KOS-based tools for archaeological dataset interoperability: NKOS Workshop, ECDL 2010

C. Binding, K. May[1], D. Tudhope, A. Vlachidis

Hypermedia Research Unit, University of Glamorgan

[1]English Heritage

Arts & Humanities Research Council

STAR Project

Semantic Technologies for Archaeological Resources

- AHRC funded project
- In collaboration with English Heritage
- http://hypermedia.research.glam.ac.uk/kos/star/

STAR Aims and Background

- *Investigate the potential of semantic terminology tools for widening access to digital archaeology resources, including disparate data sets and associated grey literature*

- *Open up the grey literature to scholarly research by investigating the combination of linguistic and KOS-based methods in the digital archaeology domain.*

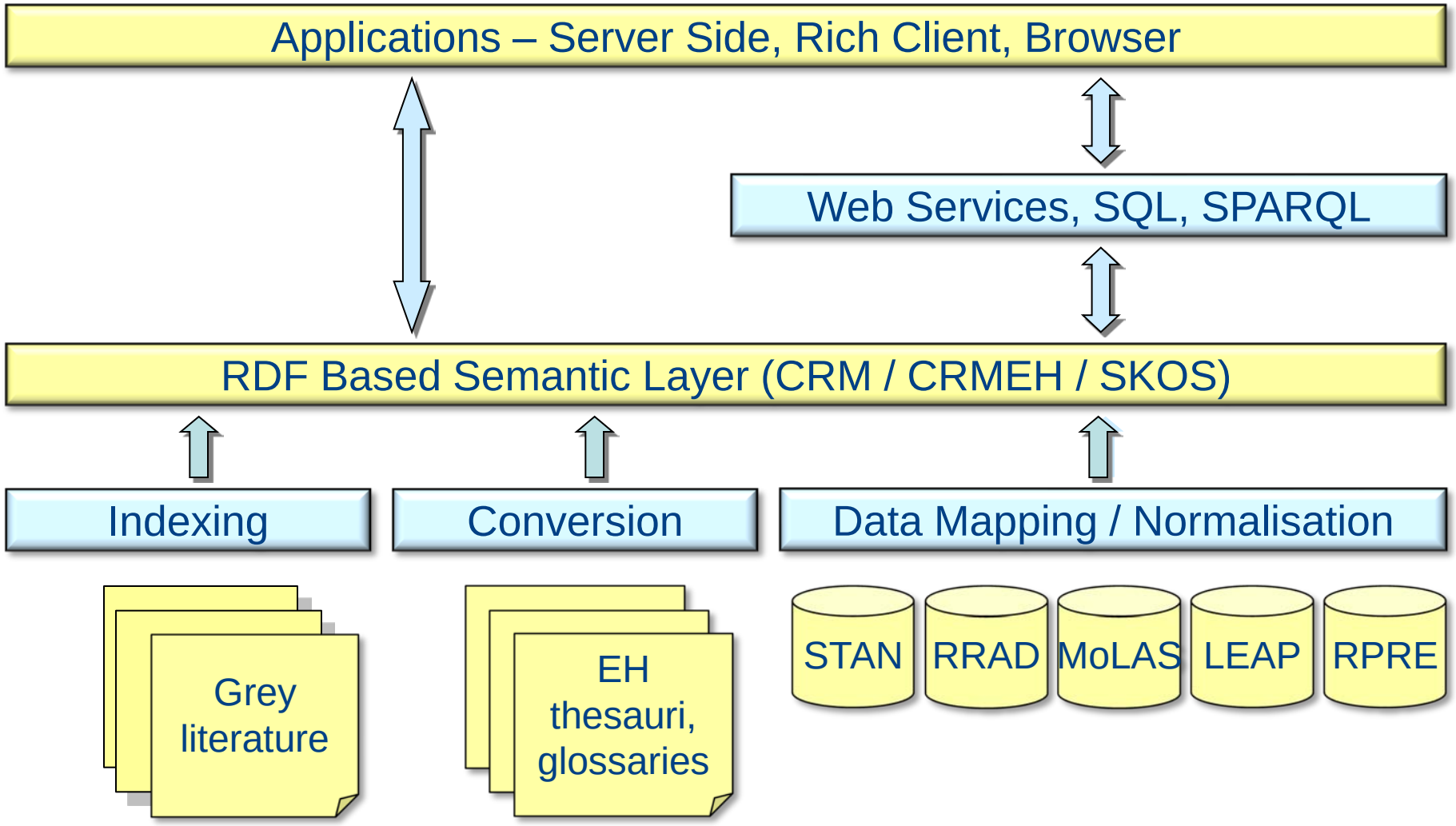
- *Develop new methods for enhancing linkages between digital archive database resources and to associated grey literature, exploiting the potential of a high level, core ontology.*

- Current situation one of fragmented datasets and applications, with different terminology systems

- Need for integrative metadata framework

  EH have designed an upper ontology based on CRM standard

STAR Project - General Architecture

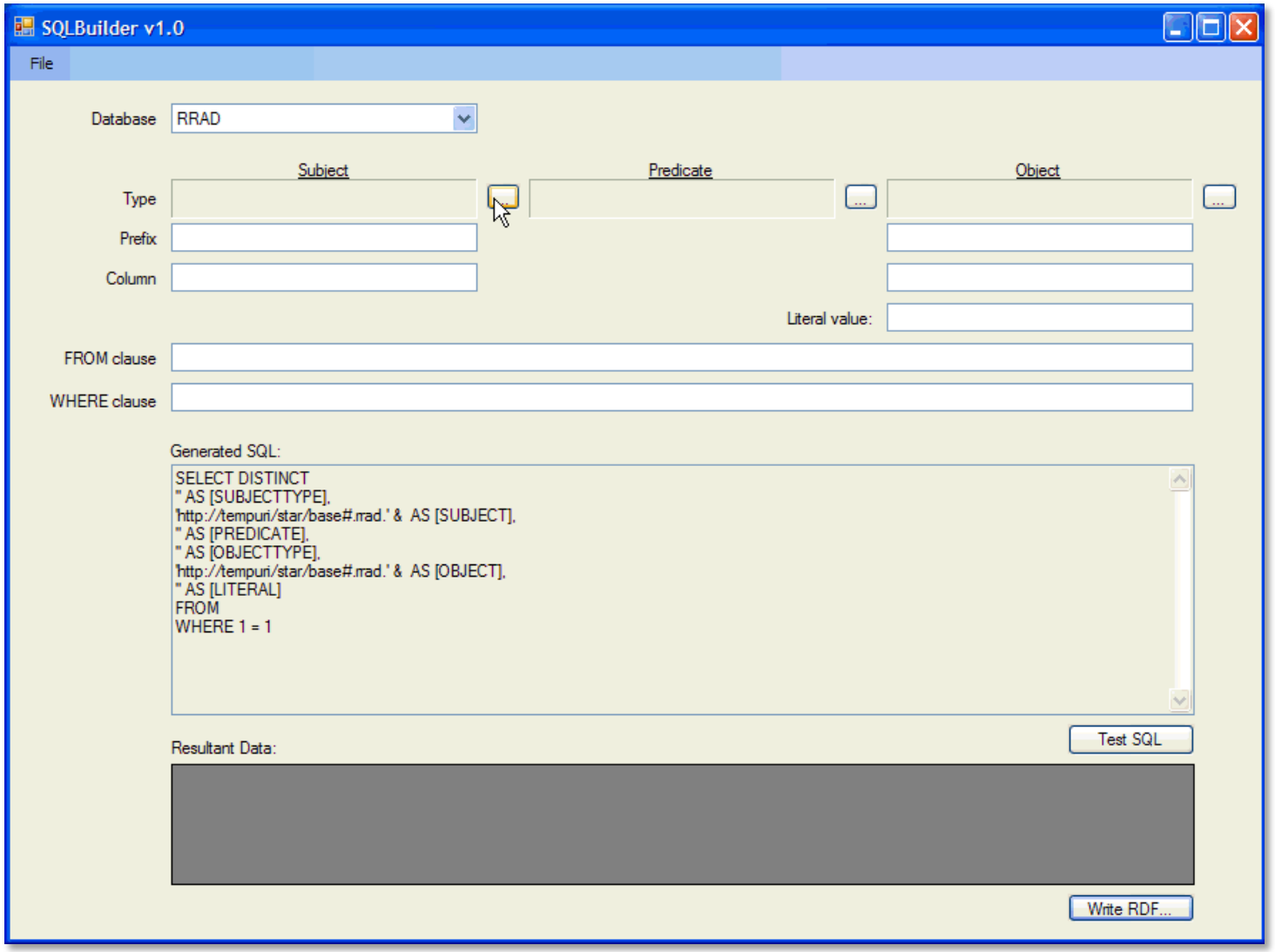


CRM Event Based model - Property chains

- CRM event model – events not explicit in datasets OR mappings
  - Additional work required to satisfy logical mappings
- E.g. Sample *taken from* Context:

crmeh:EHE0018.Sample [*crm:E18.PhysicalStuff*]
→ *crm:P113B.was_removed_by*
→ crmeh:EHE2006.ContextSamplingEvent [*crm:E80.PartRemoval*]
→ *crm:P112F.diminished*
→ crmeh:EHE0008.ContextStuff [*crm:E18.PhysicalStuff*]
→ crmeh:EHP3.occupied
→ crmeh:EHE0007.Context [*crm:E53.Place*]

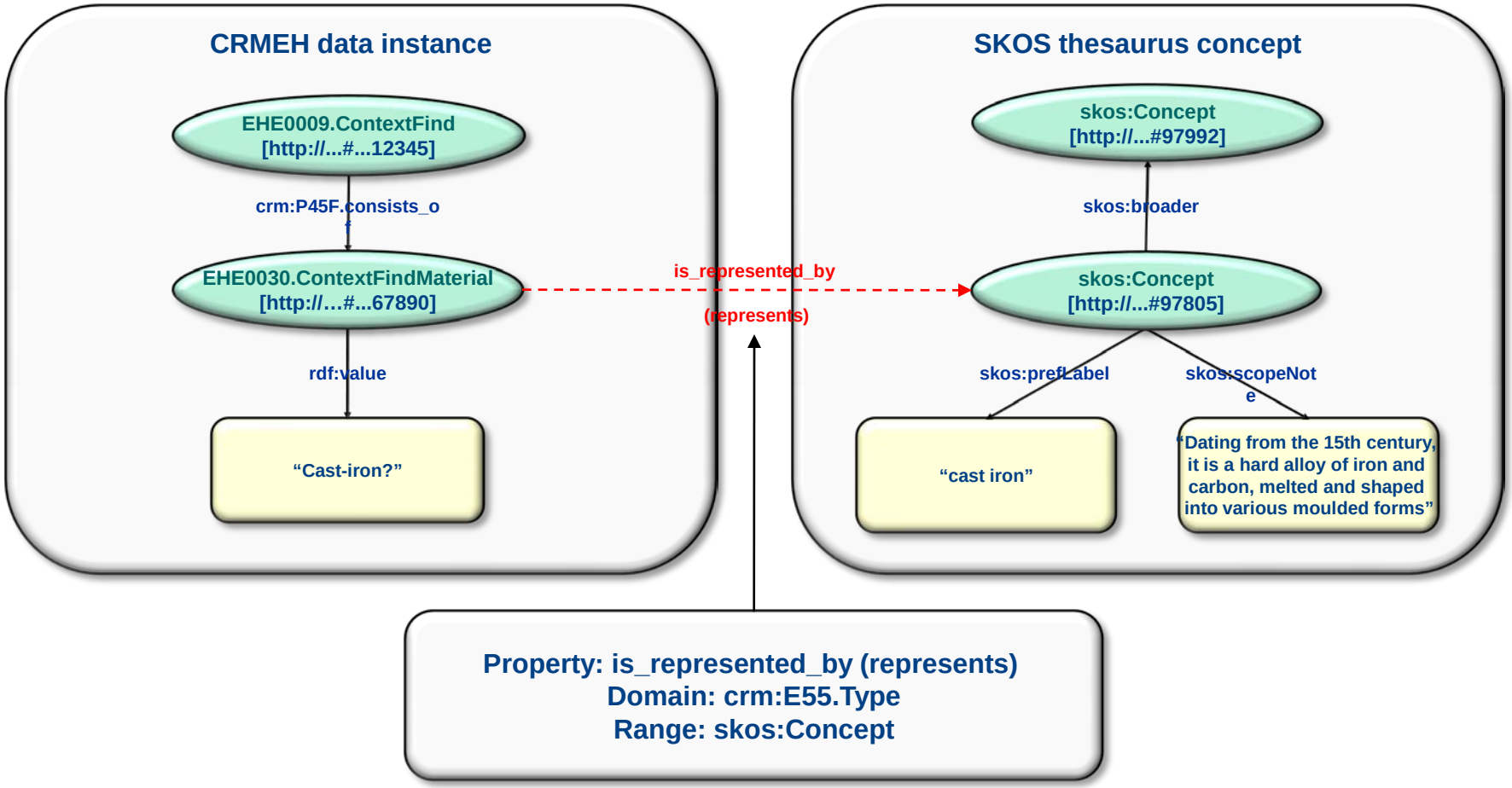
RDF Data Extraction Tool



Resultant extracted data (RDF/XML)

```
<?xml version="1.0"?>
<rdf:RDF xmlns:crmeh="http://tempuri/star/crmeh#" xmlns:crm="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" xml:base="http://tempuri/star/base#">
    <crmeh:EHE0007.Context rdf:about="http://tempuri/star/base#ehe0007.rrad.context.contextno.1">
        <crm:P3F.has_note>
            <crmeh:EHE0046.ContextNote rdf:about="http://tempuri/star/base#ehe0046.rrad.context.description.1">
                <rdf:value>Upper ploughsoil over whole site no Sub-division for the convenience of finds processing '1' contains finds contexts '3759', '3760' and '3763'.</rdf:value>
            </crmeh:EHE0046.ContextNote>
        </crm:P3F.has_note>
    </crmeh:EHE0007.Context>
    <crmeh:EHE0007.Context rdf:about="http://tempuri/star/base#ehe0007.rrad.context.contextno.2">
        <crm:P3F.has_note>
            <crmeh:EHE0046.ContextNote rdf:about="http://tempuri/star/base#ehe0046.rrad.context.description.2">
                <rdf:value>Sub-division of '60' in villa area.</rdf:value>
            </crmeh:EHE0046.ContextNote>
        </crm:P3F.has_note>
    </crmeh:EHE0007.Context>
    <crmeh:EHE0007.Context rdf:about="http://tempuri/star/base#ehe0007.rrad.context.contextno.3">
        <crm:P3F.has_note>
            <crmeh:EHE0046.ContextNote rdf:about="http://tempuri/star/base#ehe0046.rrad.context.description.3">
                <rdf:value>Destruction layer of building material over main villa structure. Consists of much wallstone and roof slate (stone). Also mortar; some tile; op sig, wall plaster and tesserae. Lying immediately beneath the lower ploughsoil (2) it mounds up over the middle of the building, thinning out towards the edges. It is cut by a series of deep plough marks, remnant of medieval ridge and furrow which does not show at ground level. To the east of corridor 24, the destruction material contains more mortar than on the west - also some box flue tile</rdf:value>
            </crmeh:EHE0046.ContextNote>
        </crm:P3F.has_note>
    </crmeh:EHE0007.Context>
    <crmeh:EHE0007.Context rdf:about="http://tempuri/star/base#ehe0007.rrad.context.contextno.4">
        <crm:P3F.has_note>
            <crmeh:EHE0046.ContextNote rdf:about="http://tempuri/star/base#ehe0046.rrad.context.description.4">
                <rdf:value>A shallow linear depression orientated on an east-west axis running across the width of excavation. Original recorded coordinates: 0980/0980</rdf:value>
            </crmeh:EHE0046.ContextNote>
        </crm:P3F.has_note>
    </crmeh:EHE0007.Context>
```

STAR implementation - linking CRM instances to SKOS concepts



STAR – Web Services and Client Applications

- Windows applications
- Browser components
- Full text search
- Browse concept space
- Navigate via expansion
- Cross search archaeological datasets

STAR Client Applications

STAR Web Services

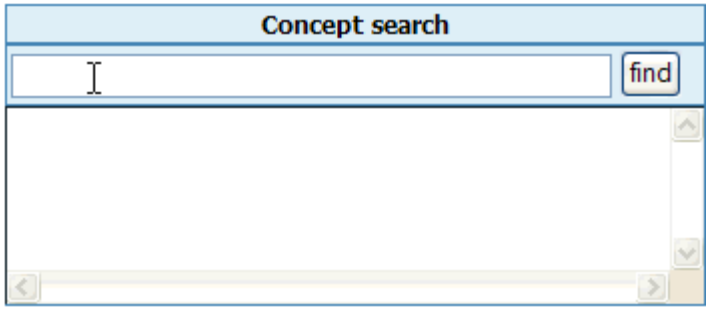
English Heritage thesauri (SKOS)

Grey literature indexing

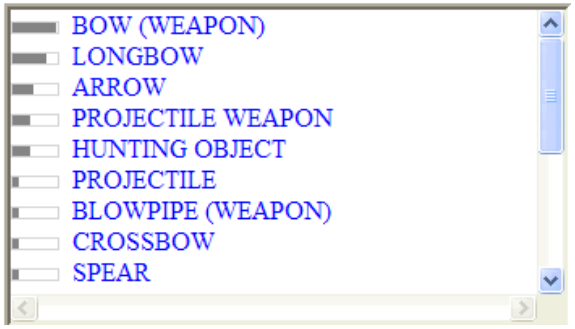
Archaeological Datasets (CRM)

STAR Datasets

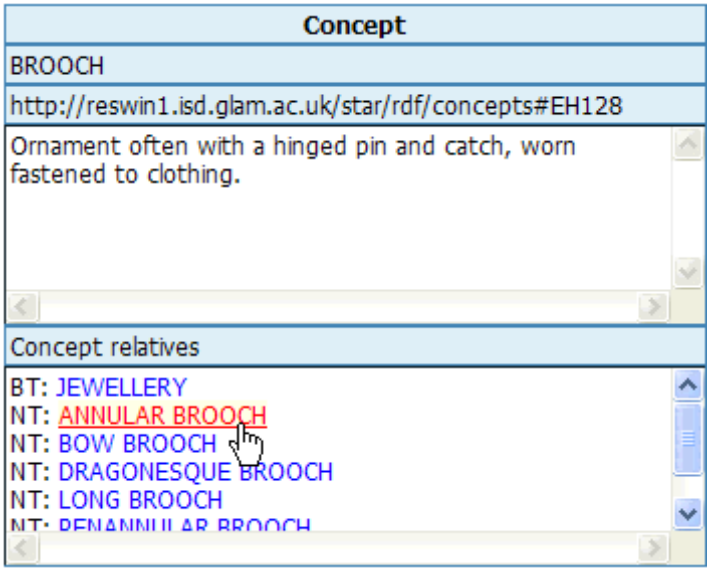
STAR – Web Client Components



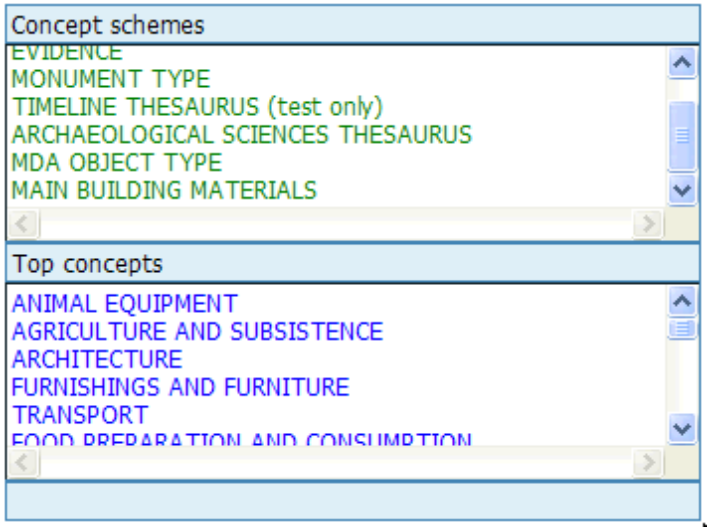
Search across multiple thesauri



Navigate via semantic expansion

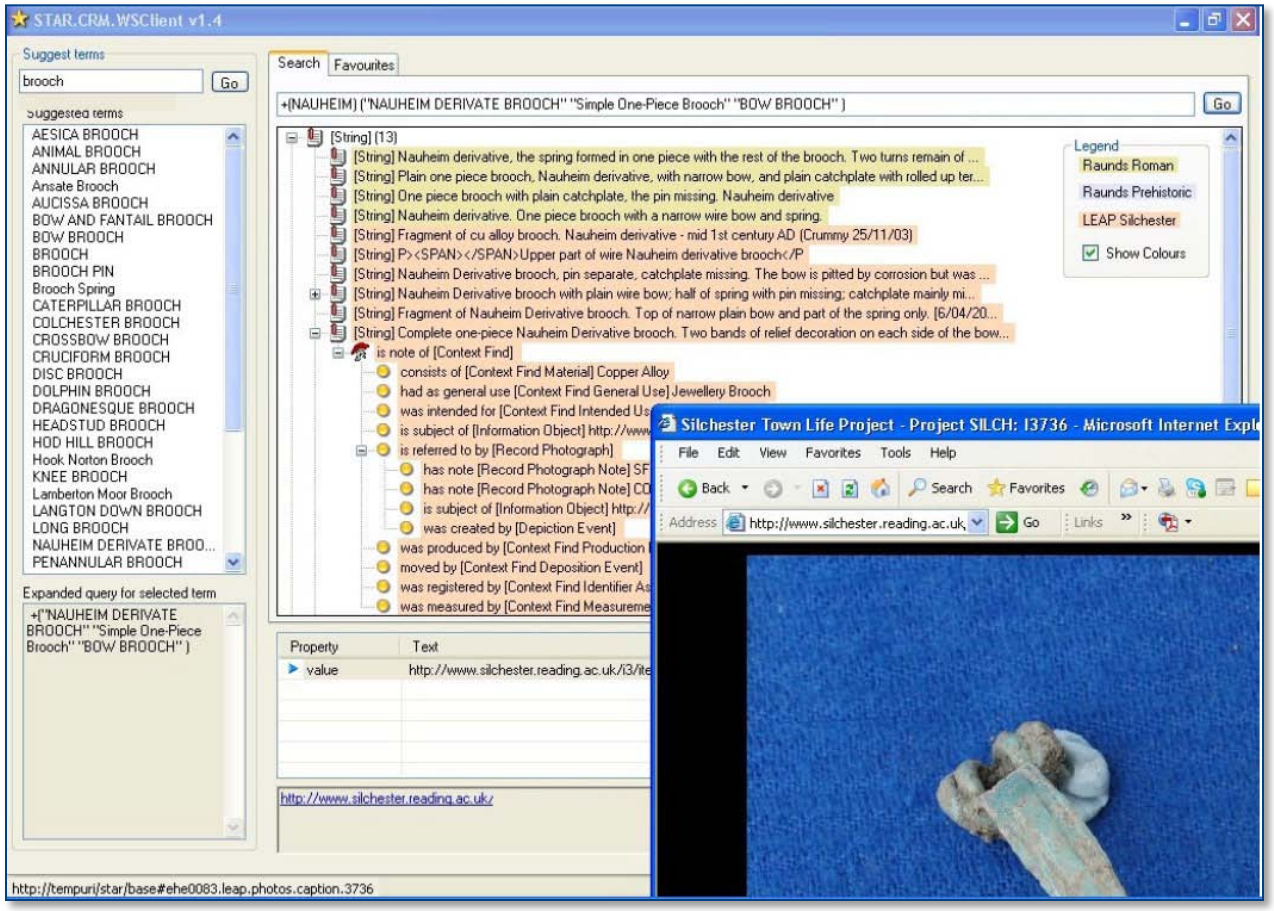


Display concept details



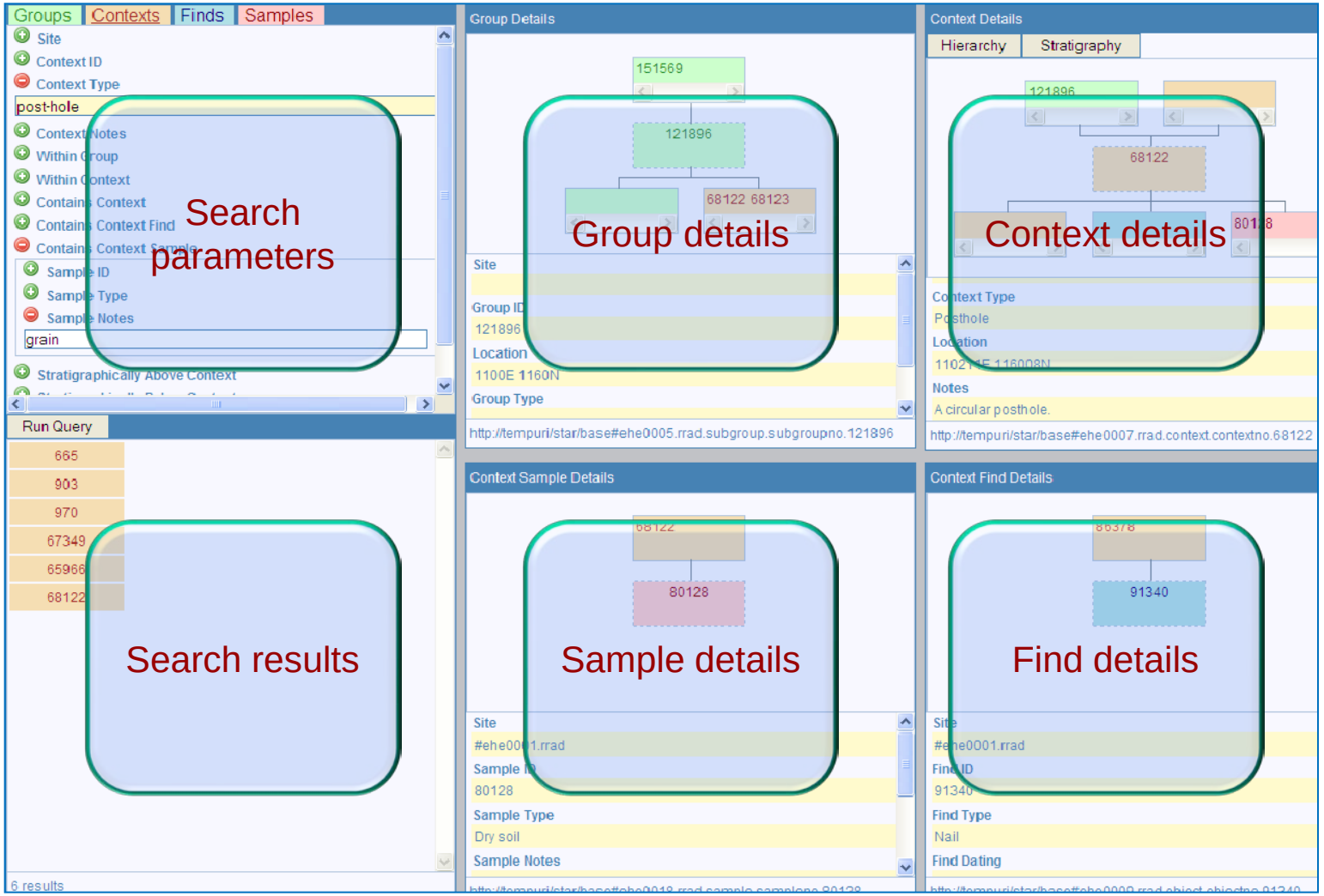
Browse available thesauri

Preliminary prototype application

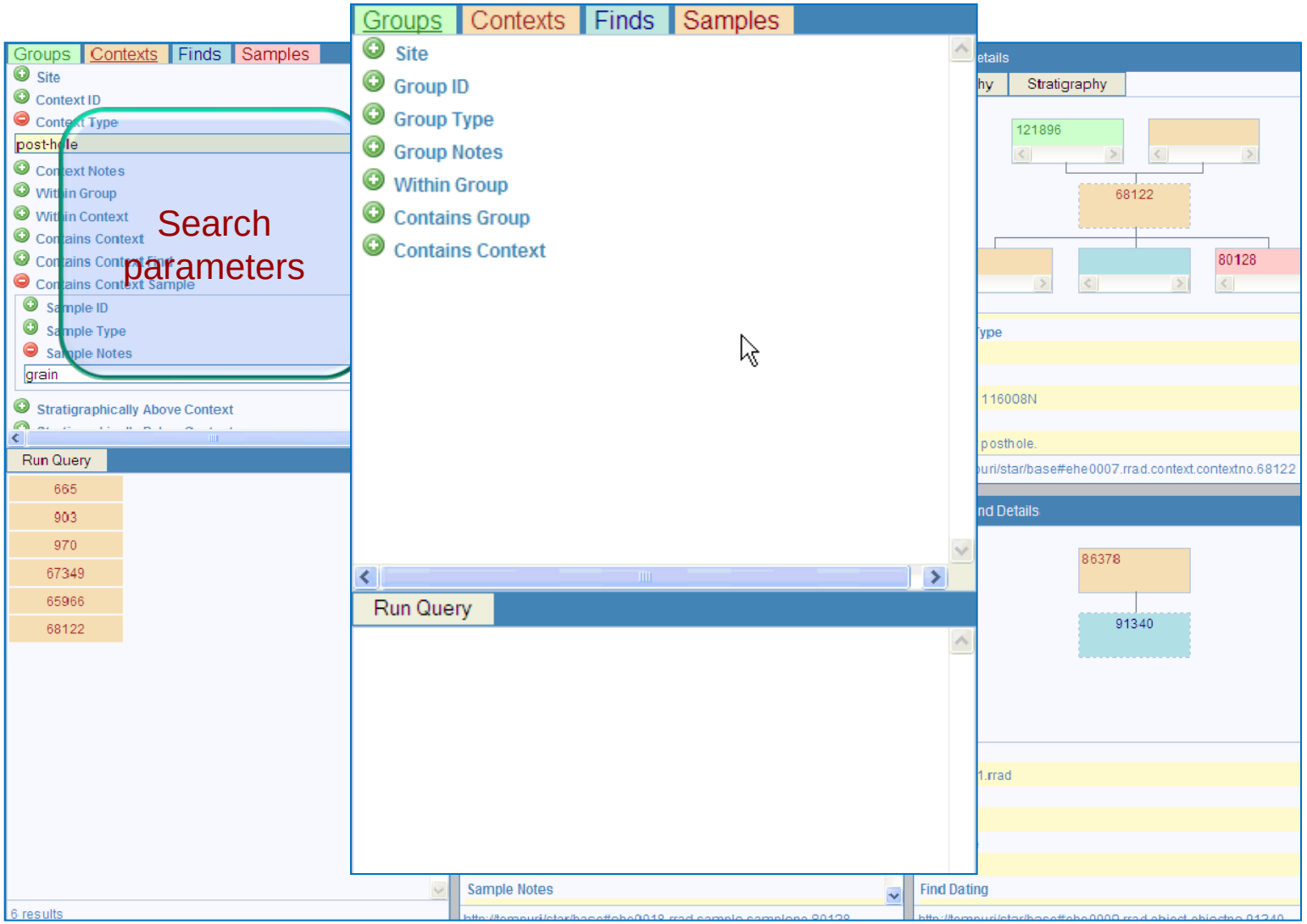


- Incorporated SKOS based thesaurus query expansion in search
- Colour coding of results by source dataset
- Browse results and drill down
- Open links to external data if available

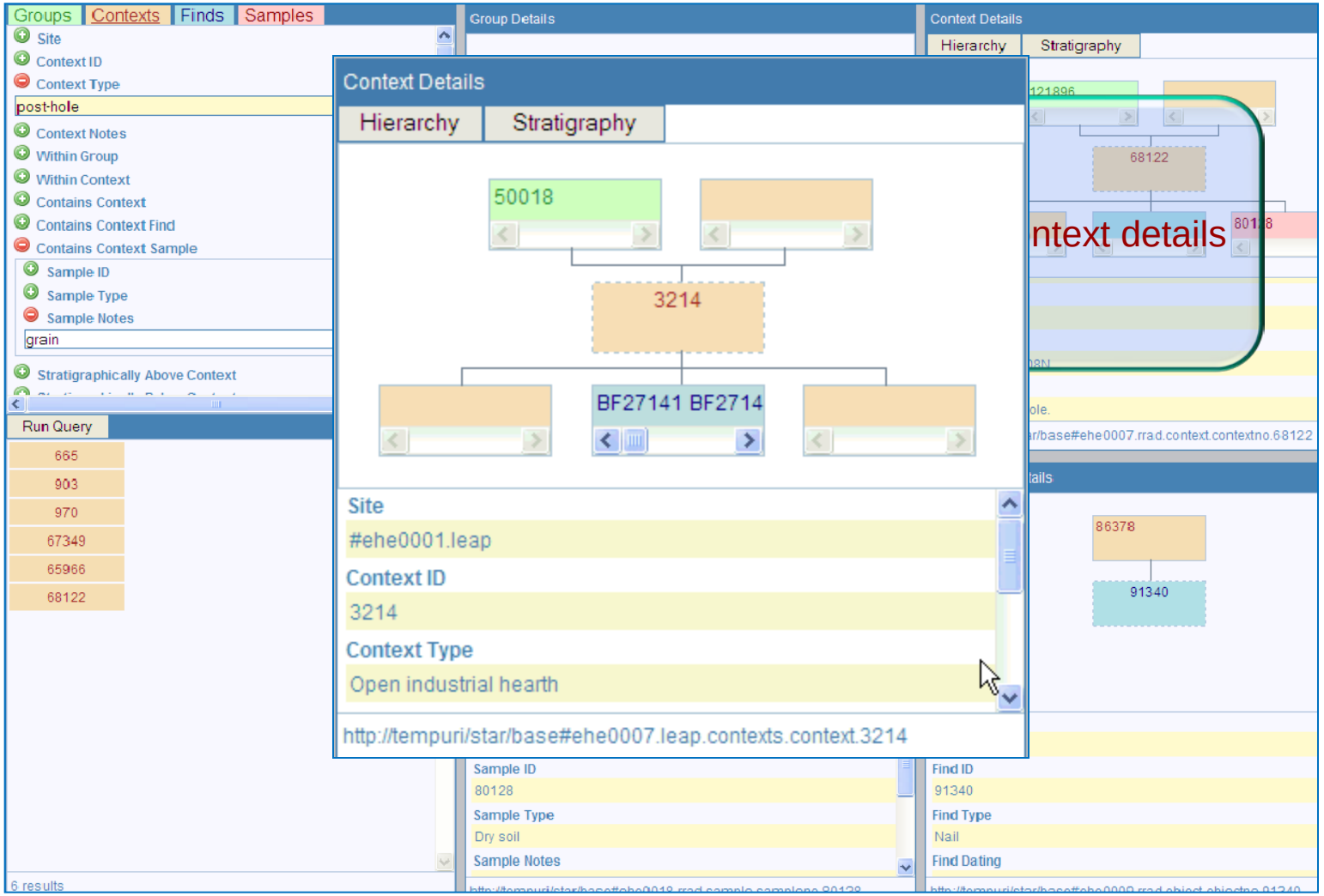
STAR web browser based search interface



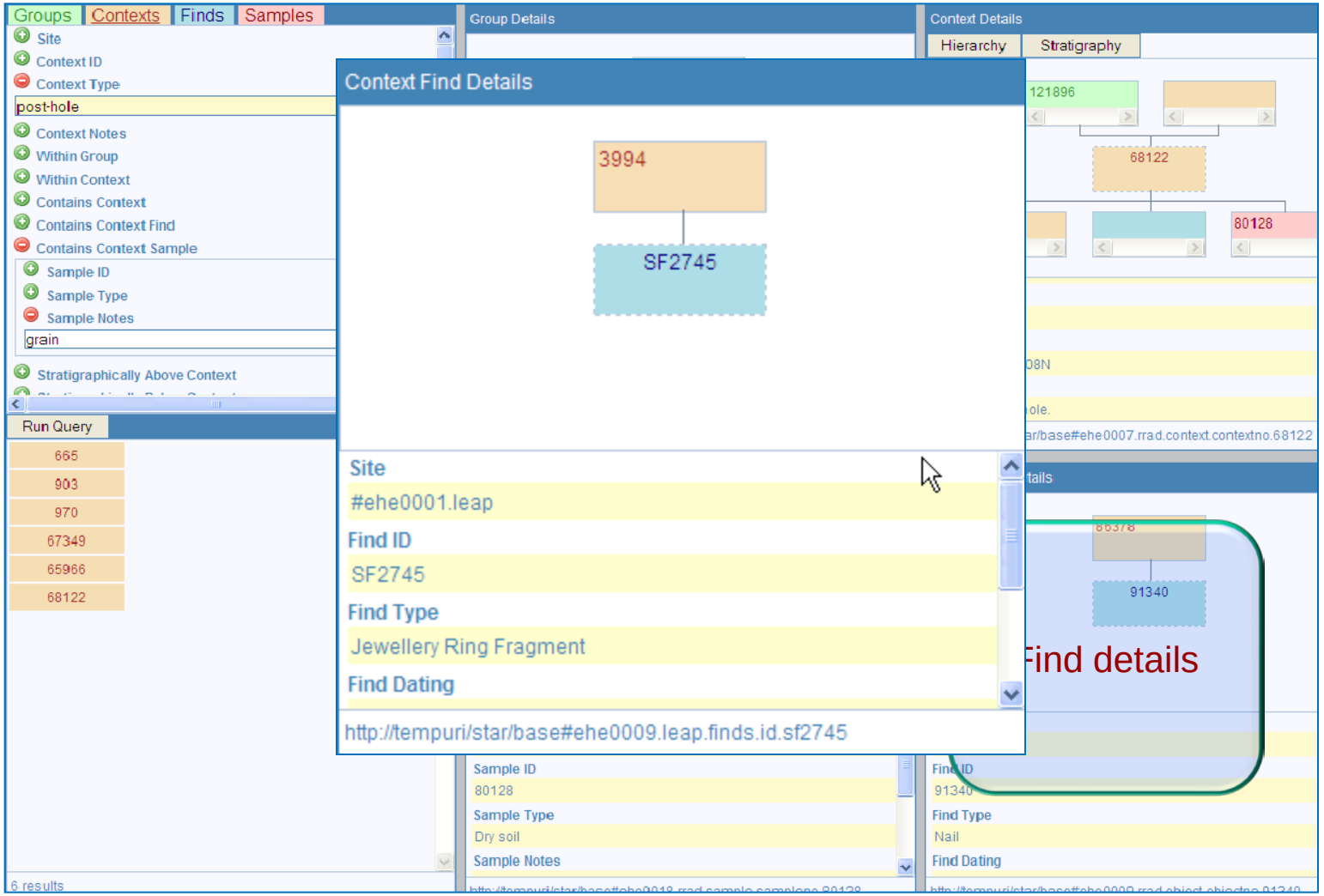
Initial search



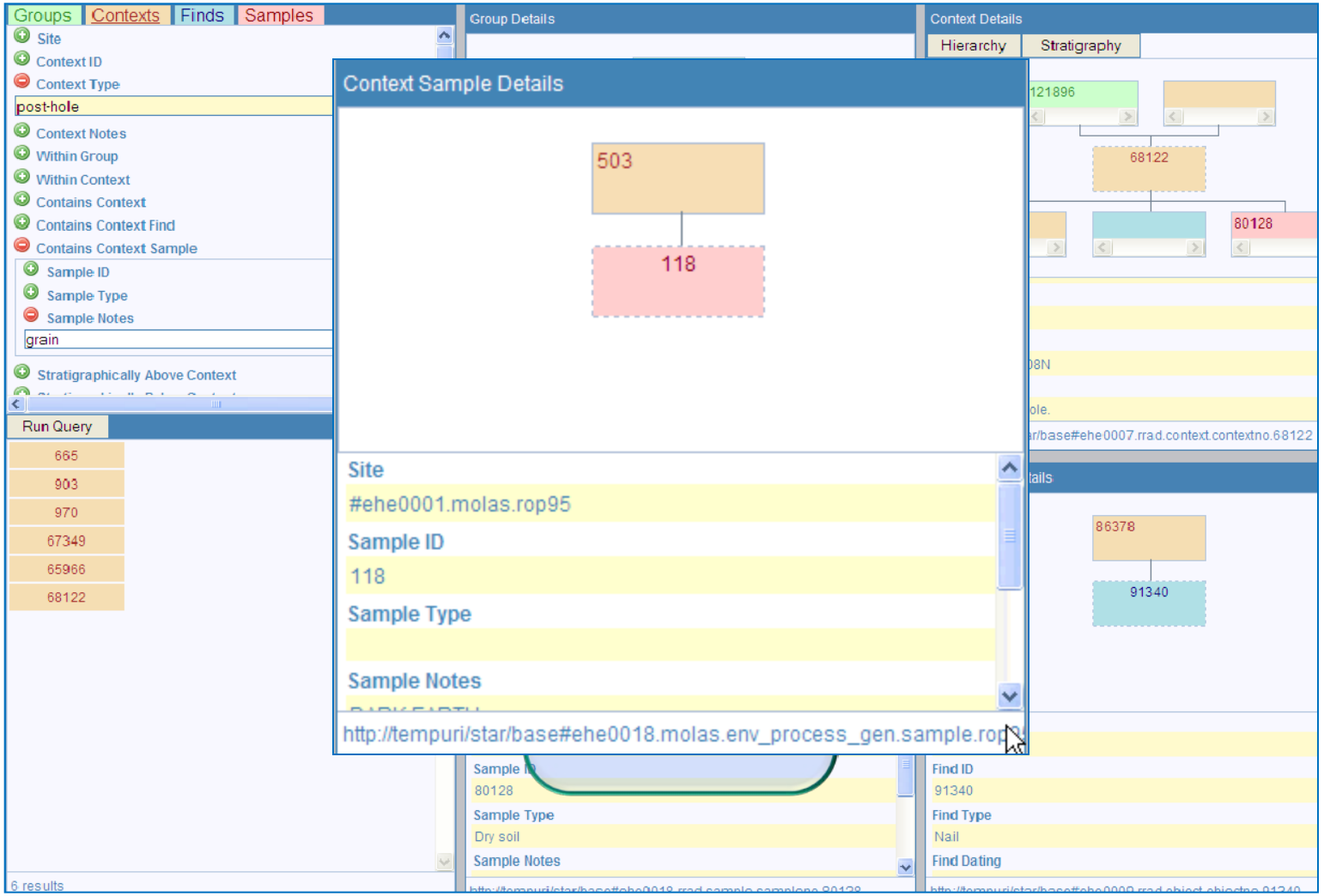
Context details



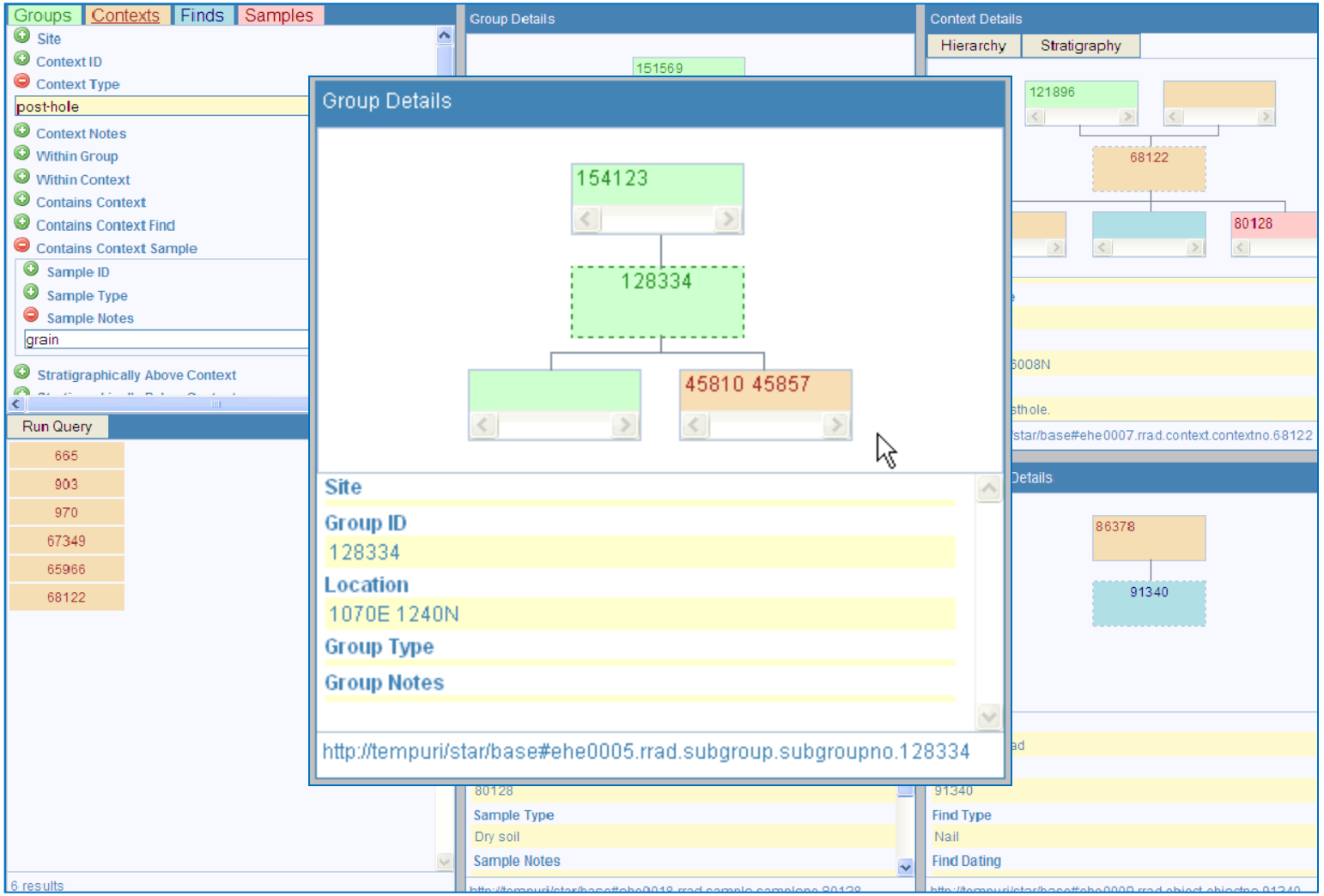
Context find details



Context sample details



Group details



Grey Literature Information Extraction (Andreas Vlachidis)

An archaeological evaluation was carried out by ECC FAU on behalf of Essex Police on the site of a proposed new police station at Smiths Farm, on the southeastern outskirts of Great Dunmow, Essex. The site was formerly rough pasture. The Chelmsford Road, which is thought to be the line of a Roman road, runs immediately to the east of the site. Five 30m x 2m trenches were excavated within the footprint of the proposed building and the area of associated carpark. Only one archaeological feature was revealed, a ditch containing prehistoric pottery dating to the Late Bronze Age or Early Iron Age along with burnt flints and flint flakes. No other archaeological features were identified, although a number of prehistoric pottery sherds and flint flakes were discovered on the surface of the natural geology. Although the results of the evaluation do not suggest intensive landscape use during the Late Bronze/ Early Iron Ages it is clear from this and other nearby investigations that a focus for the low level activity seen may well lie in the general vicinity. The absence of Roman or medieval remains indicates that this site was well outside the settlements of these periods. The low quantity and quality of the remains encountered on the site suggests that there is only a minor archaeological implication for the location of the proposed police

LATE BRONZE AGE OR EARLY IRON AGE	Term: LATE BRONZE AGE, skos: 134734 EARLY IRON AGE: 134735	E49_Time_Appellation #text	5
ROMAN OR MEDIEVAL	Term: ROMAN, skos: 134738 MEDIEVAL: 134745		
PREHISTORIC PERIOD	Term: PREHISTORIC, skos: 134718	#text	2

EARLY IRON AGE

Broad Term:
IRON AGE
Top Term:
CULTURAL PERIOD

- Looking to extract CRM-EH period, context, find, sample entities
- Aim to cross search within data

Example STAR use of URIs (NLP)

```
<crmeh:EHE0007.Context rdf:about="http://tempuri/star/base#suffolkc1-3166.22923">
<dc:source rdf:resource="http://tempuri/star/base#suffolkc1-3166" />
<dc:source rdf:resource="http://tempuri/star/base#ehe0001.oasis" />
    <crm:P2F.has_type>
        <crm:E55.Type>
            <rdf:value>walls</rdf:value>
            <crmeh:EXP10F.is_represented_by
              rdf:resource="http://tempuri/star/concept#ehg003.93"/>
            <crmeh:EXP10F.is_represented_by
              rdf:resource="http://tempuri/star/concept#70426"/>
        </crm:E55.Type>
    </crm:P2F.has_type>
    <crm:P3F.has_note>
        <crm:E62.String>
            <rdf:value>...structure with a ...</rdf:value>
        </crm:E62.String>
    </crm:P3F.has_note>
</crmeh:EHE0007.Context>
```

STELLAR

- STELLAR aims to generalise and extend the data extraction tools produced by STAR and enable third party data providers to use them."

- The extracted data will be represented in standard formats that allow the datasets to be cross searched and linked by a variety of Semantic Web tools, following a linked data approach.

Objectives

- Develop best practice guidelines for mapping and extraction of archaeological datasets into RDF/XML representation conforming to the CIDOC CRM-EH standard ontology
- Develop an enhanced tool for non-specialist users to map and extract archaeological datasets into RDF/XML representation conforming to CIDOC CRM-EH
- Develop best practice guidelines and tools for generating corresponding Linked Data
- Evaluate the mapping tool and the Linked Data provision
- Engage with the archaeological community to inform research and disseminate outcomes

STELLAR - Data Processing Stages

- Parsing source data
  - Excel spreadsheets
  - Delimited data files (CSV)
- Cleansing / Manipulation
  - Trim space
  - Force uppercase / lowercase
  - replace text
  - add prefix / suffix
  - url encoding
  - Semantic enrichment
- Mapping
  - Columns to template placeholders
- Transformation
  - Apply templates to tabular data
- Validation
  - Validate output for coherence

STELLAR – templates and placeholders

SITECODE	CONTEXT	ACC_NO	MATE	E_DATE	MAX_DIAM	COMMENTS	OBJECTID
BA84	597	2657	COPP	1302	20	SF NO 258; UNOFFICIAL STERLING	1649
BA84	569	2652	COPP	1285	23	SF NO 427; FRENCH	1650
BA84	1108	2656	COPP	1280	19	SF NO 418; BARBAROUS PRIVATE ISSUE	1651
BA84	2406	2663	COPP	1415	27	SF NO 884; TOURNAI STOCK JETTON	1652

$OBJECT_URI$ = "#e19." + ACC_NO
$OBJECT_IDENTIFIER$ = ACC_NO
$OBJECT_NOTE$ = COMMENTS
$MEASUREMENT_VALUE$ = MAX_DIAM
$PLACE_URI$ = "#e53." + SITECODE + "-" + CONTEXT

Tabular data columns mapped to template placeholders

```
<?xml version="1.0"?>
<rdf:RDF
   xmlns:crm="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

<crm:E19.PhysicalObject rdf:about="$OBJECT_URI$">
   <crm:P1F.is_identified_by>
      <crm:E41.Appellation>$OBJECT_IDENTIFIER$</crm:E41.Appellation>
   </crm:P1F.is_identified_by>

   <crm:P3F.has_note>
      <crm:E62.String>$OBJECT_NOTE$</crm:E62.String>
   </crm:P3F.has_note>

   <crm:P43.has_dimension>
      <crm:E54.Dimension>
         <crm:P2F.has_type rdf:resource="http://tempuri/diameter" />
         <crm:P91F.has_unit rdf:resource="http://tempuri/mm" />
         <crm:P90F.has_value>
            <crm:E60.Number>$MEASUREMENT_VALUE$</crm:E60.Number>
         </crm:P90F.has_value>
      </crm:E54.Dimension>
   </crm:P43.has_dimension>
</crm:E19.PhysicalObject>

<crm:E9.Move>
   <crm:P25F.moved rdf:resource="$OBJECT_URI$" />
   <crm:P26F.moved_to rdf:resource="$PLACE_URI$" />
</crm:E9.Move>

</rdf:RDF>
```

Data template with defined placeholders

```
<?xml version="1.0"?>
<rdf:RDF
   xmlns:crm="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

<crm:E19.PhysicalObject rdf:about="#e19.2652">
   <crm:P1F.is_identified_by>
      <crm:E41.Appellation>2652</crm:E41.Appellation>
   </crm:P1F.is_identified_by>

   <crm:P3F.has_note>
      <crm:E62.String>SF NO 427; FRENCH</crm:E62.String>
   </crm:P3F.has_note>

   <crm:P43.has_dimension>
      <crm:E54.Dimension>
         <crm:P2F.has_type rdf:resource="http://tempuri/diameter" />
         <crm:P91F.has_unit rdf:resource="http://tempuri/mm" />
         <crm:P90F.has_value>
            <crm:E60.Number>23</crm:E60.Number>
         </crm:P90F.has_value>
      </crm:E54.Dimension>
   </crm:P43.has_dimension>
</crm:E19.PhysicalObject>

<crm:E9.Move>
   <crm:P25F.moved rdf:resource="#e19.1650" />
   <crm:P26F.moved_to rdf:resource= "#e53.BA84-569" />
</crm:E9.Move>

</rdf:RDF>
```

Resultant output (per row)

Linked data issues in STELLAR

- What parts of dataset useful to map/extract for linked data purposes? What best left as native dataset? Cost/benefit ?

- Lack of current domain name for some organisations within project premature/temporary URI definition?

- Lack of resource commitment for some organisations within project Where data reside? Demands on server?

- Relationship between local dataset glossaries and ‘centre’
  - map to unified central super glossary?
  - maintain local glossaries with (SKOS) mappings to centre?

- Resultant application?
  Linked data in itself not offer search capability

Mapping issues

- Potentially multiple mappings can exist to a broad conceptual framework (BRICKS experience)
  - depends on purpose for mapping and focus of concern
  - STELLAR design mapping patterns (ontology → datasets)

- Datasets encountered not have a schema, not necessarily well structured and probably need semantic cleaning

  ===> infer schema for tool purposes from dataset

Mapping issues ctd.

Considering various cases of mapping RDB to CRM

- DB column to CRM entity
- DB column to multiple CRM entities
- DB column to CRM entity with intermedite (event) nodes
- DB columns (in same table) concatenated to single CRM entity
- DB complete Table maps to a single CRM entity
- DB entities map to CRM entity plus set of properties

CRM is complex => need for simpler views

- CRM useful for integration and semantic interoperabiliy
  Not mirror RDF structure in RDF

- It is not necessary to present the user with full CRM or CRM-EH views
  for mapping work and search (browsing) interfaces

- Search / browsing / mapping interfaces can present a simplified CRM view to user, while retaining the benefits of CRM standard for inter-operability

References

1. STAR Project. http://hypermedia.research.glam.ac.uk/kos/star/
2. STAR Project Publications. http://hypermedia.research.glam.ac.uk/publications/#kos
3. STELLAR Project . http://hypermedia.research.glam.ac.uk/kos/stellar