Eprints Application Profile

Open Scholarship 2006

University of Glasgow

Wednesday Oct 18th

14.30 - 17.00

Julie Allinson (UKOLN, Uni. of Bath)
Andy Powell (Eduserv Foundation)

Agenda

- background, rationale and functional requirements
- the model
- the application profile and vocabularies
- dumb-down issues
- next steps
- discussion

eduserv

UKOLN

Background, rationale and functional requirements

Julie

Background and rationale

www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_

- JISC-funded
- scope defined by JISC
- overall aim
  - to offer a solution to metadata issues identified in Eprints UK project, and by others (e.g. PerX project)
  - to provide a richer metadata profile for the Intute repository search service
- coordinated by Andy Powell (Eduserv Foundation) and Julie Allinson (UKOLN, Repositories Research Team)
  - Working Group / Feedback Group
  - Wiki for documentation
  - Email list for discussion

  www.jiscmail.ac.uk/lists/EPRINTS-APPLICATION-PROFILE.html

Scope

- Metadata:
  - In scope: DC elements plus any additional elements necessary
  - Out of scope: other metadata formats
- Identifiers:
  - In scope: Identifiers for the eprint and full-text(s); related resources etc
  - Out of scope: Other uses of identifiers
- Controlled vocabularies:
  - In scope: Hospitable to the use of a variety of subject access solutions
  - Out of scope: decisions on terminology solutions
- Complex objects:
  - In scope: Understanding of existing work; prioritising requirements
  - Out of scope: decisions on how to model complex objects
- Additional search entry points
  - In scope: additional properties to fulfil requirements
- Citations and references
  - In scope: Bibliographic citations references citing other works
  - Out of scope: Citation analysis solutions

Issues with simple DC (1)

- what's the problem with using simple DC to describe eprints?
- the ePrints UK project identified technical barriers to successful aggregation of metadata from institutional repositories
  - issues with the quality of metadata
  - the consistency of metadata
  - the handling of complex objects
  - the lack of a common approach to linking to full-text
- the ePrints UK guidelines on 'Using simple Dublin Core to describe eprints' were not widely implemented

Issues with simple DC (2)

- difficult to differentiate 'works/expressions' from 'manifestations/items' – which does dc:identifier identify?
  - in ePrints UK guidelines, dc:identifier used to identify 'work/expression' and dc:relation identifies 'manifestation/item'
  - dc:relation may be used for other resources (e.g. cited works) - ambiguity in the metadata record
  - software applications can't move reliably from the metadata record to the full-text
- other issues:
  - no means of knowing if full-text is freely available online or subject to access restrictions
  - can't distinguish between people and organistations
  - dates are ambiguous
  - subject vocabularies are not identified

Stakeholders

- Intute repository search project (JISC-funded)
- Prospero interim repository project (JISC-funded)
- repository software developers (GNU eprints, DSpace, Fedora)
- repository managers/administrators
- also:
  - users of the search service
  - depositors
  - JISC
  - other funding bodies
  - other UK regional and national services
  - DCMI community
  - global repositories community

Deliverables

- Functional Requirements Specification
- Entity-Relationship Model
- Eprints Application Profile
- Cataloguing/Usage Guidelines
- Plan for Community Acceptance and Take-up

Functional requirements

- why?
  - to find out what already exists, and
  - what the community wants
  - to engage the community in uptake
- how?
  - existing practice/application profiles/standards
  - scenarios and use cases
  - eprints UK project conclusions
  - working group, feedback group, wider community engagement

Primary use case

- primary use case
  - to develop an application profile for eprints to be used by the Intute UK repositories search service to aggregate content from repositories
- scenarios
  - aggregator search service needs consistent metadata
  - user wants to search or browse by a range of elemente, including journal, conference or publication title
  - user wants to be sure they have the latest version
  - repository wants to group together different versions
  - aggregator wants to offer added-value services

Requirements (1)

- provide a richer set of metadata than is possible with simple DC
- facilitate the creation and sharing of consistent metadata
- application profile should be sustainable, extensible and robust enough to support future added-value services
- implement an unambiguous method of identifying full-text(s)
- enable identification of metadata-only records
- offer a preliminary recommendation for version identification
- support navigation between different 'versions'
- support identification of the most appropriate or latest Copy of a discovered version
- support search of any, or all, elements, particularly of title, author, description, keyword
- support browse by any element, as required
- support title changes between expressions and the main Eprint (Scholarly Work)
- facilitate identification of open access materials

Requirements (2)

- support subject browse based on knowledge of controlled vocabulary
- support filtering of search results and browse tree - for example, by type, publisher, date range, status and version.
- enable movement from search results and browse tree to available copies
- support filtering of available copies by format
- enable movement from search results and browse tree to OpenURL link server
- support citation analysis between expressions
- be compatible with dc-citation WG recommendations
- provide for an authoritative form of Agent names, to include personal names (authors) and corporate names (publishers, funders)
- enable the author name, as it appears on an eprint, to be captured
- enable identification of the research funder and project code
- enable identification of affiliation of an eprint

Requirements (3)

- enable identification of the repository or other service making available the copy of an eprint
- enable identification of the repository or other service making available the metadata about an eprint
- support disambiguation of publication title
- enable identification of copyright holders of different expressions
- identify the date when a piece of work, or a particular copy, was/will be made publicly available
- identify the date of modification of a copy, in order to locate the latest version
- support the capture of multiple language versions of an abstract, for translations
- be compatible with library cataloguing approaches
- support extensibility of the profile for other types of material

the requirements demanded a more complex metadata model …

eduserv

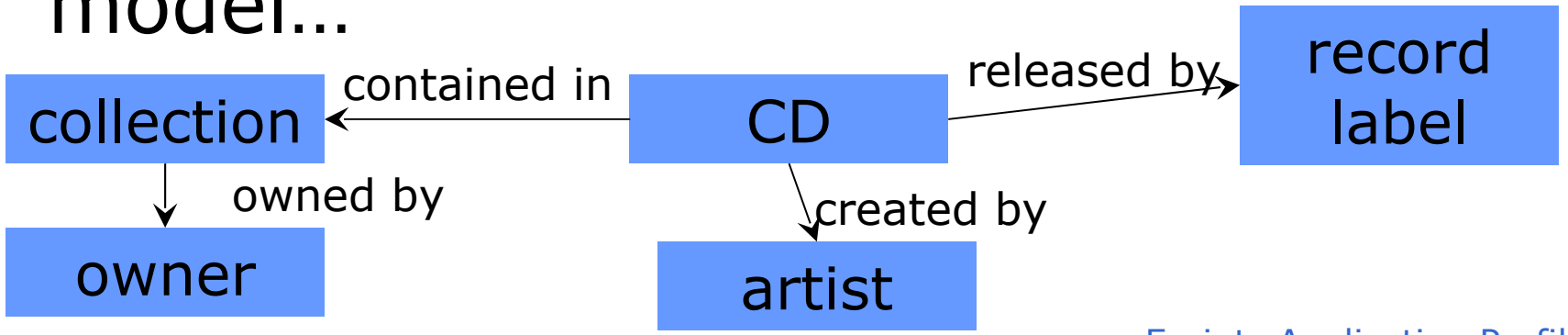
UKOLN

The eprints application model

Andy

What is an application model?

- the set of **entities** that we want to describe
- and the key **relationships** between those entities
- e.g. a CD collection entity/relationship model…



Why have an application model?

- entities appear in the application model because we want to provide descriptions of them
- AND we only want to describe each instance of an entity only once
- the application model can be documented using UML class diagrams or E/R diagrams or in plain text or …

Model vs. model

- **IMPORTANT** - the application model and the DCMI Abstract Model are completely separate
- the application model says what things are being described
- the DCAM says what the descriptions look like

A note about FRBR

- Functional Requirements for Bibliographic Records
- an application model for the entities that bibliographic records are intended to describe
- FRBR models the world using 4 key entities
  - Work, Expression, Manifestation and Item

FRBR entities

- **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** in the form of alpha-numeric, musical, or choreographic notation, sound, image, object, movement, etc., or any combination of such forms. An expression is the specific intellectual or artistic form that a work takes each time it is "realized."
- **A manifestation is the physical embodiment of an expression** of a work. The entity defined as manifestation encompasses a wide range of materials, including manuscripts, books, periodicals, maps, posters, sound recordings, films, video recordings, CD-ROMs, multimedia kits, etc.
- **An item is a single exemplar of a manifestation**. The entity defined as item is a concrete entity.

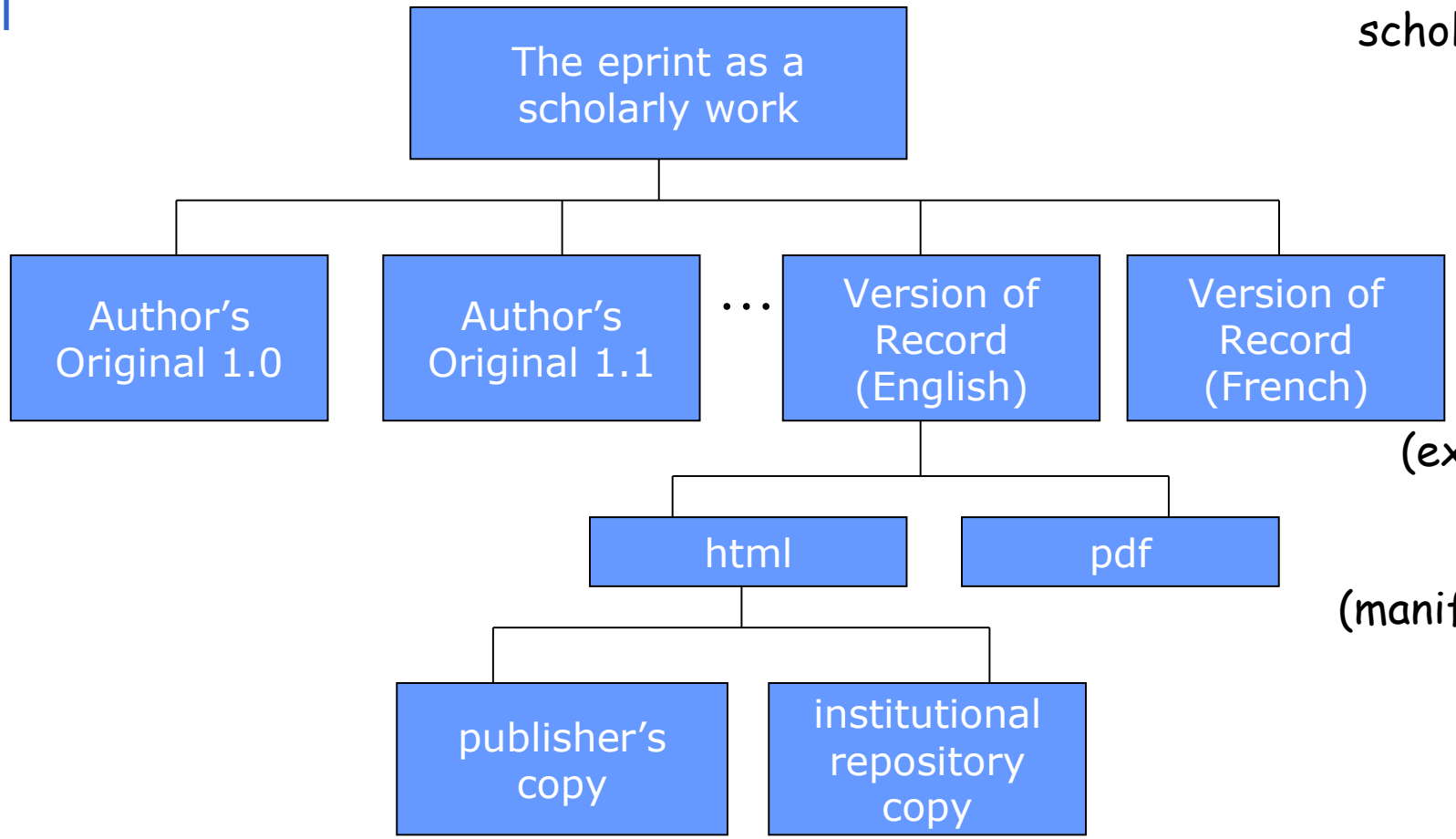
FRBR relationships

- FRBR also defines additional entities that are related to the four entities above - 'Person', 'Corporate body', 'Concept', 'Object', 'Event' and 'Place' - and relationships between them
- the key entity-relations appear to be:
  - Work -- is realized through --> Expression
  - Expression -- is embodied in --> Manifestation
  - Manifestation -- is exemplified by --> Item
  - Work -- is created by --> Person or Corporate Body
  - Manifestation -- is produced by --> Person or Corporate Body
  - Expression -- has a translation --> Expression
  - Expression -- has a revision --> Expression
  - Manifestation -- has an alternative --> Manifestation

FRBR and eprints

- FRBR is a useful model in the context of eprints because it allows us to answer questions like
  - what is the URL of the most appropriate copy (an item) of the PDF format (a manifestation) of the pre-print version (an expression) for this eprint (the work)?
  - are these two copies related? if so, how?

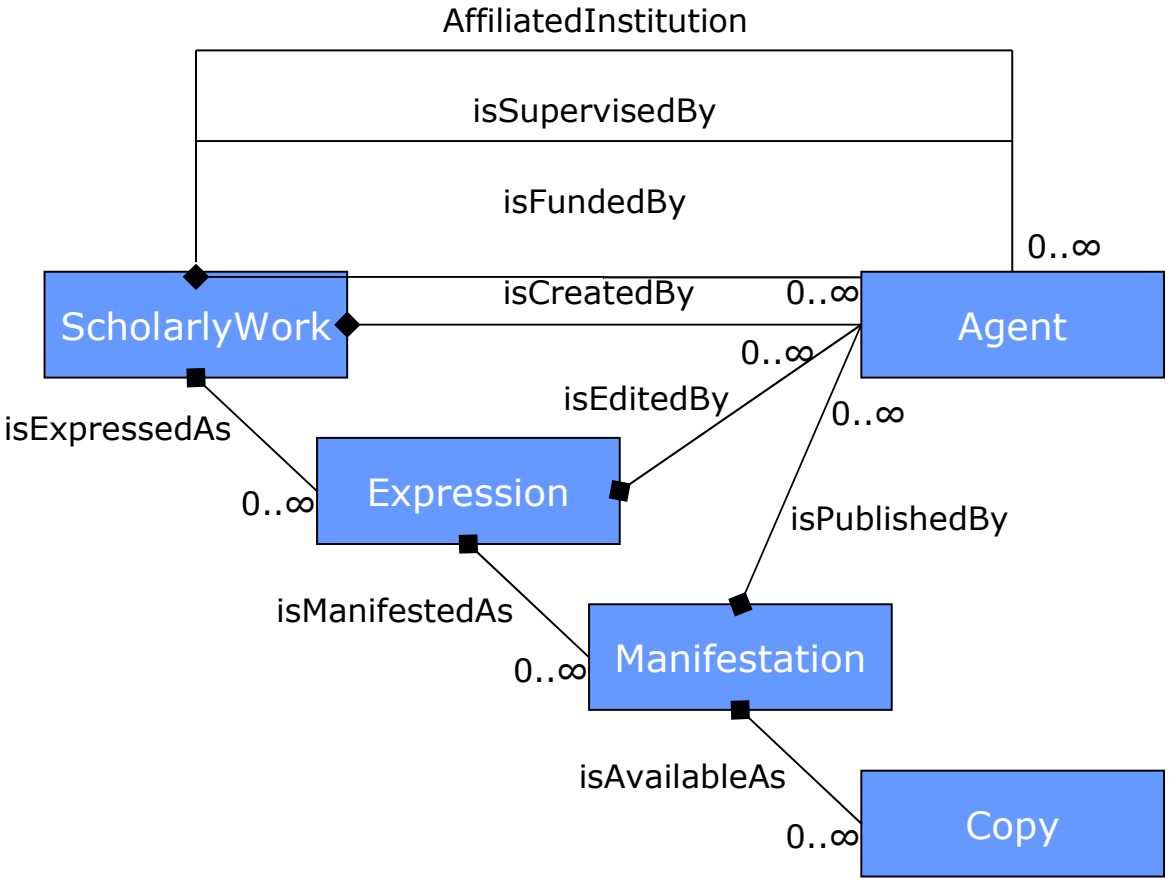
FRBR for eprints



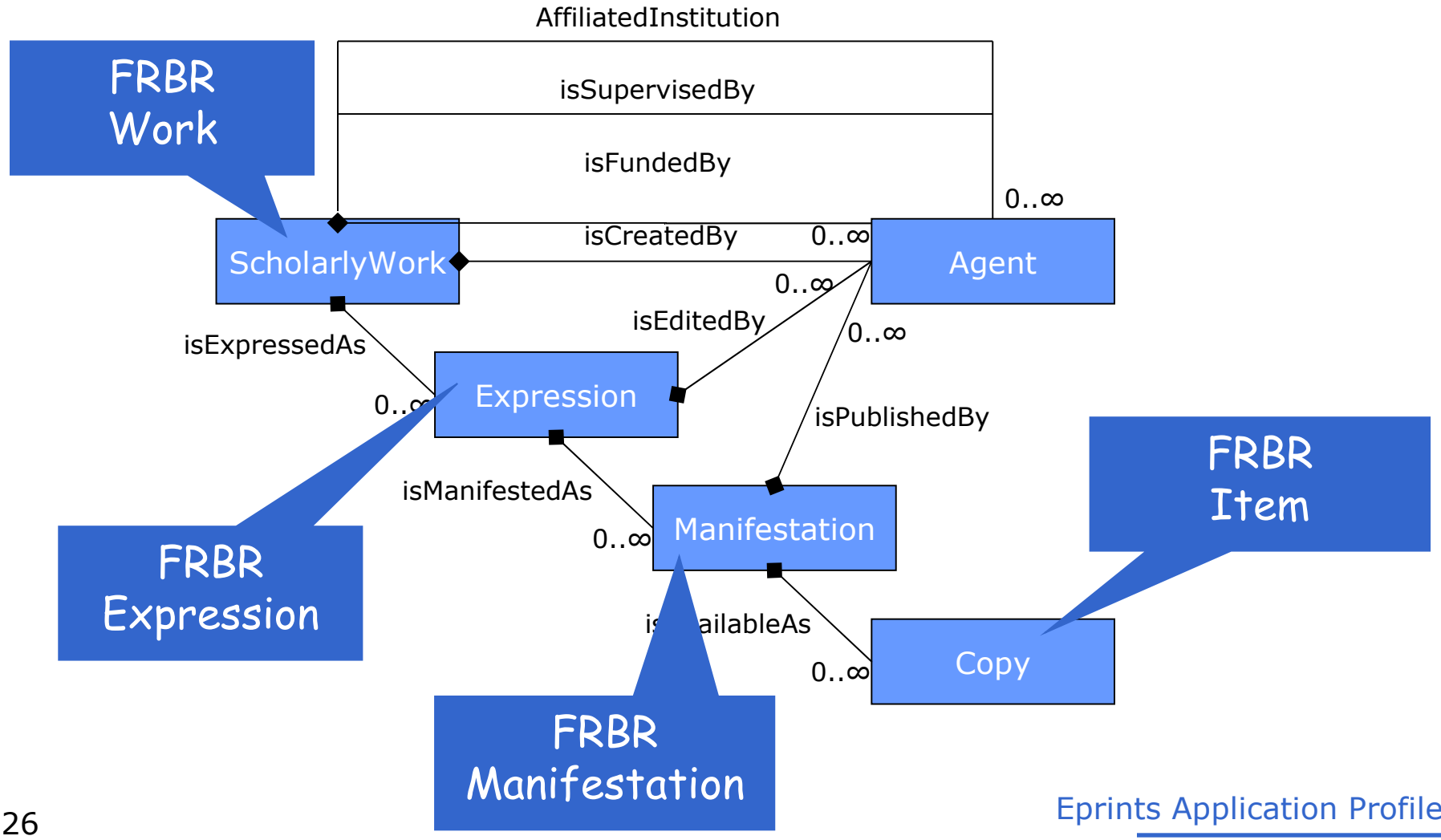
Eprints application model

- based on FRBR
- but some of the labels have been changed - to make things more intuitive, e.g.
  - Work → ScholarlyWork
  - Item → Copy

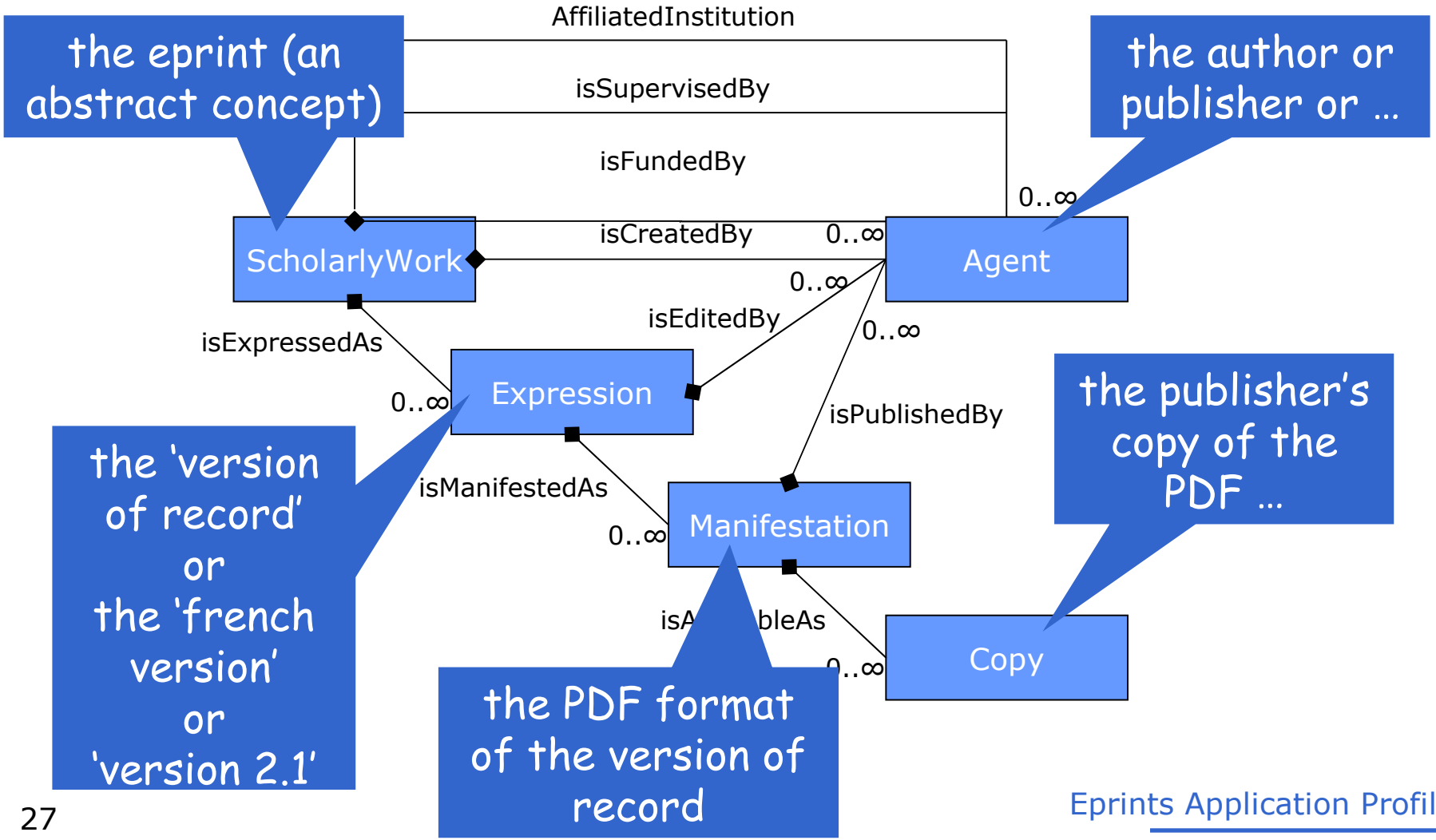
Eprints application model



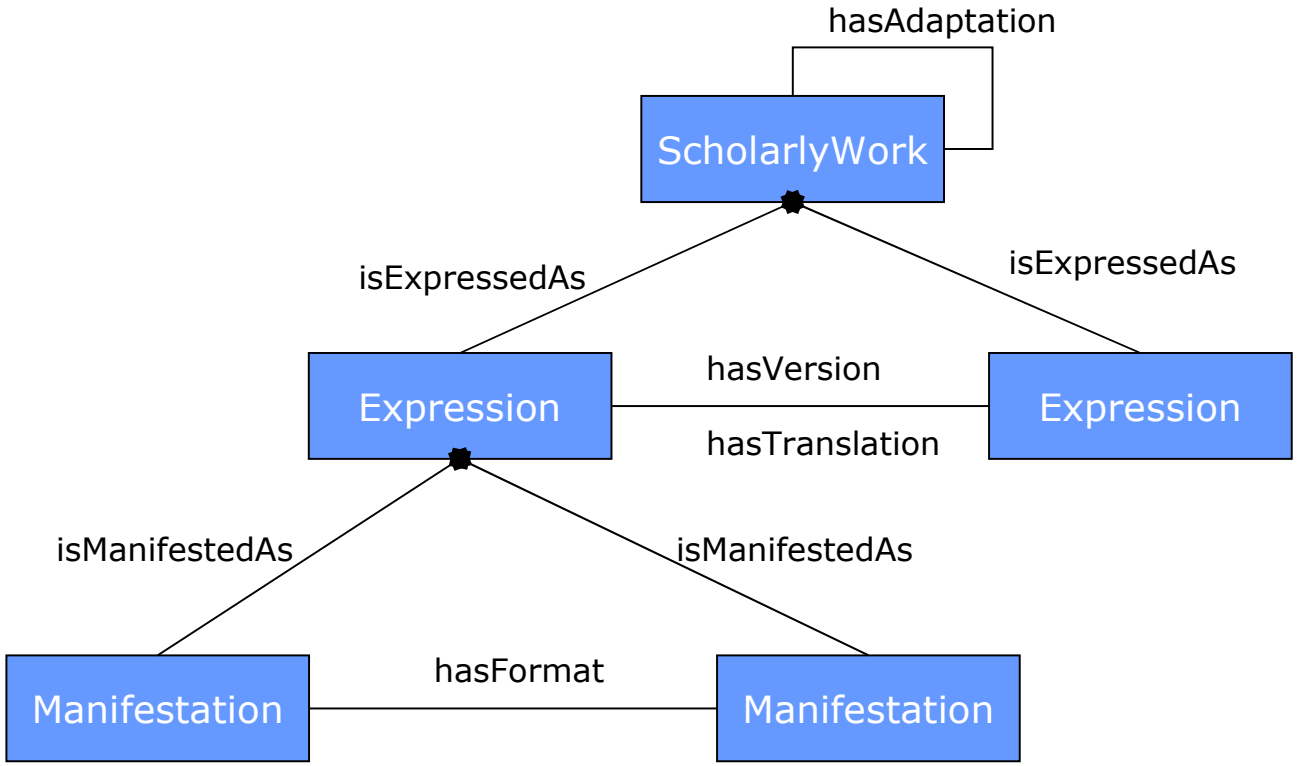
Eprints model and FRBR



Eprints model and FRBR



Vertical vs. horizontal relationships



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

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

Copy:
date available
access rights
licence
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

Final thoughts on the model

- this model makes it easier 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)
- adopting a simple underlying model now may be expedient in the short term but costly to interoperability in the long term
  - the underlying model need to be as complex as it needs to be, but not more so!
- a complex underlying model may be manifest in relatively simple metadata and/or end-user interfaces
- existing eprint systems may well capture this level of detail currently – but use of simple DC stops them exposing it to others!

eduserv

UKOLN

The application profile and vocabularies

Julie

The application profile and vocabularies

- available at

  www.ukoln.ac.uk/repositories/digirep/index/EPrints_Application_Profile
- contains recommendations, cataloguing/usage guidelines and examples
- little is mandatory, prescriptive statements are limited
- structured according to the model
  - ScholarlyWork
  - Expression
  - Manifestation
  - Copy
  - Agent

The application profile

- contains:
  - simple DC properties (the usual suspects …)
    - identifier, title, abstract, subject, creator, publisher, type, language, format
  - qualified DC properties
    - access rights, licence, date available, bibliographic citation, references, date modified
  - new properties
    - grant number, affiliation institution, status, version, copyright holder
  - properties from other schemes
    - funder, supervisor, editor (MARC relators)
    - name, family name, given name, workplace homepage, mailbox, homepage (FOAF)
  - clearer use of existing relationships
    - has version, is part of
  - new relationship properties
    - has adaptation, has translation, is expressed as, is manifested as, is available as

The vocabularies

- Eprints EntityType Vocabulary Encoding Scheme
  - ScholarlyWork
  - Expression
  - Manifestation
  - Copy
  - Agent
- Eprints Status Vocabulary Encoding Scheme
  - PeerReviewed
  - NonPeerReviewed
- Eprints AccessRights Vocabulary Encoding Scheme
  - Open Access
  - Restricted Access
  - Closed Access
- Eprints Type Vocabulary Encoding Scheme

ePrints type vocabulary encoding scheme

- http://purl.org/dc/dcmitype/Text
  - http://purl.org/eprint/type/ScholarlyText
    - http://purl.org/eprint/type/Book
    - http://purl.org/eprint/type/BookItem
    - http://purl.org/eprint/type/BookReview
    - http://purl.org/eprint/type/ConferenceItem
      - http://purl.org/eprint/type/ConferencePaper
      - http://purl.org/eprint/type/ConferencePoster
    - http://purl.org/eprint/type/JournalItem
      - http://purl.org/eprint/type/JournalArticle
    - http://purl.org/eprint/type/NewsItem
    - http://purl.org/eprint/type/Patent
    - http://purl.org/eprint/type/Report
    - http://purl.org/eprint/type/SubmittedJournalArticle
    - http://purl.org/eprint/type/Thesis
    - http://purl.org/eprint/type/WorkingPaper

Key

= sub-class

Example

- expressed in DC-Text
- uses terms from the following schemes:

```
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix eprint: <http://purl.org/eprint/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
```

- the description set contains descriptions, or links to the descriptions, for each entity.

```
DescriptionSet (
    …
```

Example : description of a scholarly work

```
Description (
    Resource URI ( <http://eprints.soton.ac.uk/22934/> )
    Statement (
      Property URI ( dc:type )
      ValueURI ( <http://purl.org/eprint/entitytype/ScholarlyWork> )
    )
    Statement (
      Property URI ( dc:identifier )
      Value String ( "http://eprints.soton.ac.uk/22934/" )
      Syntax Encoding Scheme URI ( dcterms:URI )
    )
    Statement (
      Property URI ( dc:title )
      Value String ( "Structurally integrated brushless PM motor for miniature propeller thrusters" )
    )
    Statement (
      Property URI ( dc:creator )
      Value String ( "Abu Sharkh, S.M.A. (Suleiman)" )
      DescriptionRef ( AbuSharkhSM )
    )
    Statement (
      Property URI ( dc:creator )
      Value String ( "Lai, S.H." )
    )
```

EntityType - ScholarlyWork

Each entity has an identifier

Points to a related description within the description set

Example : description of a scholarly work contd.

```
Statement (
    Property URI ( dcterms:abstract )
    Value String ( "The design, analysis and performance of a brushless PM motor that … " ))
  Statement (
    Property URI ( dc:subject )
    Vocabulary Encoding Scheme URI ( dcterms:LCSH )
    Value String ( "T Technology--TC Hydraulic engineering. Ocean engineering" ))
  Statement (
    Property URI ( dc:subject )
    Vocabulary Encoding Scheme URI ( dcterms:LCSH )
    Value String ( "T Technology--TK Electrical engineering. Electronics Nuclear engineering" ))
  Statement (
    Property URI ( dc:subject )
    Vocabulary Encoding Scheme URI ( dcterms:LCSH )
    Value String ( "T Technology--TL Motor vehicles. Aeronautics. Astronautics" )
  )
  Statement (
    Property URI ( eprint:affiliatedInstitution )
    Value String ( "University of Southampton" )
    DescriptionRef ( sotonuni )
  )
  Statement (
    Property URI ( eprint:isExpressedAs )
    Value URI ( <http://dx.doi.org/10.1049/ip-epa:20040736> )
  )
)
```

The referenced expression has a DOI

Eprints Application Profile

Example : description of an expression

```
Description (
    Resource URI ( <http://dx.doi.org/10.1049/ip-epa:20040736> )
    Statement (
      Property URI ( dc:type )
      ValueURI ( <http://purl.org/eprint/entitytype/Expression> )
    )
Statement (
      Property URI ( dc:type )
      Value URI ( <http://purl.org/eprint/type/JournalArticle> )
    )
    Statement (
      Property URI ( dc:identifier )
      Value String ( "http://dx.doi.org/doi:10.1049/ip-epa:20040736" )
      Syntax Encoding Scheme URI ( dcterms:URI )
    )
    Statement (
      Property URI ( dcterms:available )
      Syntax Encoding Scheme URI (dcterms:W3CDTF )
      Value String ( "2004" )
    )
   Statement (
      Property URI ( eprint:status )
      Vocabulary Encoding Scheme ( eprint:status )
      ValueURI ( <http://purl.org/eprint/status/PeerReviewed> )
    )
```

EntityType - Expression

Each expression has at least one Type value

The Status VES is used to indicate if an expression is peer reviewed

Example : description of an expression contd.

```
Statement (
      Property URI ( dcterms:copyrightHolder )
      Value String ( "Institution of Engineering and Technology" )
    )
    Statement (
      Property URI ( dcterms:bibliographicCitation )
      Value String ( "IEE Proceedings - Electric Power Applications, 151, (5), 513-519 (2004)" )
      Value String ( "&ctx_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&rft.genre=article
      &rft.atitle=Structurally+integrated+brushless+PM+motor+for+miniature+propeller+thrusters
      &rft.jtitle=IEE+Proceedings+-+Electric+Power+Applications&rft.volume=151&rft.issue=5
      &rft.spage=513&rft.date=2004&rft.issn=1350-2352
      &rft.aulast=Sharkh&rft.auinit=S+M+A
      &rfr_id=info:sid/eprints.soton.ac.uk"
        Syntax Encoding Scheme URI ( <info:ofi/fmt:kev:mtx:ctx> ) )
    )
    Statement (
      Property URI ( eprint:isManifestedAs )
      DescriptionRef ( manifestation1 )
    )
    Statement (
      Property URI ( dc:language )
      Value String ( "en" )
    )
  )
```

A text bibliographic citation and OpenURL Context Object can be supplied

Example : description of a manifestation

```
Description (
    DescriptionId ( manifestation1 )
    Statement (
      Property URI ( dc:type )
      ValueURI ( <http://purl.org/eprint/entitytype/Manifestation> )
    )
    Statement (
      Property URI ( dc:format )
      Vocabulary Encoding Scheme URI ( dcterms:IMT )
      Value String ( "application/pdf" )
    )
    Statement (
      Property URI ( dc:publisher )
      Value String ( "Institution of Engineering and Technology" )
    )
    Statement (
      Property URI ( eprint:isAvailableAs )
      Value URI
      ( <http://scitation.aip.org/getpdf/servlet/GetPDFServlet?filetype=pdf&id=IEPAER00015100000500051
      3000001&idtype=cvips&prog=normal> )
    )
  )
```

Each entity has an EntityType value

Example : description of a Copy

```
Description (
    Resource URI
     ( <http://scitation.aip.org/getpdf/servlet/GetPDFServlet?filetype=pdf&id=IEPAER00015100000500051
     3000001&idtype=cvips&prog=normal> )
    Statement (
      Property URI ( dc:type )
      Value URI ( <http://purl.org/eprint/entitytype/Copy> )
    )
    Statement (
      Property URI ( dcterms:licence )
      Value URI ( <http://www.ietdl.org/journals/doc/IEEDRL-home/info/subscriptions/terms.jsp> )
    )
    Statement (
      Property URI ( dcterms:accessRights )
      Value URI ( <http://purl.org/eprint/accessRights/RestrictedAccess> )
    )
    Statement (
      Property URI ( dcterms:isPartOf )
      Value URI ( <http://www.theiet.org/> )
      Value String ( "Institution of Engineering and Technology" )
    )
    Statement (
      Property URI ( dcterms:isPartOf )
      Value URI ( <http://www.ietdl.org/> )
      Value String ( "IET Digital Library" )
    )
)
```

EntityType - Copy

This Copy is restricted access

This Copy is supplied by the IET Digital Library

Example : description of an Agent (organisation)

```
Description (
  DescriptionId ( sotonuni )
  Statement (
    Property URI ( dc:type )
    Value URI ( <http://purl.org/eprint/entitytype/Organization> )
  )
  Statement (
    Property URI ( foaf:name )
    Value String ( "University of Southampton" )
  )
  Statement (
    Property URI ( foaf:homepage )
    Value URI ( "http://www.soton.ac.uk/" )
  )
)
```

EntityType - Organization

The FOAF standard provides agent information

Example : description of an Agent (person)

```
Description (
    DescriptionId ( AbuSharkhSM )
    Statement (
      Property URI ( dc:type )
      Value URI ( <http://purl.org/eprint/entitytype/Person> )
    )
    Statement (
      Property URI ( foaf:givenname )
      Value String ( "Suleiman" )
    )
    Statement (
      Property URI ( foaf:familyname )
      Value String ( "Abu Sharkh" )
    )
    Statement (
      Property URI ( foaf:homepage )
      Value URI ( <http://www.soton.ac.uk/ses/people/AbuSharkhSM.html> )
    )
    Statement (
      Property URI ( foaf:workplaceHomepage )
      Value URI ( <http://www.soton.ac.uk/> )
    )
  )
)
```

EntityType - Agent

eduserv

UKOLN

Dumb-down issues

Andy

Dumb-down

- so… how do we get from these complex descriptions to simple DC descriptions?
- first need to decide what the resulting simple DC description is going to be about?
- eprints application profile metadata contain descriptions about all the entities in the model…

Dumb-down to what?

- some options for dumbing-down:
  - one simple DC description about the ScholarlyWork
  - one simple DC description about each Copy
  - separate simple DC descriptions about every entity
  - separate simple DC descriptions about the ScholarlyWork and each Copy

ScholarlyWork and Copy

- we have chosen to dumb-down to separate simple DC descriptions of the ScholarlyWork and each Copy
- rationale:
  - simple DC about the ScholarlyWork corresponds to previous guidance about using simple DC to describe eprints
  - simple DC about each Copy useful for getting to full-text, e.g. by Google

Dumb-down algorithm

- not covered here…
- see detailed documentation in the Wiki

http://www.ukoln.ac.uk/repositories/digirep/index/Mapping_the_Eprints_Application_Profile_to_Simple_DC

Community acceptance

Julie

Next steps ...

- Application Profile as a start
- Community acceptance plan outlines further work towards community take-up
  - xml schema
    - awaiting new Dublin Core XML guidelines
  - deployment by developers
    - statements from Eprints.org, DSpace, Fedora, Intute and EDINA
  - deployment by repositories, services
    - early adopters from established projects and repositories
    - UK initially
    - benefits of global acceptance
  - dissemination
    - DC-2006 workshop and new DC taskforce
    - this workshop
    - ongoing, e.g. Dlib, Ariadne, Open Repositories 2007, discussion list etc.

eduserv

UKOLN

Discussion