a dublin core application profile for describing scholarly works

JISC Repositories and Preservation Programme Meeting
5th July 2007, London

Julie Allinson

Repositories Research Officer

UKOLN, University of Bath

UKOLN is supported by:

www.ukoln.ac.uk

www.bath.ac.uk

UKOLN A centre of expertise in digital information management

the order of things

- where are we coming from? - background, scope and functional requirements …

- what and why? - the model, application profile and vocabularies

- where are we going?

where are we coming from?

background and scope

background and scope

- overall aim:
  - to offer a solution to issues with using simple DC for interoperability
  - to provide a richer metadata profile for the Intute repository search project
- development
  - summer 2007
  - funded (and scoped) by JISC
  - co-ordinated by Andy Powell and Julie Allinson; with Pete Johnston and others
- scope
  - Dublin Core properties as far as possible, plus other necessary elements
  - identifiers for the eprint and full-text(s); and for related resources
  - support subject access solutions (without mandating any)
  - additional properties to fulfil search/browse requirements
  - bibliographic citations and references citing other works

www.ukoln.ac.uk

UKOLN A centre of expertise in digital information management

terminology

- eprints, research papers and scholarly works are used synonymously for
  - a ''scientific or scholarly research text''
    (as defined by the Budapest Open Access Initiative
    http://www.earlham.edu/~peters/fos/boaifaq.htm#literatur
)
  - e.g. a peer-reviewed journal article, a preprint, a working paper, a thesis, a book chapter, a report, etc.
- the application profile is known as the eprints application profile by the DCMI community
- but it's often called the scholarly works application profile (SWAP) in the UK repositories community (to demonstrate its software independence!)

www.ukoln.ac.uk

UKOLN A centre of expertise in digital information management

what's wrong with simple DC?

defining the problem

```
<metadata>

  <dc:title> multiple titles, what language?
  <dc:creator> normalised form? person or org?
  <dc:publisher> normalised form? person or org?
  <dc:identifier> full-text or metadata? is it a uri?
  <dc:date> of what? modification? publication?
  <dc:format> is this a MIME type?
  <dc:subject> local keyword or controlled scheme?
  <dc:contributor> what did they contribute?
  <dc:language> is this an RFC 3066 value?
  <dc:relation> what relationship? is this a uri?
  <dc:rights> what does this tell me?
  <dc:source> is this a citation? or something else?

</metadata>
```

what do we need metadata to do?

functional requirements

functional requirements for describing scholarly works

- a richer metadata set
- consistent, good quality metadata
- unambiguous method of identifying full-text(s)
- distinguish open access materials from restricted
- support browse based on controlled vocabularies
- make use of OpenURL link servers and support citation analysis
- identify the research funder and project code
- identify the repository or other service making available the copy
- say when a copy of a scholarly work will be made available
- better search and browse options
- consider version identification and finding the most appropriate copy of a version
- support for added-value services

the requirements demanded a more complex model ...

www.ukoln.ac.uk

UKOLN A centre of expertise in digital information management

what and why?

the model, application profile and vocabularies

www.ukoln.ac.uk

UKOLN   A centre of expertise in digital information management

model : what's that?

- it's an entity-relationship model
- it says what 'things' we want to describe
  - the set of **entities**
  - and the key **relationships** between those entities
- several models already exist, e.g.
  - FRBR (Functional Requirements for Bibliographic Records)
  - CIDOC CRM for cultural heritage information
  - Common European Research Information Format (CERIF)
- FRBR provides the basis for our model
  - it's a model for the entities that ***bibliographic records*** are intended to describe and the relationships between them
  - it's working in a similar space to our modelling of ***scholarly works***
  - and it could have wider applicability

FRBR and eprints entities

- there are 4 key FRBR entities: Work, Expression, Manifestation and Copy
  - A **work** is a distinct intellectual or artistic creation. A work is an abstract entity
  - An **expression** is the intellectual or artistic realization of a work
  - A **manifestation** is the physical embodiment of an expression of a work.
  - An **item** is a single exemplar of a manifestation. The entity defined as item is a concrete entity.
- FRBR also defines additional entities - 'Person', 'Corporate body', 'Concept', 'Object', 'Event' and 'Place'
- and the relationships between entities

We use Scholarly Work to distinguish our refinement.

We use 'Copy' as a more appropriate entity for digital information

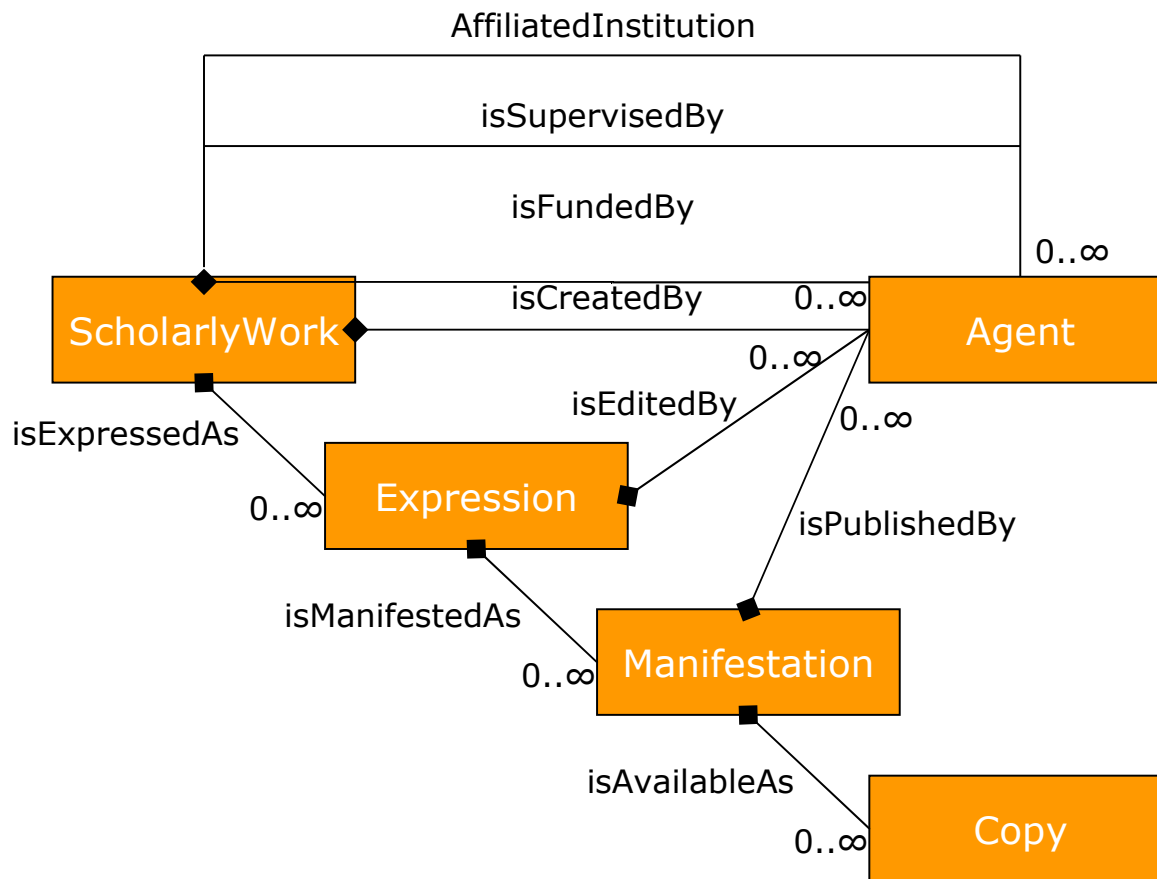We use 'Agent' to describe a Person or Organisation

www.ukoln.ac.uk

UKOLN   A centre of expertise in digital information management

the model in pictures

from model to where?

- the model defines the entities and relationships

- each entity and its relationships are described using an agreed set of attributes / properties

- this is where the model ends
  - it doesn't tell us where to get those properties from,
  - what vocabularies to use,
  - how to construct our descriptions,
  - or how to encode all of this

www.ukoln.ac.uk

UKOLN   A centre of expertise in digital information management

Dublin Core Abstract Model

- using Dublin Core was in-scope from the beginning
- the DCMI Abstract Model (DCAM) guides us on what our descriptions 'look like'
- it 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
- and each description contains statements about each attribute
  - using property-value pairs

www.ukoln.ac.uk

UKOLN   A centre of expertise in digital information management

application profile

- relationships and attributes are captured as metadata properties in the application profile
  - contains recommendations, cataloguing/usage guidelines and examples
  - little is mandatory (identifier and title)
  - structured according to the entities in the model
  - re-uses properties from existing schemes
    - dc, dcterms, foaf, MARC relators
  - introduces new 'eprint' properties
  - supported by various value vocabularies

www.ukoln.ac.uk

UKOLN   A centre of expertise in digital information management

example properties

ScholarlyWork:
title (dc)
subject (dc)
abstract (dcterm
affiliated institu
identifier (dc)
funder (marc)
grant number (
has adaptation

Expression:
title (dc)
date available (dcterms)
status (new)
version number (
language (dc)
genre / type (dc)
copyright holder (new)
bibliographic citation (dc
identifier (dc)
has version (new)
has translation (new)

Agent:
name (foaf)
type of agent (new)
date of birth (foaf)
mailbox (foaf)
homepage (foaf)
identifier (dc)

Manifestation:
format (dc)
date modified (dcterms)

Copy:
date available (dcterms)
access rights (dcterms)
licence (dcterms)
identifier (dc)

www.ukoln.ac.uk

UKOLN   A centre of expertise in digital information management

enough with the theory

what does this actually mean for repositories?

www.ukoln.ac.uk

UKOLN   A centre of expertise in digital information management

revisiting the functional requirements

the model and application profile mean we can support this …

- a richer metadata set
- consistent, good quality metadata
- unambiguous method of identifying full-text(s)
- distinguish open access materials from restricted
- support browse based on controlled vocabularies
- make use of OpenURL link servers and support citation analysis
- identify the research funder and project code
- identify the repository or other service making available the copy
- say when a copy of a scholarly work will be date available
- better search and browse options
- consider version identification and finding the most appropriate copy of a version
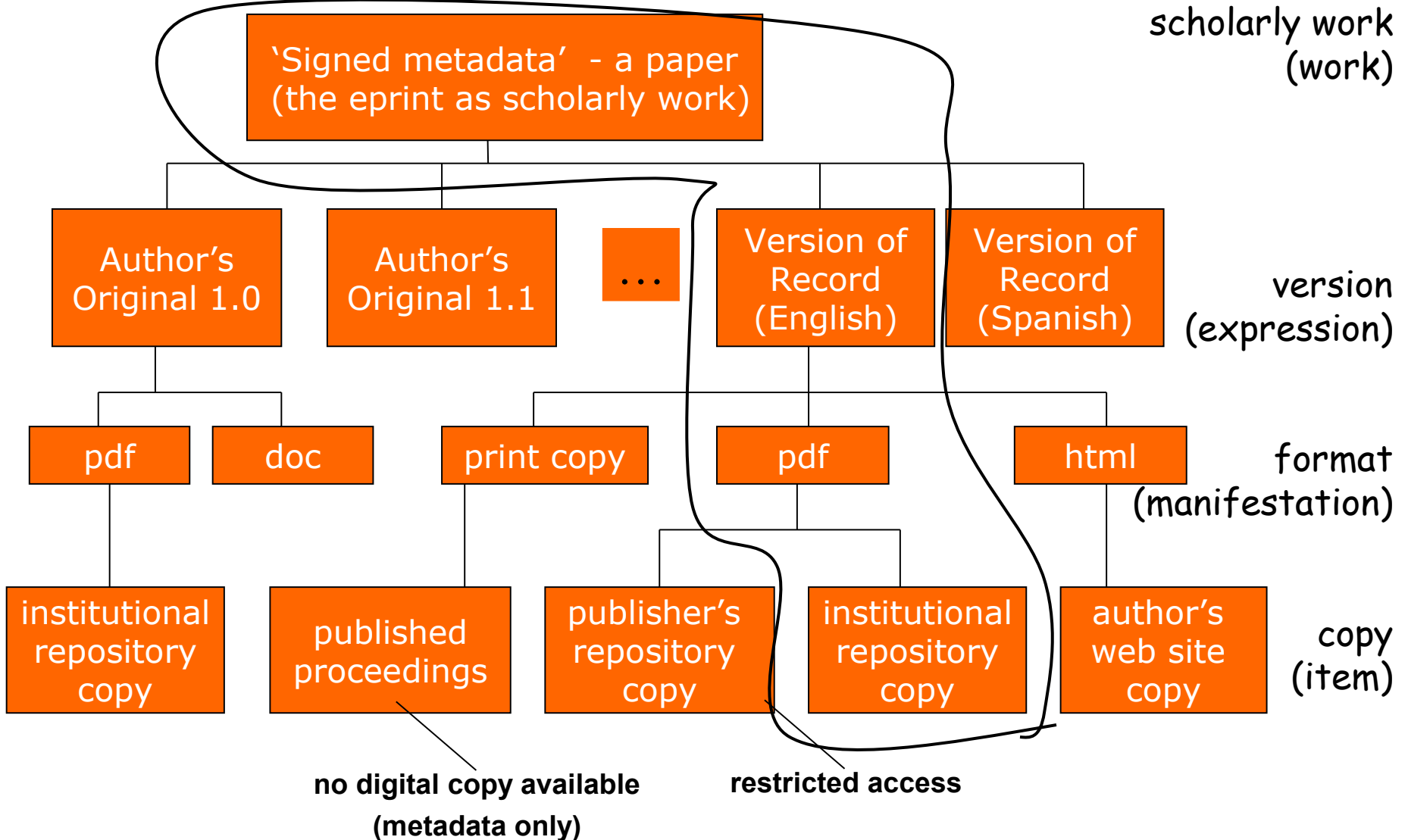- support for added-value services

www.ukoln.ac.uk

UKOLN A centre of expertise in digital information management

an example

scholarly work
(work)

'Signed metadata' - a paper
(the eprint as scholarly work)

Author's
Original 1.0

Author's
Original 1.1

...

Version of
Record
(English)

Version of
Record
(Spanish)

version
(expression)

pdf

doc

print copy

pdf

html

format
(manifestation)

institutional
repository
copy

published
proceedings

publisher's
repository
copy

institutional
repository
copy

author's
web site
copy

copy
(item)

**no digital copy available
(metadata only)**

restricted access

thoughts on the approach …

- this approach is guided by the functional requirements identified and the primary use case of richer, more functional, metadata
- it makes it possible to group together descriptions
- and therefore to rationalise 'traditional' and 'modern' citations
  - traditional citations tend to be made between eprint 'expressions'
  - hypertext links tend to be made between eprint 'copies' (or 'items' in FRBR terms)
- a complex underlying model may be manifest in relatively simple metadata and/or end-user interfaces
- the application profile is for metadata ***exchange***, it is not a blueprint for local metadata (but it can help)
- existing eprint systems may well capture this level of detail – but use of simple DC stops them exposing it to others!

www.ukoln.ac.uk

UKOLN   A centre of expertise in digital information management

what about interoperability?

- xml format and schema allows eprint description sets to be encoded, shared over oai-pmh, searched using SRU/W etc.
- for this exchange to happen we need
  - deployment by developers
  - deployment by repositories
  - consumption and use by services
- dumb-down
  - we still need to be able to create simple DC descriptions
  - we have guidelines for dumbing-down to separate simple DC descriptions of the ScholarlyWork and each Copy
    - simple DC about the ScholarlyWork corresponds to previous guidance
    - simple DC about each Copy useful for getting to full-text, e.g. by Google

www.ukoln.ac.uk

UKOLN   A centre of expertise in digital information management

where are we going?

- community acceptance and implementation are ongoing …
- more application profiles funded by JISC following a similar approach …