Supporting a Research Repository information infrastructure

presented by Wouter Klapwijk

Stellenbosch Symposium 2006

Academic libraries: proactive partners in Learning and Research

Outline of presentation

i. Supporting e-Research @ SU: general approach and motivation

ii. The *CIB DSpace* project

iii. Functional requirements of the research community

iv. Components of a standards-based research repository infrastructure

v. Looking ahead

Conventions used in this presentation

Suggestions / comments

Comment

❖

Questions / requests

Question

➢

i. Supporting e-Research @ SU: general approach and motivation

i. Supporting e-Research @ SU: general approach and motivation

The South African National Research and Development Strategy

- Published in August 2002
- Focuses on stimulating the development of Science and Technology in a South African context to the benefit and growth of the Nation
- Identifies the need to create ‘centres and networks of excellence’ in Science and Technology (including the social sciences)
- Initially 6 (now 7) centres established to stimulate sustained distinction in research
- Impact on key national and global areas of research

i. Supporting e-Research @ SU: general approach and motivation

DST-NRF Centres of Excellence Programme

- Centres are hosted or co-hosted at various HE institutions:

1. Centre of Excellence for Biomedical TB Research
2. Centre of Excellence for Invasion Biology
3. Centre of Excellence in Strong Materials
4. Centre of Excellence in Birds as Keys to Biodiversity Conservation
5. Centre of Excellence for Catalysis
6. Centre of Excellence in Tree Health Biotechnology
7. Centre of Excellence for Epidemiological Modeling and Analysis

i. Supporting e-Research @ SU: general approach and motivation

DST-NRF Centres of Excellence Programme

- Centres are hosted or co-hosted at various HE institutions:

1. Centre of Excellence for Biomedical TB Research
2. **Centre of Excellence for Invasion Biology**
3. Centre of Excellence in Strong Materials
4. Centre of Excellence in Birds as Keys to Biodiversity Conservation
5. Centre of Excellence for Catalysis
6. Centre of Excellence in Tree Health Biotechnology
7. Centre of Excellence for Epidemiological Modeling and Analysis

i. Supporting e-Research @ SU: general approach and motivation

DST-NRF Centres of Excellence Programme

- Centres are inter-institutional in nature in terms of partnerships and collaboration – e.g. http://academic.sun.ac.za/cib/Partners.htm
- Researchers are located at non-HE institutions as well
- Researchers at the various partnership institutions need to populate repositories with research outputs by themselves

Comment

- Possible federation of repositories between HE institutions and the NRF with institutional identity management - 3A's (authentication, authorization and access)

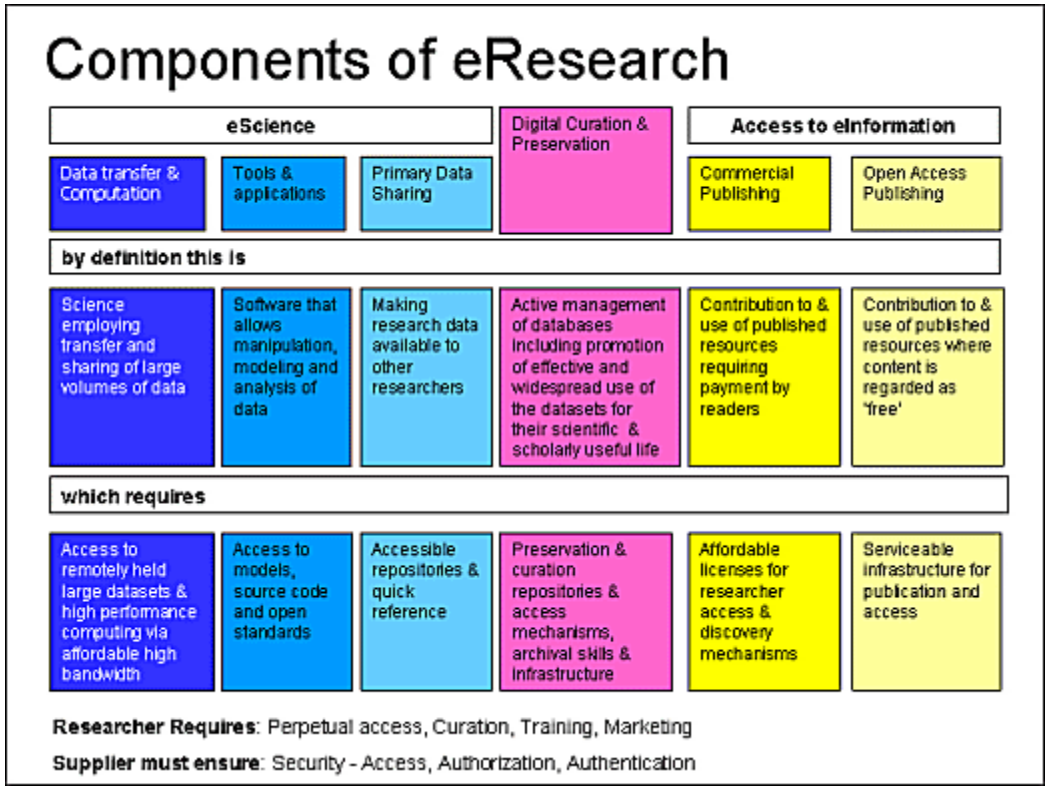
i. Supporting e-Research @ SU: general approach and motivation

South African Research Information Services (SARIS) investigation and recommendations

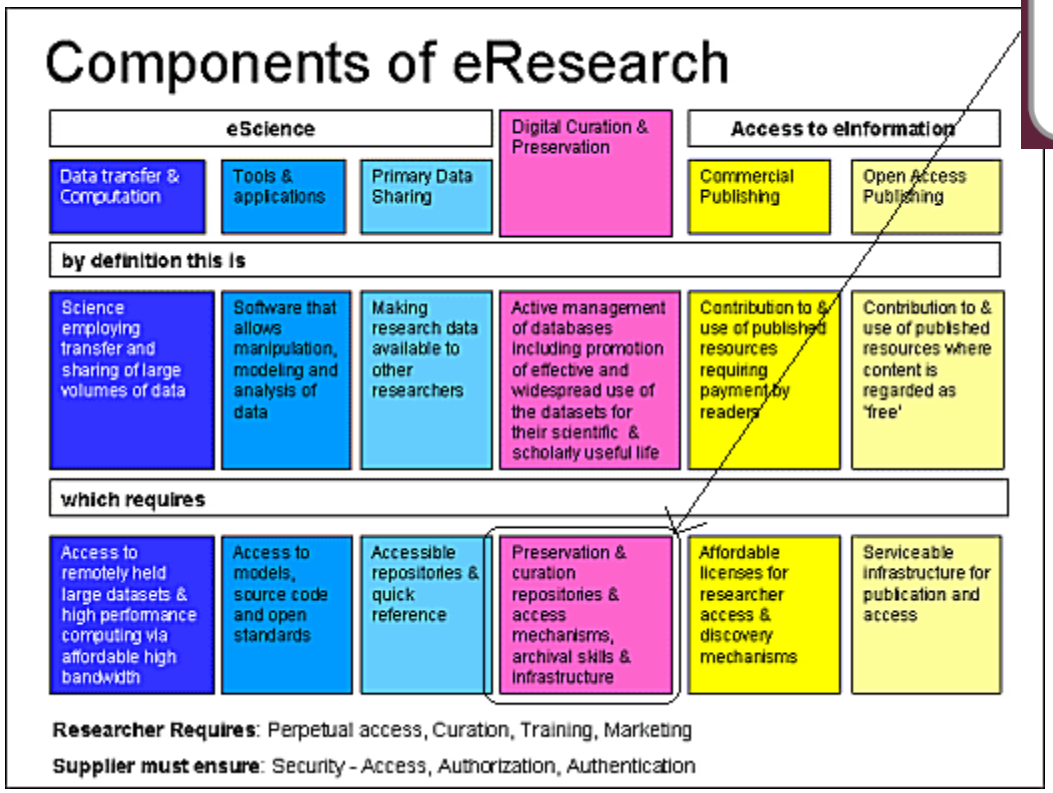
---

- Investigated the requirements for a national information service framework
- Calls for coordination of e-Research support systems between HE institutions (to ensure cost-effectiveness and efficiency)
- Identifies components of e-Research
- Proposes a e-Research framework - specifically *e-Research Development & Innovation* group of activities are of interest
- To a certain extent aligned with current initiatives at Stellenbosch University

i. Supporting e-Research @ SU: general approach and motivation



i. Supporting e-Research @ SU: general approach and motivation



i. Supporting e-Research @ SU: general approach and motivation

Formulation of a *Research Support Framework* at Stellenbosch University

---

- The Framework is still in it's infancy
- Must address Administrative as well Technical e-Research support criteria:
  - research management
  - institutional repositories, identity management, federated search, etc.
  - home connectivity, etc.
- Must integrate with National strategies:
  - Proposed eR3SA framework
  - SANReN
- Must maximize usage of existing campus systems and infrastructure

ii. The ***CIB DSpace*** project

ii. The *CIB DSpace* project

C•I•B = **C**entre of Excellence for **I**nvasion **B**iology

DSpace = digital repository software used to house the research outputs

---

ABOUT

- Initiation of the project is due to the DST and NRF requirements to create databases of research outputs
- The CIB approached the SU Department of Information Technology for the facilities management of the Research Database
- Collaboration between the SU Department of Information Technology and the Library and Information Services Dept to create a research repository database in consultation with the CIB (terms encapsulated in a project charter)
- Acknowledges the Library and Information Services responsibility as a leading Knowledge Management advisor at SU

iii. Functional requirements of the research community

iii. Functional requirements of the research community

More specific, research community = CIB (“the client”)

- Must address the client’s needs
- Client prepared a set of Use Cases (not in UML notation), included:
  - workflow specifications (add, modify, retrieve)
  - levels of access:

    admin, staff, research team, students, public
  - access rules
  - permissions
  - data model relationships (simple entity relationship diagrams)
  - Select List values
  - metadata schemas (Dublin Core, ISO-19115)

iii. Functional requirements of the research community

Permissions according to levels

(extracted from the CIB Use Cases)

Level	Metadata	Authors	Projects	Theses	Publications	Datasets
L0						
L1						
L2						
L3						
L4						

No Access

Access

To be confirmed

View if it is owner

Examples of Access Rules

(extracted from the CIB Use Cases)

Datasets

a) for the first 3 years only the owner
b) then both the owner and the research team members for 2 years
c) the public after 5 years (owners must be informed)

Publications

a) Level 1 and Level 3 users must have access (refer previous slide)
b) public (Level 4 users) can be given a URL to the relevant publisher's journal or website

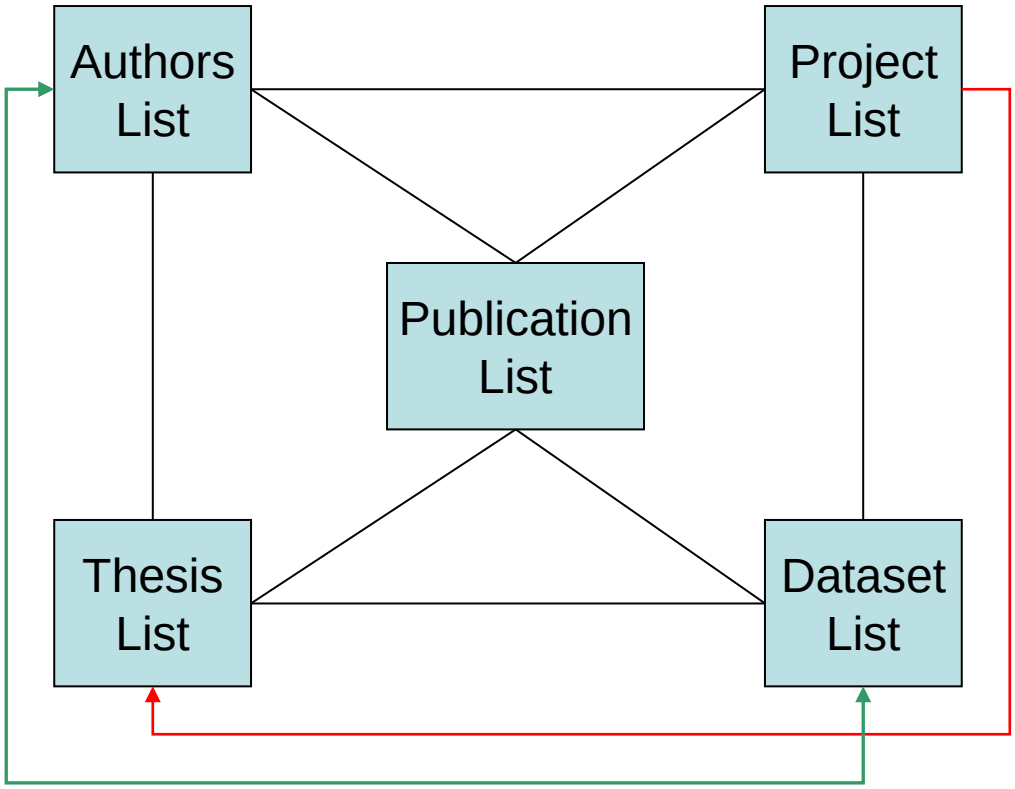
Comment

- Not really Open Access, not really a Closed Repository either

iii. Functional requirements of the research community

Simplistic data model:
Many to many relationship “diagram”

(extracted from the CIB Use Cases)



iii. Functional requirements of the research community

Conflict between user requirements and system design

- No “off the shelf” digital repository software available with required functionality
- But, risks are involved if development is outsourced:
  - costly
  - final project unlikely to be standards-based
  - won’t be open source
  - not necessarily interoperable with other campus systems
  - community of support?

Question

- Is flexible digital repository software available to meet differing user requirements?

iii. Functional requirements of the research community

Conflict between user requirements and system design

- No “off the shelf” digital repository software available with required functionality
- But, risks are involved if development is outsourced:
  - costly
  - final project unlikely to be **standards-based** ◄▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪
  - won’t be open source
  - not necessarily interoperable with other campus systems
  - community of support?

Question

- Is flexible digital repository software available to meet differing user requirements?

iv. Components of a standards-based research repository

iv. Components of a standards-based research repository

Some key fundamentals to consider

1. The Platform: OS + Digital Repository software
2. The Metadata Schema required
3. The Digital Preservation Plan
4. The Dissemination Policy
5. The Integration Plan

iv. Components of a standards-based research repository

Some key fundamentals to consider

1. **The Platform: OS + Digital Repository software**
2. The Metadata Schema required
3. The Digital Preservation Plan
4. The Dissemination Policy
5. The Integration Plan

iv. Components of a standards-based research repository

The Platform

1) Depends on specs of Digital Repository software ►

2) Hardware (IBM, Dell, SUN, etc.) – can be cheap.

3) OS Support and Upgrade cycle

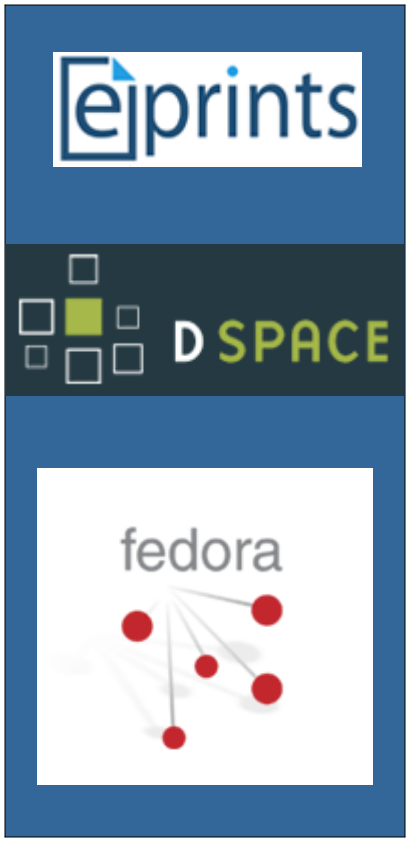
▼

ubuntu 6.10

Stellenbosch University DSpace Wiki page:

http://www.lib.sun.ac.za/dspacewiki

(courtesy of Laurence Munro)



iv. Components of a standards-based research repository

Some key fundamentals to consider

1. The Platform: OS + Digital Repository software
2. **The Metadata Schema required**
3. The Digital Preservation Plan
4. The Dissemination Policy
5. The Integration Plan

The Metadata Schema

- Five **types** of metadata:

1) **Descriptive** [title, author, summary, topic]
2) **Technical & Structural** [file size, software needed]
3) **Administrative** [record number, date, source]
4) **Rights** [copyright ownership, use privileges]
5) **Management** [price paid, circulation restrictions]

The Metadata Schema

- Five **types** of metadata:

1) **Descriptive** [title, author, summary, topic] ◄▪▪▪▪▪▪▪
2) **Technical & Structural** [file size, software needed]
3) **Administrative** [record number, date, source]
4) **Rights** [copyright ownership, use privileges]
5) **Management** [price paid, circulation restrictions]

The Metadata Schema

Metadata is the glue of any digital repository strategy.

Within a digital repository, “metadata accompanies and makes reference to each digital object and provides associated descriptive, structural, administrative, rights management, and other kinds of information”.

- Clifford Lynch (D-Lib Magazine, 1999)

The Metadata Schema

- Some metadata **schemas**:

1) **MARC21** (ISO 2709)
2) **MODS** (Metadata Object Description Schema) – essentially MARC21 recast in an XML-native framework
3) **Dublin Core** (ISO15836:2003)
4) **TEI** (Text Encoding Initiative) – for complex markup of literary texts
5) **EAD** (Encoded Archival Description) – a format for expressing electronic archival finding aids
6) **METS** (Metadata Encoding and Transmission Standard)
7) **FGDC** (ISO-19115) – for describing digital geospatial data

The Metadata Schema

- Some metadata **schemas**:

1) **MARC21** (ISO 2709)
2) **MODS** (Metadata Object Description Schema) – essentially MARC21 recast in an XML-native framework
3) **Dublin Core** (ISO15836:2003)
4) **TEI** (Text Encoding Initiative) – for complex markup of literary texts
5) **EAD** (Encoded Archival Description) – a format for expressing electronic archival finding aids
6) **METS** (Metadata Encoding and Transmission Standard)
7) **FGDC** (**ISO-19115**) – for describing digital geospatial data ◄

The Metadata Schema

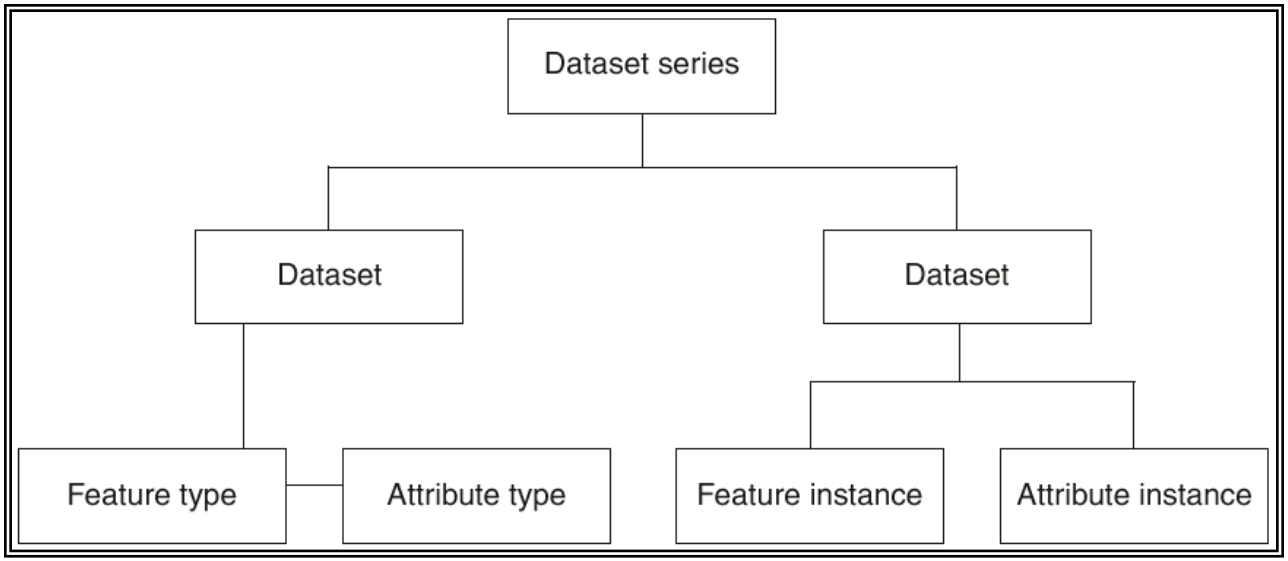
- Spatial versus non-spatial data
- Dublin Core is a standard for cross-domain information resource description (only)
- Not suitable for describing geospatial digital objects

Spatial metadata encapsulates knowledge "about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data."

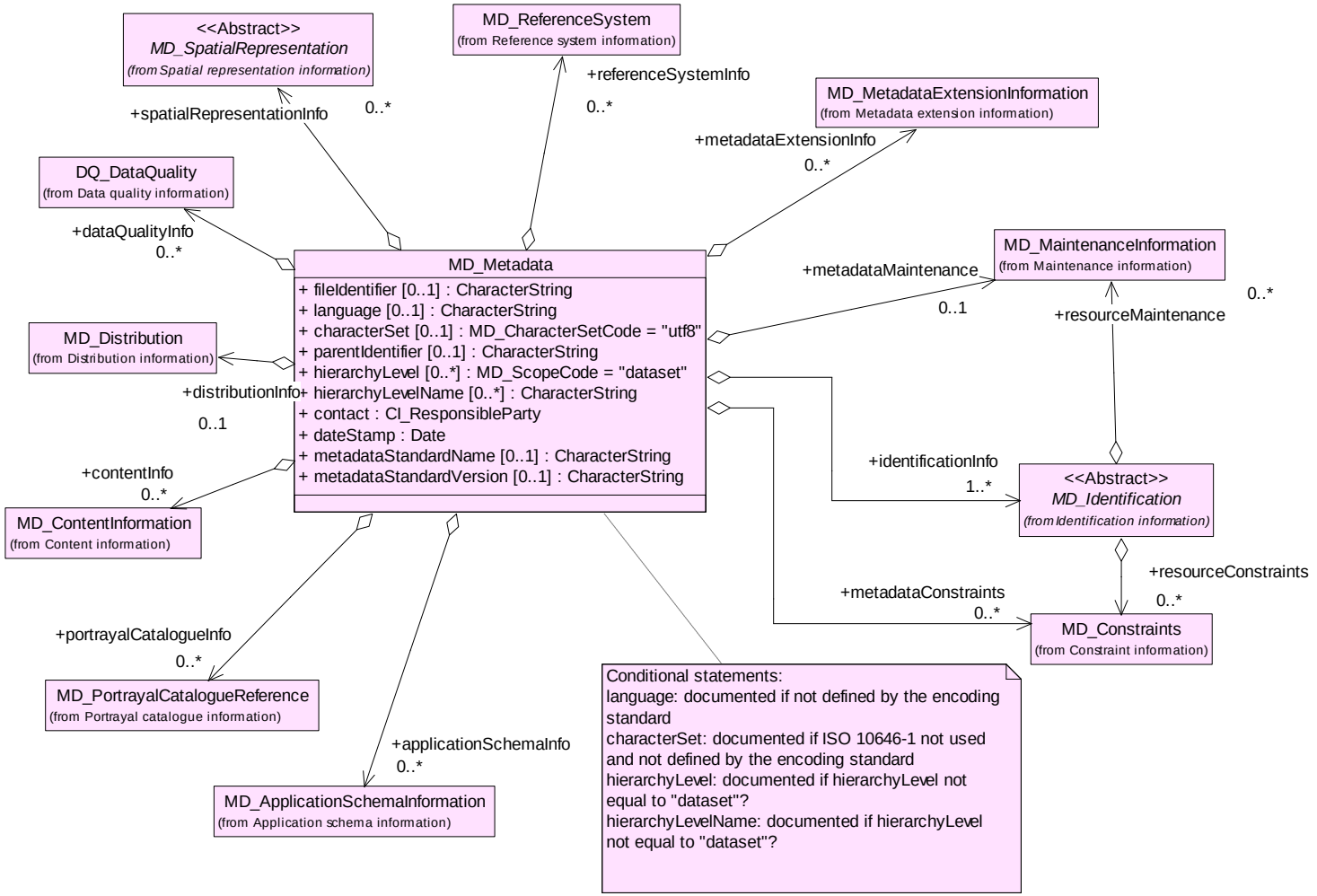
Source: ISO/TC 211 --Geographic information/Geomatics – Committee Draft 19115.2

The Metadata Schema

- Features of ISO-19115:
  - Expressed in a UML structure (formal modeling)
  - Can be profiled (different types)
  - Multi-leveled metadata schema – as opposed to Dublin Core



iv. Components of a standards-based research repository



The Metadata Schema

- What did we do at Stellenbosch University?
  - Created a ISO-19115 profile – validated by our Cataloguing Department
  - Profile was enhanced by a complete data dictionary (data type, occurrences, obligation, etc.)
  - SANS 1878: South African spatial metadata standard
  - SANS 1880: South African geospatial data dictionary (SAGDaD)
  - Used DSpace Dublin Core elements (qualified and unqualified) where possible – to ensure harvesting for future purposes
  - [Dspace-general] listserv: MIT are considering tackling geospatial metadata as part of their SIMILE project
  - Decision: “crosswalk” XML-exported data to future schema using XSLT

iv. Components of a standards-based research repository

The Metadata Schema

- Ask yourself:

  "Are your researchers expecting geo-referencing of items in DSpace?"

  "The ability to search on spatial coordinates?"

- Google Earth – Keyhole Markup Language: http://ir.sun.ac.za/cib

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://earth.google.com/kml/2.0">
 <Placemark>
  <description>New York City</description>
  <name>New York City</name>
  <Point>
    <coordinates>-74.006393,40.714172,0</coordinates>
  </Point>
 </Placemark>
</kml>
```

iv. Components of a standards-based research repository

Some key fundamentals to consider

1. The Platform: OS + Digital Repository software
2. The Metadata Schema required
3. **The Digital Preservation Plan**
4. The Dissemination Policy
5. The Integration Plan

The Digital Preservation Plan

Digital preservation activities can be divided into:

1) those that promote the long-term maintenance of a bitstream (the zeros and ones = 10101010101)

2) those that provide continued accessibility of its contents (the persistent linking mechanisms)

The Digital Preservation Plan

Those that promote the long-term maintenance of a bitstream

- DP in the context of Institutional Repositories:

  a) Preserv project – http://preserv.eprints.org (assessment of IR content)
  b) LOCKSS project – http://www.lockss.org (ASERL / e-journals)
  c) TOM (Typed Object Model) - http://tom.library.upenn.edu (format migration system)

- OAIS compliance – a reference model for an **O**pen **A**rchival **I**nformation **S**ystem (SIP, AIP, DIP)

The Digital Preservation Plan

Those that promote the long-term maintenance of a bitstream

- What we are doing at Stellenbosch University:
  - Taking a pragmatic approach while monitoring new developments
  - Continuing our participation and interest in the LOCKSS-SA project

- The LOCKSS-SA project
  - A South African initiative to pilot the feasibility of the LOCKSS system to preserve SA Open Access e-journal content
  - Wiki page: http://www.lib.sun.ac.za/lockss-sa

The Digital Preservation Plan

Those that provide continued accessibility of its contents

- The Handle System
  - http://www.handle.net
  - A persistent linking mechanism – register your institutional prefix (.e.g. 10019)
  - Cannot share a prefix over more than one DSpace instance on the same server/ip address.
  - Sub-prefixes?

iv. Components of a standards-based research repository

Some key fundamentals to consider

1. The Platform: OS + Digital Repository software
2. The Metadata Schema required
3. The Digital Preservation Plan
4. **The Dissemination Policy**
5. The Integration Plan

The Dissemination Policy

- Open Access or Closed Repositories?
- Information Sciences community might prefer Open Access, but the researchers might stipulate Closed Access
- Granularity of harvesting your repository content – all communities and collections or only specific?
- Optimal use of the OAI-PMH protocol
- Register your IR in ALL relevant Open Access Registries

iv. Components of a standards-based research repository

Some key fundamentals to consider

1. The Platform: OS + Digital Repository software
2. The Metadata Schema required
3. The Digital Preservation Plan
4. The Dissemination Policy
5. **The Integration Plan**

The Integration Plan

- Does your digital repository software provide a set of API's to interface with?
- How interoperable is it with campus authentication systems (e.g. LDAP)?
- Are you going to consider multiple Data Provider repositories with a Service Provider using a OAI Harvester?

v. Looking ahead

v. Looking ahead

Some aspects we are considering

- Exploring the functionality of Fedora
- Multiple IR instances
- OAI Harvester
- Cooperation with academic and non-academic departments
- Integration with our new web site structure – under construction

THANK YOU