Describing Information Units: Metadata, Cataloging, and Beyond

CS 431 – February 15, 2006

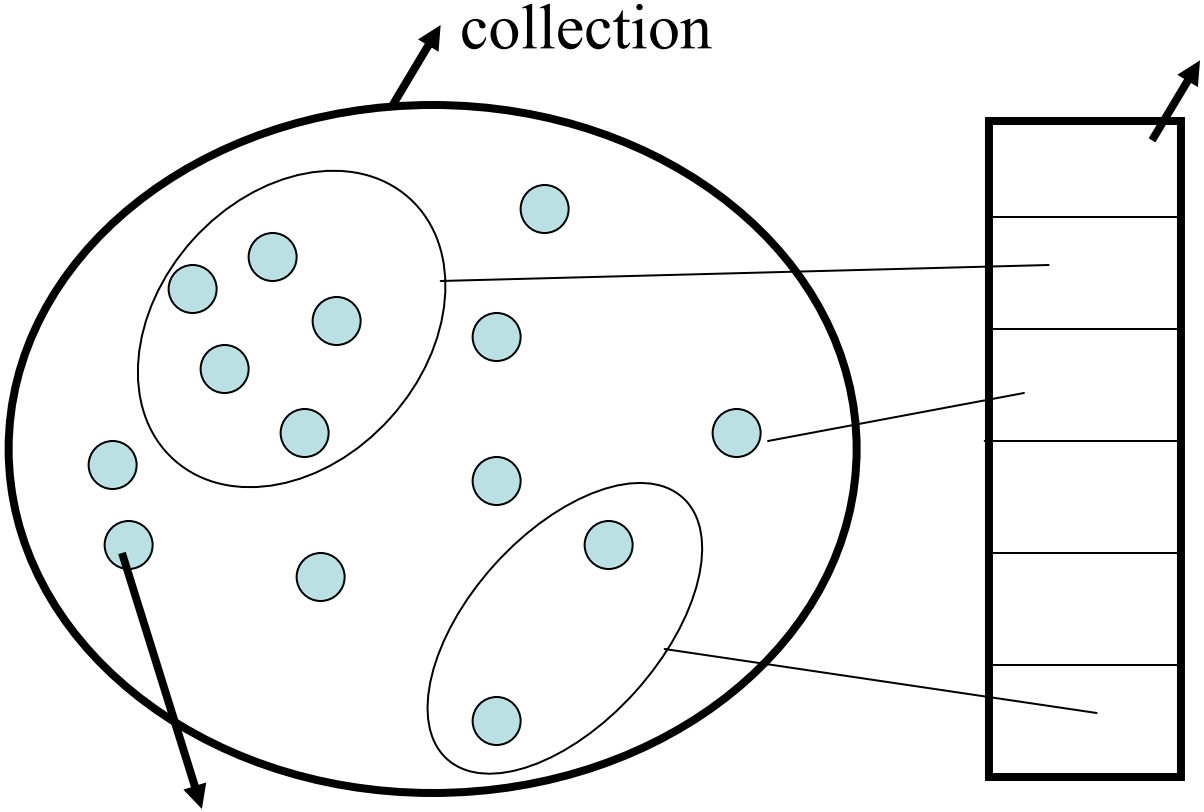
Carl Lagoze – Cornell University

Acknowledgments

- Andy Powell, Head of Development, Eduserv Foundation, UK

Bibliographic model

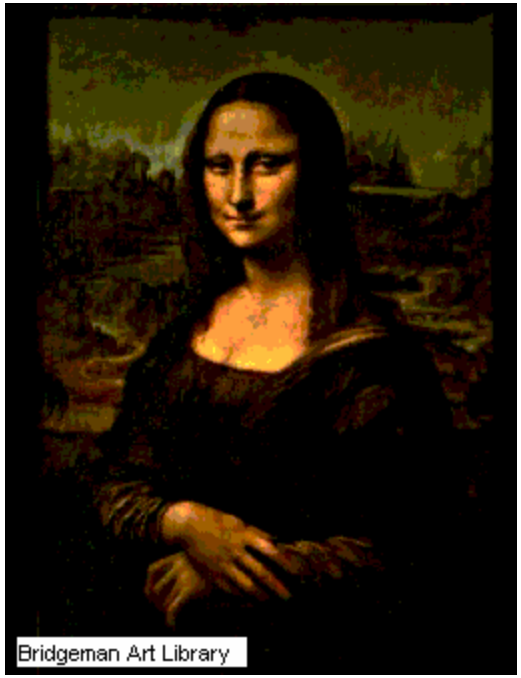
establishes equivalence classes to organize information objects for human understanding and management



Metadata surrogates:
- Identity
- Description
- Structure

object = piece of content

Reality is Complex



Created by:
Leonardo da Vinci

Created on:
1506

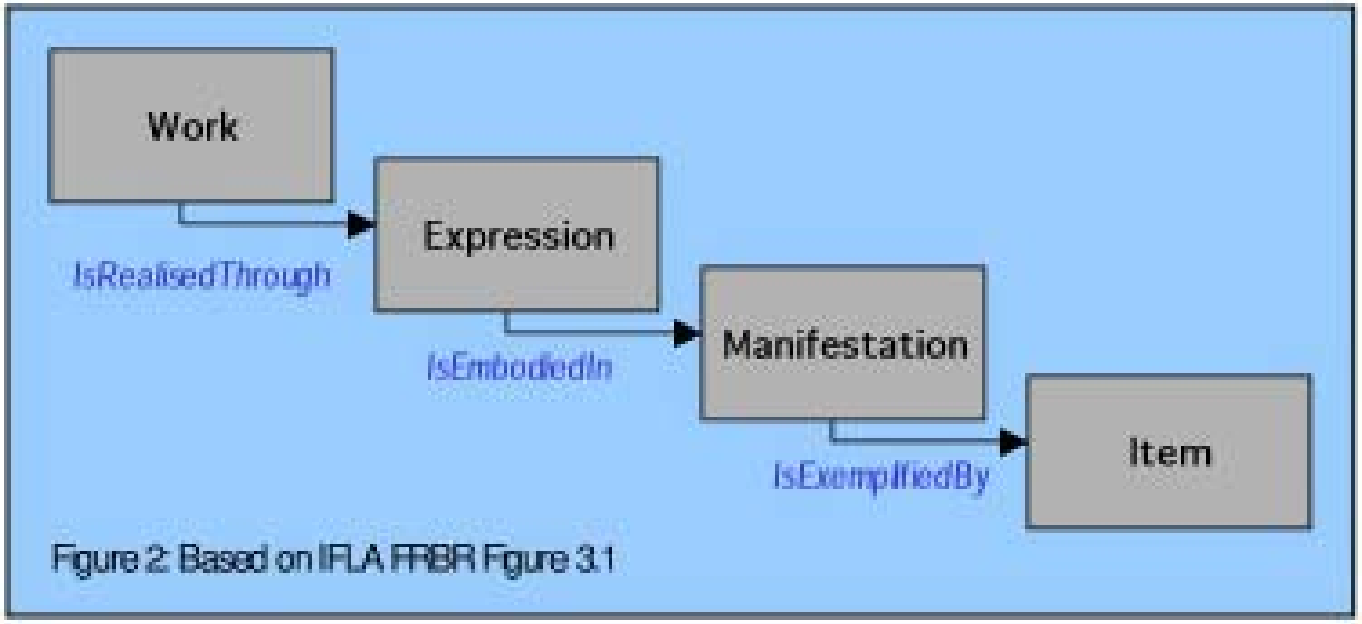
Relationship?

Created by:
George Castaldo

Created on:
1994

Objects are Related

IFLA Entity Model



Attributes Change Over Time

Metadata in the form of the Catalog

MARC STANDARDS

Library of Congress
Network Development and MARC Standards Office

In the beginning…..

History of the catalog…

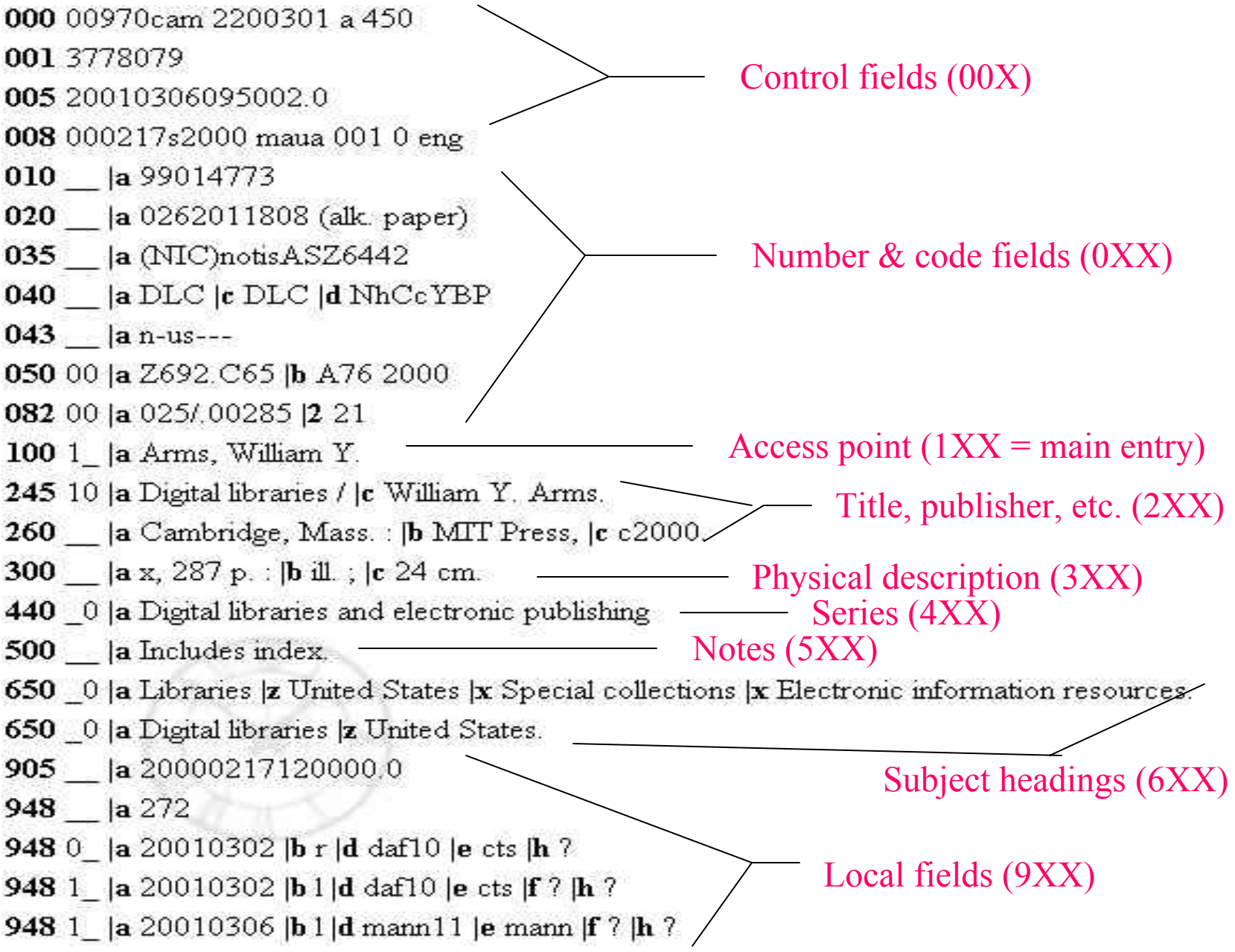
- LC card distribution begins in 1890s
- MARC developed (by Henriette Avram) at LC in the 1960s
- OCLC (first bibliographic utility using MARC) in the early 1970s
- AACR2 (takes effect in 1981) pushes libraries into the online catalog era

... to metadata

- Second (third?) generation library management systems bring on web-based catalogs in 1990s
- AACR2 and MARC extended to remote resources in mid-1990s
- Metadata other than MARC begins to filter into libraries

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)

Brief View | Long View | MARC View

Digital libraries / William Y. Arms.

Database: Cornell University Library
Author/Creator: Arms, William Y.
Title: Digital libraries / William Y. Arms.
Published: Cambridge, Mass. : MIT Press, c2000.
Description: x, 287 p. : ill. ; 24 cm.
Subjects: Libraries--United States--Special collections--Electronic information resources.
Digital libraries--United States.
Series: Digital libraries and electronic publishing
Notes: Includes index.
ISBN: 0262011808 (alk. paper)

Location: Engineering Library (Carpenter Hall)
Call Number: Z692.C65 A76 2000
Status: Not Charged

Location: Engineering Library (Carpenter Hall)
Call Number: Z692.C65 A76 2000
Copy Number: 2
Status: Not Charged

From Holdings Record

Name Authority Record

Tag	Ind1	Ind2	Field Data
000			0531nz___2200157n__4500
001			2791960
005			20011012145755.0
008			871014n|_acammaab|_____|a_aaa_|||_c
010			$a n 87870185
035			$a n 87870186
040			$a InU $c DLC $d DLC
100	1	0	$a Arms, William Y.
400	1	0	$w nnaa $a Arms, W. Y.
400	1	0	$a Arms, W. Y. $q (William Y.)
670			$a His Report on the performance problems of the RLIN computer system, 1982: $b t.p. (William Y. Arms)
670			$a LC data base, 8/26/86 $b (hdg.: Arms, W. Y. ; usage: William Y. Arms; W. Y. Arms)

Authorized heading

Cross-references

Source where data found

Series Record

Tag	Ind1	Ind2	Field Data
000			00494nz___2200169n__4500
001			4929402
005			20011018022500.0
008			990311n|_acaabaaa|_____|n_aaa_|||__
010			$a n 99019988
035			$a (DLC)n 99019988
040			$a DLC $b eng $d DLC
130		0	$a Digital libraries and electronic publishing
643			$a Cambridge $b MIT Press
644			$a f $5 DLC
645			$a t $5 DPCC $5 DLC
646			$a s $5 DLC
670			$a Digital libraries, 2000: $b CIP ser. t.p. (Digital libraries and electronic publishing

Authorized heading

Place/Publisher

Treatment codes

Source where data found

Subject Record

Tag	Ind1	Ind2	Field Data
000			00729nz____2200193n__4500
001			746090
005			19990424120000.0
008			860211i|_anannbab|_____|b_ana_|||__
010			$a sh 85076502
035			$a (NIC)notisCDU9359
040			$a DLC $c DLC $d DLC
053			Z662 $b Z997
150			$a Libraries
360			$i subdivision $a Library $i under names of individual persons, families, and corporate bodies; also subdivision $a Libraries $i under names of individual corporate bodies; also headings beginning with the word $a Library; $i and names of individual libraries
550			$w g $a Documentation
550			$w g $a Public institutions
550			$a Librarians
681			$i Notes under $a Library reports; Library surveys
905			$a 19990424120000.0

LC Classification

Authorized heading (topic)

See also ref.

See also from (broader)

See also from (related)

Information in other headings

Description & Access

- Anglo-American Cataloging Rules
- AACR2 divided into two major parts:
  - Description
    - Organized by format, with specific rules for describing each type of materials
  - Headings, Uniform Titles, and References
    - Choice of access points
    - Headings for persons, geographic names, corporate bodies, etc.
    - References to guide readers to the correct heading

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/

Subject Analysis

- Can be either term based (alphabetically arranged) or alphanumeric (arranged by topic)
- US research libraries generally use the Library of Congress Subject Headings (LCSH) and Classification (LCC)

Dewey Classification

- Dewey Decimal Classification System (DDC) first published in 1876 by Melvil Dewey
- Most widely used classification system in the world (used in 135 countries)
- In this country used primarily by public and school libraries

Dewey, continued

- DDC is divided into ten main classes, then ten divisions, each division into ten sections
- The first digit in each three-digit number represents the main class.
  - “5<u>0</u>0” = natural sciences and mathematics.
- The second digit in each three-digit number indicates the division.
  - “5<u>0</u>0” is used for general works on the sciences
  - “5<u>1</u>0” for mathematics
  - “5<u>2</u>0” for astronomy
  - “5<u>3</u>0” for physics

More Dewey

- The third digit in each three-digit number indicates the section.
  - “530”is used for general works on physics
  - “531” for classical mechanics
  - “532” for fluid mechanics
  - “533” for gas mechanics
- A decimal point follows the third digit in a class number, after which division by ten continues to the specific degree of classification needed.

Library of Congress Classification

- 21 basic classes, based on single alphabetic character (K=law, N=art, etc.)
- Subdivided into two or three alpha characters (KF=American Law, ND=painting, etc.)
- Further subdivision by specific numeric assignment
- Author numbers and dates arrange works by a particular author together and in chronological order

Ranganathan: Colon Classification

- S. R. Ranganathan
  - developed Colon Classification System in the 1930's based on the concept of "facets"
  - notion of "universal principals inherent in all knowledge
  - observed that pre-planned hierarchical categorization systems were too restrictive to developing new forms of information

More Ranganathan

- Facets
  - Personality—what the object is primarily "about." This is considered the "main facet."
  - Matter—the material of the object
  - Energy—the processes or activities that take place in relation to the object
  - Space—where the object happens or exists
  - Time—when the object occurs
- Example - In the 18th Century Style: Building Furniture Inspired by the 18th Century Tradition
  - Personality—furniture
  - Matter—wood
  - Energy—design
  - Space—America
  - Time—18th century

LCSH

- Language of controlled subject index terms
- arranged as a thesaurus
  - broader, narrower, see-also
- Not faceted
- LCSH Live - http://lcsh.orhost.org/

LCSH Example

- Digital libraries
  - see from “Electronic libraries”
  - see from “Virtual libraries”
  - see broader term: “Libraries”
  - see also “Information storage and retrieval systems”

MESH (Medical Subject Headings)

- Maintained by National Library of Medicine (NLM)
- MESH Browser http://www.nlm.nih.gov/mesh/MBrowser.html
- Pubmed
  - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed

Wikipedia

- http://en.wikipedia.org/wiki/Wikipedia:Browse

Classification is Problematic

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

The *fiction* of classification

…there is no classification of the universe that is not fictional and conjectural.

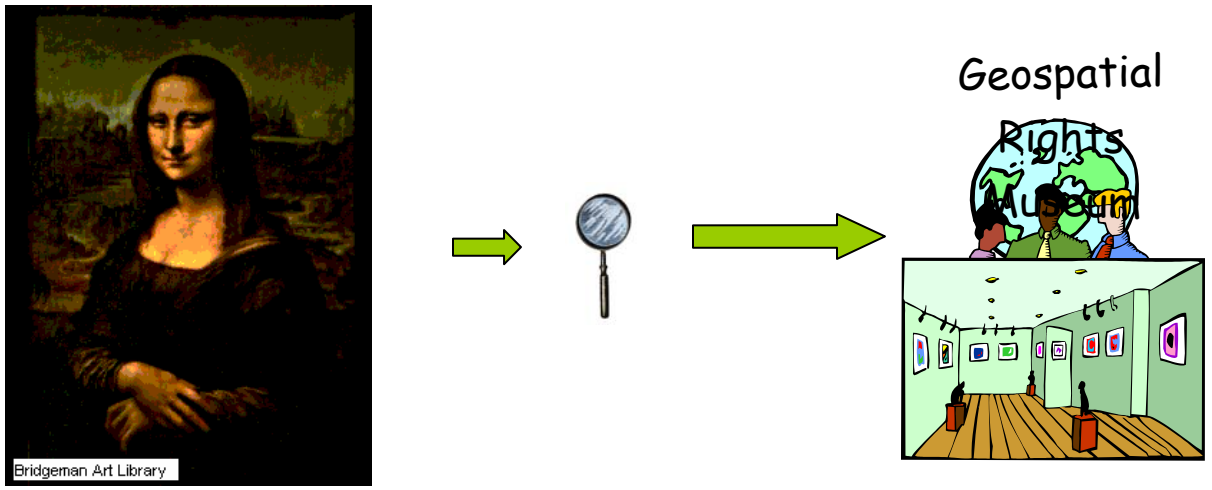
Jorge Luis Borges

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

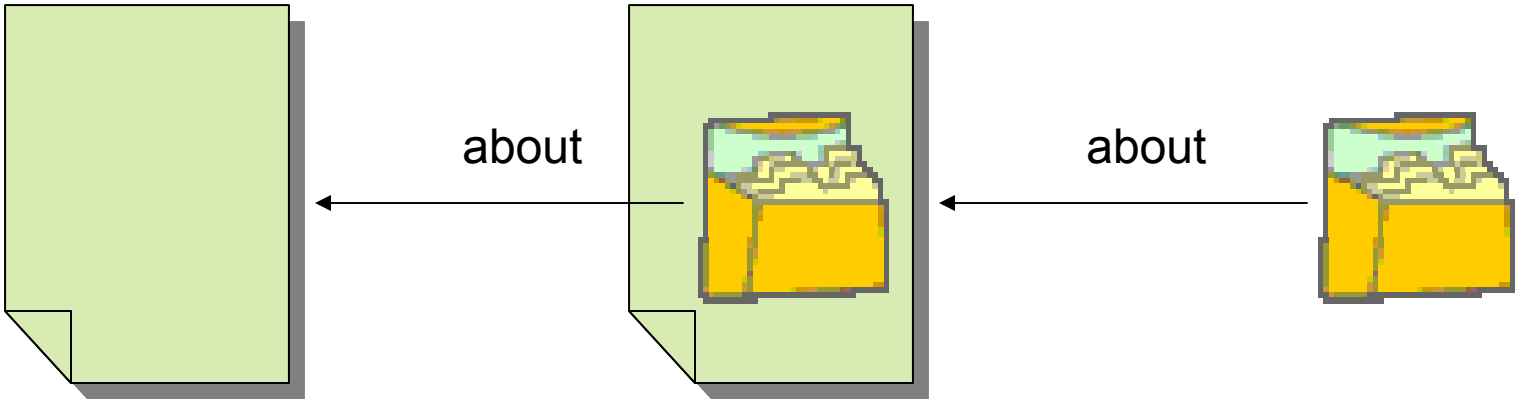
- Providing a more “simple” solution
- Accepting that multi-lens view of reality
- Accepting the multiple functions of description
- Adapting to the changing resource contexte

“Metadata is data about data”

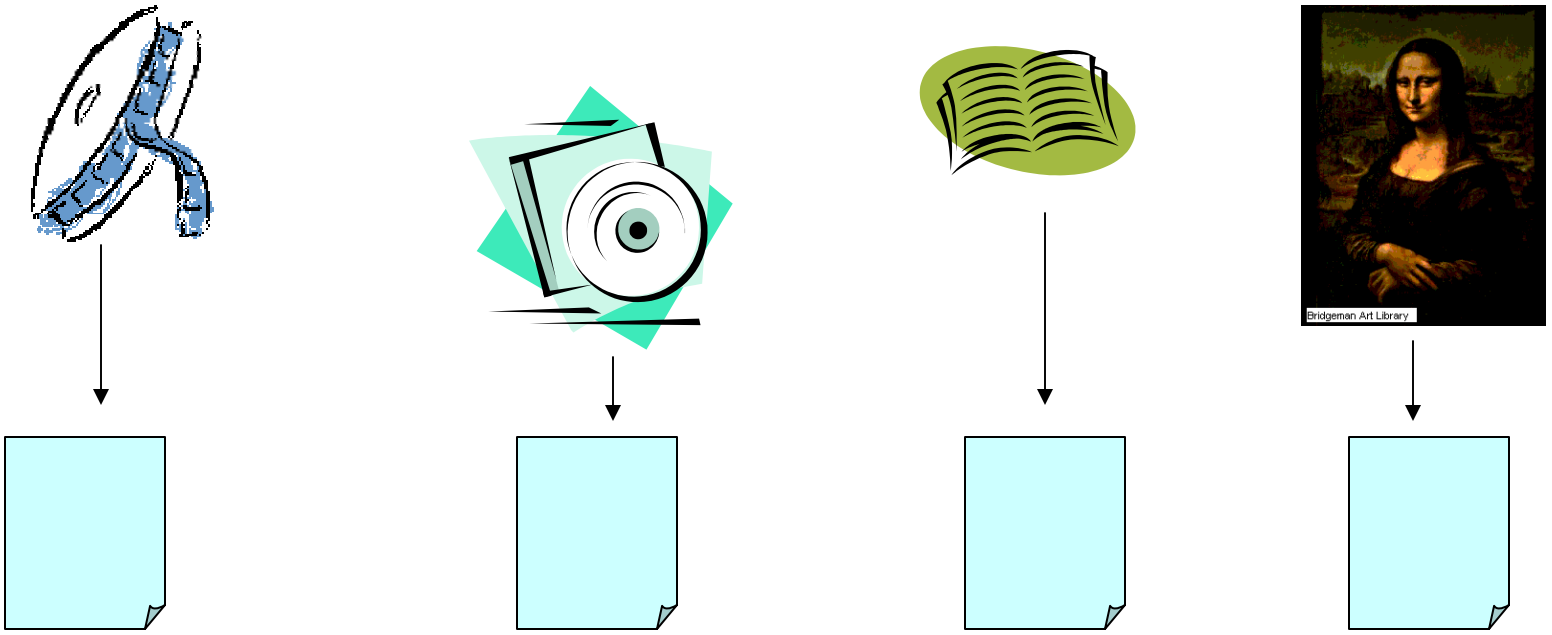
Are metadata and data distinguishable?

- 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



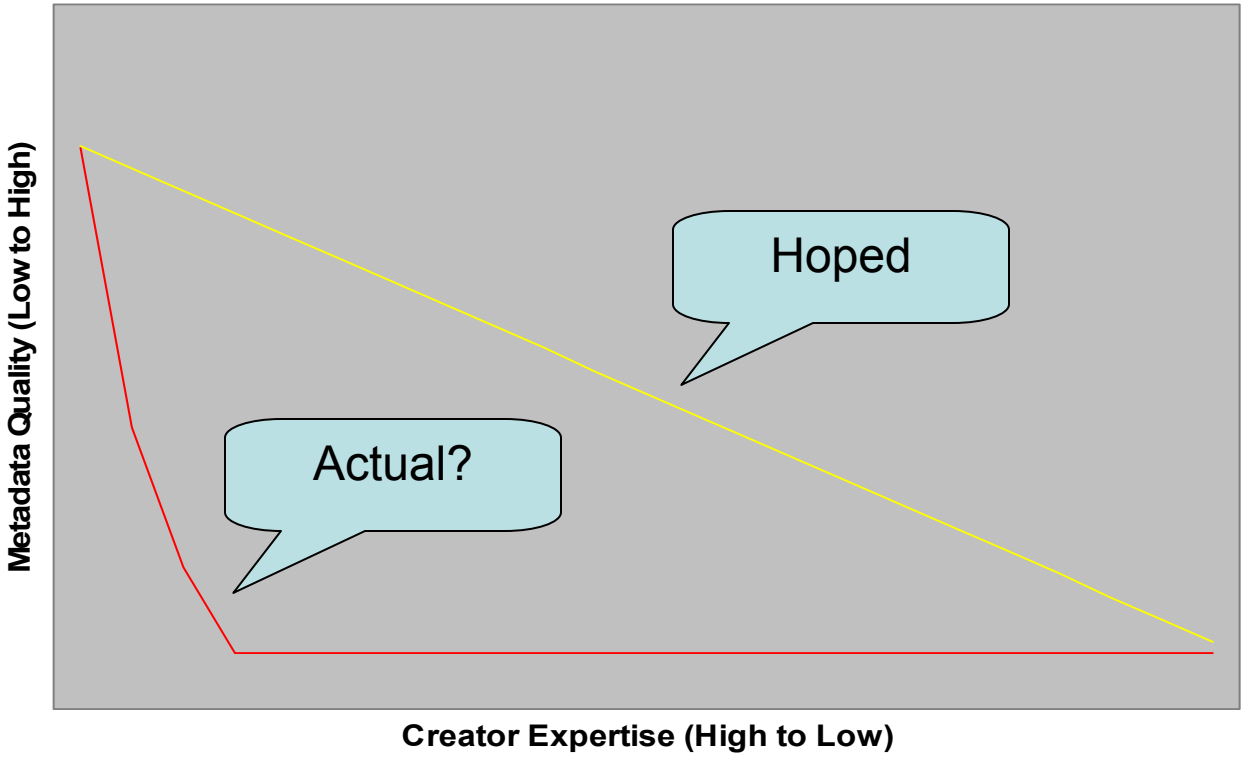
Contexts for utility of metadata

- non-machine process-able information
  - complex objects
  - services
  - data
- information hiding
- restricted domains
- Framework for automated services (e.g., citation matching)
- beyond description and discovery

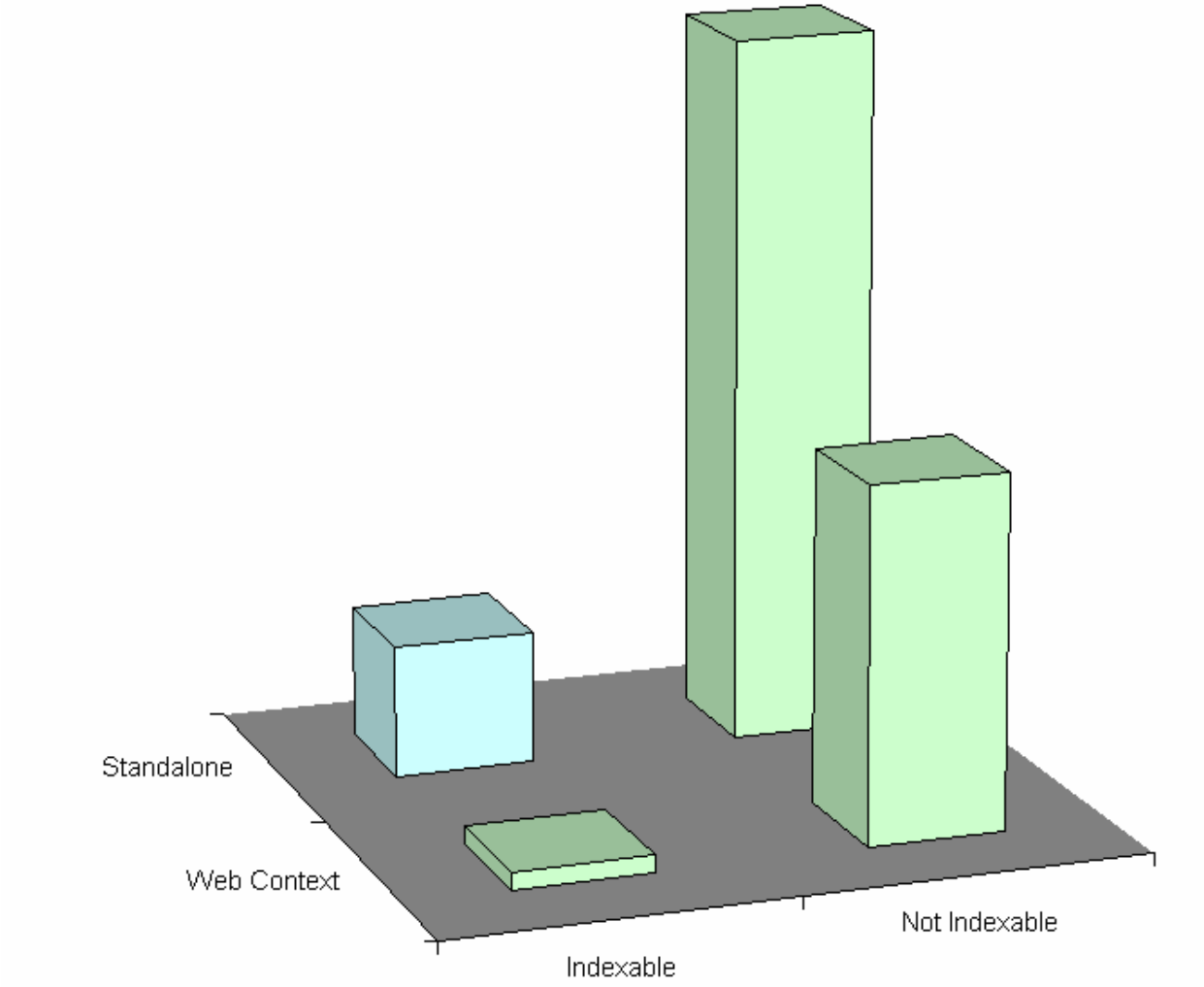
Metadata Takes Many Forms

resource discovery	document administration	rights management
content rating	security and authentication	archival status
products and services	database schemas	process control or description

Metadata Quality as function of Creator Expertise



Metadata Triage



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
- 1997 – notion of “qualification”
  - Building more complex descriptions on basic elements
  - Dumb up and down
- 2004 ISO standard elements
  - http://dublincore.org/documents/dces/

The fifteen Dublin Core Elements

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

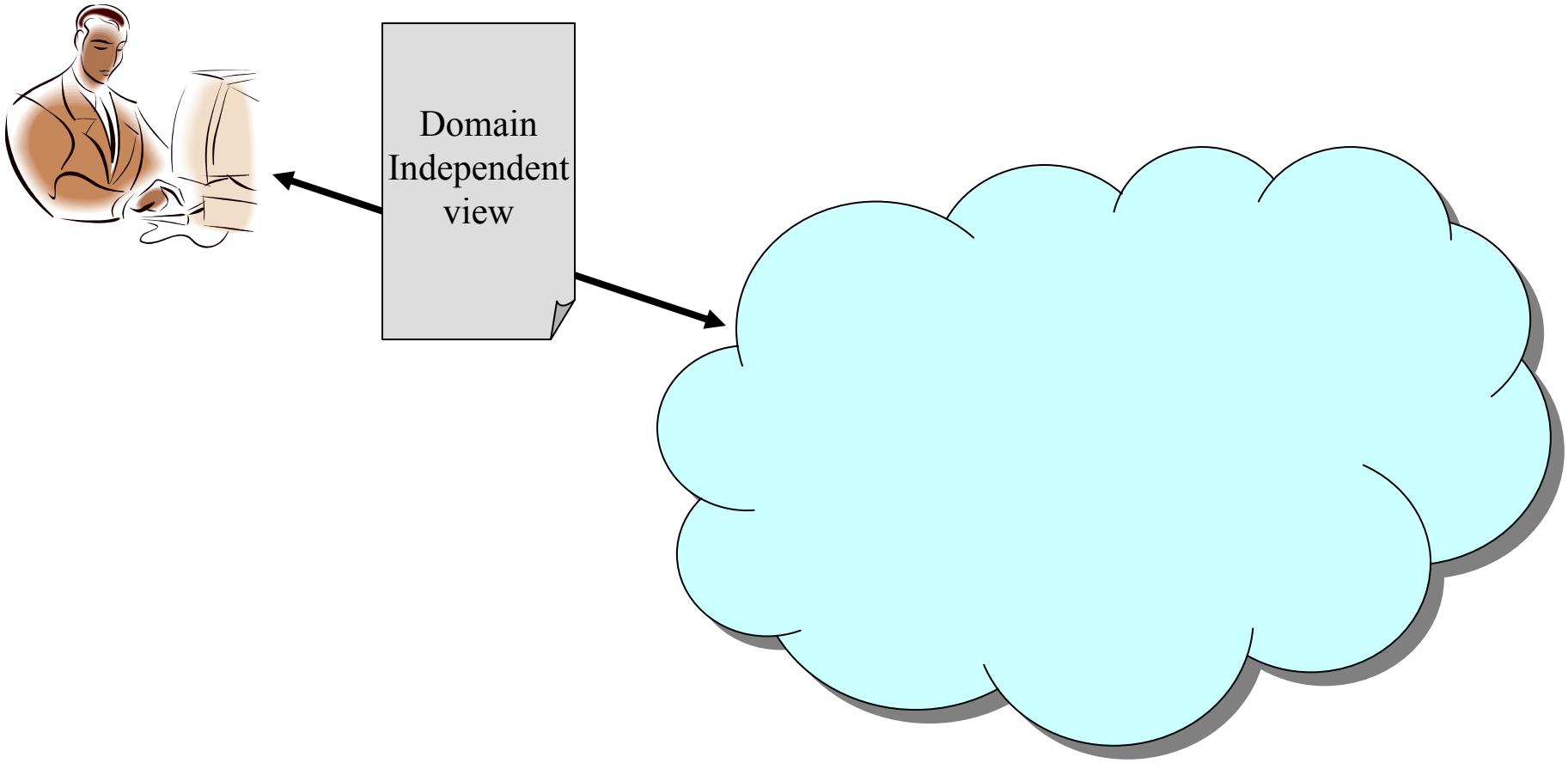
http://dublincore.org/usage/terms/dc/current-elements/
http://dublincore.org

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”
  - Value description – from “subject” to “LCSH subject”
  - Language

What is the Dublin Core (1)

- A simple set of properties to support resource discovery on the web?



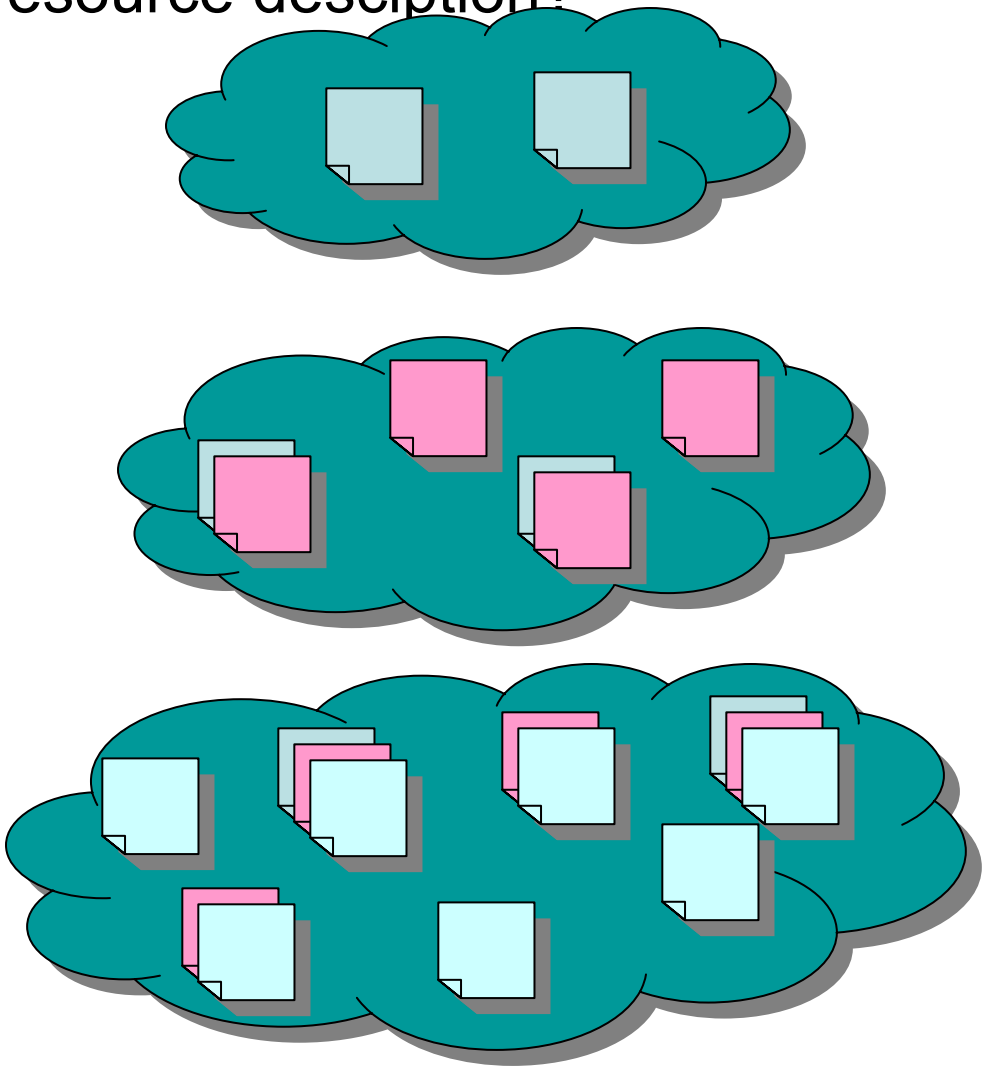
Why hasn't metadata worked as a general solution for web search?

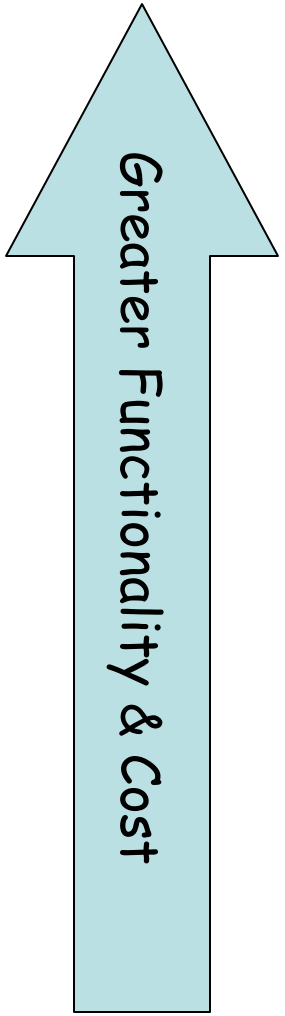
- Its all about trust
- People are lazy
- Metadata is hard
- No perceived benefit
  - "Reverse tragedy of the commons"
- No agreement on one way to describe things

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

What is Dublin Core (2)?

- Qualification view- An extensible ontology for resource desciption?

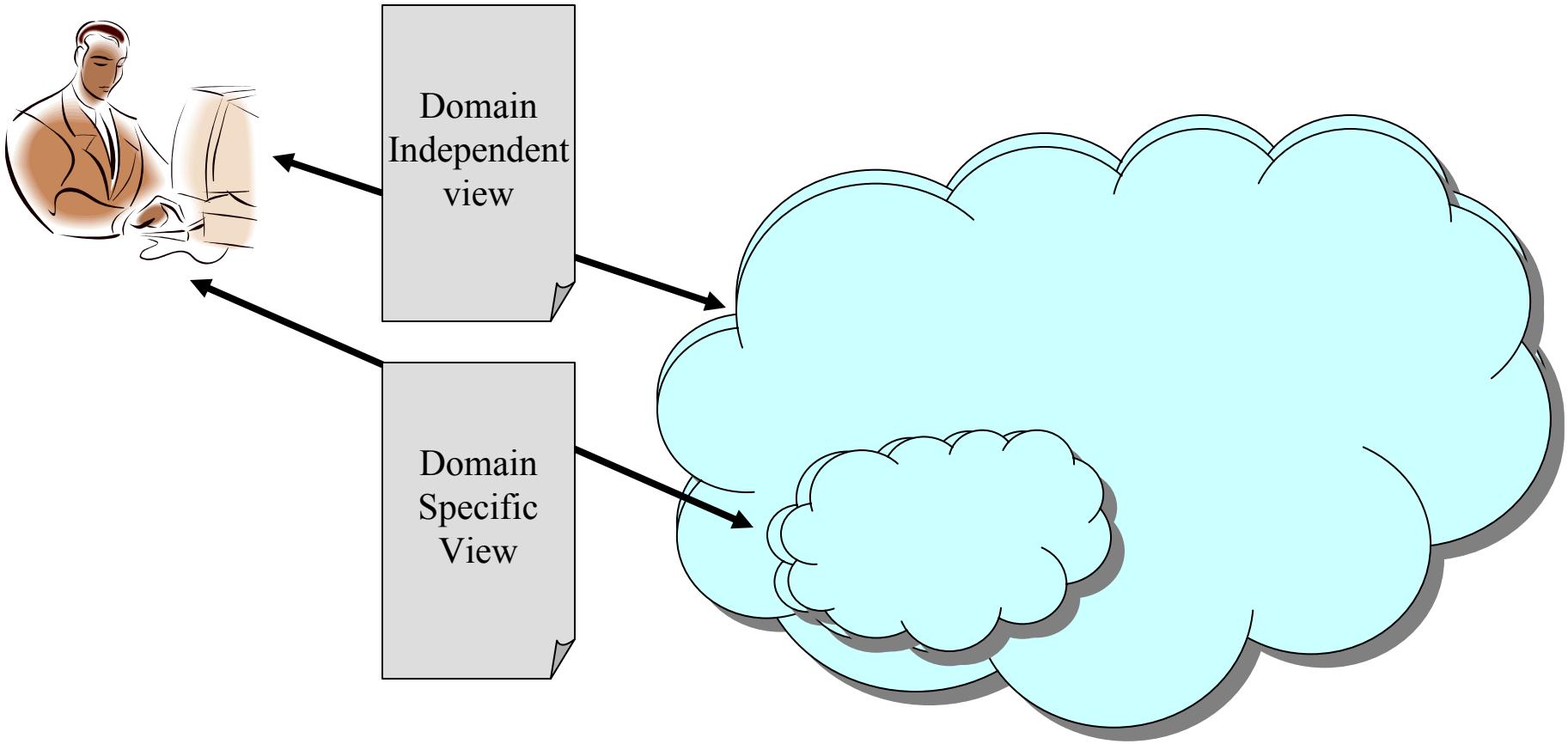




Progressive Metadata Models: Drill-Down Searching Paradigm

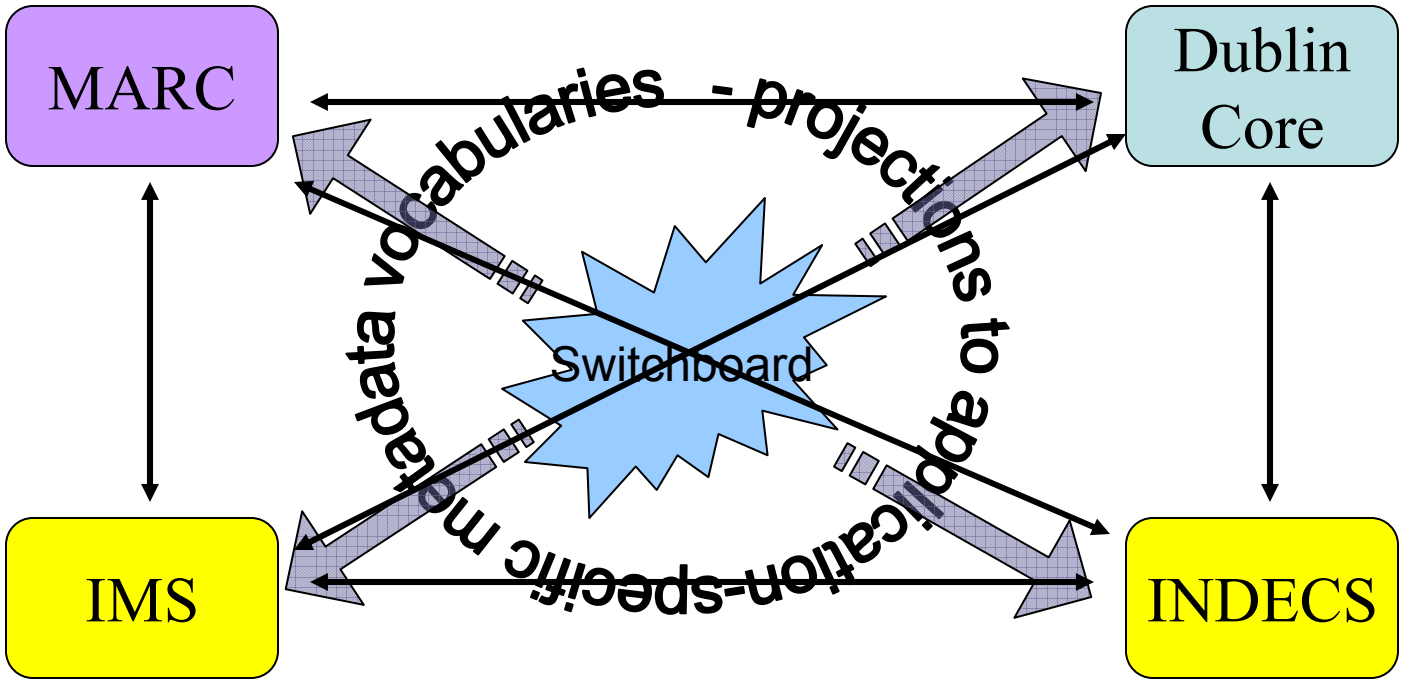
- Moving along a specificity spectrum
- Inter-domain vs. intra-domain terms, models, query mechanisms

Drill-down search paradigm



What is the Dublin Core (3)?

- A cross-domain switchboard for interoperable metadata?



What is Dublin Core (4)?

- A vocabulary for resource description
  - Maintained by an agency
  - Assigned unique names (URIs)
  - Evolves over time
  - http://dublincore.org/documents/dcmi-terms/
- A model for resource description
  - DCMI abstract model
    - http://dublincore.org/documents/abstract-model/
  - Why an abstract model?
    - Because encoding evolves over time and is technologically based

DCMI resource model

- each *resource* that we want to describe has zero or more *properties*
- a *property* is a specific aspect, characteristic, attribute or relation used to describe a *resource*
- each *property* has one or more *values*
- each *value* is a *resource* (the physical or conceptual entity that is associated with a *property* when it is used to describe a *resource*)

but what is a resource?

- W3C/IETF definition of resource is

"…anything that has identity. Familiar examples include an electronic document, an image, a service (e.g., "today's weather report for Los Angeles"), and a collection of other *resources*. Not all *resources* are network "retrievable"; e.g., human beings, corporations, and bound books in a library can also be considered *resources*."

- i.e. a *resource* is "anything"
  - physical things (books, cars, people)
  - digital things (Web pages, digital images)
  - conceptual things (colours, points in time)

UKOLN

Constrain for DCMI

- but… this seems to be too wide for the things we can describe with DC!
  - can we really describe people using DC?
  - do people have titles and subjects?
- no… in general we only use DC to describe a sub-set of all *resources*
- anything covered by the DCMIType list…
  - Collection, Dataset, Event, Image (Still or Moving), Interactive Resource, Service, Software, Sound, Text, Physical Object

DCMI resource model (2)

- each *resource* may be a member of one or more *classes*
- each *class* may be related to one or more other *classes* by a refines (sub-class) relationship
  - the two *classes* share some *semantics* such that all *resources* that are members of the *sub-class* are also members of the related *class*

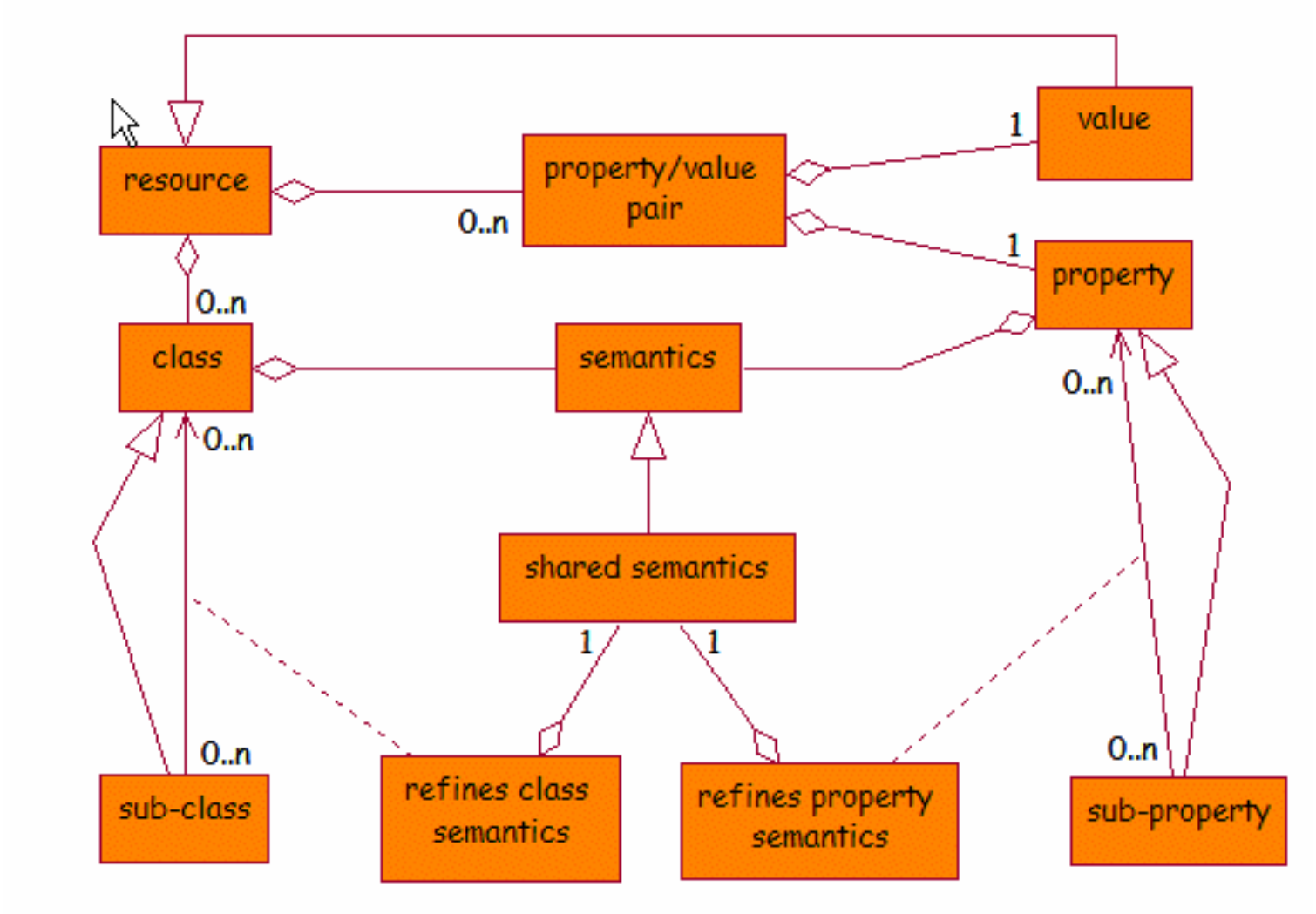
where the *resource* is the *value* of a *property*, the *class* is referred to as a *vocabulary encoding scheme*

UKOLN

DCMI resource model (3)

- each *property* may be related to exactly one other *property* by a refines (sub-property) relationship
  - the two *properties* share some *semantics* such that all valid *values* of the *sub-property* are also valid *values* of the related *property*

DCMI Resource Model



UKOLN

DCMI description model

- a *description* is made up of
  - one or more *statements* (about one, and only one, *resource*) and
  - zero or one *resource URI* (a URI reference that identifies the *resource* being described)
- each *statement* is made up of
  - a *property URI* (that identifies a *property*),
  - zero or one *value URI* (that identifies a *value* of the *property*),
  - zero or one *encoding scheme URI* (that identifies the *class* of the *value*) and
  - zero or more *value representations* of the *value*

DCMI description model (2)

- each *property* is an attribute of the *resource* being described
- each *property URI* may be repeated in multiple *statements*
- the *value representation* may take the form of a *value string*, a *rich value* or a *related description*

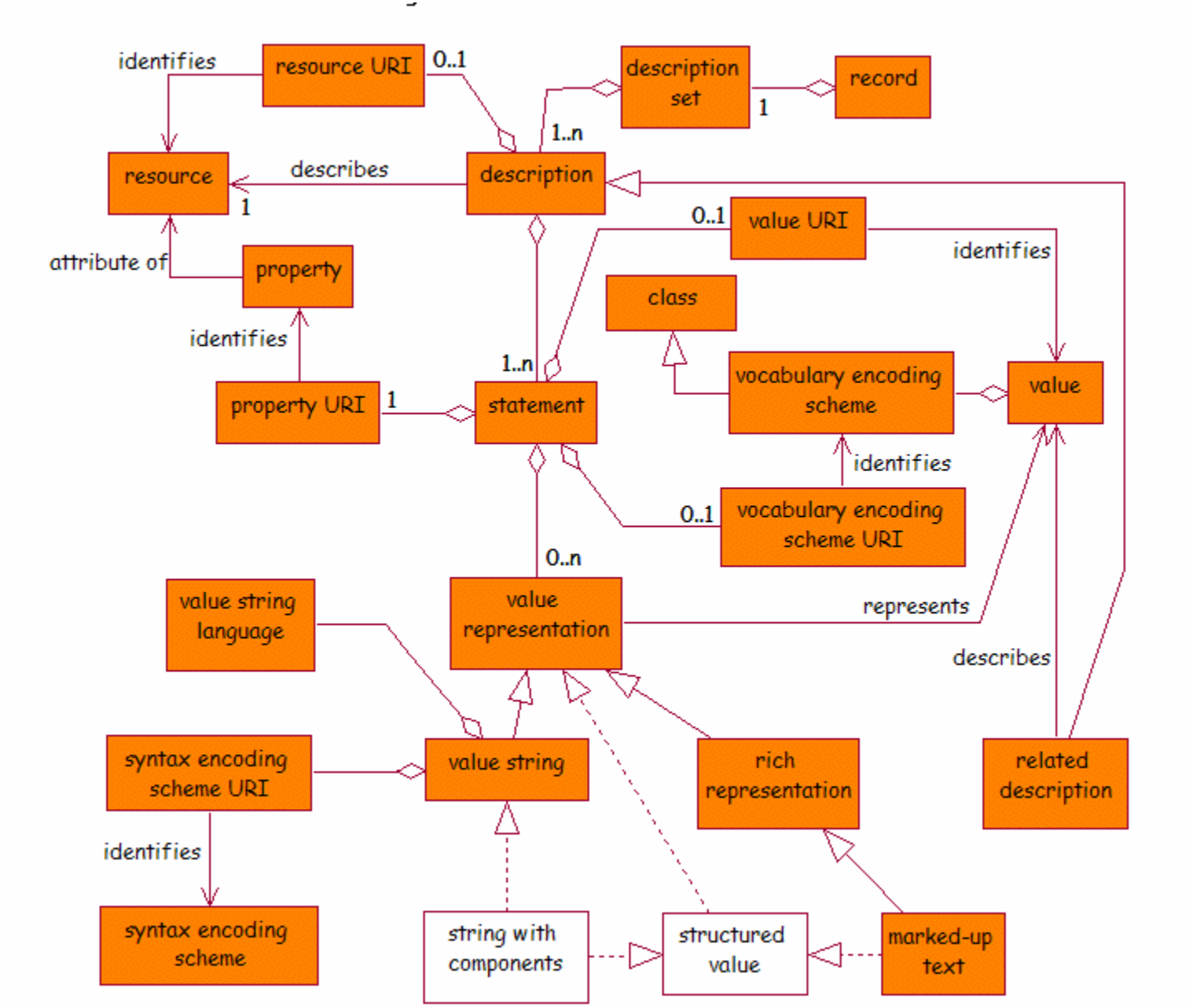
DCMI description model (3)

- each *value string* is a simple, human-readable string that represents the *value* of the *property*
- each *value string* may have an associated *encoding scheme URI* that identifies a *syntax encoding scheme*
- each *value string* may have an associated *value string language* that is an ISO language tag (e.g. en-GB)

DCMI description model (4)

- each *rich value* is some marked-up text, an image, a video, some audio, etc. or some combination thereof that represents the *resource* that is the *value* of the *property*
- each *related description* is a description of (i.e. some metadata about) the *resource* that is the *value* of the *property*

DCMI Description Model



The 1:1 principle

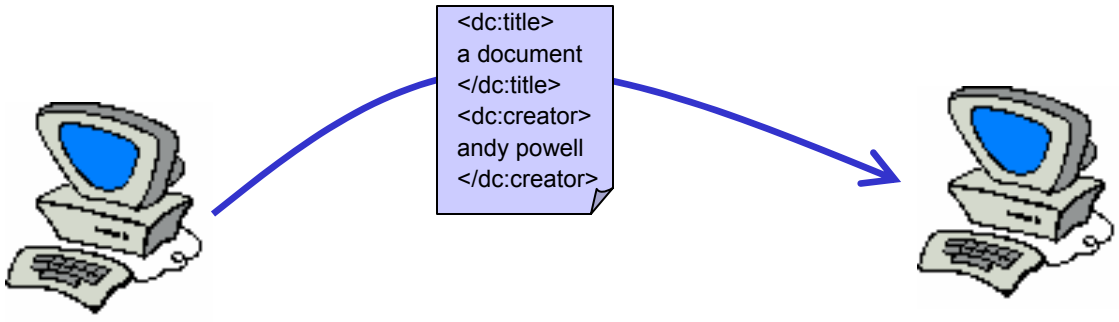
- notice that the model indicates that each *property* used in a *description* must be an attribute of the *resource* being described
- this is commonly referred to as the 1:1 principle - the principle that a DCMI metadata *description* describes one, and only one, resource
- however…

Description sets

- real-world metadata applications tend to be based on loosely grouped sets of *descriptions* (where the described *resources* are typically related in some way)
- known here as *description sets*
- for example, a *description set* might comprise *descriptions* of both a painting and the artist

DCMI records

- *description sets* are instantiated, for the purposes of exchange between software applications, in the form of metadata *records*
- each *record* conforms to one of the DCMI encoding guidelines (XHTML meta tags, XML, RDF/XML, etc.)



Values (again!)

- a *value* is the physical or conceptual entity that is associated with a *property* when it is used to describe a *resource*
  - the *value* of the DC Creator *property* is a person, organisation or service - a physical entitiy
  - the *value* of the DC Date *property* is a point in time - a conceptual entity
  - the *value* of the DC Coverage *property* may be a geographic region or country - a physical entity
  - the *value* of the DC Subject *property* may be a concept - a conceptual entity - or a physical object or person - a physical entity
- each of these entities is a *resource*

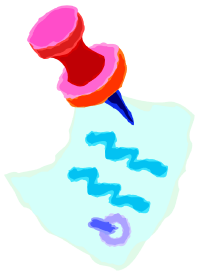
Simple DC record

- a simple DC record is a *record* that:
  - conforms to the abstract model,
  - comprises only a single *description*,
  - uses only the 15 *properties* in the Dublin Core Metadata Element Set,
  - makes no use of *value URIs*, *encoding schemes*, *rich values* or *related descriptions*.

UKOLN

A couple of notes…

- there is no guaranteed linkage between a *simple DC record* and the *resource* being described because the *resource URI* is optional
- such a linkage may be made by encoding the URI of the *resource* as the *value string* of the DC Identifier element, however this is not mandatory – **everything in DC is optional**
- while the *value string* of a *property* may look like a URI, there is nothing in the simple DC model that indicates this is the case



…at their own risk, implementations may choose to guess which *value strings* are URIs and which are not…

UKOLN

Qualified DC model

- a qualified DC record is a *record* that:
  - conforms to the DCMI abstract model,
  - contains at least one *property* taken from the DCMI Metadata Terms recommendation

A couple of notes…

- it is still the case that there is no guaranteed linkage between a *qualified DC record* and the *resource* being described!
- a linkage may be made by encoding the URI of the *resource* as the *value string* of the DC Identifier element, however this is not mandatory – **everything in DC is optional**

…where the *value* of a *property* is a URI, we can now indicate that it is a URI by using the 'URI' encoding scheme…

UKOLN

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 dumb-down and value dumb-down.
- each of these processes can be be approached in one of two ways
  - informed dumb-down
  - uninformed dumb-down

UKOLN

Dumb and dumberer

- **informed dumb-down** takes place where the software performing the dumb-down algorithm has knowledge built into it about the *property* relationships and *values* being used within a specific DCMI metadata application
- **uninformed dumb-down** takes place where the software performing the dumb-down algorithm has no prior knowledge about the *properties* and *values* being used

Dumb-down algorithm

	element	value
uninformed	ignore any *property* that isn't in the Dublin Core Metadata Element Set	use *value URI* (if present) or *value string* as new *value string*
informed	recursively resolve sub-property relationships until one of the 15 properties in the DCMES is reached, otherwise ignore	use knowledge of the *related descriptions* or the *value string* to create a new *value string*

…and in all cases:

- ignore any *related descriptions* and *rich values*,
- ignore any *encoding scheme URIs*.

UKOLN