Repository and CRIS interoperability issues within a 'connector lite' environment

OR2019: Repository/CRIS Workshop (I):
Realising Technical Interoperability and/or Integration

George Macgregor
University of Strathclyde
https://purl.org/g3om4c
@g3om4c
Institutional history

- **University of Strathclyde** – Established 1796 as "place of useful learning" by John Anderson
  - Now 21,470 FTE students & 3,200 staff
- Among 20 top research-intensive universities in the UK
- Research income 2016: £60 million

Assessed by UK Research Excellence Framework (REF) to have #1 research in physics
- 3rd in Electrical & Electronic Engineering (1st in Scotland)
- 4th in Engineering (Aerospace, Mechanical, Marine, etc)

Strong outside science & engineering…
Making modelling count - increasing the contribution of shelf-seas community and ecosystem models to policy development and management


Official URL: https://doi.org/10.1016/j.marpol.2015.07.015

Abstract

Marine legislation is becoming more complex and marine ecosystem-based management is specified in national and regional legislative frameworks. Shelf-seas community and ecosystem models (hereafter termed ecosystem models) are central to the delivery of ecosystem-based management, but there is limited uptake and use of model products by decision makers in Europe and the UK in comparison with other countries. In this study, the challenges to the uptake and use of ecosystem models in support of marine environmental management are assessed using the UK capability as an example. The UK has a broad capability in marine ecosystem modelling, with at least 14 different models that support management, but few examples exist of ecosystem modelling that underpin policy or management decisions. To improve understanding of policy, and management issues that can be addressed using ecosystem models, a workshop was convened that brought together advisors, assessors, biologists, social scientists, economists, modellers, statisticians, policy makers, and funders. Some policy requirements that can be addressed without further model development were identified including: attribution of environmental change to underlying drivers, integration of models and observations to develop more efficient monitoring programmes, assessment of indicator performance for different management goals, and the costs and benefits of legislation. Multi-model ensembles are being developed in cases where many models exist, but model structures are very diverse making a standardised approach of combining outputs a significant challenge, and there is need for new methodologies for describing, analysing, and visualising uncertainties. A stronger link to social and economic systems is needed to increase the range of policy-related questions that can be addressed. It is also important to improve communication between policy and modelling communities so that there is a shared understanding of strengths and limitations of ecosystem models.
Repository-CRIS configuration

- Connected Repository-CRIS ecosystem at Strathclyde: CRIS content writing to Strathprints
- ‘Connectors’ developed for main repository platforms
  - Maintenance overhead variable across platforms….?
- Data exchange between CRIS & IR handled by proprietary ‘connector’
- Complex two-way exchange of data with IRs
  - WebDAV and XSLT for object translation
  - Submit operations, update and delete operations handled by the ‘connector’
  - EPrints CRUD interface via plugin
  - Principal metadata elements accommodated, embargoes, etc.
- Validated content written to EPrints via cron job running every minute
Metadata schema

• Research output metadata from Pure
• Serialized as MODS/XML for exchange and ‘long-term’ storage
  – ICYDK, Metadata Object Description Schema (MODS)
  – Maintained by the Network Development & MARC Standards Office of the Library of Congress
• MODS/XML used as basis for transformation by XSLT to EPrints schema writing metadata to IR (Strathprints)
  – MODS used for EPrints but also other connected repository platforms, e.g. DSpace
MODS/XML metadata – XSLT transformation to EPrints XML metadata schema

Possible to modify XSLT to accommodate *some* IR requirements

Restrictions on data available from Pure – must be in MODS/XML serialization to be available to IR

693  <!-- Modification by @g3om4c to accommodate RIOXX elements for OpenAIRE compliance via CORE -->
694  <xsl:key name="openaire" match="v3:extension/*(not(self::rioxf:funder))" use="generate-id(preceding-sibling::rioxf:funder)"/>
695  <xsl:template match="v3:extension">
696    <rioxx2_project_input>
697      <xsl:for-each select="rioxf:funder">
698        <item>
699          <project>
700            <xsl:value-of select="key('openaire', generate-id())[self::rioxf:projectid]"/>
701            <project/>
702            <xsl:for-each select="rioxf:projectid">
703              <xsl:value-of select="",/>
704            </xsl:for-each>
705          </project>
706        </item>
707      </xsl:for-each>
708    </rioxx2_project_input>
709  </xsl:template>
710  </xsl:stylesheet>
‘Connector lite’

- Connection between Pure & connected IR
  - ‘Connector heavy’
  - No control over content sent to external repository
  - Conflict: the CRIS & IR dichotomy
    - CRIS ≠ repository

- Long standing requirement for partial connection between CRIS & IR
  - So-called ‘connector lite’
  - Possible late 2016
  - Configuration
Feeding the beast: workloads in a hybrid IR / CRIS environment

George Macgregor, Institutional Repository Coordinator, University of Strathclyde

This blog post is about the tension between serving the needs of an institutional repository (IR) and a Current Research Information System (CRIS). A recent blog post discussed improvements in the discoverability and usage of Strathprints. In this blog I mentioned "connector lite". Here is what I said:

Incidentally, you might be asking, "What is "connector-lite"? Prior to "connector-lite" Strathprints was receiving huge volumes of (mostly) unnecessary (and unwanted) metadata, all of it being supplied automatically (and uncontrollably) by our connected CRIS for the purposes of research management. All of this metadata were arriving in such volumes that the growth in full-text remained proportionally low, impeding our ability to grow Strathprints as a full-text destination. Implementing a solution of only a partial connection with the CRIS was therefore a long held ambition – and it was something I identified as essential as soon as I was appointed; but, for technical reasons at the CRIS side, it was an ambition that only became technically possible in early 2016. But when it became possible its implementation was seized upon as a way of restricting the flow of metadata. Repositories are, after all, supposed to facilitate access to full-text.

This scenario would not be an issue for institutions using a stand-alone IR, or indeed a stand-alone CRIS – but for institutions using both there is an inherent tension between the IR and the CRIS. So the following questions arise: Why does this tension exist? Why use both systems?

Back to basics: CRIS and IR systems

The principal reason why many institutions have elected to run an IR and CRIS in parallel is because they are essentially

Promoting content discovery of open repositories: reviewing the impact of optimization techniques (2016-2019)


Interoperability challenges
Connector openness: protocols

• Proprietary CRIS = Proprietary standards
  – Restrictions on data which can be pulled via Pure connector
    • Limits repository functionality scope
    • Limits compliance scope
    • Catch 22 & the ‘compliance conundrum’
  – Maintaining the connector more onerous depending on your repository (?)
  – Connector over-writes pre-existing repository content with every CRIS-to-repository interaction

• But proprietary CRIS ≠ proprietary standards
  – Greater support for open standards
  – SWORD V2 support preferable; simplification of base interoperability but resisted by vendor (see Symplectic Elements)

• Elsevier: Apathy & not knowing the business
  – Lack of domain experience by developers leads to inappropriate ‘solutions’
  – Reaching for convenient solutions rather than appropriate ones
Connector openness: metadata

- MODS/XML as foundation metadata v. good!
  - Good specificity for core descriptive metadata
- But:
  - Incorrect data modelling in areas
  - Native MODS/XML does not support all data types necessary for IRs
  - Does Pure include these data types anyway? No.
    - Project data
    - Publication states
    - Relations: associative links to related content / data unsupported, e.g. data
      - `<mods:extension>` - what should be used within this bucket?!
      - Agreement necessary about which schema to be used
Metadata application profiles

- Adopting metadata application profiles problematic in connected CRIS/IR configuration
  1. Pure = limited native support
  2. EPrints = excellent native support
  3. But bridging the gap requires hacking
- Unsatisfactory & linked to vendor apathy / lack of openness
- [Spectre of semantic interoperability & data quality problems
  - Quality of CRIS administrative data an additional interoperability impediment]

Case study: OpenAIRE compliance of Strathprints

- (Minor) RIOXX support at Pure connector side, enabling XSLT changes to support Strathprints data import

CORE aggregation
- CORE aggregates Strathprints RIOXX content & maps to OpenAIRE for Lit. Repos

OpenAIRE compliance
- CORE send mapped metadata to OpenAIRE

Only certain RIOXX data elements supported by Pure, e.g. no FundREF id
Concluding thoughts

• Interoperability could be better!
• Improved support by Pure for open standards & protocols needed
  – Reaching for domain standards first rather than what is convenient for local developers
  – Opening up more data for repository integration
  – Minimising overhead for institutions & avoiding disenfranchisement
• Agreement on exposure of metadata to repository systems
  – Likely disparate practice across CRIS systems & repositories
• Getting one’s own house in order: semantic interoperability and data quality issues in campus-wide CRIS systems