Pattern based mapping and extraction via the CIDOC CRM

Douglas Tudhope\textsuperscript{1}, Ceri Binding\textsuperscript{1}, Keith May\textsuperscript{2}, Michael Charno\textsuperscript{3}
\textsuperscript{(1}University of South Wales, \textsuperscript{2}English Heritage, \textsuperscript{3}Archaeology Data Service\textsuperscript{)}

douglas.tudhope@southwales.ac.uk
ceri.binding@southwales.ac.uk
keith.may@english-heritage.org.uk
michael.charno@york.ac.uk

Hypermedia Research Unit, University of South Wales

http://hypermedia.research.southwales.ac.uk/
The long road to interoperability...

- Achieving interoperability requires more than just a common data model – as data compatibility occurs on 2 levels – semantic and syntactic. Ontologies / data structures deal with the semantic but not necessarily the syntactic.
  - “The CRM relies on existing syntactic interoperability and is concerned only with adding semantic interoperability” (CIDOC CRM documentation)
- Deciding on CIDOC CRM as an integrating framework is a sensible first step on the road to interoperability – but after that there’s often still a long way to go, particularly for legacy datasets.
Pattern based mapping and extraction via the CIDOC CRM

Another dataset, another schema...

- Perform a cross search on small finds and materials?
What to model, how to model

- Do these rows represent:
  - Data from paper forms? Yes
  - Electronic database records? Yes
  - The small finds themselves? Yes
  - The results of a series of archaeological assessments? Yes

- How to approach modelling?
  - As immaterial records
  - As physical objects
  - As properties associated with a series of events (e.g. identifier assignment, production material)

- How much to include?
  - All data in all rows
  - Data to answer specific research questions and use cases
  - Administrative data (describing contemporary events) – excavation, assessment, who/when/why
  - Implicit data - *known knowns, known unknowns*... (D. Rumsfeld)
    - E.g. production event - where we may know nothing else except there must have been one
Using CIDOC CRM as an integrating framework

A small pattern to model the relationships between an object and a material. The pattern applies to all the records. It would be nice to reuse it in future.

So the problem is now solved? Not quite...
Pattern based mapping and extraction via the CIDOC CRM

Implementation issues

- Raw data
  - Data formats
  - Data cleansing
  - Data mapping
  - Character encoding
  - Terminology concordance
- XML syntax
  - Brackets and tags
  - Namespaces
  - Data types
- RDF syntax
  - Entities and properties
  - URI identifiers
  - Naming conventions
  - Modelling patterns

- Wider issues
  - Scope
  - Consistency
  - Repeatability
  - Coverage
  - Scalability
  - Performance
  - Versioning
  - Licensing
  - Curation
  - Discoverability
  - Documentation

for real implementations things can get complicated very quickly
Pattern based mapping and extraction via the CIDOC CRM

**STELLAR Project**

- Produced tools and techniques to manage (some of) this complexity & to maintain consistency at scale

---

**Examples of tools and techniques**

- Tools for mapping and extracting data
- Techniques for managing complexity

---

**Results**

- Evaluation and feedback from users
- Continuous improvement of tools and techniques

---

**Conclusion**

- The importance of maintaining consistency at scale
- Future directions for research and development
Pattern based mapping and extraction via the CIDOC CRM

STELLAR data conversions

Delimited Data
- Data from file

Internal template
- SQL query results

Database
- SQL commands
- SQL query results

Data from file

RDF data
- SQL query results

User-defined template
- SQL query results

Other textual Data formats
STELLAR templates

- Templates implement predefined data patterns, facilitate consistent data conversion and handle lower-level syntactic issues.
- Template fields provide a layer of abstraction - allowing us to deal with the data at a higher level, and in a modular fashion.
- Templates can create inverse relationships, fully formed paths and shortcuts - enabling more flexible querying without necessarily requiring extensive reasoning capability.
- Doesn’t have to be one way or the other - can model both shortcut paths and more detailed representations within same data.
- Can orient to higher level ‘query model’ by developing specialised custom shortcuts (e.g. stratigraphic relationships).
- Can model CRM E55 type hierarchies and express SKOS concepts – again not one thing or the other, not violating compatibility of either model.
Using STELLAR templates to produce RDF

- Templates are just text files. May be copied, edited, exchanged, disseminated
- XML/RDF syntax and namespace details are handled within the template
- User input is simple tabular delimited textual data with named fields that will be recognised by the template, e.g.:

  id, material
  123, copper
  234, gold
  345, silver

- Predefined patterns of entities, properties and inverse properties are created by the template. Tabular data populates placeholders ($$) at runtime
- Output is consistent and repeatable
Resultant RDF - example

- Placeholders replaced with (XML encoded) data from named columns

```xml
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:crm="http://www.cidoc-crm.org/cidoc-crm/>

<crm:E22_Man-Made_Object rdf:about="http://myexample/E22_123">
<crm:E12_Production rdf:about="http://myexample/E12_123"/>
<crm:E57_Material rdf:about="http://myexample/E57_copper">

<rdf:Description rdf:about="http://myexample/E22_123">
<crm:P108i_was_produced_by rdf:resource="http://myexample/E12_123"/>
</rdf:Description>

<rdf:Description rdf:about="http://myexample/E57_copper">
<crm:P45i_is_incorporated_in rdf:resource="http://myexample/E22_123"/>
<crm:P126i_was-employed_in rdf:resource="http://myexample/E12_123"/>
</rdf:Description>

<rdf:Description rdf:about="http://myexample/E12_123">
<crm:P108_has_produced rdf:resource="http://myexample/E12_123"/>
<crm:P126_employed rdf:resource="http://myexample/E57_copper"/>
</rdf:Description>
</rdf:RDF>
```
Pattern based mapping and extraction via the CIDOC CRM

Data conversion extract (NMW) using STELLAR templates

```xml
<crm:E22_Man-Made_Object rdf:about="http://tmp/nmw/E22_1000">
  <rdfs:label xml:lang="en">81.79H/1.1</rdfs:label>
  <crm:P1_is_identified_by rdf:resource="http://tmp/nmw/E42_1000" />
  <crm:P140i_was_attributed_by rdf:resource="http://tmp/nmw/E15_1000" />
  <crm:P108i_was_produced_by rdf:resource="http://tmp/nmw/E12_1000" />
  <crm:P2_has_type rdf:resource="http://tmp/nmw/E55_denarius" />
  <crm:P70i_is_documented_in rdf:resource="http://tmp/nmw/E31_crawford" />
  <crm:P70i_is_documented_in rdf:resource="http://tmp/nmw/E31_crawford_222%2f1" />
  <crm:P43_has_dimension rdf:resource="http://tmp/nmw/E54_1000_weight" />
  <crm:P46i_forms_part_of rdf:resource="http://tmp/nmw/E76_nmw+roman" />
  <crm:P128_carries rdf:resource="http://tmp/nmw/E34_1000_reverse" />
</crm:E22_Man-Made_Object>

<crm:E15_Identifier_Assignment rdf:about="http://tmp/nmw/E15_1000">
  <rdfs:label xml:lang="en">81.79H/1.1</rdfs:label>
  <crm:P140_assigned_attribute_to rdf:resource="http://tmp/nmw/E22_1000" />
  <crm:P37i_assigned rdf:resource="http://tmp/nmw/E42_1000" />
</crm:E15_Identifier_Assignment>

<crm:E42_Identifier rdf:about="http://tmp/nmw/E42_1000">
  <rdfs:label xml:lang="en">81.79H/1.1</rdfs:label>
  <crm:P1l_identifies rdf:resource="http://tmp/nmw/E22_1000" />
  <crm:P37i_was_assigned_by rdf:resource="http://tmp/nmw/E15_1000" />
</crm:E42_Identifier>

<crm:E12_Production rdf:about="http://tmp/nmw/E12_1000">
  <rdfs:label xml:lang="en">81.79H/1.1</rdfs:label>
  <crm:P108_has_produced rdf:resource="http://tmp/nmw/E22_1000" />
  <crm:P126_employed rdf:resource="http://tmp/nmw/E57_silver" />
  <crm:P4_has_time-span rdf:resource="http://tmp/nmw/E52_-143%2f-143" />
</crm:E12_Production>
```
"Gruff" visualisation – entities and properties

(Object 52.194 from the NMW Tudor numismatics collection)
Pattern based mapping and extraction via the CIDOC CRM

Templates create shortcuts and hide complexity...

<table>
<thead>
<tr>
<th>context_id</th>
<th>strat_lower_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>456</td>
</tr>
</tbody>
</table>

Intermediate entities, inverse properties and shortcuts all generated by the template
Pattern based mapping and extraction via the CIDOC CRM

... just as user interfaces hide complexity

STAR project query builder - generates and issues SPARQL queries in the background
CRM Shortcuts

- Fully elaborated property paths in CRM event based model can be verbose
- CRM allows for certain ‘shortcut’ properties
- Reasoners could not automatically substitute between fully formed path and shortcut path without additional machine readable information
- Templates can model both alternative paths simultaneously
Pattern based mapping and extraction via the CIDOC CRM

CRM Shortcuts

- P140_assigned_attribute_to
  - E13_Attribute_Assignment
    - P141_assigned
      - E1_CRM_Entity
  - E1_CRM_Entity

- P34_concerned
  (was_assessed_by)
  - E14_Condition_Assessment
    - P35_has_identified
      (identified_by)
      - E3_Condition_State
    - P37_assigned
      (was_assigned_by)
      - E15_Identifier_Assignment

- P140_assigned_attribute_to
  - E1_CRM_Entity
    - P1_is_identified_by
      (identifies)
    - P44_has_condition
      (is_condition_of)
      - E18_Physical_Thing
    - P48_has_preferred_identifier
      (is_preferred_identifier_of)
      - E42_Identifier
      - E41_Appellation

- P142_used_constituent
  (was_used_in)
  - P37_assigned
    (was_assigned_by)
    - E2_CRM_Entity
  - E18_Physical_Thing
  - E3_Condition_State
  - E41_Appellation
  - E42_Identifier
Summary

- Different mappings can potentially pose significant problems for semantic interoperability (cf BRICKS).
- Reasoning is an important possibility for CRM and there will be cases where clearly needed.
- However do not need to create unnecessary alternative paths for similar data
- Pragmatic approach: combine developments in reasoning with efforts at consensus on patterns for CRM mappings and guidelines.
**Summary**

- Mapping and extraction process is inherently complex, needs tools to maintain consistency at scale, and repeatable workflow.
- Templates can simultaneously model multiple alternate paths (e.g. shortcuts) and alternate representations (e.g. E55 Type / SKOS Concept).
- Templates handle lower level syntax issues and implement predefined patterns of data - improving consistency and hiding complexity – if we can just agree on the patterns (!).
Future?

- Agreement on implementation details?
- Agreement on mapping patterns and guidelines?
- Possible to state purpose of a mapping exercise?
- Registries of mapping patterns?
- Core metadata for mapping patterns?
Pattern based mapping and extraction via the CIDOC CRM

Douglas Tudhope\textsuperscript{1}, Ceri Binding\textsuperscript{1}, Keith May\textsuperscript{2}, Michael Charno\textsuperscript{3}
\textsuperscript{1}University of South Wales, \textsuperscript{2}English Heritage, \textsuperscript{3}Archaeology Data Service

douglas.tudhope@southwales.ac.uk
ceri.binding@southwales.ac.uk
keith.may@english-heritage.org.uk
michael.charno@york.ac.uk

Hypermedia Research Unit, University of South Wales

http://hypermedia.research.southwales.ac.uk/
Archaeology Data Service (ADS) Linked Data

Pattern based mapping and extraction via the CIDOC CRM

This page shows information obtained from the SPARQL endpoint at http://data.archaeologydataservice.ac.uk/sparql/repositories/archives. You can query the endpoint directly with a SPARQL client or at our SPARQL query interface. This data is also available as RDF/XML and Turtle.

View the data in the following other browsers:

- Disco
- Tabulator

University of York legal statements | ADS terms and conditions
**CRM Shortcuts**

- **E70_Thing**
  - P39 measured (was_measured_by)
  - P40 observed_dimension (was_observed_in)
  - P43 has_dimension (is_dimension_of)

- **E16_Measurement**
  - E17_Type_Assignment
    - P41 classified (was_classified_by)
    - P42 assigned (was_assigned_by)
  - P2 has_type (is_type_of)

- **E1_CRM_Entity**

- **E54_Dimension**

- **E17_Type_Assignment**

- **E55_Type**

- **E8_Acquisition**
  - P24 transferred_title_of (changed_ownership_through)
  - P52 has_current_owner (is_current_owner_of)
  - P51 has_former_or_current_owner (is_former_or_current_owner_of)

- **E18_Physical_Thing**

- **E39_Actor**

- **E8_Acquisition**
  - P22 transferred_title_to (acquired_title_through)
### CRM Shortcuts (3 of 4)

- **E24_Physical_Man-Made_Thing**
  - P65_shows_visual_item
    - (is_shown_by)
  - P62_depicts
    - (is(depicted by))

- **E18_Physical_Thing**
  - P58_has_section_definition
    - (defines_section)
  - P59_has_section
    - (is_located_on_or_within)

- **E46_Section_Definition**
  - P87_is_identified_by
    - (identifies)
  - P59_has_section
    - (is_located_on_or_within)

- **E1_CRM_Entity**
  - P138_represents
    - (has_representation)

- **E19_Physical_Object**
  - P25_moved
    - (moved_by)
  - P55_has_current_location
    - (is_current_location_of)
  - P53_has_former_or_current_location
    - (is_former_or_current_location_of)

- **E36_Visual_Item**
  - P62_depicts
    - (is(depicted by))

- **E46_Section_Definition**
  - P59_has_section
    - (is_located_on_or_within)

- **E9_Move**
  - P26_moved_to
    - (was_destination_of)
CRM Shortcuts (4 of 4)

- **E10_Transfer_of_Custody**
  - P30_transferred_custody_of (custody_transferred_through)
  - P50_has_current_keeper (is_current_keeper_of)
  - P49_has_former_or_current_keeper (is_former_or_current_keeper_of)

- **E18_Physical_Thing**
  - P29_custody_received_by (received_custody_through)

- **E74_Group**
  - P143 Joined (wasJoined_by)

- **E74_Group**
  - P144_joined_with (gained_member_by)

- **E19_Physical_Object**
  - P56_bears_feature (is_found_on)

- **E26_Physical_Feature**
  - P59_has_section (is_located_on_or_within)

- **E53_Place**
  - P53_has_former_or_current_location (is_former_or_current_location_of)

- **E39_Actor**
  - P107_has_current_or_former_member (is_current_or_former_member_of)