

Pattern based mapping and extraction via the CIDOC CRM

Douglas Tudhope¹, Ceri Binding¹, Keith May², Michael Charno³

(¹University of South Wales, ²English Heritage, ³Archaeology 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/>

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



Another dataset, another schema...

- Perform a cross search on small finds and materials?

The image displays seven overlapping Notepad windows, each showing a different CSV file with various column headers. Red circles highlight specific headers in each window:

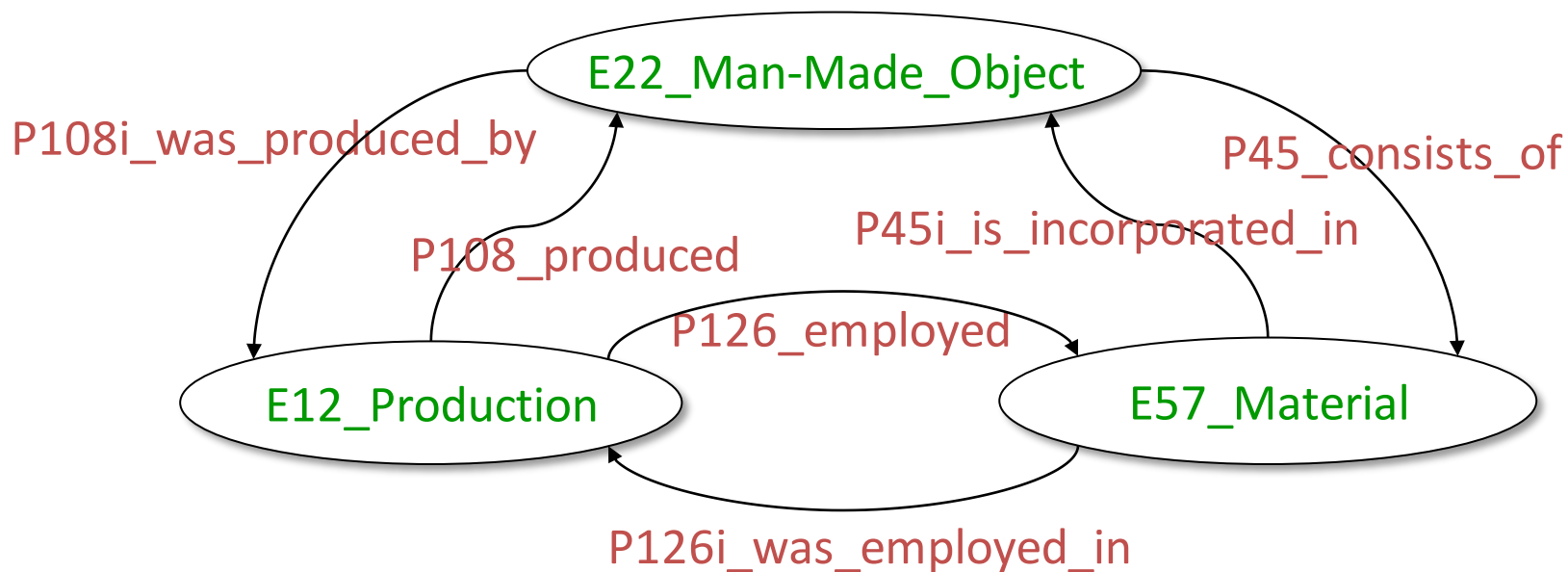
- finds.csv**: Headers include ID, PID, REC, TYPE, SFNO, MATERIAL, NAME, CONTEXT, AOD, EASTINGS, NORTHINGS, START_DATE, END_DATE, DATE_TERM, QUANTITY, SPOTDATE.
- WSW_030_Small_finds.csv**: Headers include ID, Small find no., context, Period, Material, Object Name, Finds category, No of frags, weight, Dimensions (Length), Dimensions (width), Dimensions (Depth), Diameter, X-ray.
- FTF97_FIND.csv**: Headers include FINDNO, CLASS, WEIGHT, NOS, CTX, SECTION, LAYER.
- arc_whs98_finds.csv**: Headers include EVENT_NAME, EVENT_CODE, CONTEXT, SPECIAL_NUM, MATERIAL, COUNT, WEIGHT, PERIOD, TYPE, COMMENTS.
- FND_RF.csv**: Headers include SITECODE, CONTEXT, ACC_NO, MATER, OBJECT, C, P, I, C, PUB, PERI, COMMENTS, E_DATE, DATE_MOD, FIND_GROUP, OBJECTID, SIT E_C.
- database_small_finds_register.csv**: Headers include Small Find Number, photo, Context No, Layer, material, type, Period, Date Recorded, Level, Eastings, Northings, Height (mm), width (mm), Thickness (mm), Diameter (mm), weight (g), sent to, Date Sent, Date Returned, Notes.
- FLN_009_SF.csv**: Headers include SF No., context, Material, Category, Find type, No., wt/g, Notes, Date.

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...

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

STELLAR Project

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

STELLAR

An application for converting delimited (CSV) format data to valid RDF data conforming to a chosen 'template'.

Delimited Data File test_crneh_contexts_strat_lower_id.csv

Template name

Namespace prefix

Validator Type the two words:

Results

[nia33qr.rdf](#)
(the download link will remain available on the server for 30 minutes, after which it will be automatically deleted)

Statistics

- 5 unique subject URIs
- 8 unique object URIs
- 1 unique literals using 1 language
- 3 unique class URIs:
 - `<http://purl.org/crneh#EHE0007_Context>` [1]
 - `<http://purl.org/crneh#EHE0061_ContextUID>` [1]
 - `<http://purl.org/crneh#EHE1001_ContextEvent>` [2]
- 14 statements, using 9 predicate URIs:
 - `<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>` [4]
 - `<http://www.w3.org/2000/01/rdf-schema#label>` [1]
 - `<http://www.w3.org/1999/02/22-rdf-syntax-ns#value>` [3]
 - `<http://purl.org/NET/crm-ow#P87_is_identified_by>` [1]
 - `<http://purl.org/NET/crm-ow#P871_identifies>` [1]
 - `<http://purl.org/NET/crm-ow#P71_look_place_at>` [2]
 - `<http://purl.org/NET/crm-ow#P71_witnessesD>` [2]
 - `<http://purl.org/NET/crm-ow#P120_occurs_before>` [1]
 - `<http://purl.org/NET/crm-ow#P120i_occurs_after>` [1]

STELLAR.Console v1.0

(type HELP for instructions)

STELLAR.Console>help
For information on a particular command type HELP command

- DBNAMES List databases in a directory
- DBTABLES List tables in a database
- DBCOLUMNS List columns in a database table
- DBROWCOUNT Count rows in a database table
- TAB2DB Import tab delimited file to database table
- CSU2DB Import comma delimited file to database table
- SQL2CSU Run SQL, export result to CSU file
- SQL2TAB Run SQL, export result to tab delimited file
- CSU2RDF Convert comma delimited file to RDF file (via template)
- TAB2RDF Convert tab delimited file to RDF file (via template)
- SQL2RDF Run SQL, export result to RDF file (via template)
- TEMPLATES Display list of possible templates to use
- CSUSTATS Display statistics for comma delimited file
- RDFSSTATS Display statistics for RDF file
- EXIT Exit the application

STELLAR.Console>

Stellar.Win v1.0: C:\Projects\IntrasisLinkedData\data\phase_testing.stellar

File Help
New Open Save Help

Conversion description: Test conversion from Intrasis PHASE data to RDF

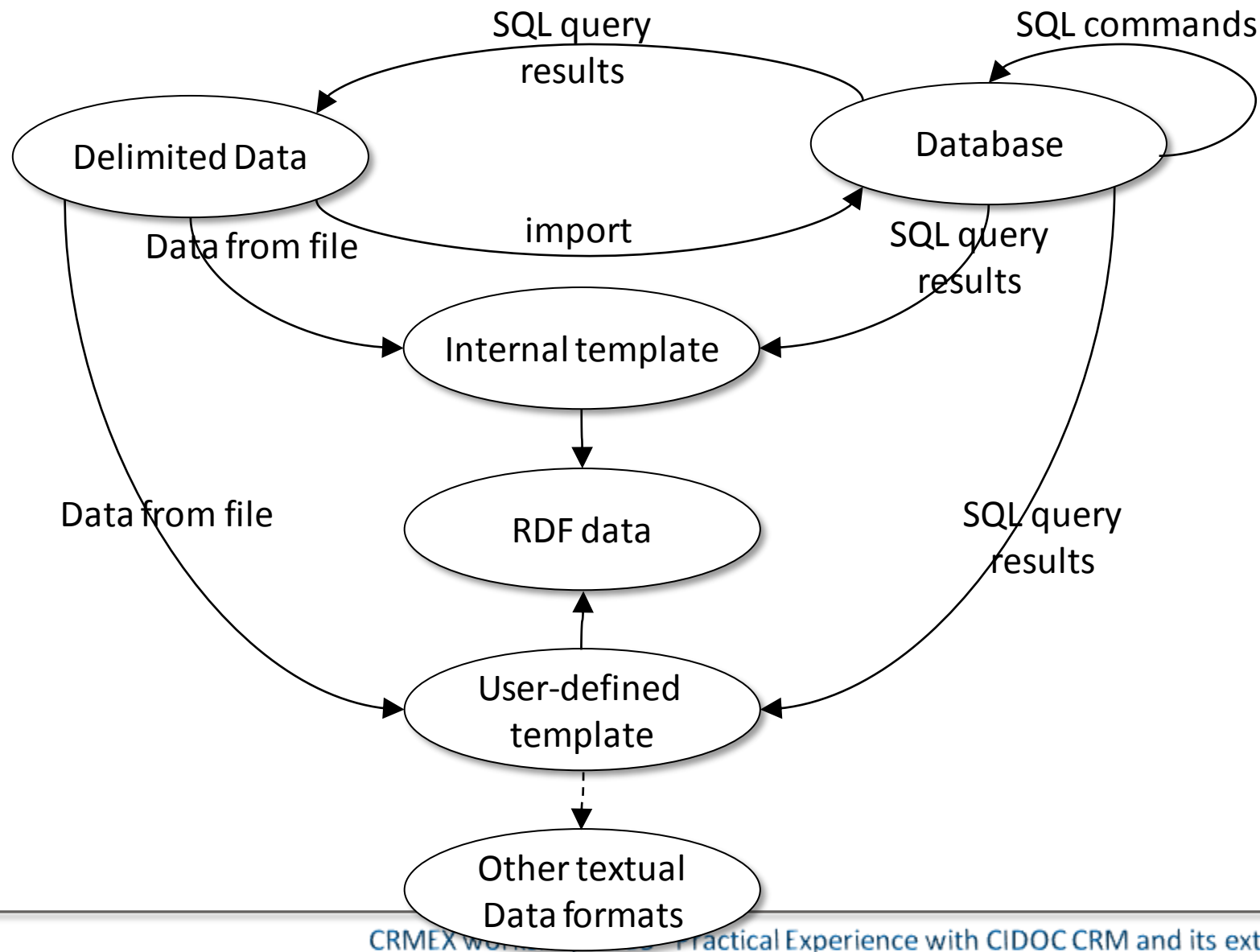
Equivalent STELLAR Console command: DELIMITED /data: C:\Projects\IntrasisLinkedData\data\Phase_testing.csv" /delimiter: ; /tag: C:\Program Files\STELLAR\templates\CRMEH_PHASES.stg"

Converted data file: C:\Projects\IntrasisLinkedData\data\Phase_te Choose... Edit...

Converted output: <?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/2000/01/rdf-schema#" xmlns:scsd="http://www.w3.org/2001/XMLSchema#" xmlns:owid="http://www.w3.org/2002/07/owid#" xmlns:skos="http://www.w3.org/2004/02/skos/core#" xmlns:skosd="http://www.w3.org/2006/05/skos-xld#" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:crneh="http://eulangen-crm.org/101001/" xmlns:crnehH="http://purl.org/crneh#">
<crneh:EHE001_GroupEventRecord
rdf:about="http://mp/data/EHE001_PHASE_10001489">
 <rdf:label xml:lang="en">10001489:</rdf:label</crneh:EHE001_GroupEventRecord>

Options file: C:\Projects\IntrasisLinkedData\data\OPTIONS Choose... Edit...

STELLAR data conversions



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

//HEADER template is output once at start of processing

```
HEADER(options) ::= <<
  <?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/">
>>
```

//RECORD template is output once per data row

```
RECORD(options, data) ::= <<
  <crm:E22_Man-Made_Object rdf:about="http://myexample/E22_$(data.id)" />
  <crm:E12_Production rdf:about="http://myexample/E12_$(data.id)" />
  <crm:E57_Material rdf:about="http://myexample/E57_$(data.material)" />

  <rdf:Description rdf:about="http://myexample/E22_$(data.id)">
  <crm:P45_consists_of rdf:resource="http://myexample/E57_$(data.material)" />
  <crm:P108i_was_produced_by rdf:resource="http://myexample/E12_$(data.id)" />
  </rdf:Description>

  <rdf:Description rdf:about="http://myexample/E57_$(data.material)">
  <crm:P45i_is_incorporated_in rdf:resource="http://myexample/E22_$(data.id)" />
  <crm:P126i_was_employed_in rdf:resource="http://myexample/E12_$(data.id)" />
  </rdf:Description>

  <rdf:Description rdf:about="http://myexample/E12_$(data.id)">
  <crm:P108_has_produced rdf:resource="http://myexample/E22_$(data.id)" />
  <crm:P126_employed rdf:resource="http://myexample/E57_$(data.material)" />
  </rdf:Description>
>>
```

//FOOTER template is output once at end of processing

```
FOOTER(options) ::= <<
  </rdf: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 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:P45_consists_of rdf:resource="http://myexample/E57_copper" >
    <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>

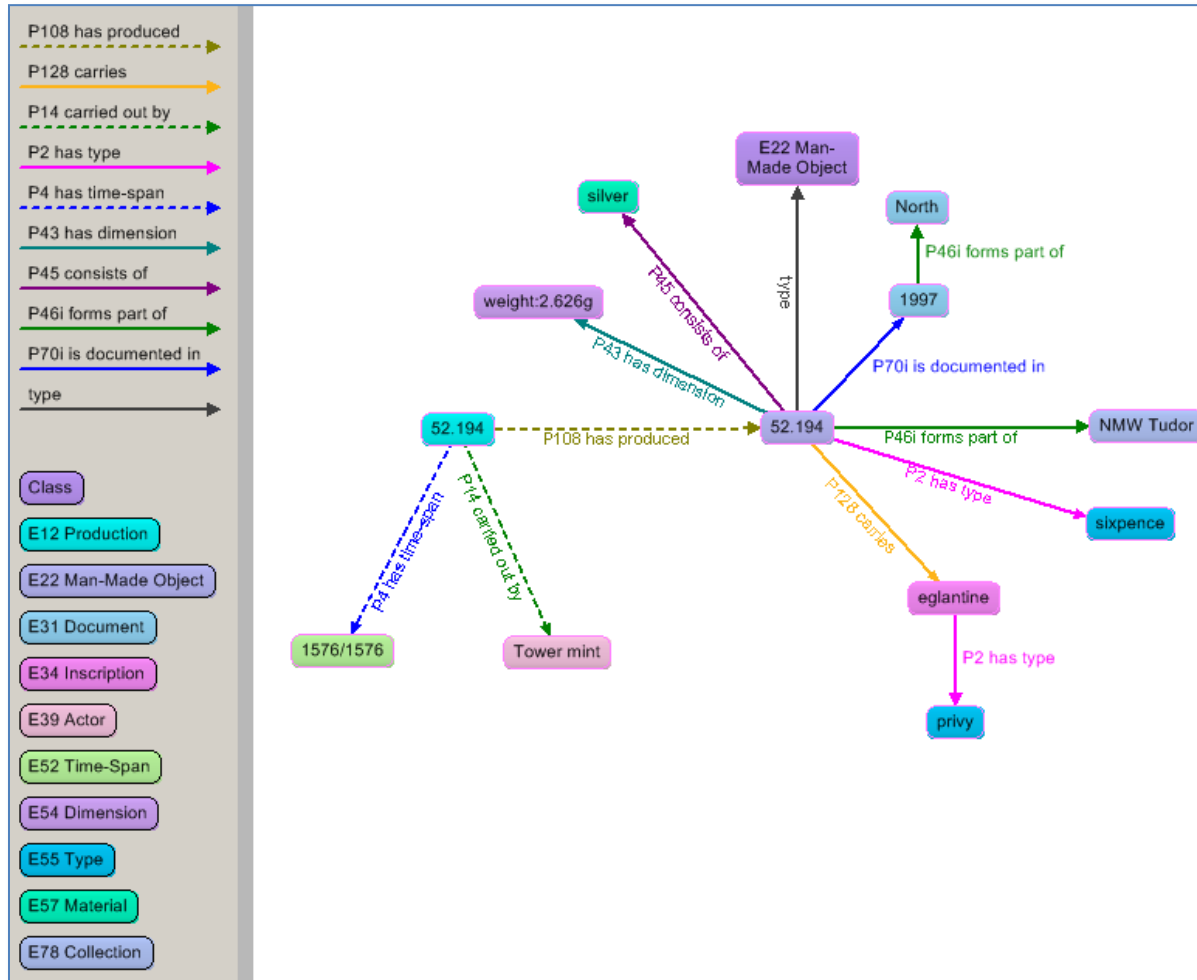
  <rdf:Description rdf:about="http://myexample/E12_123">
    <crm:P108_has_produced rdf:resource="http://myexample/E12_123" />
    <crm:P126_employed rdf:resource="http://myexample/E57_copper" />
  </rdf:Description>

</rdf:RDF>
```

Data conversion extract (NMW) using STELLAR templates

```
<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:P45_consists_of rdf:resource="http://tmp/nmw/E57_silver" />
  <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/E78_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:P37_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:P1i_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:P14_carried_out_by rdf:resource="http://tmp/nmw/E39_rome+mint" />
  <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)

Templates create shortcuts and hide complexity...

| context_id | strat_lower_id |
|------------|----------------|
| 123 | 456 |

Template input

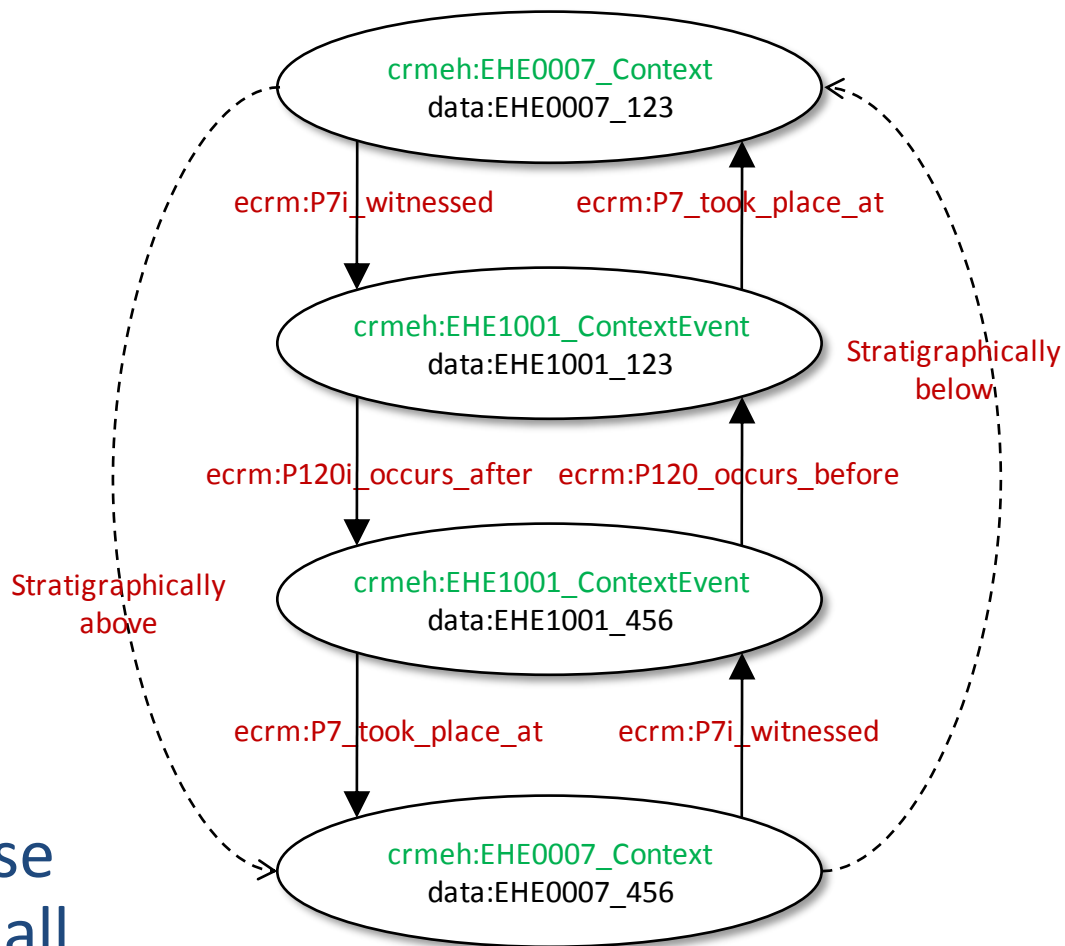
```

<crmeh:EHE0007_Context rdf:about="http://stellar/EHE0007_123"/>
<crmeh:EHE0007_Context rdf:about="http://stellar/EHE0007_456"/>
<crmeh:EHE1001_ContextEvent rdf:about="http://stellar/EHE1001_123"/>
<crmeh:EHE1001_ContextEvent rdf:about="http://stellar/EHE1001_456"/>
<rdf:Description rdf:about="http://stellar/EHE1001_123">
  <ecrm:P7_took_place_at rdf:resource="http://stellar/EHE0007_123"/>
</rdf:Description>
<rdf:Description rdf:about="http://stellar/EHE0007_123">
  <ecrm:P7i_witnessed rdf:resource="http://stellar/EHE1001_123"/>
</rdf:Description>
<rdf:Description rdf:about="http://stellar/EHE1001_456">
  <ecrm:P7_took_place_at rdf:resource="http://stellar/EHE0007_456"/>
</rdf:Description>
<rdf:Description rdf:about="http://stellar/EHE0007_456">
  <ecrm:P7i_witnessed rdf:resource="http://stellar/EHE1001_456"/>
</rdf:Description>
<rdf:Description rdf:about="http://stellar/EHE1001_456">
  <ecrm:P120_occurs_before rdf:resource="http://stellar/EHE1001_123"/>
</rdf:Description>
<rdf:Description rdf:about="http://stellar/EHE1001_123">
  <ecrm:P120i_occurs_after rdf:resource="http://stellar/EHE1001_456"/>
</rdf:Description>

```

Template output

Intermediate entities, inverse properties and shortcuts all generated by the template



Data modelled

... just as user interfaces hide complexity

The screenshot shows the STAR project query builder interface. At the top, there are tabs for 'Groups', 'Contexts', 'Finds', and 'Samples'. Below these, a list of context IDs is displayed, including 1164, 3201, 3260, 3272, 3273, and 5757. A search filter is visible, showing 'Context ID' with the value '2036' entered. The interface is designed to be user-friendly, hiding the underlying complexity of the CIDOC CRM data.

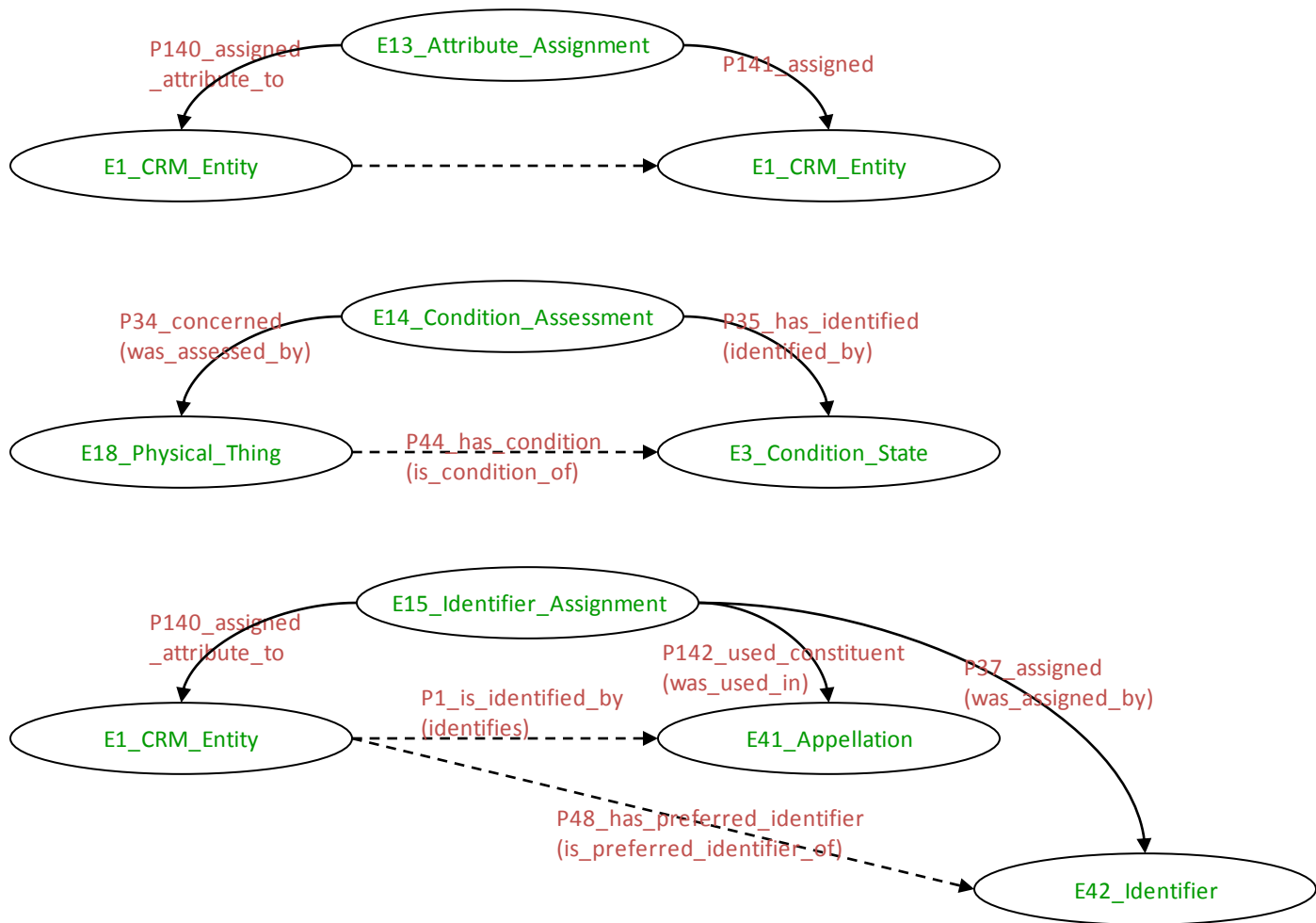
The screenshot shows the 'Context Details' interface. It features a 'Hierarchy/Stratigraphy' diagram with nodes 2036, 1164, and 1681 4599. Below the diagram, the context information is displayed, including the Site '#ehe0001.leap', Context ID '1164', and Context Type 'Partition wall between rooms 1 and 2 aligned SW-NE, House 1'. The URL <http://tempuri/star/base#ehe0007.leap.contexts.context.1164> is also shown.

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

CRM Shortcuts



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 Tudhope¹, Ceri Binding¹, Keith May², Michael Charno³

(¹University of South Wales, ²English Heritage, ³Archaeology 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


ARCHAEOLOGY
DATA SERVICE


http://data.archaeologydataservice.ac.uk/10.5284/1000365/EHE0007_1010

1010

| Property | Value |
|------------------------|---|
| ?:P2_has_type | < http://data.archaeologydataservice.ac.uk/10.5284/1000365/E55_EHE0007_deposit > |
| ?:P3_has_note | fill of posthole |
| ?:P7i_witnessed | < http://data.archaeologydataservice.ac.uk/10.5284/1000365/EHE1001_1010 > |
| ?:P87_is_identified_by | < http://data.archaeologydataservice.ac.uk/10.5284/1000365/EHE0061_1010 > |
| ?:P89_falls_within | < http://data.archaeologydataservice.ac.uk/10.5284/1000365/EHE0003_hcb+03 > < http://data.archaeologydataservice.ac.uk/10.5284/1000365/EHE0005_146 > < http://data.archaeologydataservice.ac.uk/10.5284/1000365/EHE0007_1009 > |
| ?:label | 1010 |
| ?:type | < http://purl.org/crmeh#EHE0007_Context > |



Metadata [\[show\]](#)

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](#).

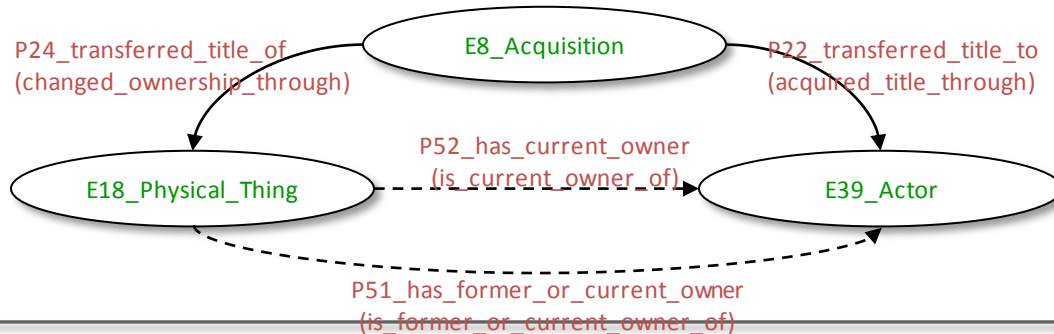
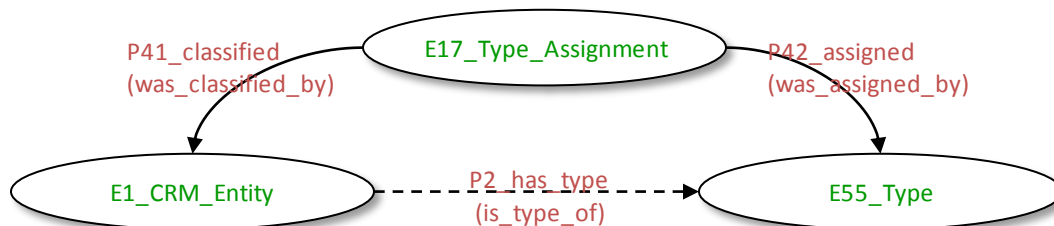
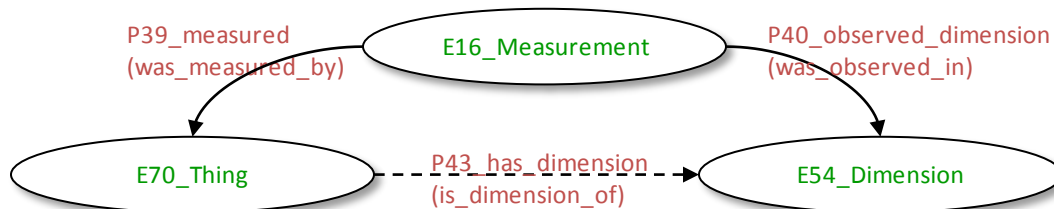
View the data in the following other browsers:

- [Disco](#)
- [Tabulator](#)

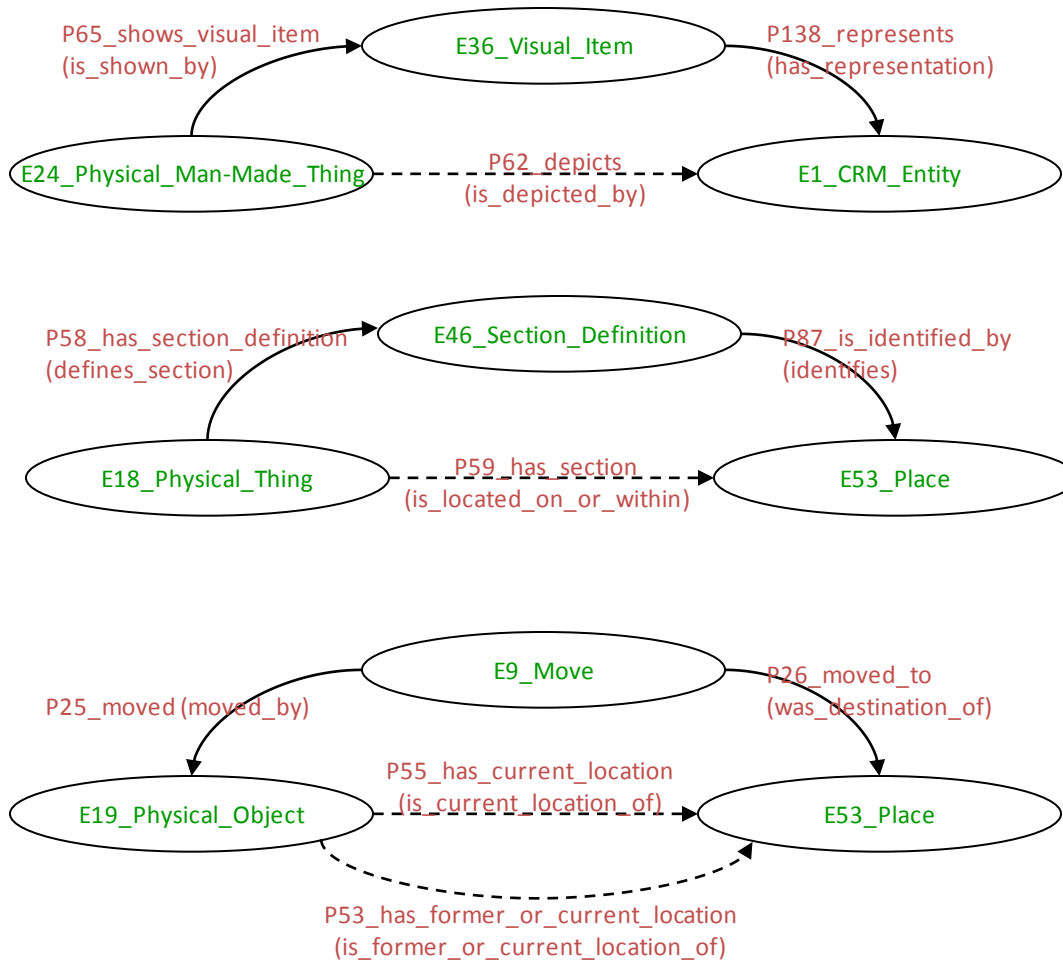
[University of York legal statements](#) | [ADS terms and conditions](#)


ARCHAEOLOGY
DATA SERVICE


CRM Shortcuts



CRM Shortcuts (3 of 4)



CRM Shortcuts (4 of 4)

