CERN workshop on Innovations in Scholarly Communication (OAI5)

18-20 April 2007

# Tutorial 4 - Object models and object representation

**Julie Allinson and Mahendra Mahey**

Repositories Research Officers, UKOLN, University of Bath

**with Herbert van de Sompel**

Los Alamos National Laboratory

Supported by JISC MLA Museums, Libraries and Archives Council

# Order of play

- 09.00: Overview and introduction to the scenario
  - (Julie Allinson)
- 09.10: Exploring options for metadata modelling (JA)
- 09.45: Content packaging for complex objects
  - (Mahendra Mahey and Herbert van de Sompel)
- 10.30: Break
- 10.50 Content packaging for complex objects continued
  - (MM)
- 11.20: Concluding remarks and questions
- 11.30: Finish

# Overview and introduction

Julie Allinson

# Unpacking the tutorial title

- Object models and object representation
- We inherited this title!
- Focus is on **metadata**, simple, compound or complex **digital objects** and **content packaging** for **interoperability** across **scholarly communications,** with specific emphasis on (**institutional**) **repositories**

UKOLN

# Looking at …

- What are we talking about?
  - Dublin Core, MODS, DIDL, IMS CP, METS and more
- Why do we need to know about it?
- Where and when is it used?
- Who needs to know about it?
- How do we use it?

# What is a digital object?

- Digital objects
  - are anything that might be stored by a digital repository …
  - can be any media or semantic type, e.g. an image, an article, XML metadata record, PDF etc.
  - have (unique) identifier(s)
    - to be considered **Resources** as per the W3C web architecture they must be identified by a URI
  - convey information, digitally, i.e. they are not abstract concepts, or physical objects
    - these are things that metadata is also used to describe
    - or representations of these
  - could also be called information objects, or information resources (in W3C speak)
- To be useful for scholarly communication they should have associated metadata
  - metadata can be a digital object in its own right

UKOLN

# Compound and complex digital objects

- Aggregations of *related* digital objects gathered together to form a *logical* whole.
- The relationship may be purely structural (e.g. a book and its chapters)
- Complexity is added when we begin to think beyond the structural,
- to a richer set of relationships between digital objects
- and relationships with other kinds of resources (people, organisations, concepts, events etc.)
- Metadata and/or content packaging help us to express structure and relationships

UKOLN

# Examples – digital objects

- A PDF scholarly paper
- A JPEG image
- A scientific dataset

- Each uniquely identified by a URI
- and accessed from a repository or web server

- Simple digital objects?

UKOLN

# Examples – compound objects

- A book and its chapters
  - XML-encoded chapters and table of contents
  - metadata describing each chapter
  - content packaging wrapper enclosing all of the above
- An image in different resolutions
  - RAW image file
  - JPEG print size
  - thumbnail

# Towards a repository ecology – compound objects complex-ified

- A scholarly paper, with different versions, metadata, metadata describing the agents (author, publisher etc.), references to supplementary materials etc.
- An issue of an overlay journal built from distributed papers
- An eScience publication combining text and primary research data, simulations, statistical analysis etc.
- Examples can become more and more complex, if we want them to be
- The repository ecology – a way of examining how systems and services interact to support scholarly communication
  - digital objects are flowing around this ecology

UKOLN

# Scenario for the tutorial

- A single scenario to capture **some** of the issues raised here:
  - A conference paper with different versions
  - calling on additional, external, resources
  - re-used in other compound resources
- See handout

# Some areas for consideration

- Boundaries
  - what is and isn't part of a particular compound digital object
  - what are the relationships within the object
  - and beyond its boundary
- Context and use
  - This might dictate particular requirements
    - e.g. a preservation service needs access to the full datastream and a specific set of preservation metadata
    - a repository may want only the descriptive metadata and a reference
- Expertise and local requirements
  - This might dictate the choice of standard used

# Exploring options for metadata modelling

Julie Allinson

# Overview

- basic metadata semantics
- a metadata framework for interoperability
  - syntax
  - vocabularies
  - application profiles and application models
- metadata for compound / complex objects
  - the scenario

UKOLN

# Metadata : what?

- Data about data? … this isn't very helpful

- "Metadata consists of statements we make about resources to help us find, identify, use, manage, evaluate, and preserve them".

(Marty Kurth, tutorial on DC Semantics, 2006 http://dc2006.ucol.mx/program.htm)

UKOLN

# Why?

- Without metadata, data is useless
  - 'orange' – a fruit, a company, a colour, a password, an identifier, arbitrary text string?
  - '01234567890' – a telephone number, an identifier?
  - non-text resources

    e.g. an image

describes

**metadata:**
Joan Miro
'Chicago',
detail

- Functions, a selection
  - resource discovery (oai-pmh, rss, z39.50)
  - identifying and differentiating resources
  - contextual information
  - authenticating and evaluating
  - sharing information
  - geographic locations

UKOLN

# When? Where? Who?

- Metadata is everywhere
- and is used all the time
  - in business, in education, in libraries, on the Internet …
  - for local purposes
  - and for wider interoperability
- there are many different 'types' of metadata
  - descriptive metadata
  - rights metadata
  - administrative metadata
  - etc.

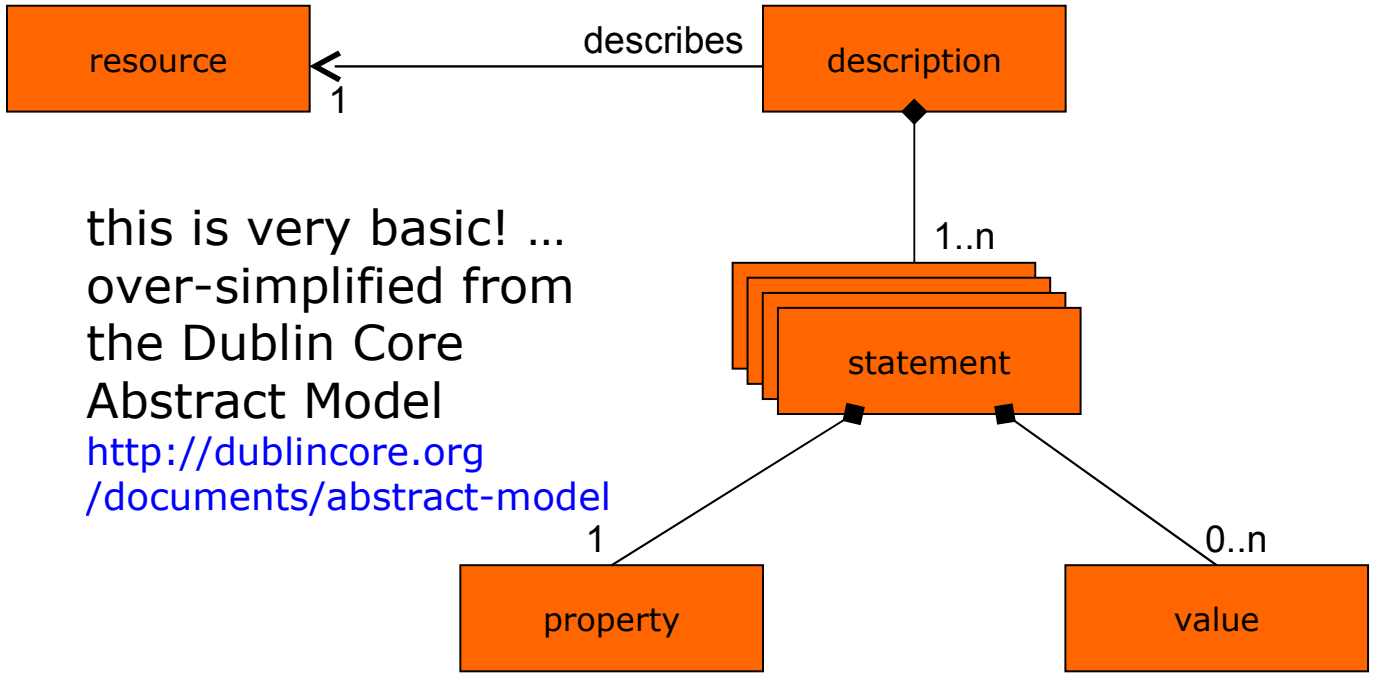
UKOLN

# How? metadata semantics …

- ***metadata*** describes ***resources***
- these ***resources*** can be ***digital***, ***physical*** or ***abstract*** things
- as a general principle metadata describes one, and only one, resource (the ***1:1*** rule)
- metadata descriptions contain ***statements*** about the resource
- a statement consists of a metadata ***property*** (aka, an *element*) and a ***value*** (*a property/value pair*)

UKOLN

# Metadata semantics diagram




this is very basic! …
over-simplified from
the Dublin Core
Abstract Model
http://dublincore.org
/documents/abstract-model

UKOLN

# Example, based on our scenario

- A conference paper [the resource]
- with the title [property] 'Signed metadata : method and application' [value]
- and the resource type [property] 'Text' or 'ConferencePaper' [value]

- **MODS metadata:**

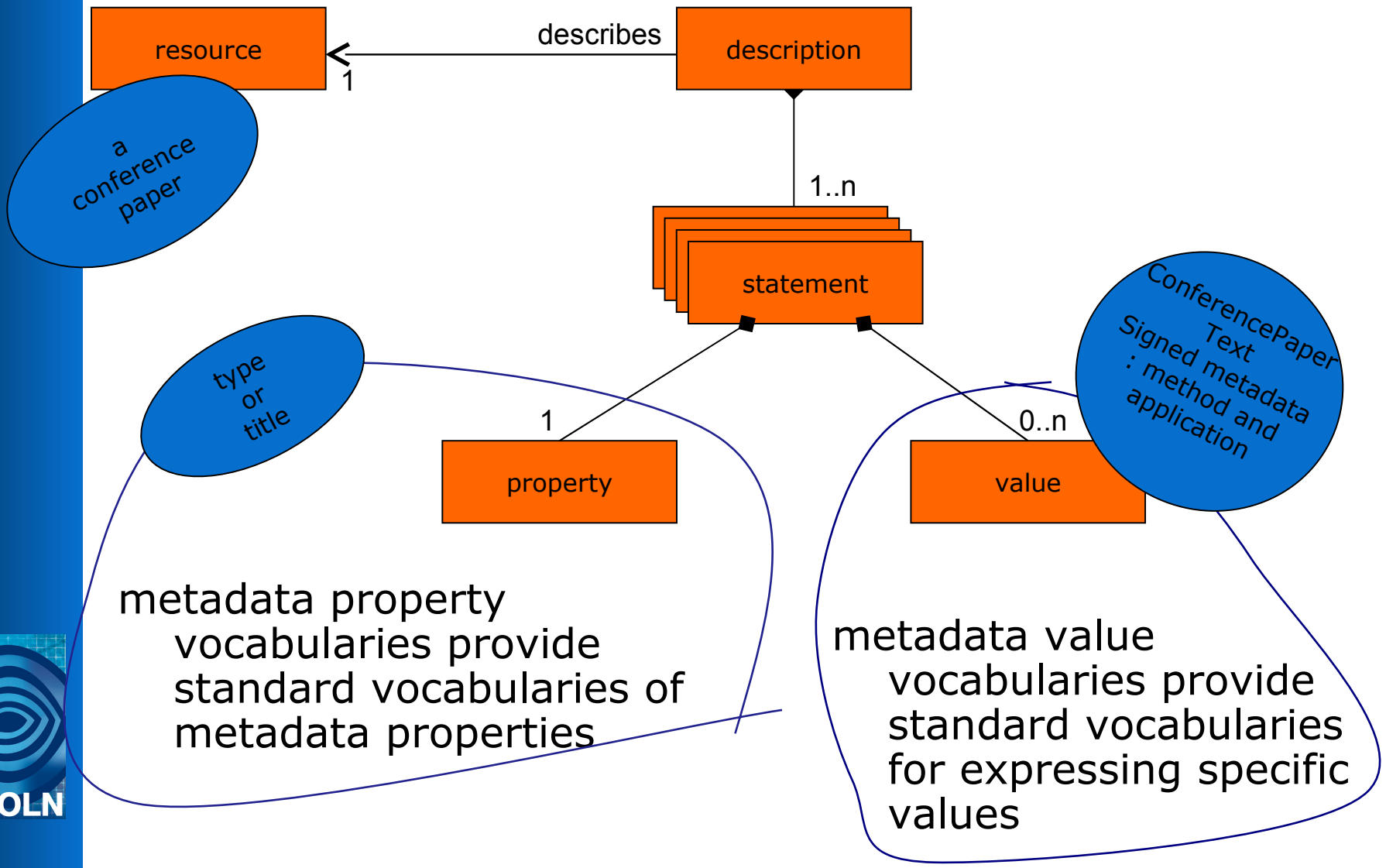
```
<titleInfo>
    <title>Signed metadata</title>
    <subTitle>method and application</subTitle>
</titleInfo>
<typeOfResource>Text</typeOfResource>
```

- **Dublin Core in XHTML:**

```
<meta name="DC.title" content="Signed metadata :
                    method and application" />
<meta name="DC.type" content="ConferencePaper" />
```

UKOLN

# Metadata example diagram



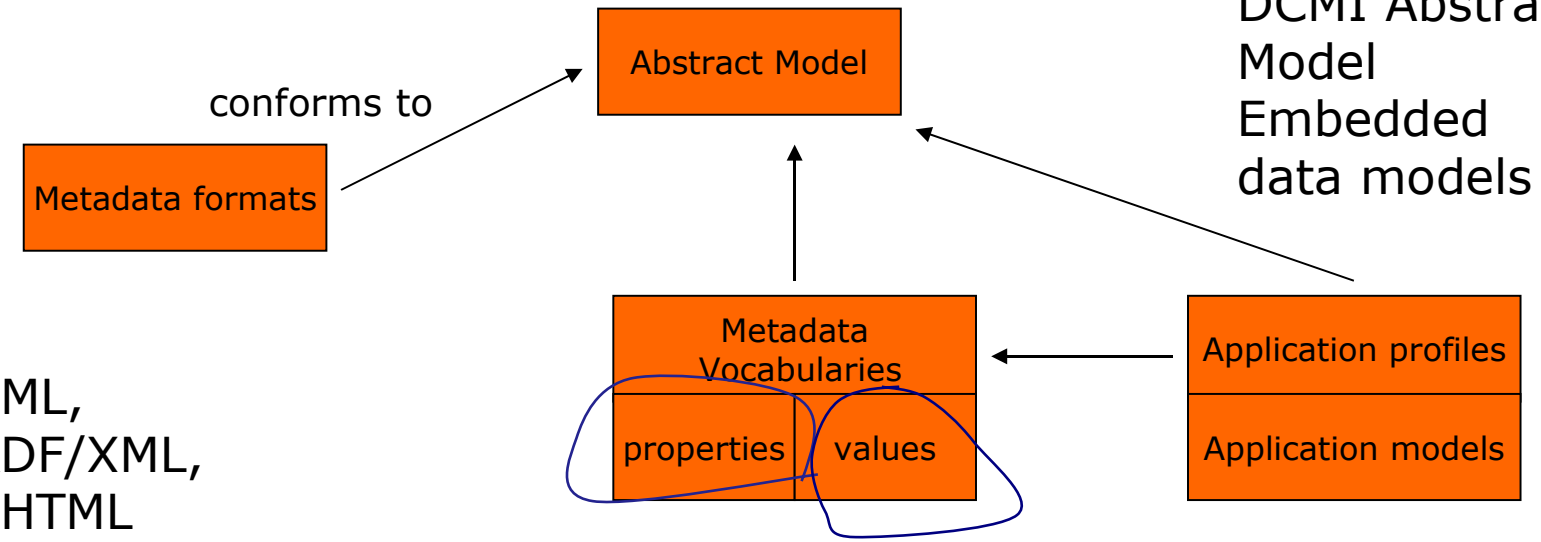

UKOLN

# Hang on, there's more …

- Metadata vocabularies are only one piece of the jigsaw
- For exchange, we also need machine-readable **Metadata Formats**
- **Application Profiles** draw together properties from one or more namespaces, for a particular purpose,
- **Abstract Models** provide a model, or a set of rules for how descriptions are constructed (this may be embedded in the metadata standard itself)
  - an abstract model can act as a mechanism for mapping between syntaxes
- together these give us the foundation for a metadata framework

UKOLN

# A metadata framework?




XML,
RDF/XML,
XHTML

DCMES,
MODS,
LOM,
ISO 19115,
e-GMS,
PREMIS,
FOAF

DCMI Type,
DDC,
LCSH,
TGN,
MARC Relator Codes

UK LOM Core,
CanCore,
RSLP Schema,
AGRIS,
Scholarly Works AP,
NDLTD (theses), DDF-MXD, DiVA

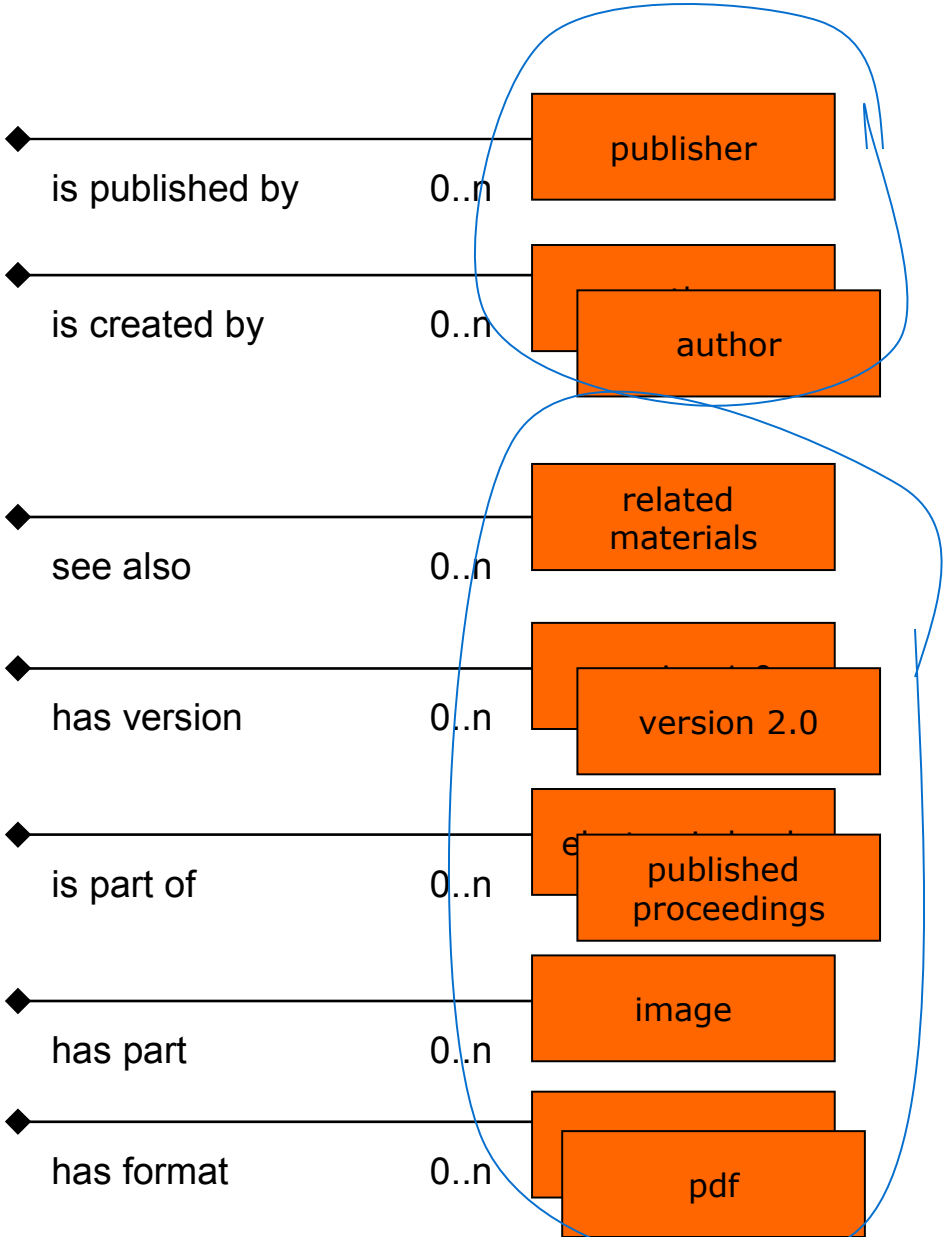
UKOLN

# Taking the next metadata step – from 'flat' to 'modelled'

- traditionally metadata has been seen a largely 'flat' set of metadata elements describing a single resource (e.g. a scholarly paper)
- but often the metadata is implicitly describing other resources (e.g. an author) but without explicitly recognising these as resources
- 'flat' metadata cannot adequately describe multiple resources and the relationships between them
- nor can it group together descriptions of resources that are closely related

UKOLN

# Some of the possible resources and relationships in our scenario




UKOLN

# Metadata is not flat!

- in Dublin Core
  - the **Dublin Core Abstract Model** introduces support for 'description sets'
  - it is for application profile developers to define the relationships they want to support
- in MODS
  - the <**modsCollection**> wrapper element can be used to group <**mods**> descriptions
  - <**relatedItem**> facilitates the capture of hasPart, isPartOf and seeAlso type relationships

UKOLN

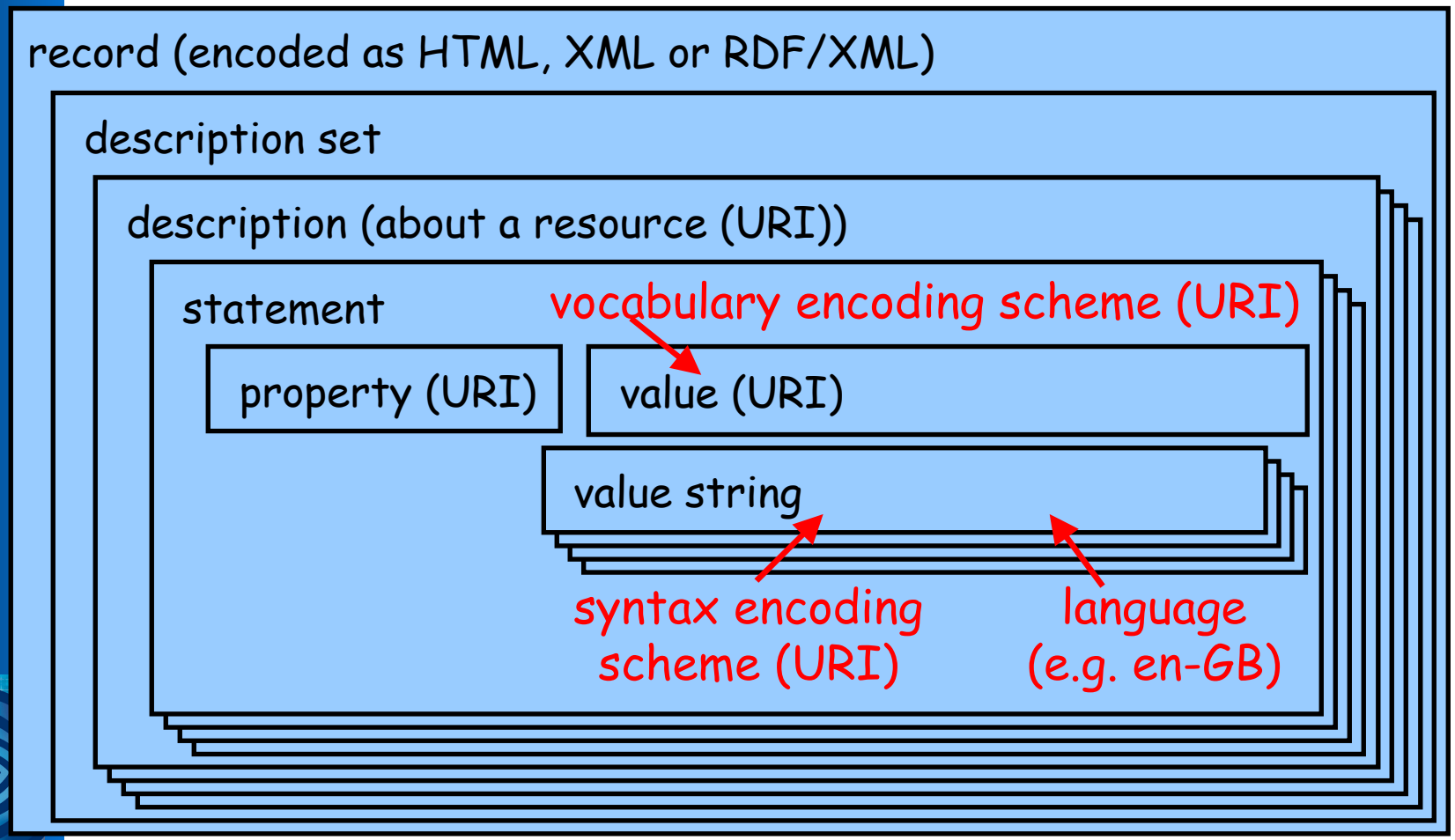
# Dublin Core as a case study

- The Dublin Core Abstract Model attempts to make explicit the model that underpins Dublin Core
- the DCAM starts from the central notion of a 'description set'
  - a set of 'descriptions' about a group of related 'resources'
  - where each description is about a single 'resource' (the 1:1 rule)
  - and where each 'description' comprises property/value pair 'statements'
  - 'description sets' are instantiated as 'records' (e.g. using XML, RDF/XML or XHTML) for the purpose of exchanging information between networked systems

UKOLN

Slide courtesy of Andy Powell, Eduserv Foundation
http://www.slideshare.net/eduservfoundation/the-dublin-core-abstract-model-a-packaging-standard

# Dublin Core Abstract Model summary




Slide courtesy of Andy Powell, Eduserv Foundation
http://www.slideshare.net/eduservfoundation/the-dublin-core-abstract-model-a-packaging-standard

# Dublin Core Abstract Model contd.

- the DCAM is open about the relationships between resources described in a description set
  - whole / part (book, chapter, section, page)
  - physical / digital (painting / digitised image)
  - object / human (document / author)
  - conceptual / physical (work / item)
- the relationships between things must be articulated in an 'application model' and captured using the properties specified in an 'application profile'

UKOLN

Slide courtesy of Andy Powell, Eduserv Foundation
http://www.slideshare.net/eduservfoundation/the-dublin-core-abstract-model-a-packaging-standard

# Scholarly Works Application Profile

- Also known as the Eprints Application Profile
- the application model says what things are being described
  - the set of **entities** that we want to describe
  - and the key **relationships** between those entities
- each entity and its relationships are described using an agreed set of properties
- the application profile describes these properties
- **model vs. Model** - the application model and the DCMI Abstract Model are completely separate
- the DCMI Abstract Model says what the descriptions 'look' like

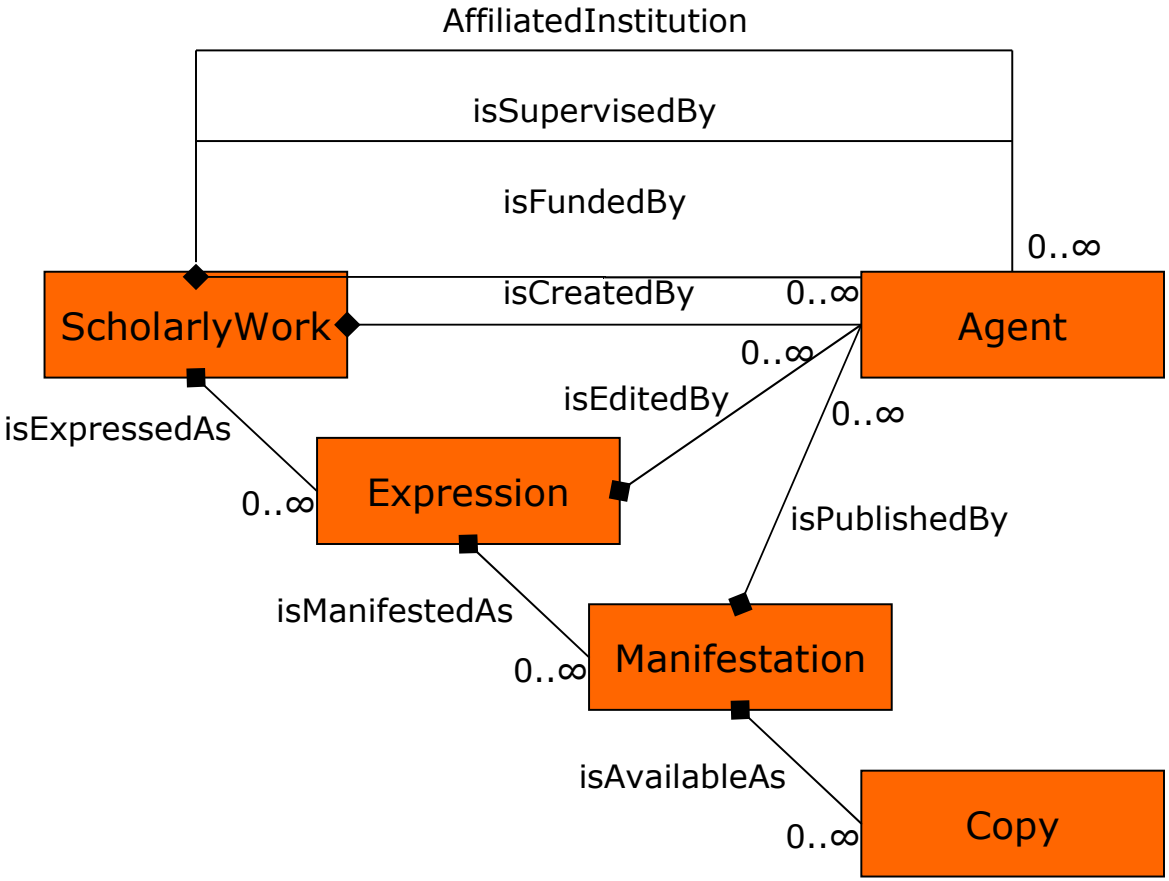
UKOLN

# FRBR for eprints

- FRBR (Functional Requirements for Bibliographic Records) provides the basis for our model
  - it's a model for the entities that ***bibliographic records*** describe
  - but we've applied it to ***scholarly works***
- FRBR is a useful model for scholarly works because it allows us to answer questions like:
  - what is the URL of the most appropriate copy (a FRBR item) of the PDF format (a manifestation) of the pre-print version (a expression) for this eprint (the work)?
  - are these two copies related? if so, how?

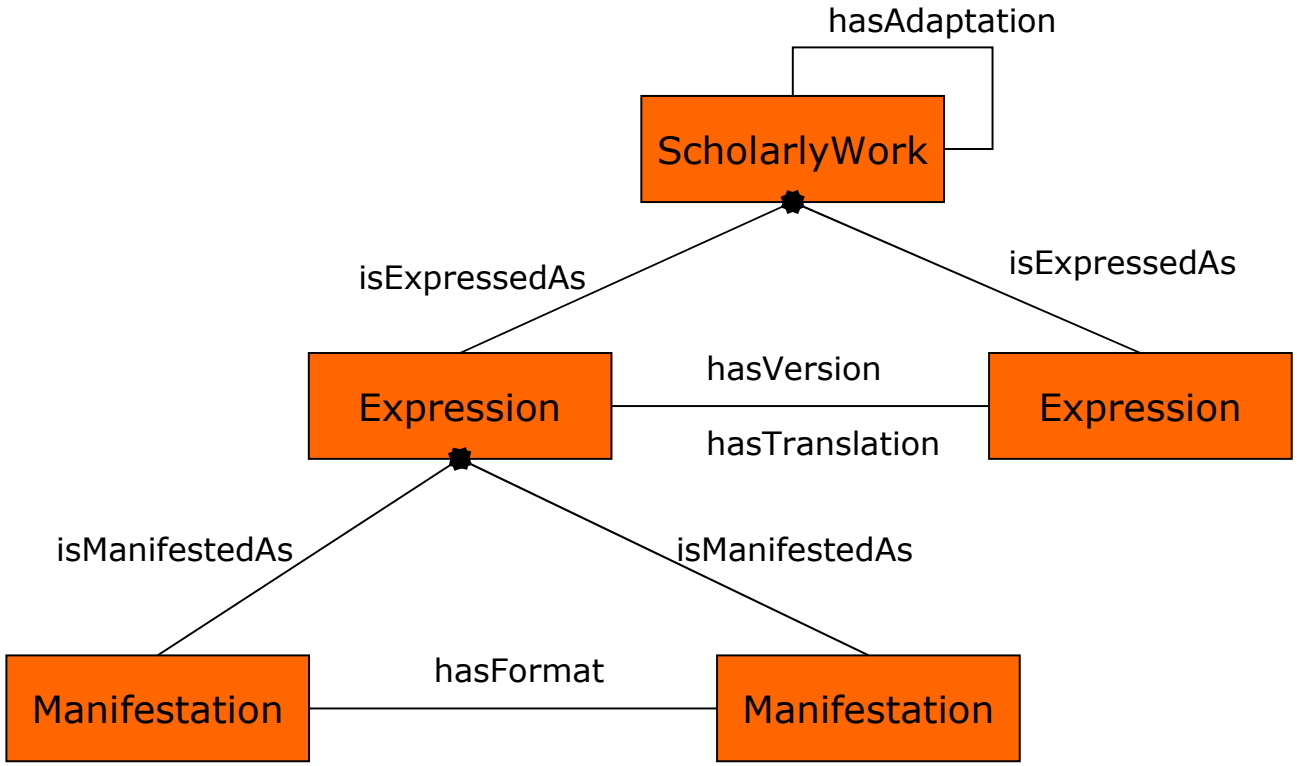
UKOLN

# the application model




UKOLN

# vertical vs. horizontal relationships




UKOLN

# Example properties

**ScholarlyWork:**
title
subject
abstract
affiliated institution
identifier

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

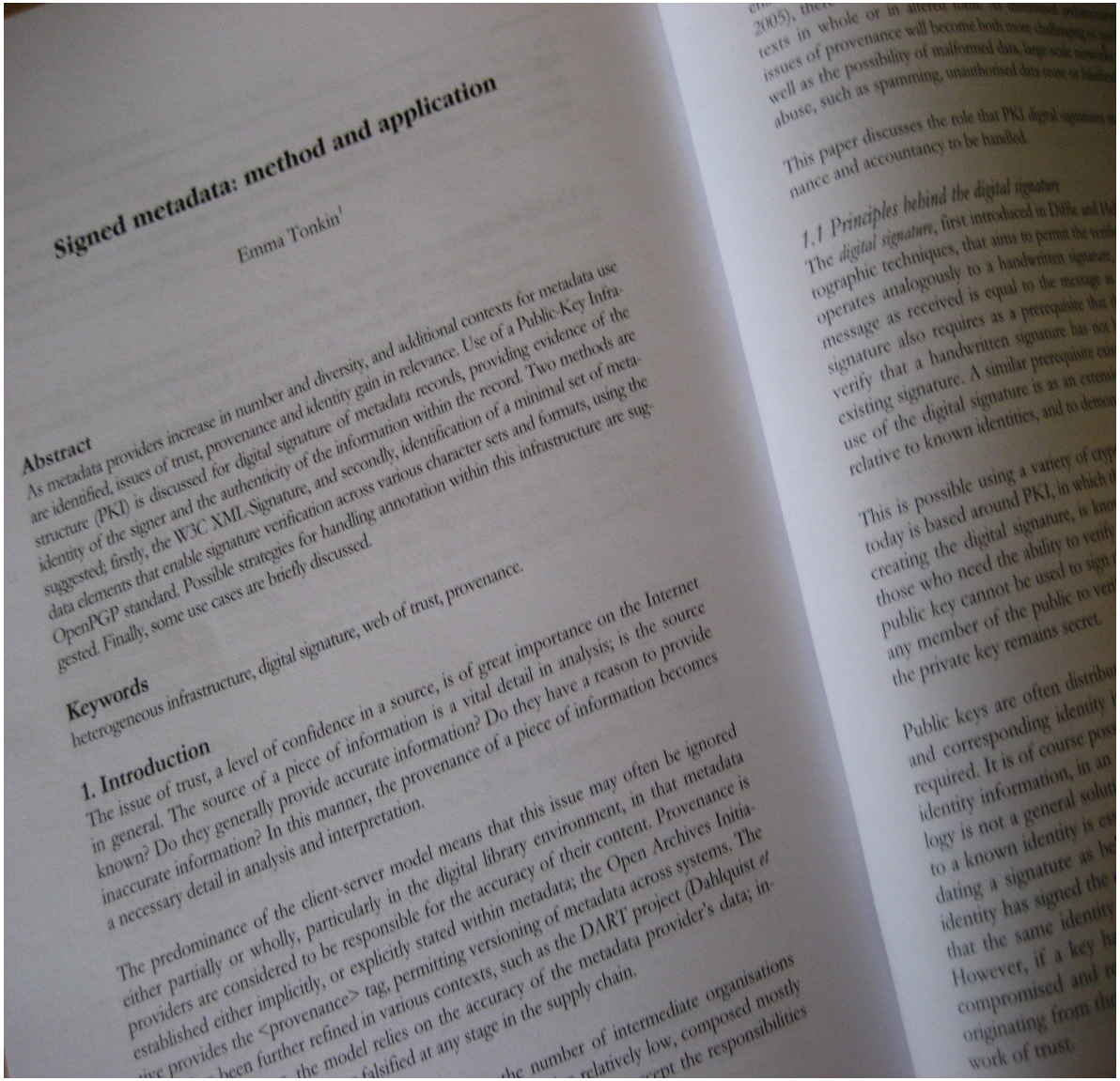
**Manifestation:**
format
date modified

**Copy:**
date available
access rights
licence
identifier

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

UKOLN

# Considering the scenario in DC



**Signed metadata: method and application**

Emma Tonkin[1]

**Abstract**

As metadata providers increase in number and diversity, and additional contexts for metadata use are identified, issues of trust, provenance and identity gain in relevance. Use of a Public-Key Infrastructure (PKI) is discussed for digital signature of metadata records, providing evidence of the identity of the signer and the authenticity of the information within the record. Two methods are suggested; firstly, the W3C XML-Signature, and secondly, identification of a minimal set of metadata elements that enable signature verification across various character sets and formats, using the OpenPGP standard. Possible strategies for handling annotation within this infrastructure are suggested. Finally, some use cases are briefly discussed.

**Keywords**

heterogeneous infrastructure, digital signature, web of trust, provenance.

**1. Introduction**

The issue of trust, a level of confidence in a source, is of great importance on the Internet in general. The source of a piece of information is a vital detail in analysis; is the source known? Do they generally provide accurate information? Do they have a reason to provide inaccurate information? In this manner, the provenance of a piece of information becomes a necessary detail in analysis and interpretation.

The predominance of the client-server model means that this issue may often be ignored either partially or wholly, particularly in the digital library environment, in that metadata providers are considered to be responsible for the accuracy of their content. Provenance is established either implicitly, or explicitly stated within metadata; the Open Archives Initiative provides the <provenance> tag, permitting versioning of metadata across systems. The ... been further refined in various contexts, such as the DART project (Dahlquist *et al* ... the model relies on the accuracy of the metadata provider's data; ... falsified at any stage in the supply chain.

... the number of intermediate organisations ... relatively low, composed mostly ... cept the responsibilities

issues of provenance will become both more challenging to ... well as the possibility of malformed data, large-scale ... abuse, such as spamming, unauthorised data reuse or ...

This paper discusses the role that PKI digital signatures ... nance and accountancy to be handled.

*1.1 Principles behind the digital signature*

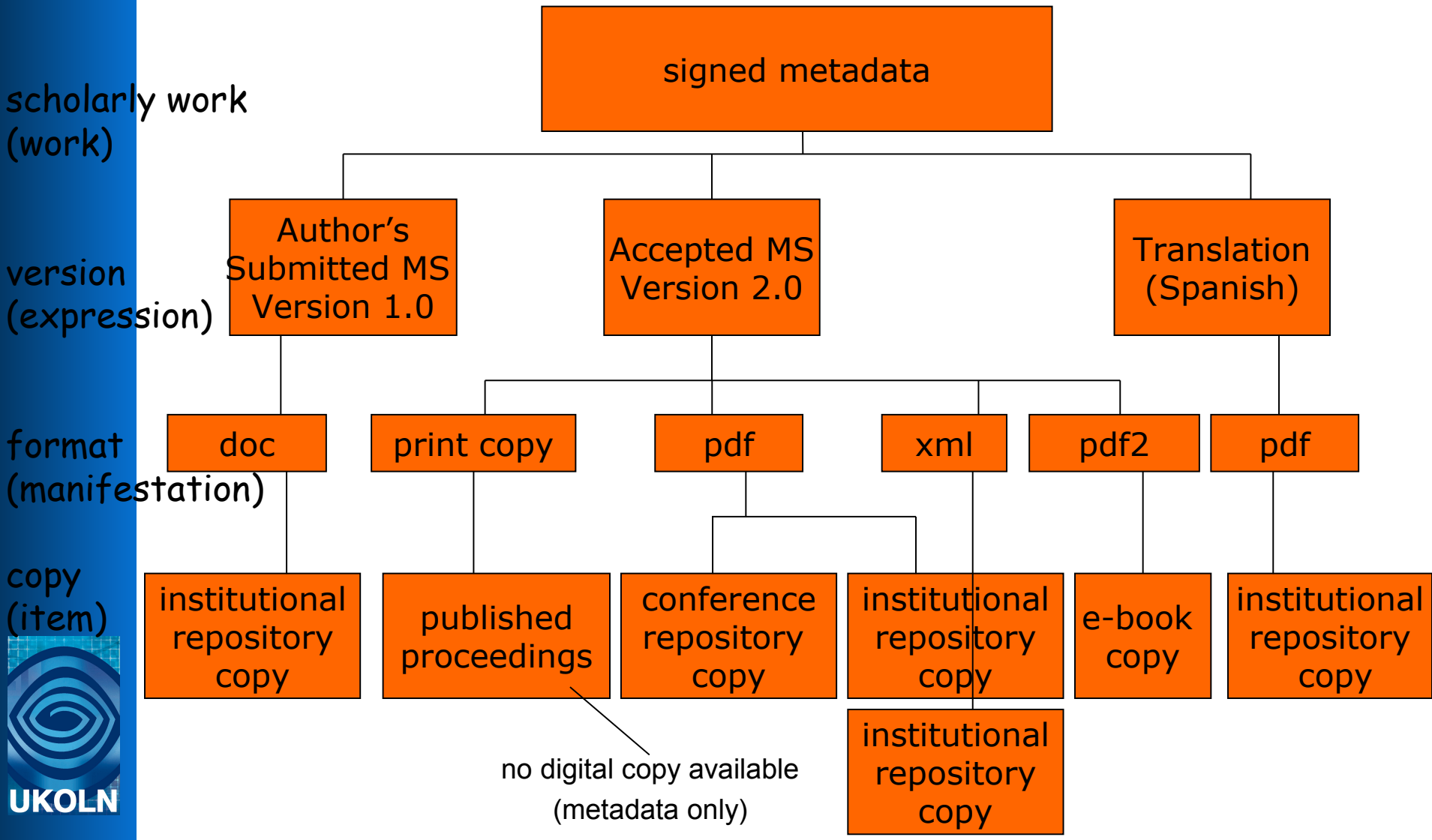
The *digital signature*, first introduced in Diffie and ... tographic techniques, that aims to permit the ... operates analogously to a handwritten signature, ... message as received is equal to the message as ... signature also requires as a prerequisite that ... verify that a handwritten signature has not ... existing signature. A similar prerequisite ... use of the digital signature is as an ... relative to known identities, and to ...

This is possible using a variety of ... today is based around PKI, in which ... creating the digital signature, is ... those who need the ability to ... public key cannot be used to ... any member of the public to ... the private key remains secret.

Public keys are often ... and corresponding ... required. It is of course ... identity information, in an ... logy is not a general ... to a known identity is ... dating a signature as ... identity has signed the ... that the same identity ... However, if a key ... compromised and ... originating from ... work of trust.

UKOLN

# Multiple expressions, manifestations and copies



# Boundaries of this profile

- created to fulfil a specific set of requirements, chiefly
  - richer metadata set & consistent metadata
  - unambiguous method of identifying full-text(s)
  - versions & most appropriate copy support
  - identification of open access materials
  - identification of the research funder and project code
- limited support for part/whole relationships
- and for related materials
- but it is modular and extensible
- and it fits well with the semantic web
- implementation with throw up new requirements

UKOLN

# MODS overview

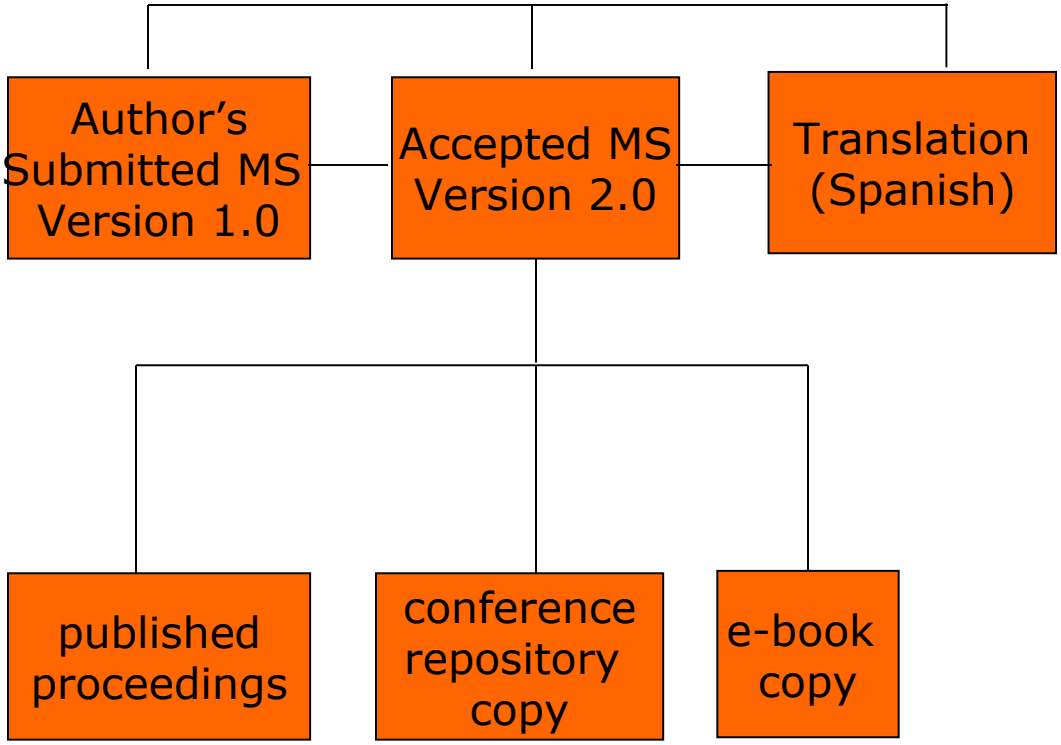
- The Metadata Object Description Schema (MODS) XML schema
- is intended to be able to carry selected data from existing MARC 21 records
- and to enable the creation of original resource description records
- includes a subset of MARC fields and uses language-based tags rather than numeric ones
- maintained by the Network Development and MARC Standards Office the Library of Congress with input from users
- a range of mappings and stylesheets are available

# Considering the scenario in MODS

<modsCollection>
can be used to wrap together several <mods> records

<mods> records for each version

<relatedItem> sub-element used within a <mods> description for see also, has part, is part of relationships




UKOLN

# MODS observations

- MODS has more built-in richness than Dublin Core
- it is closely aligned with the library community and MARC, yet is simpler and more user friendly than MARC
- it has some support for creating collections, or sets
- and for describing related materials (within a single metadata description)
- but, it doesn't have an 'abstract model' so is more difficult to map to other syntaxes
- and doesn't support the extensibility and flexibility of application profiles and the DCAM
- and is less interoperable with semantic web approaches

UKOLN

# Other approaches

- Danish DDF-MXD format
- created by DEFF to support
  - the DDF - Danish Research Database – the national CRIS
  - and the exchange of metadata
- Metadata model based on CERIF
  - Common European Research Information Format
  - designed to describe a much richer set of information captured by Research Information Systems

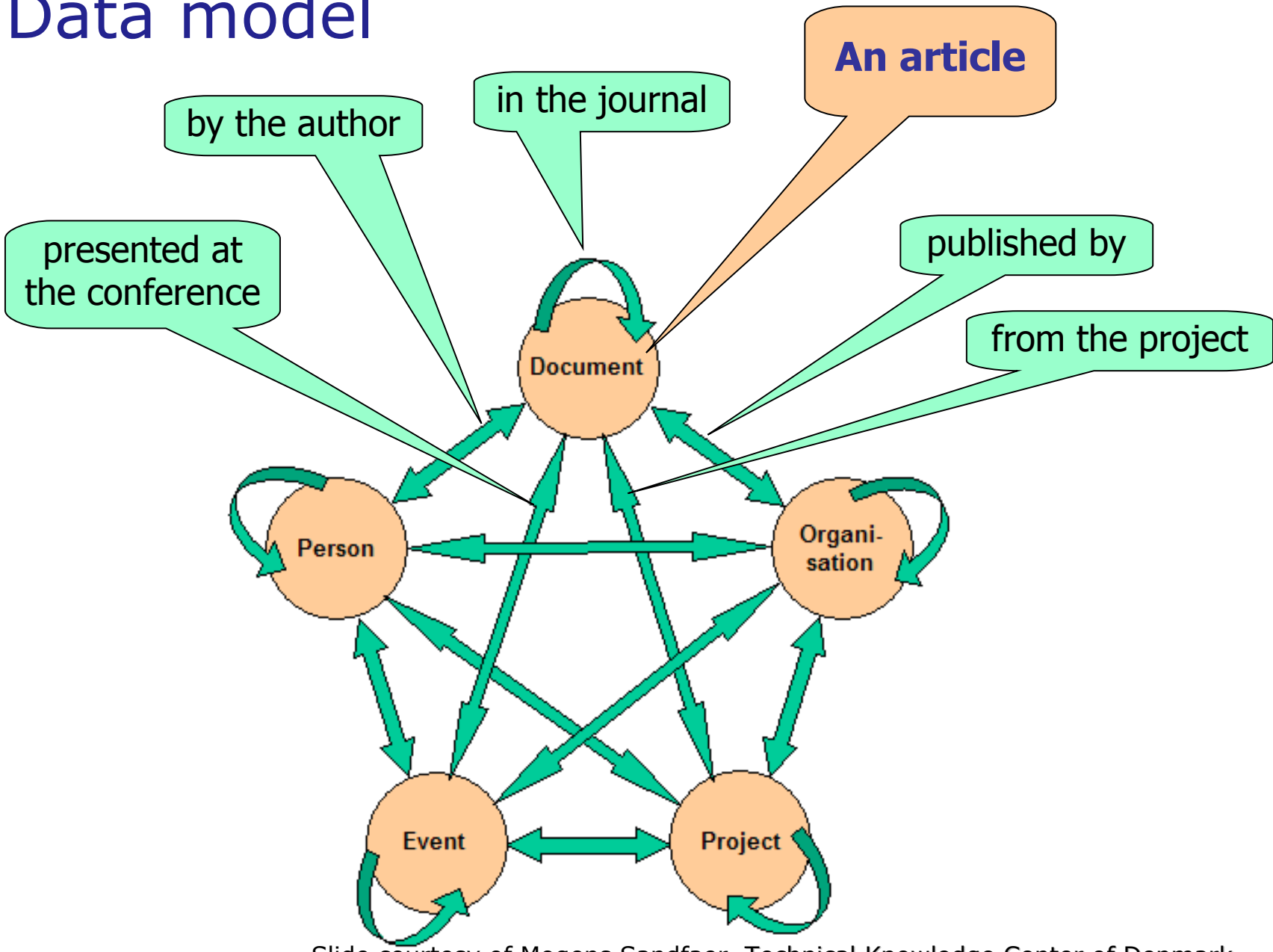
UKOLN

# Entities described by DDF-MXD

- Research database with repository
  - Persons (researchers and their competences)
  - Organisations (universities, institutes, labs etc.)
  - Projects (research and development projects)
  - Events (conferences, workshops etc.)
  - Documents (books, articles, eprints, slide shows, software, patents, data sets, simulations, learning objects, etc. etc.)


Slide courtesy of Mogens Sandfaer, Technical Knowledge Center of Denmark

# Data model




Slide courtesy of Mogens Sandfaer, Technical Knowledge Center of Denmark

UKOLN

# Thoughts about metadata

- Metadata is an essential element of the scholarly communications chain
- there are a number of existing property and value vocabularies
- application profiles can be developed for specific communities or purposes, using properties from existing vocabularies
- to facilitate interoperability, in a machine-to-machine context, metadata must be expressed in an encoding format/syntax such as XML
- adhering to an abstract model helps achieve understanding and agreement and provides a mechanism for mapping between syntaxes

UKOLN

# Towards content packaging

- Metadata is not limited to describing flat, single-entity items
- Metadata models and application profiles can be used to describe complex/compound objects
- and can offer some degree of content packaging 'by reference' (i.e. by providing a URI)
- Content packaging standards are another mechanism for gathering together multiple metadata records alongside the digital objects they describe
  - either 'by reference'
  - or 'by value' (i.e. by embedding the object within the package)

UKOLN

# Content packaging standards

Mahendra Mahey and
Herbert van de Sompel

# Final thoughts …

- Interoperability is achievable
- But communities need to work together
- Standard metadata formats, application profiles, abstract models and content packaging standards really can help
- And they can also interoperate with each other
  - For example, the RAMLET project <ieeeltsc.org/wg11CMI/ramlet/>
- Particularly if we agree things between us

UKOLN

# Final, final thoughts

- Don’t underestimate local expertise
- Don’t forget that our world is in a constant state of flux
- The future will see scholarly communication happening in an increasingly seamless and joined up way
  - For example the OAI-ORE project <www.openarchives.org/ore>
- Hopefully!

UKOLN