dcterms:URI">http://example.org/test/</dc:identifier>
  <dc:relation
   xsi:type="dcterms:URI">http://example.org/another-test/</dc:relation>
  <dc:rights
   xsi:type="dcterms:URI">http://creativecommons.org/licenses/by/1.0/</dc:rights>
</metadata>
```

http://dublincore.org/documents/2002/12/02/dc-xml-guidelines/

Encoding DC - XHTML



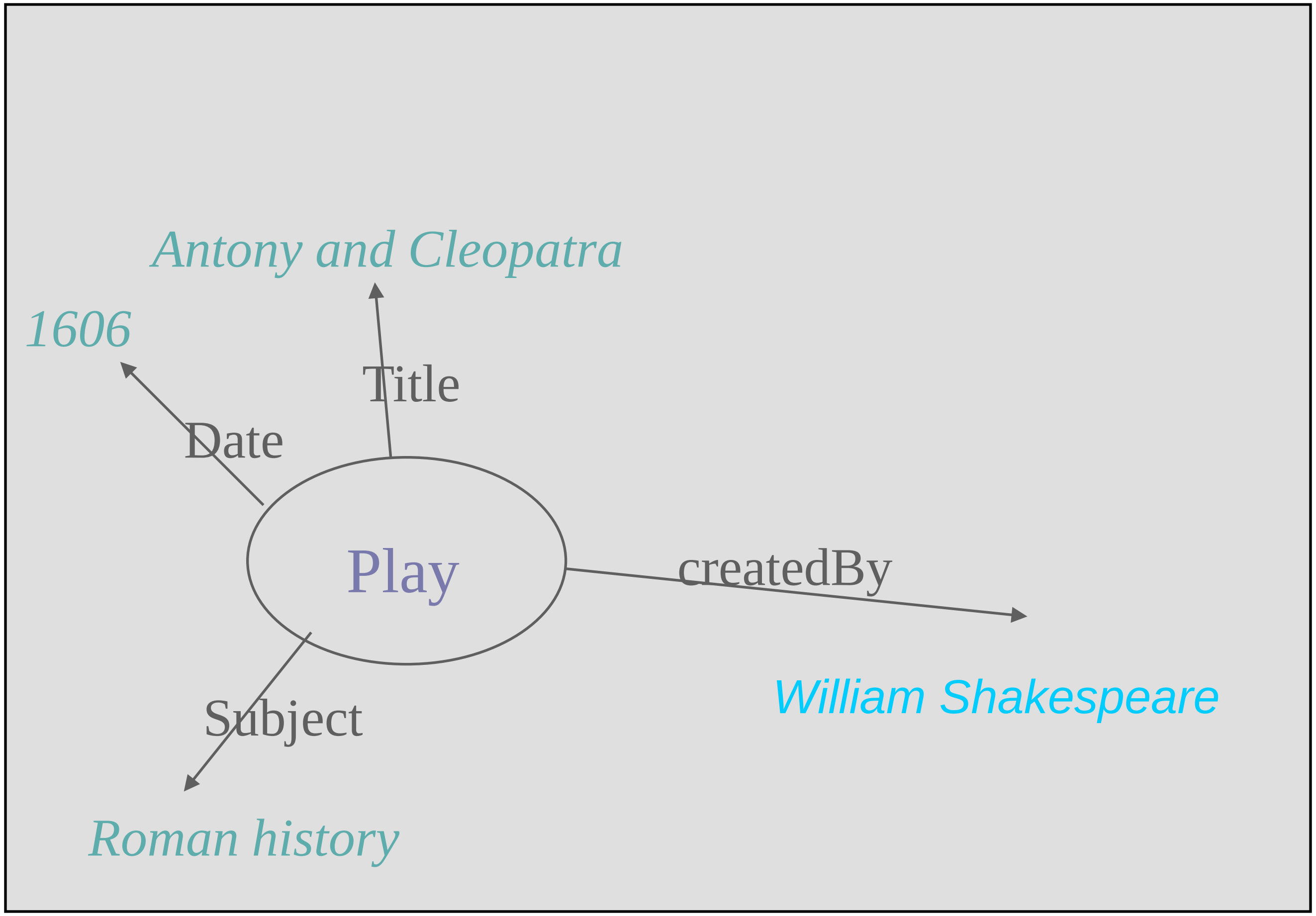
```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/ " />
<meta name="DC.title" xml:lang="en" content="A test document" />
<meta name="DC.creator" content="Andy Powell" />
<meta name="DC.type" scheme="DCTERMS.DCMIType"
        content="Text" />
<meta name="DC.subject" scheme="DCTERMS.MeSH"
        content="Formate Dehydrogenase" />
<meta name="DC.identifier" scheme="DCTERMS.URI"
        content="http://example.org/test/" />
<link rel="DC.relation"
       href="http://example.org/another-test/" />
<link rel="DC.rights"
       href="http://creativecommons.org/licenses/by/1.0/" />
```

http://dublincore.org/documents/dcq-html/

DC Vocabulary in Context: Model and modularity

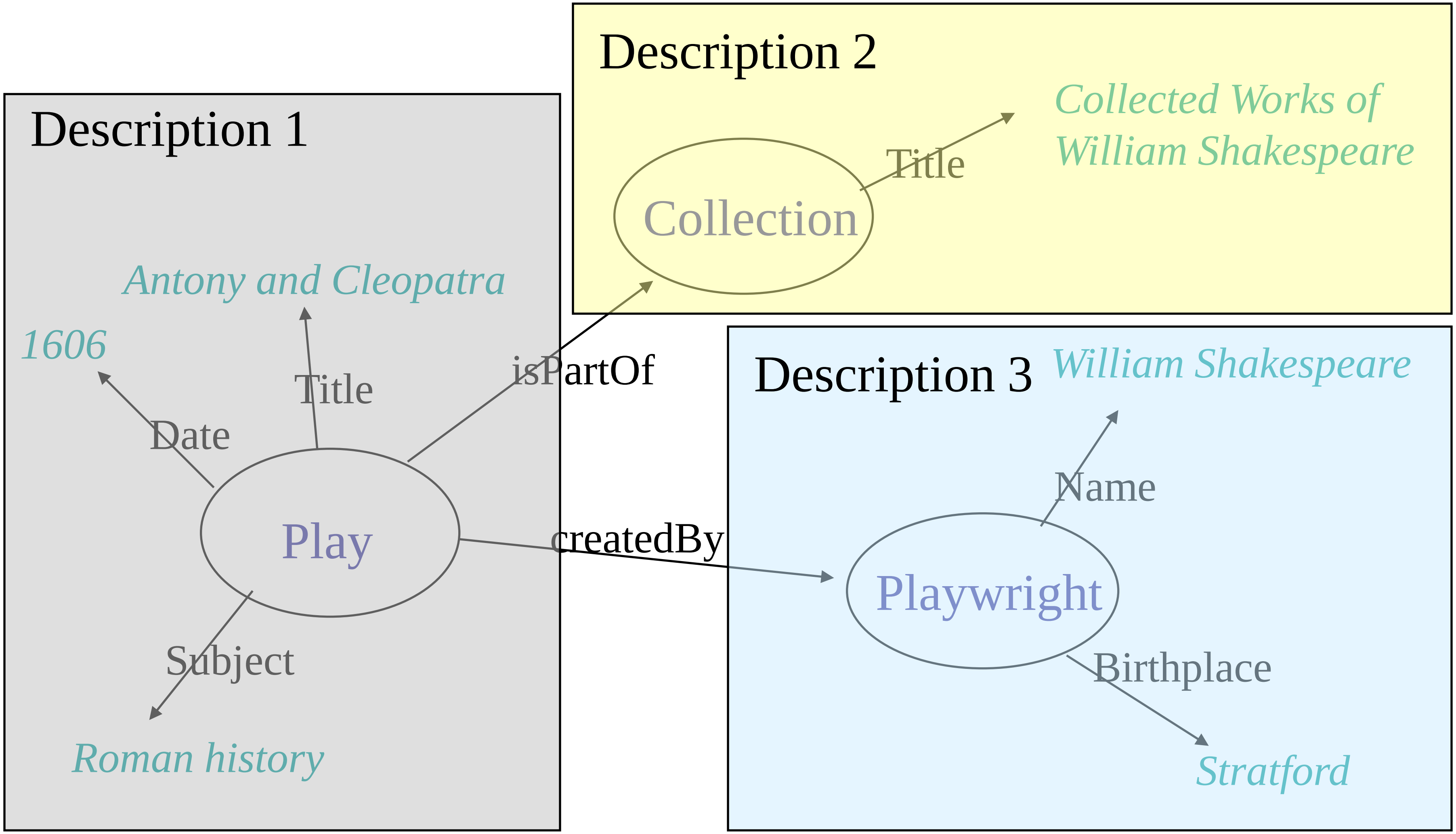
- Resources are related to each other
- There are many vocabularies

One resource, on description

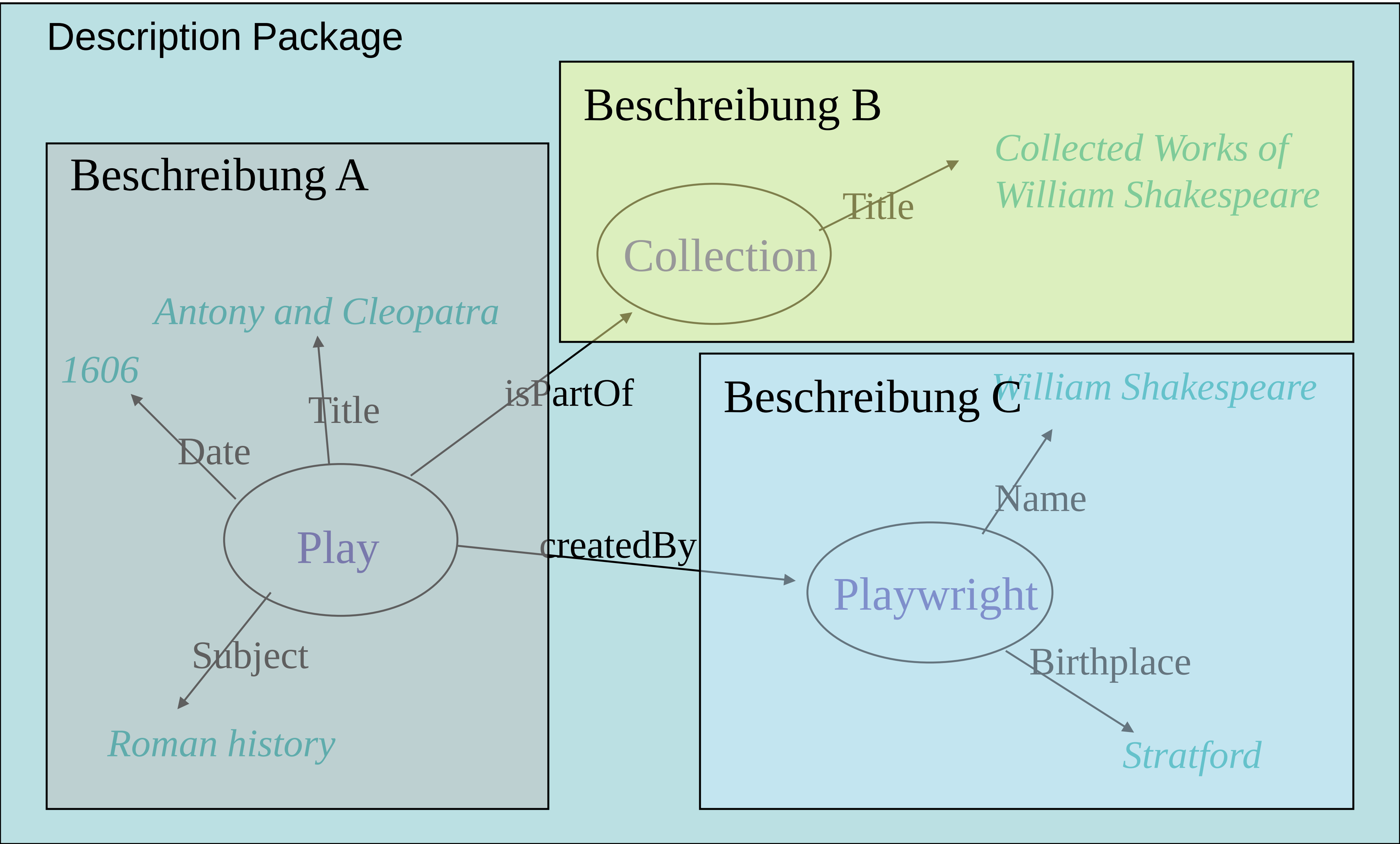


Relationship among many resources

One-to-one principle

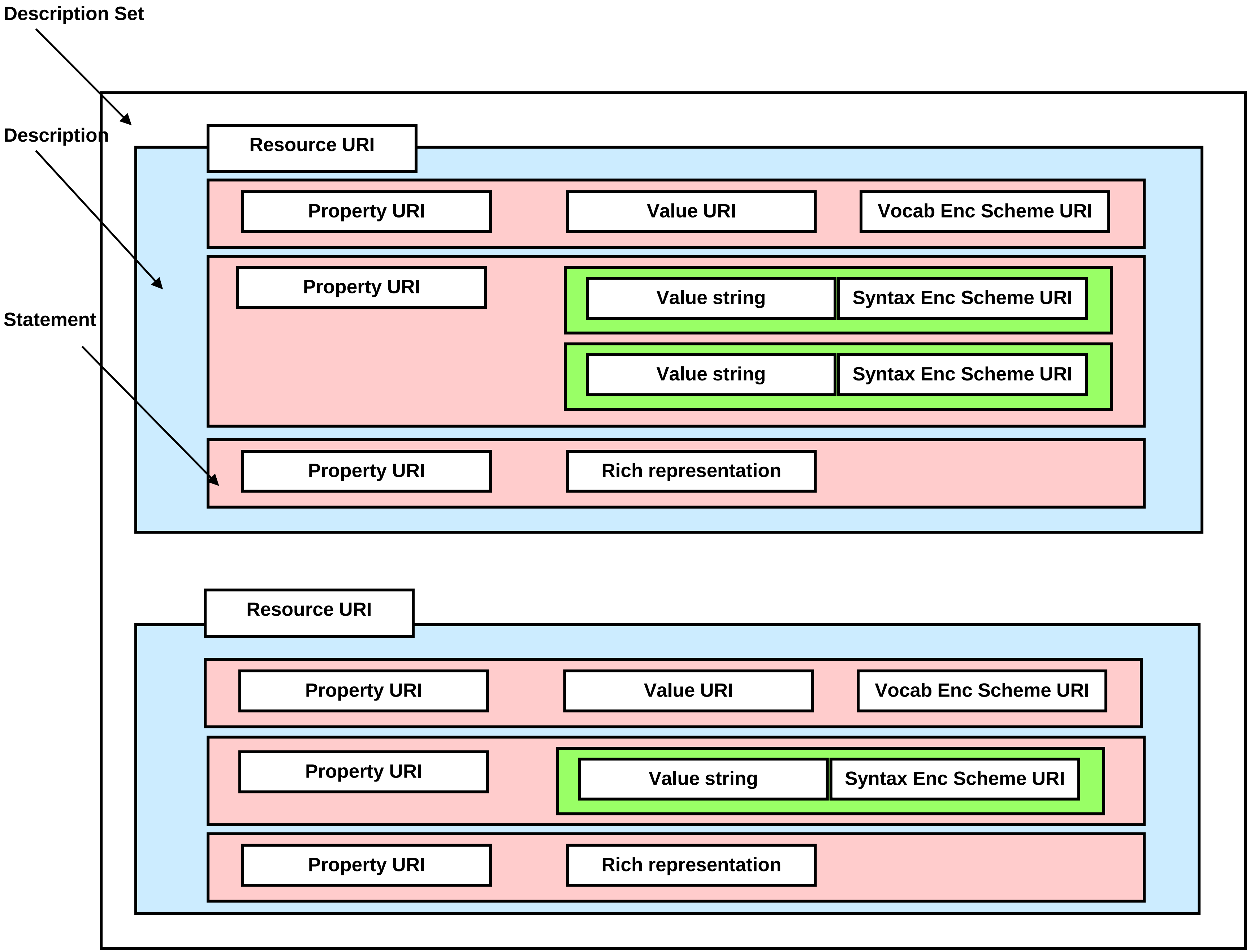


...in one record

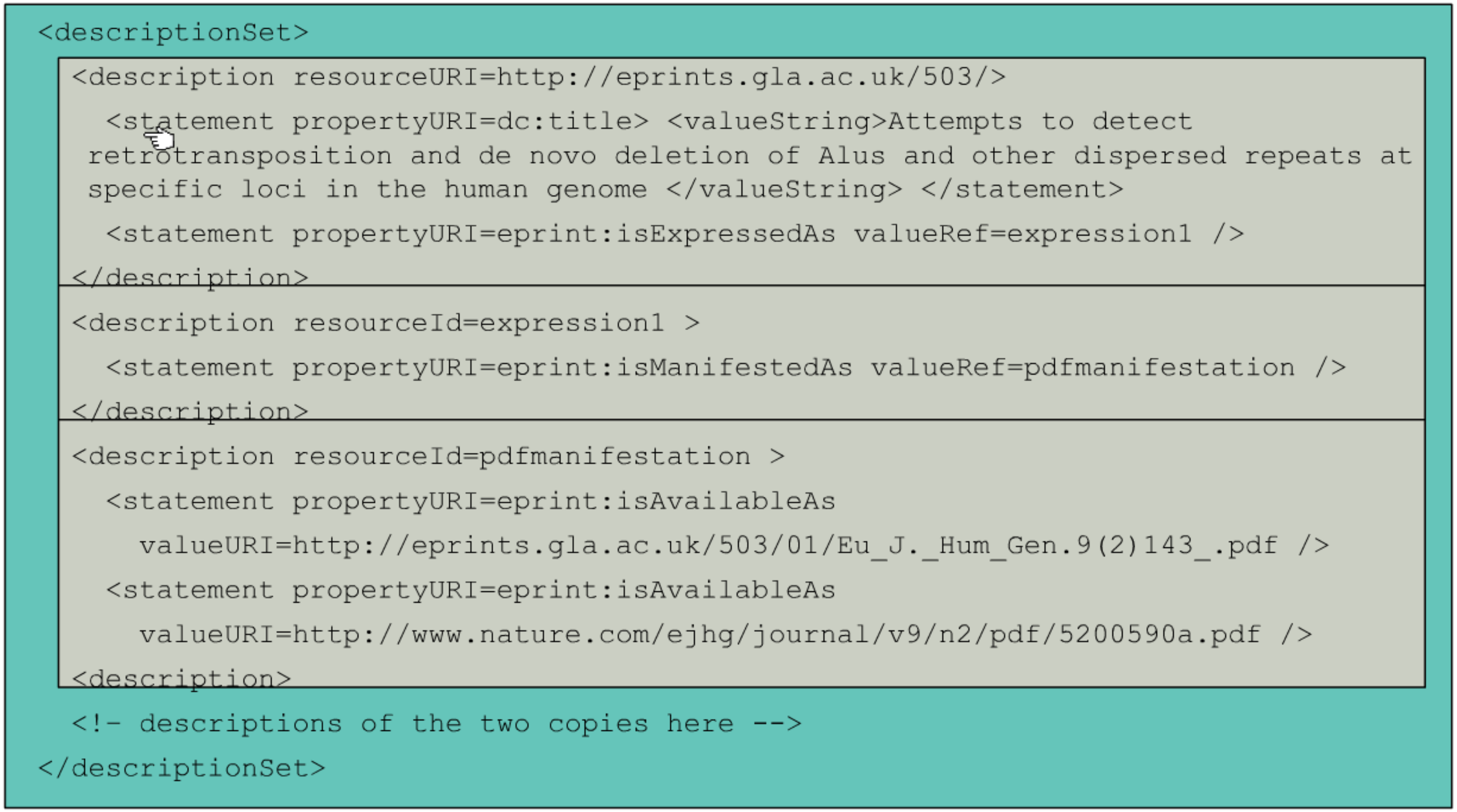


Dublin Core Abstract Model

Packaging multiple descriptions
and vocabularies together


Description Set
Description
Statement
Resource URI
Property URI
Value URI
Vocab Enc Scheme URI
Property URI
Value string
Syntax Enc Scheme URI
Value string
Syntax Enc Scheme URI
Property URI
Rich representation
Resource URI
Property URI
Value URI
Vocab Enc Scheme URI
Property URI
Value string
Syntax Enc Scheme URI
Property URI
Rich representation

Packaging a Complex Object

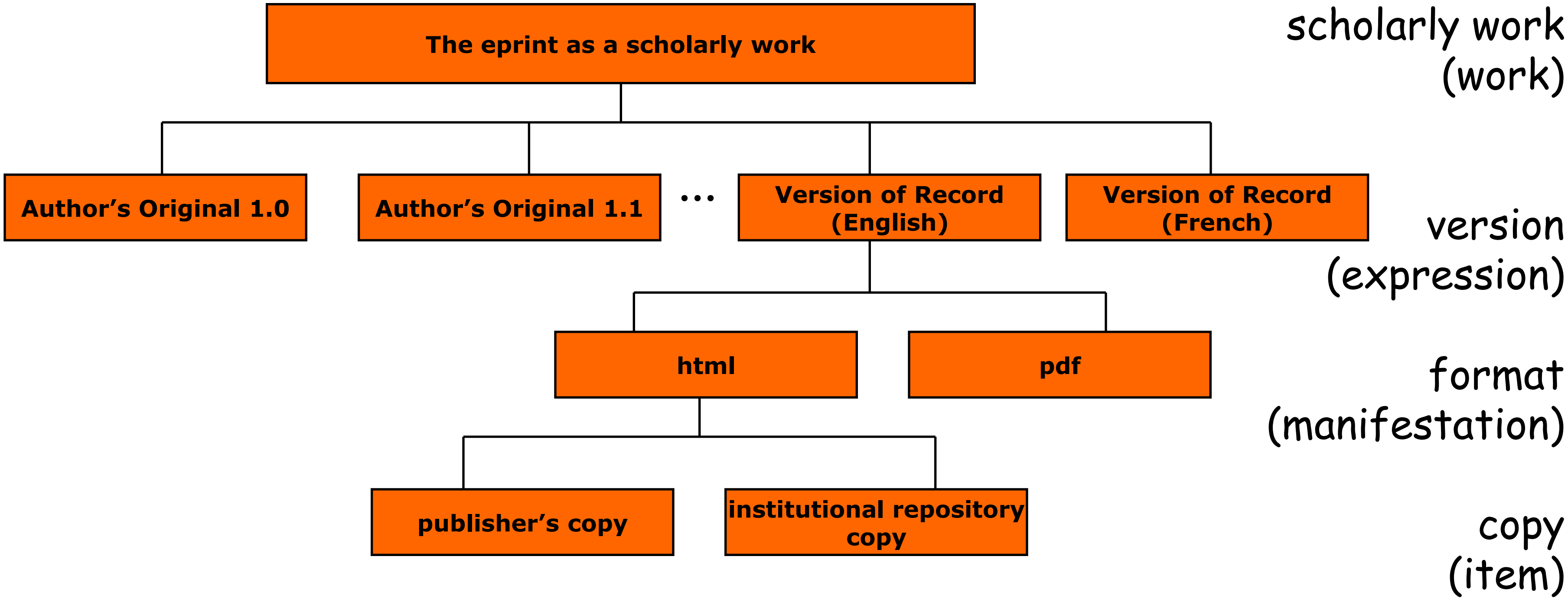


```
<descriptionSet>
  <description resourceURI=http://eprints.gla.ac.uk/503/>
    <statement propertyURI=dc:title> <valueString>Attempts to detect
   retrotransposition and de novo deletion of Alus and other dispersed repeats at
   specific loci in the human genome </valueString> </statement>
    <statement propertyURI=eprint:isExpressedAs valueRef=expression1 />
  </description>
  <description resourceId=expression1 >
    <statement propertyURI=eprint:isManifestedAs valueRef=pdfmanifestation />
  </description>
  <description resourceId=pdfmanifestation >
    <statement propertyURI=eprint:isAvailableAs
      valueURI=http://eprints.gla.ac.uk/503/01/Eu_J._Hum_Gen.9(2)143_.pdf />
    <statement propertyURI=eprint:isAvailableAs
      valueURI=http://www.nature.com/ejhg/journal/v9/n2/pdf/5200590a.pdf />
  <description>
  <!- descriptions of the two copies here -->
</descriptionSet>
```

Applying this model in the context of scholarly communciation

- Increasing availability of scholarly research in open access repositories – e.g., arXiv
  - Mirrored
  - Multi-format (pdf, laTex)
  - Co-exist in journal published form and ePrint form
- FRBR is a model for representing these relationships.

FRBR for eprints



The eprint as a scholarly work

Author's Original 1.0

Author's Original 1.1

...

Version of Record (English)

Version of Record (French)

html

pdf

publisher's copy

institutional repository copy

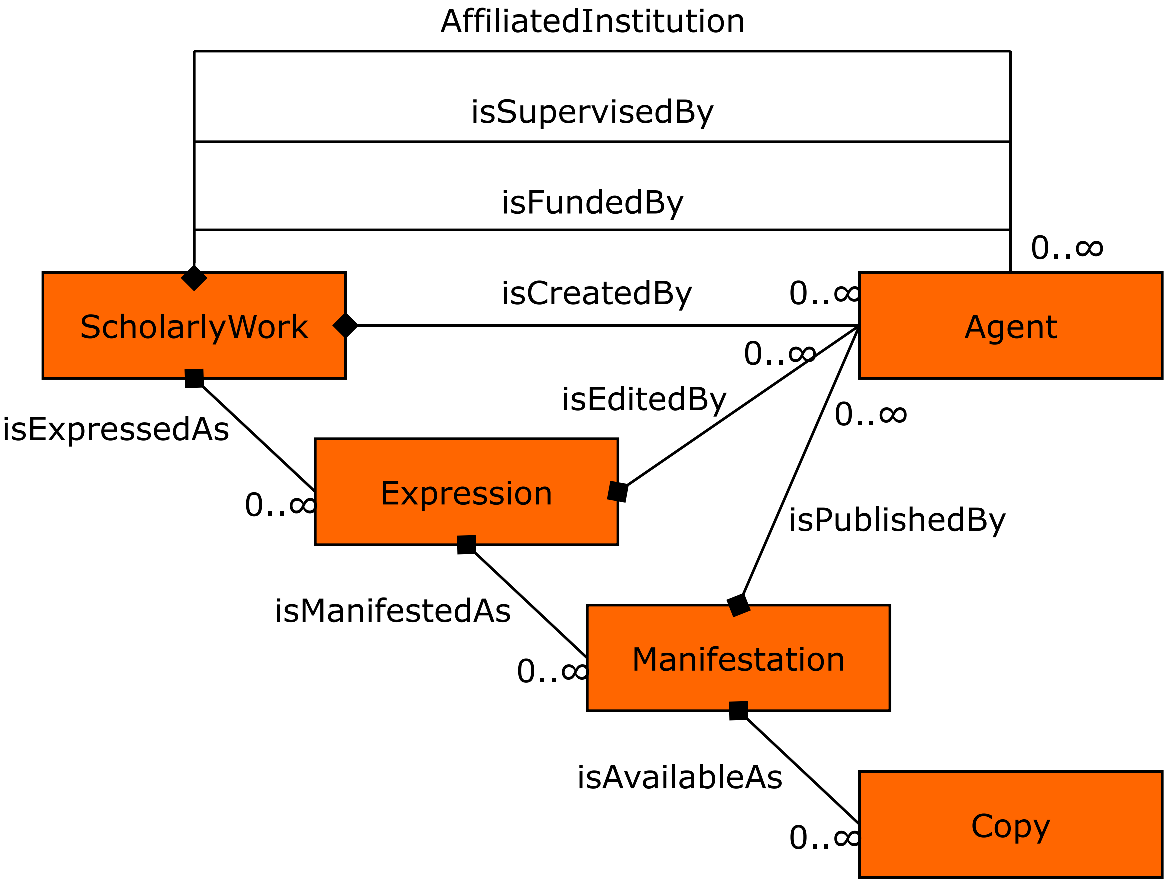
scholarly work (work)

version (expression)

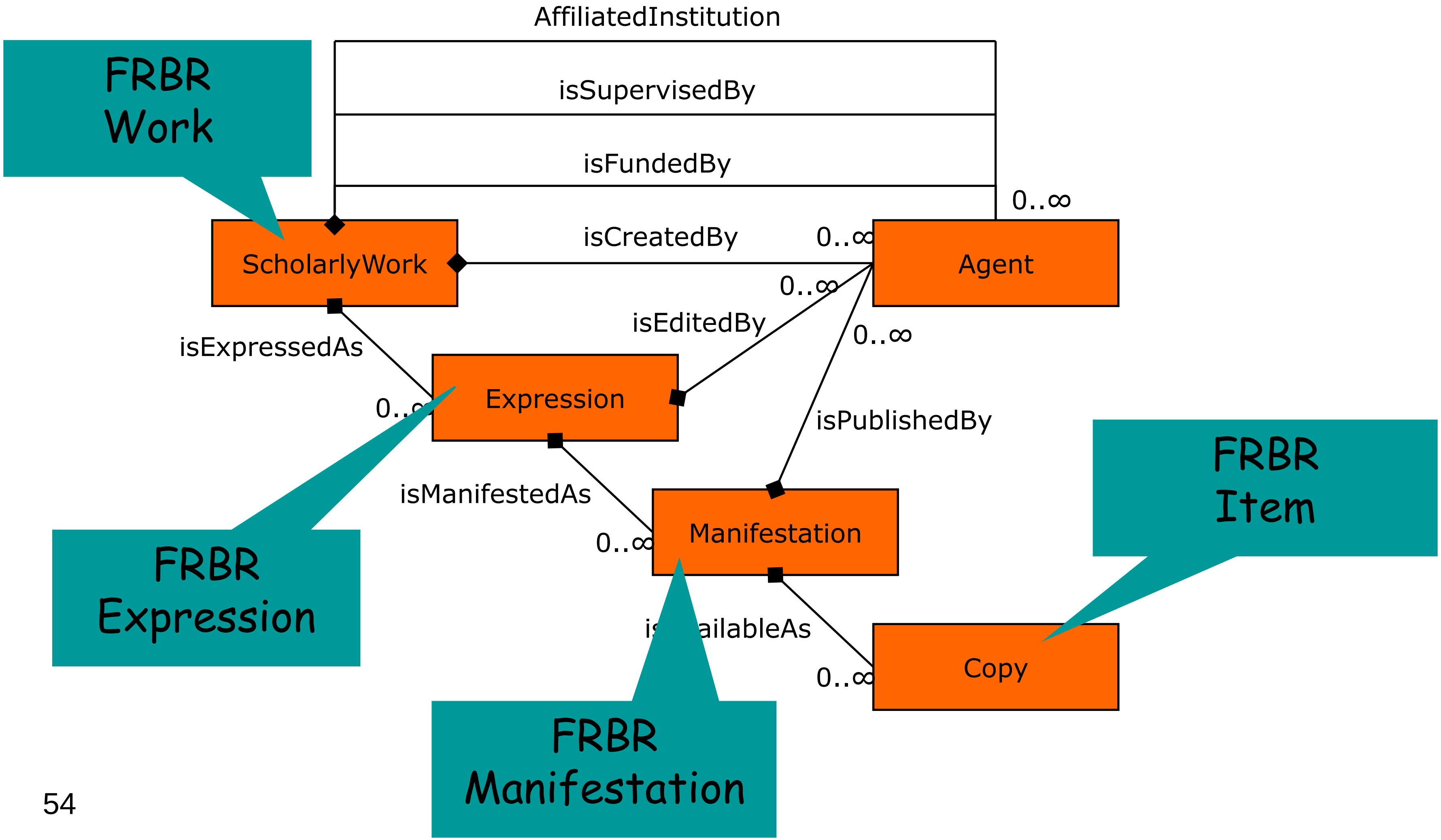
format (manifestation)

copy (item)

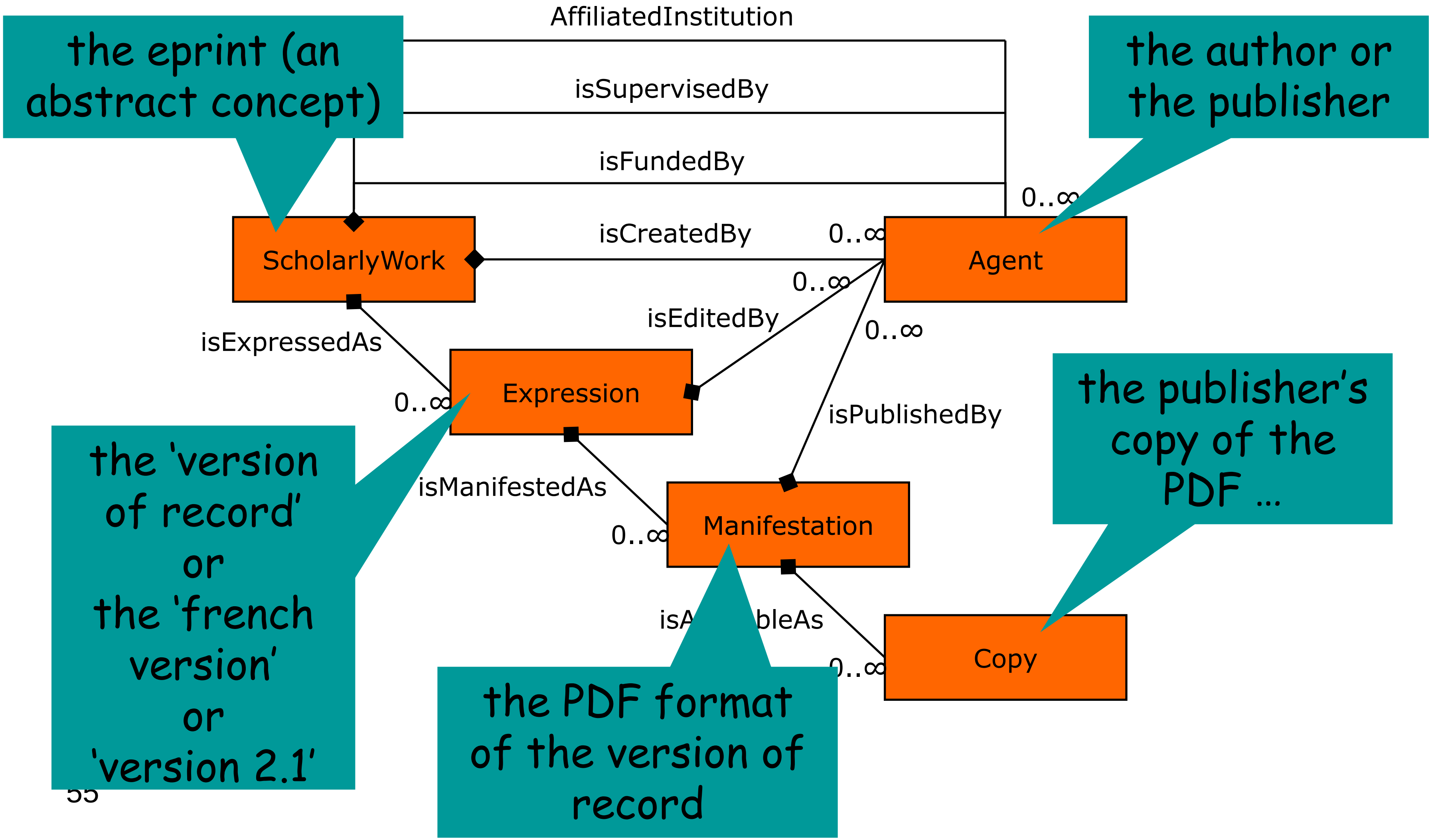
Eprints application model



Eprints model and FRBR



Eprints model and FRBR



the eprint (an abstract concept)

the author or the publisher

the 'version of record' or the 'french version' or 'version 2.1'

the PDF format of the version of record

the publisher's copy of the PDF …

Attributes

- the application model defines the entities and relationships
- each entity needs to be described using an agreed set of attributes

Example attributes

ScholarlyWork:
title
subject
abstract
affiliated institu
identifier

Expression:
title
date available
status
version number
language
genre / type
copyright holder
bibliographic citation
identifier

Manifestation:
format
date modified

Copy:
date available
access rights
licence
identifier

Agent:
name
type of agent
date of birth
mailbox
homepage
identifier

How is this complexity captured in DC?

- the DC Abstract Model provides the notion of ‘description sets’
- i.e. groups of related ‘descriptions’
- where each ‘description’ is about an instance of one of the entities in the model
- relationships and attributes are instantiated as metadata properties

Resources

- DCMI Abstract Model
  - http://dublincore.org/documents/abstract-model/
- Eprints Application Profile
  - http://www.ukoln.ac.uk/repositories/digirep/index/EPrints_Application_Profile
- Eprints DC XML
  - http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_DC_XML
- Eprints DC XML/Instances
  - http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_DC_XML/Instances