Domain Vocabularies for Semantic Interoperability of Museum Collections

Dr Andreas Vlachidis and Dr Daniele Metilli

UCL Department of Information Studies

a.vlachidis@ucl.ac.uk
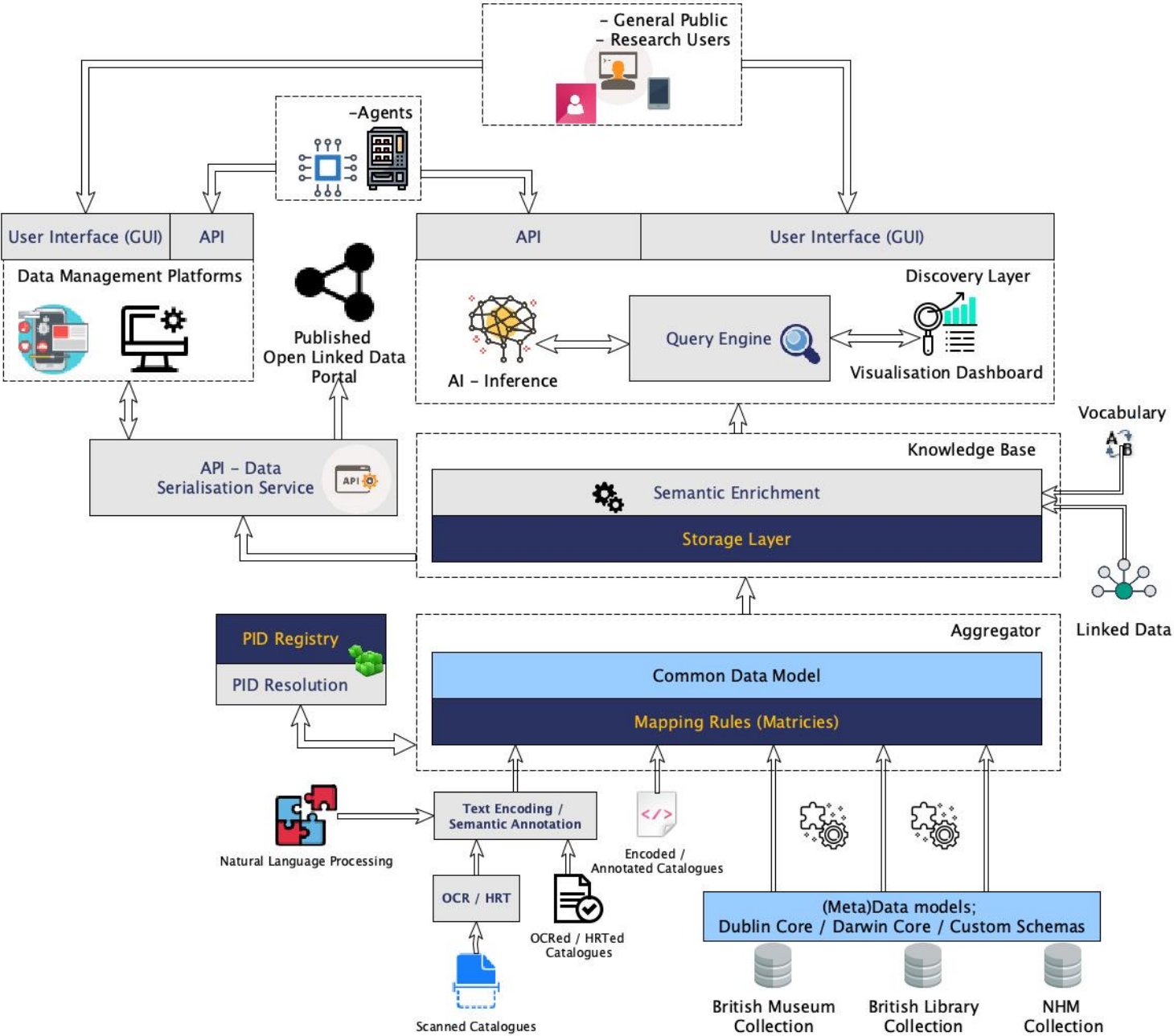
d.metilli@ucl.ac.uk

The Sloane Lab

Sloane an **exemplary case study** to unite surviving objects (current digital collections) and records across the NHM, BM and BL from the original catalogues

Reunify and extend digital collection
Searchable and amenable to further analysis and interpretation

TOWARDS A NATIONAL COLLECTION | UKRI Arts and Humanities Research Council
DISCOVERY PROJECTS

The British Museum   NATURAL HISTORY MUSEUM   TECHNISCHE UNIVERSITÄT DARMSTADT

Infrastructure

Data Model

- The **Sloane Lab data model** allows us to represent the **knowledge** collected in the project

- Based on **Semantic Web** standards and built on top of the **CIDOC CRM** reference model

- Centred on the **record** instead of the **object**, to support multiple perspectives over the same object

Fishes Entry 172 — "The large tooth of a shark"
object with a Natural History Museum record
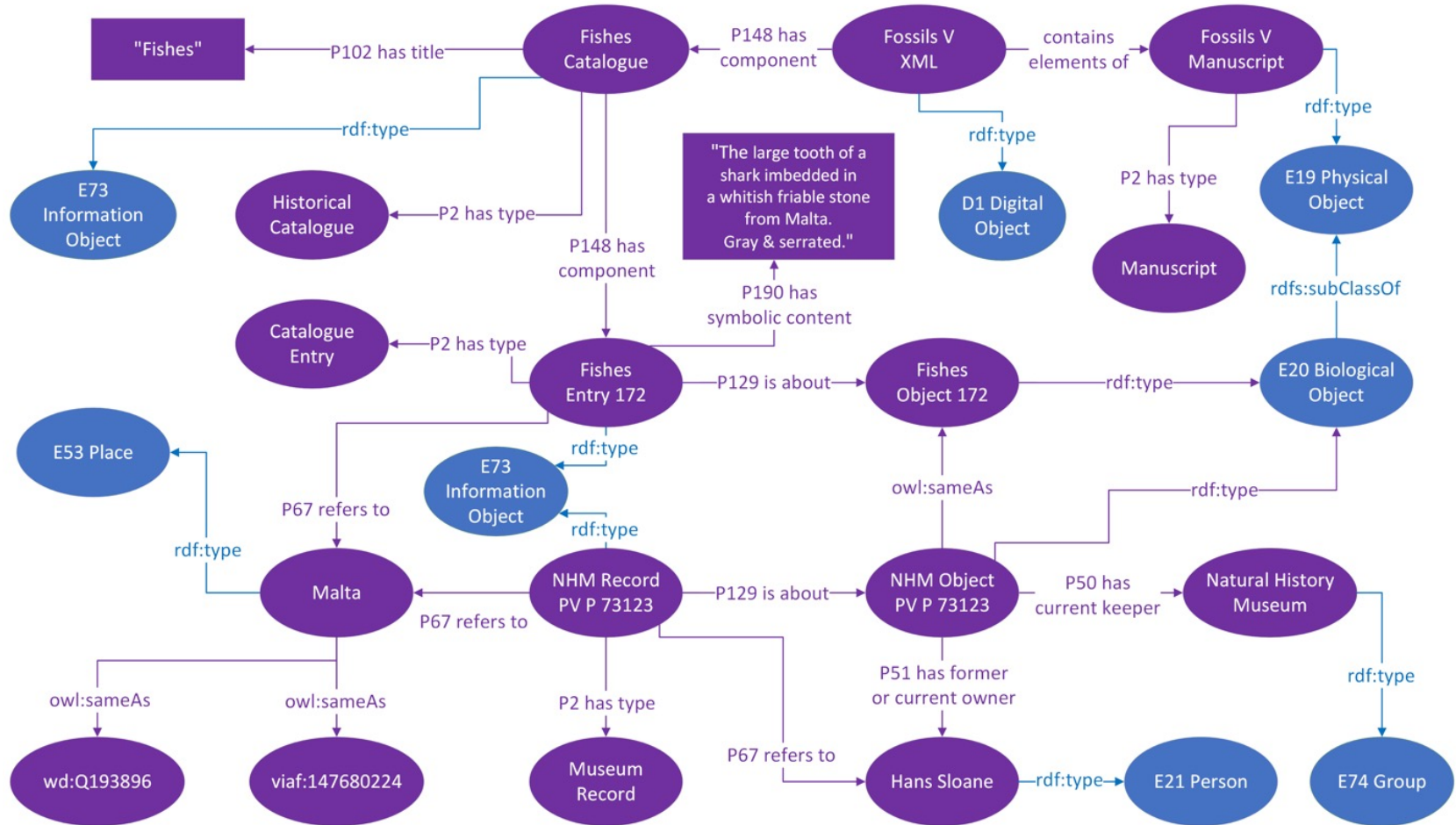
Vocabularies for Semantic Interoperability

- Authority Lists for NLP on historical catalogues

- Aim: Automatically recognise Place and Person entities in Sloane catalogues, using rule-based Name Entity Recognition techniques, supported by input from British Museum Named Authority files

- Integrate the output into the provided TEI scheme

The Framework

The NER Pipeline

Catalogue (Input)

TEI (output)
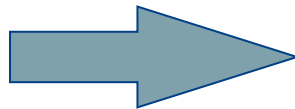
Selected Processing resources

!	Name	Type
	Document Reset PR	Document Reset PR
	ANNIE English Tokeniser	ANNIE English Tokeniser
	Authority List	ANNIE Gazetteer
	ANNIE Sentence Splitter	ANNIE Sentence Splitter
	ANNIE POS Tagger	ANNIE POS Tagger
	Place Match	JAPE Transducer
	Place Match Features	JAPE Transducer
	tempPlace2Place	JAPE Transducer
	Person Match	JAPE Transducer
	Person Match Features	JAPE Transducer
	tempPerson2Person	JAPE Transducer
	Validate	JAPE Transducer

`<?xml?>`

JAPE
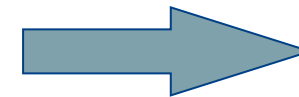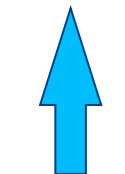
Named Authority Files

- British Museum Place names (49,382 entries)
  - 55 Attributes

- British Museum Person names (initially 215,186 entries)
  - Refined to 126,018 (cut off date 1799)
  - 80 Attributes

System ID	Place name	Place type	Authority Type	Broader Terms	Terminology	Scope Note	Source	Amendment history - user	Amendment history - date
x110470	'Ainahau	Q	d s v	Honolulu	PLACE	Royal estate of Princess Victoria Ka'iulani.	Merlin	ssl~ jstribblehill~ jstribblehill~ eherdman	21 Jul 2017~21 Apr 2017~18 Feb 2008~29 Oct 2007
x26828	Antwerp	K	d s v	Antwerp	PLACE	Antwerpen~ Anvers~ Antverpia		ssl~ jstribblehill~ tszrajber~ swood~ tszrajber~ pmain~ jstirling~ rsparkinson~ ssl~ pmain	21 Jul 2017~ 10 Feb 2017~ 03 Apr 2007~ 06 Dec 2006~18 Mar
x40131	Antwerp	E	d s v	Flemish Region	PLACE	Antwerpen~Anvers		ssl~ jstribblehill~ pmain~ pmain~ pmain~ pmain	21 Jul 2017~27 Jan 2017~12 Apr 2002~11

Challenges
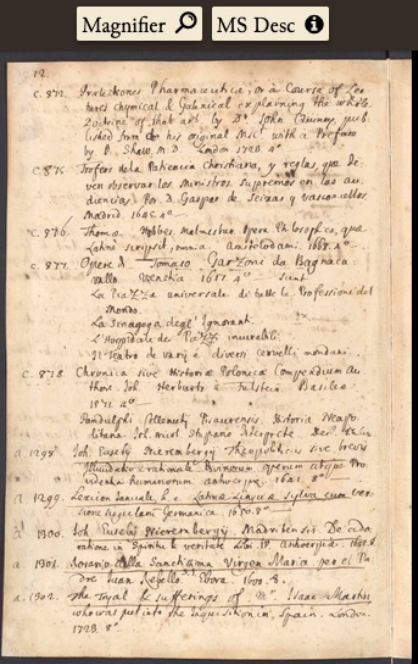
- Domain Independent
  - Know limitations of NER (maximum accuracy)
  - Toponymical Polysemy
    - When a place name is a surname!
    - What about "Duke of Edinburgh"

- Domain Dependent
  - Particularly relating to sentence structure
  - Hyphenation and the whole world matching apparatus

- Vocabulary Related
  - Authority Files are not built for NLP
    - coverage and suitability
  - Person names and Organisation and Unknown in one List
  - Multiple Entries (Madrid as a City and as a Province)
  - Person name Synonymy

Enlightenment Architectures TEI Output

British Museum - Thesaurus

- BM Thesaurus linked to CMS
  - Object Types
  - Places (Find Spot)
  - Materials
  - People

The British Museum

abacus

Object Type

abacus

Museum number

As,SLMisc.594

Description

Abacus (with 17 rows of beads, divided by cross-bar carrying Chinese characters) made of wood, metal.

Findspot

Found/Acquired: China
Asia: China

Materials

wood

metal

wood

Information	Related objects

Also known as

wood

branch

conifer

deal

hardwood

twig

Related terms

bamboo

cane

lignite

British Museum – Controlled Vocabulary

- Event
- Associate
  - Name
  - Place
  - Authority
  - Title
- Ethnic Name
- Field Collection
- Findspot
- Producer
- Production Place

	A	B	C	D
1	PO	Associated with		
2	PP	Authorised or patronised by		
3	EE	Emblem of		
4	AB	Illustration to		
5	F	Made for		
6	PI	Named in inscription		
7	II	Named in inscription & portrayed		
8	OP	Original produced by		
9	IP	Portrait of		
10	RP	Repaired by		
11	IR	Representation of		
12				

	A	B	C
1	AW	Associated with	
2	M	Made by	
3	OP	Original produced by	
4	IR	Representation of	
5			

	A	B
1	E	Eponym
2	G	Governor
3	I	Issuer
4	Y	Magistrate
5	M	Moneyer
6	K	Ruler
7		

Natural History Museum - Vocabularies

- Biological Collection Access Service (BioCASe)
  - for describing collections
  - https://www.biocase.org

- International Plant Names Index
  - Nomenclatural data for the scientific names of vascular plants from family to infraspecific ranks
  - https://www.ipni.org/

- In-house Glossaries
  - Record Type
    - specimen, artefact, tissue prep, bound volume etc.
  - Kind
    - Sheet, photograph, microscope slide, box

Challenges

- Different institutions adopt different vocabularies

- Different levels of openness and adoption of standards, different attitudes toward making vocabularies public

- External resources (e.g., AAT) are readily available but not controlled by partner institutions

- Reconciliation of different authority files and vocabularies is complex

Conclusions

- Vocabularies provide a common conceptual layer of high-level semantics for the purposes of unification, alignment and harmonisation

- Important role of museum-based vocabulary resources towards the realisation of unified collections

- Big opportunities for semantic enrichment, linking and interoperability, but also many challenges

Domain Vocabularies for Semantic Interoperability of Museum Collections

Dr Andreas Vlachidis and Dr Daniele Metilli

UCL Department of Information Studies

a.vlachidis@ucl.ac.uk

d.metilli@ucl.ac.uk