



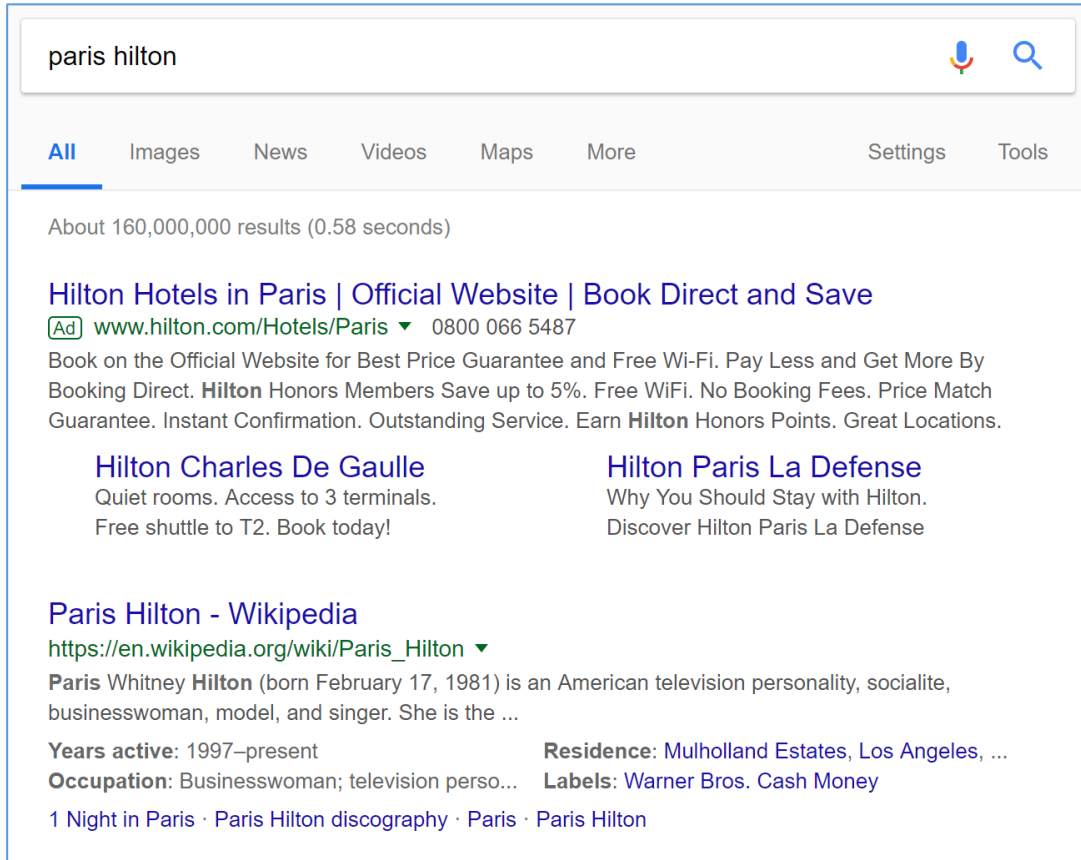
ARIADNEplus Vocabulary Mapping Strategy

Ceri Binding | **University of South Wales** | Monday 27th March 2023 | Centre Marc Bloch, Berlin

Discover
Connect
Collaborate



Why use controlled vocabularies?



A screenshot of a Google search for "paris hilton". The search bar shows "paris hilton" with a microphone and search icon. Below the search bar are tabs for "All", "Images", "News", "Videos", "Maps", and "More", along with "Settings" and "Tools". The results show "About 160,000,000 results (0.58 seconds)". The first result is an advertisement for "Hilton Hotels in Paris | Official Website | Book Direct and Save" with a link to "www.hilton.com/Hotels/Paris". Below the ad are two hotel listings: "Hilton Charles De Gaulle" and "Hilton Paris La Defense". The second result is "Paris Hilton - Wikipedia" with a link to "https://en.wikipedia.org/wiki/Paris_Hilton". The Wikipedia snippet describes Paris Whitney Hilton as an American television personality, socialite, businesswoman, model, and singer. It also lists her years active (1997–present), residence (Mulholland Estates, Los Angeles), occupation (Businesswoman; television perso...), and labels (Warner Bros. Cash Money).

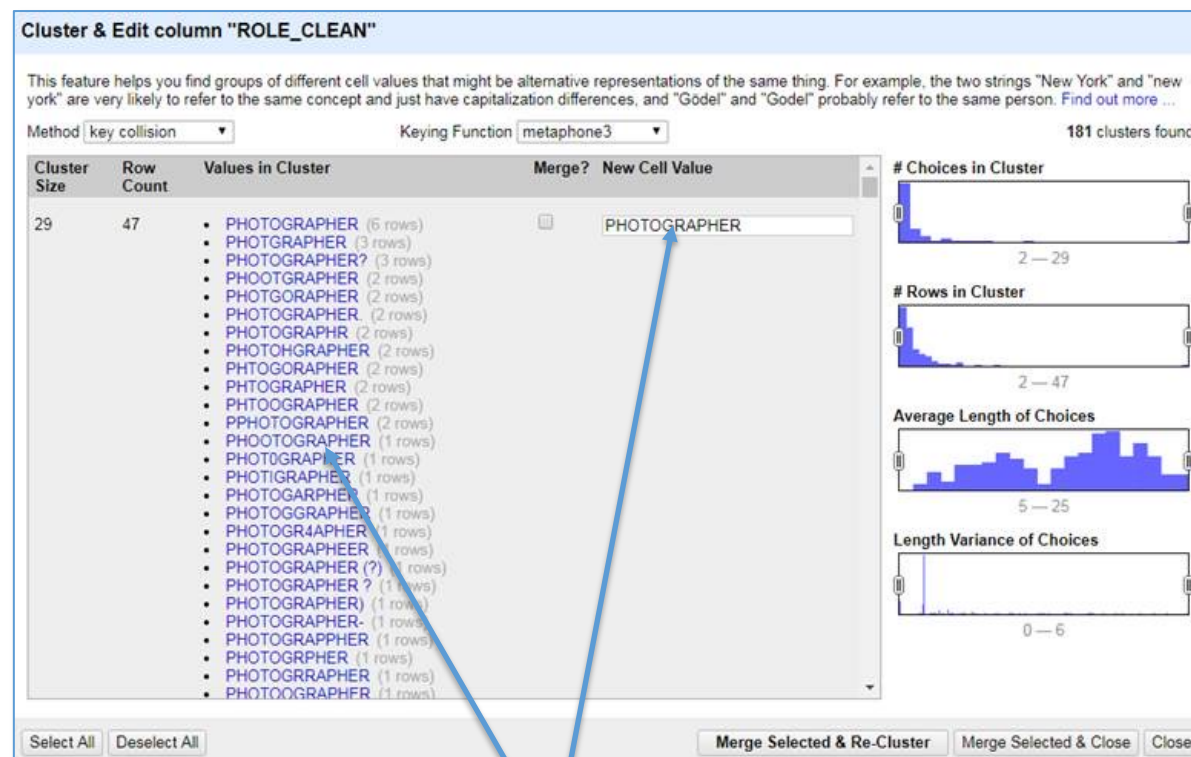
- Words are ambiguous
 - If we use words to index, the indexing is ambiguous
 - If we use words to search, the search criteria are ambiguous
 - *"Paris Hilton"* - person or hotel?
 - The meaning depends on additional context
- Words are language-specific
 - Paris (English)
 - ≠ Parijs (Dutch)
 - ≠ Parigi (Italian)
 - ≠ باريس (Arabic)
 - For typical web search each returns *different* results, though they all *mean* the same thing

Issues preventing effective subject integration

- Spelling errors affect recall
 - *"posthole" != "posthlole", "cess pit" != "cess pitt"*
- Alternate word forms or punctuation affect recall
 - *"posthole" != "post-hole" != "post hole" != "post holes"*
 - *"gulley" != "gullies", "boundary" != "boundaries"*
- Synonyms affect recall
 - *"fresco" != "mural", "cask" != "barrel", "brimstone" != "sulphur" (or "sulfur")*
- Differing levels of specificity / granularity affect recall
 - *"weapon" != "sword" != "rapier" != "Pappenheimer"*
- Homographs affect precision
 - *"compound" (enclosure) != "compound" (material)*
 - *"lead" (object) != "lead" (material)*
 - *"pitch" (English) has over 20 different meanings*
 - Regional: *"tenement" (Scotland) != "tenement" (England)*
 - Multilingual: *"coin" (fr) != "coin" (en), "monster" (nl) != "monster" (en)*

Data cleaning: OpenRefine

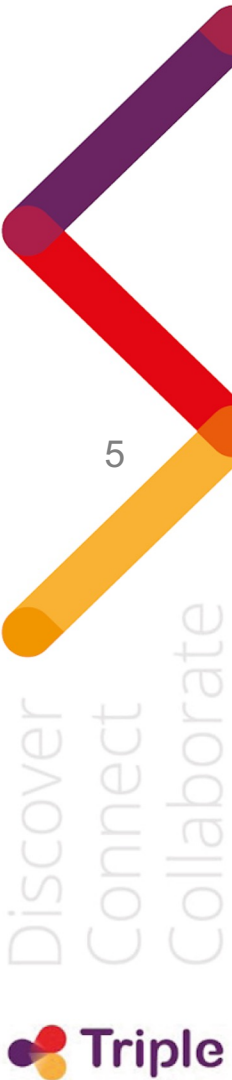
- <http://openrefine.org/>
- Flexible open source data cleaning tool
- Import, filter, transform, align and export data
- Scripted operations for repeatable bulk processing



Clustering variations by similarity,
Merging all to a single (correct!) value

Controlled vocabulary relationships

- Equivalence (ALT)
 - *post hole* ALT [*post-hole, posthole, post holes*]
 - *sulphur* ALT [*sulfur, brimstone*]
- Hierarchical (BT/NT)
 - *weapon*
 - *sword*
 - *rapier*
 - *Pappenheimer*
- Associative (RT)
 - *cup* RT *saucer*
 - *thermometer* RT *temperature*



Background: ARIADNEplus project

- Cloud based infrastructure and services
- Archaeological metadata - aggregation & integration
 - ≈ 3.5 million records representing 50+ countries
- Cross searching data by Place / Time / Subject
 - Place: WGS84 coordinates, place names
 - Time: BCE/CE years, named periods (e.g. *Iron Age*)
 - **17 named period vocabularies (Perio.do)**
 - **21 languages**
 - **1,880 named periods**
 - Subject:
 - **59 local subject vocabularies**
 - **16 languages**
 - **19,000+ local subject terms**

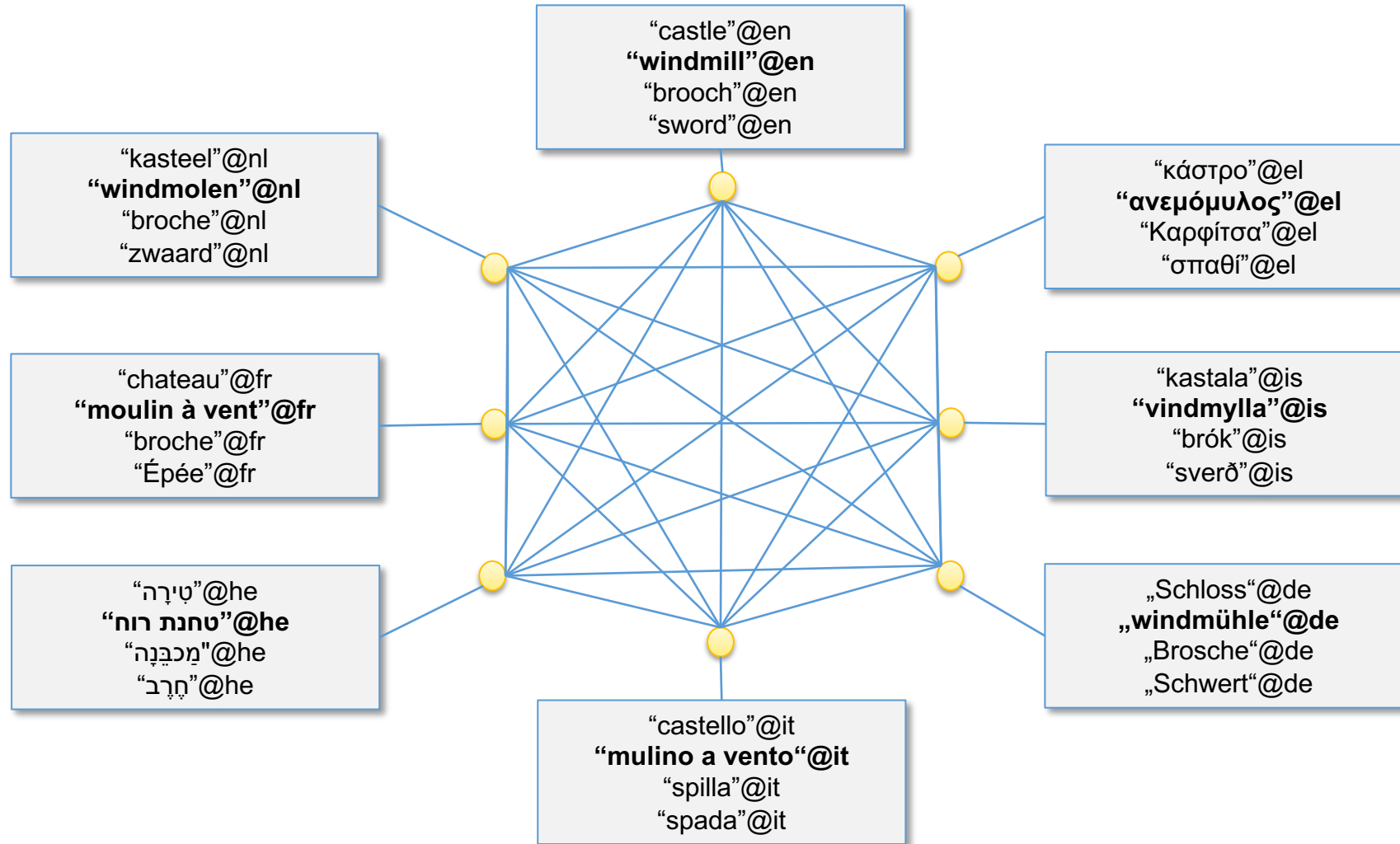


Integration - named periods

- e.g. *Bronze Age, Iron Age, Roman* – a special case of subject indexing
- *Bronze Age* refers to different dates in different locations
- [Perio.do](http://perio.do) – public domain multilingual gazetteer of named historical periods
- ARIADNEplus records aligned with specific Perio.do authority resources; enriched with start/end years

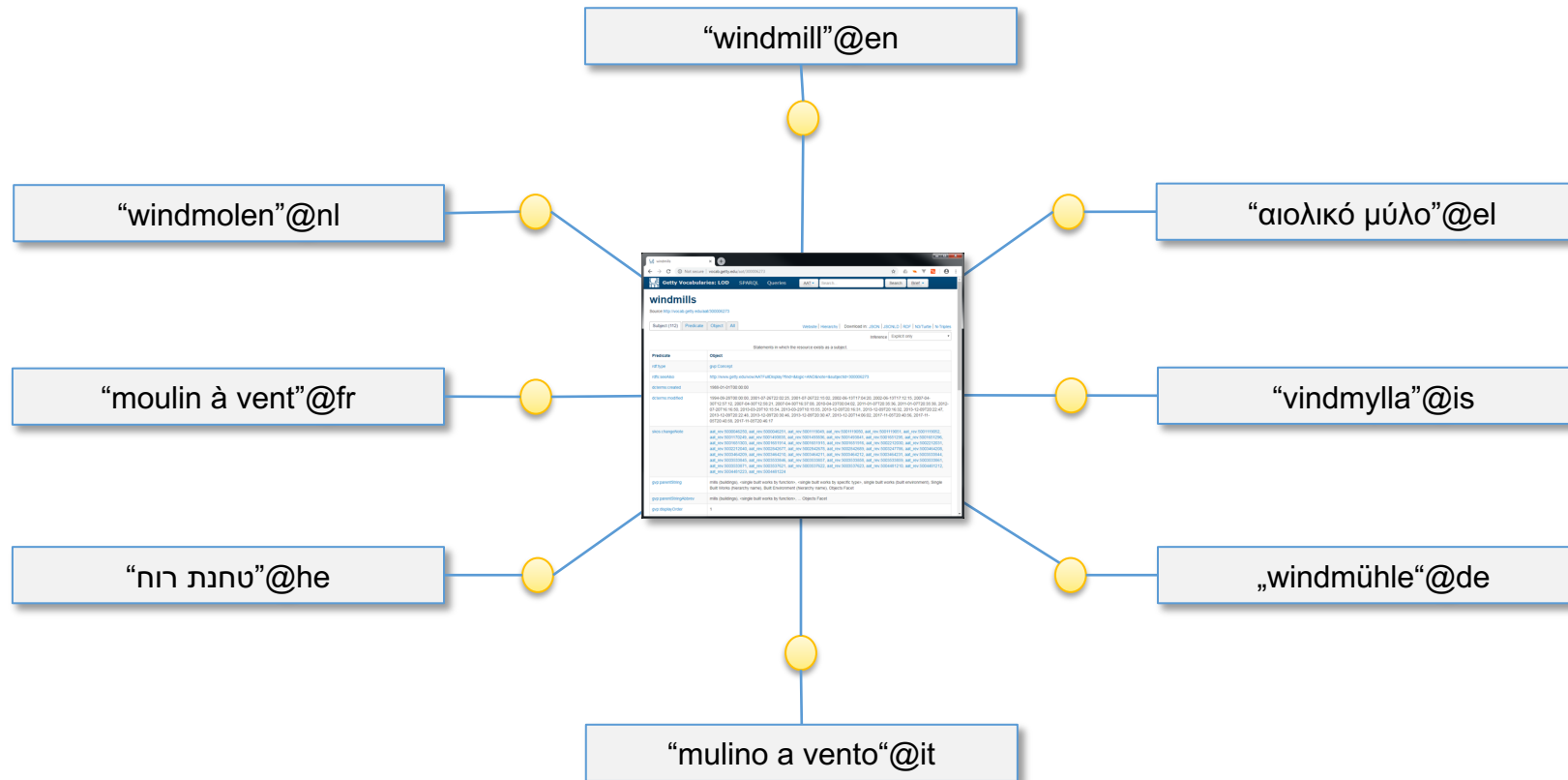


Integration – Mapping subject vocabularies



Equivalent subject terms originating from multiple vocabularies - map everything to everything?

Strategy - map local subjects to spine vocabulary



Hub architecture – more manageable. Using mappings, a search on one term can incorporate all others

Vocabulary Matching Tool

- For matching local subject terms to Getty AAT concepts
- Search & browse Getty AAT structure and create mappings
- Not automatic matching – can examine scope and context of concepts
- 6,400+ existing subject mappings from original ARIADNE project reused, revised & extended
- New subject mappings created for new partners = 19,220 total mappings, local subject → AAT

<https://vmt.ariadne.d4science.org/vmt/>

Vocabulary Matching Tool

English

Source Concept	Match Type	Target Concept	Suggest	Delete Row
Identifier		Filter column...		
Label				
Filter column...				
http://purl.org/heritagedata/schemes... Abbey Church	Close Match	abbey churches	Q	
http://purl.org/heritagedata/schemes... ABBEY	Exact Match	abbeys (monasteries)	Q	
http://purl.org/heritagedata/schemes... AGRICULTURAL BUILDING	Exact Match	agricultural buildings	Q	
http://purl.org/heritagedata/schemes... AGRICULTURAL DWELLING	Broad Match	agricultural buildings	Q	
http://purl.org/heritagedata/schemes... AGRICULTURAL HALL	Broad Match	agricultural buildings	Q	
http://purl.org/heritagedata/schemes... FARM BUILDING	Close Match	agricultural buildings	Q	
http://purl.org/heritagedata/schemes... FIELD SYSTEM	Broad Match	agricultural land	Q	
http://purl.org/heritagedata/schemes... FIELD SYSTEM	Broad Match	agricultural land	Q	
http://purl.org/heritagedata/schemes... LAND USE SITE	Broad Match	agricultural land	Q	
http://purl.org/heritagedata/schemes... LYNCHET	Broad Match	agricultural land	Q	
http://purl.org/heritagedata/schemes... CURVILINEAR ENCLOSURE	Broad Match	agricultural settlements	Q	
http://purl.org/heritagedata/schemes... DITCHED ENCLOSURE	Broad Match	agricultural settlements	Q	
http://purl.org/heritagedata/schemes... DOUBLE DITCHED ENCLOSURE	Broad Match	agricultural settlements	Q	
http://purl.org/heritagedata/schemes... ENCLOSED SETTLEMENT	Broad Match	agricultural settlements	Q	
http://purl.org/heritagedata/schemes... ENCLOSURE	Broad Match	agricultural settlements	Q	
http://purl.org/heritagedata/schemes... AGRICULTURE AND SUBSISTENCE	Broad Match	agriculture	Q	
http://purl.org/heritagedata/schemes... AIR RAID SHELTER	Exact Match	air raid shelters	Q	
http://purl.org/heritagedata/schemes... AIRCRAFT	Close Match	aircraft	Q	

390 rows

IMPORT JSON EXPORT JSON EXPORT CSV + ADD NEW ROW CLEAR ROWS SHOW HELP

ARIADNEplus
Created by University of South Wales

ARIADNEplus is a Horizon 2020 project funded by the European Commission (Grant Agreement No. 823914).
This application retrieves some information originating from Getty Art & Architecture Thesaurus (AAT)® which is made available under the ODC Attribution License. See <http://vocab.getty.edu/> for further details.

Type of match between concepts

SKOS mapping properties define the type of match between concepts
Don't rely on label matches; consider meaning and scope of concepts

skos:exactMatch



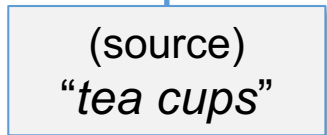
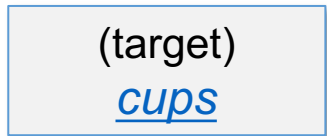
Where there is a high degree of confidence that the concepts may be used interchangeably

skos:closeMatch



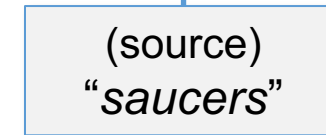
Where scope or hierarchy of concepts suggests slight conceptual differences

skos:broadMatch



Use “**some/all**” test for generic hierarchical relationships:
some *cups* are *tea cups*;
all *tea cups* are *cups*

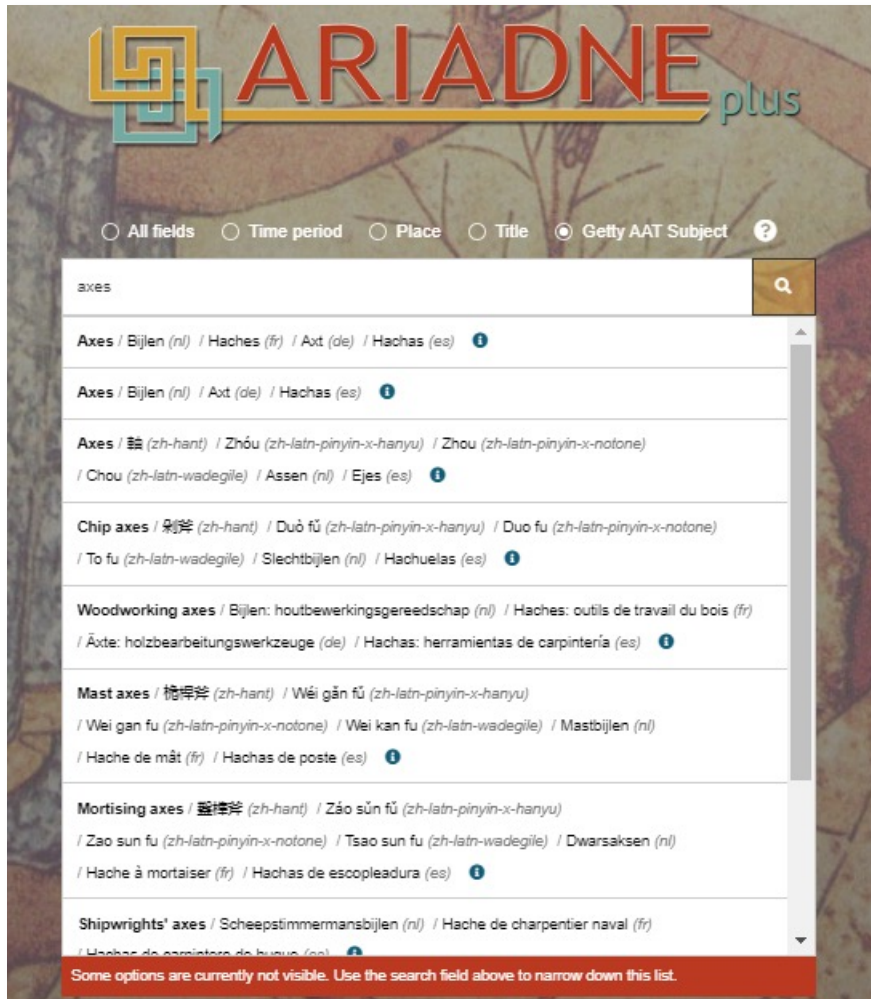
skos:relatedMatch



Where some other association exists between concepts

[skos:narrowMatch also exists, but was not used]

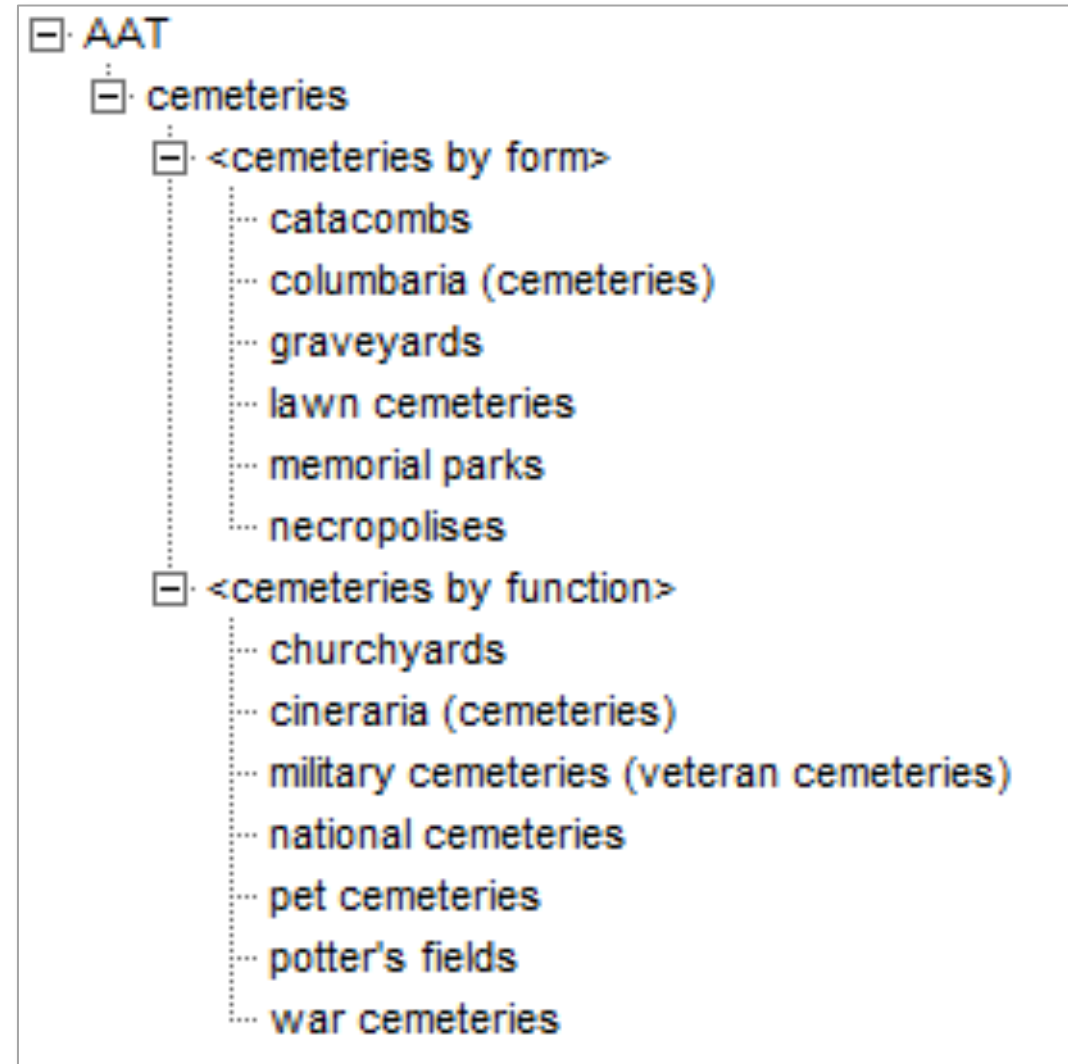
AAT built in to ARIADNEplus search



- (optionally) searching on AAT concepts
- Exploits multilingual entry vocabulary of AAT
- Suggests concept matches
- View context of suggested concepts – scope note, broader & narrower concepts

Use of vocabulary structure

- A search on AAT [cemeteries](#) can be expanded to include items indexed using any descendant concept ([catacombs](#), [graveyards](#) etc.)
- The search can also be expanded to include the multilingual entry vocabulary for the concept and any descendants ("cimetières"@fr, "begraafplaatsen"@nl, "Friedhof"@de etc.)



Using local subject → AAT mappings

ARIADNE PORTAL

Back to search results

Riom (Puy-de-Dôme), Parc Européen Economique - rue Ludwig von Beethoven, rapport de diagnostic

Description

L'opération se situe au sein d'un secteur archéologiquement sensible.

Les tranchées réalisées, ont permis de reconnaître des indices d'occupations humaines se rapportant à deux périodes chronologiques : le Néolithique et la période Médiévale à Moderne.

Les paléoenvironnements locaux ont été restitués grâce à l'observation des séquences sédimentaires (nature et géométrie des dépôts).

Pour le Néolithique, quatre fosses (F0 à F3) montrent une occupation du secteur au cours de la première moitié du 4e mill. av. J.-C. (Néolithique moyen 2) qui peut être corrélée à celle, plus dense, fouillée en 1992 légèrement plus au Sud-Ouest (site de PEER 2).

Pour la période Médiévale à Moderne, des creusements successifs (F1 à F5) du fossé méridional bordant l'axe de communication Riom/Varennes-sur-Morge/Thuret, matérialisé actuellement par la D 211, ont été observés dans le sondage S1. Il est vraisemblable que cet axe soit relativement ancien : le mobilier du fossé F5 tend à montrer que ce dernier est en place au moins depuis la période médiévale.

Metadata

Original ID: 29407

Landing page: <https://dola.inrap.fr/riom/ark:/04298/0129407>

Language: French

Resource type: [Inventaire](#)

Subject - AAT:

- 1 pits (earthworks) (en)
- 1 quartz (mineral) (en)
- 1 industry (object groupings) (en)
- 1 fauna (en)
- 1 teeth (animal components) (en)

Original:

- Céramique médiévale
- Céramique moderne
- Céramique néolithique
- Dent
- Faune
- Fossé
- Fosse
- Géomorphologie
- Industrie lithique
- Quartz
- Voie

Dating:

- 1 Néolithique: -6500 to -2201
- 1 Époque médiévale: 0500 to 1500
- 1 Néolithique moyen: -5300 to -4501

Resource links

View resource at provider

Resource is part of

Dola

Thematically similar

Thematically similar resources based on terms in common of:

Subject & Time period

Le Boulou (cat. El Voló) Rue Chapelle Saint Luc - El Cortal d'en Quiró. Projet de lotissement : Les jardins d'Auréli, Occitanie, Pyrénées-Orientales, rapport de diagnostic

Entre alluvions du Castelnou et ancienne terrasse de la Tet, quelques traces d'occupations humaines, Occitanie, Pyrénées-Orientales, Castelnou - Thuir, RD612 tranche 3, rapport de diagnostic

Extension septentrionale de l'occupation du Néolithique moyen, Provence-Alpes-Côte d'Azur, Bouches-du-Rhône, Trets, La Builière 2, rapport de fouilles

Saint-Paul (11), Le Gasquet-Parc photovoltaïque Le Gasquet, rapport de diagnostic

4 - Exploitation d'un territoire en bord de Seine: de l'enceinte monumentale du Néolithique moyen II à la ferme fossuée médiévale, une néropole exceptionnelle du Néolithique moyen II en "Bassée auboise", Pont-aux-Bois, Aube "Ferme de l'île" [Grand Est], volume 4: Etude funéraire, rapport de fouille

Berstet-Rumersheim (Bas-Rhin) rue des Acacias, rapport de diagnostic, un habitat du Néolithique récent et les traces d'une houlonnrière d'époque contemporaine

Saint-Jean-Le-Vieux (Ain) - Au Mollard - La Vigne Orset v., rapport de diagnostic, Tranche 2

Subject indexing derived from local subject → AAT mappings

- Enriched metadata records with derived AAT subjects
- Can exploit AAT entry vocabulary
- Facilitates multilingual cross-search e.g. search for *sword* returns *sværd*, *svärd*, 剣 etc.
- Facilitates hierarchical semantic search e.g. search for weapons – can return axes, spears, swords, rapiers, rifles, pistols etc.

Benefits of the approach

- Prompted data providers to clean and improve their data, in some cases producing new vocabularies
- Mappings allowed data owners to express how their own data relates to a common spine vocabulary (AAT)
- Matching tool and mappings are reusable
- Improved precision and recall in search
 - Use of multilingual entry vocabulary of AAT
 - Use of poly-hierarchical structure of AAT
- Improved visibility of resources originating from smaller data provider organisations



Thank you

ARIADNEplus was a project funded by the European Commission under the H2020 Programme, contract no. H2020-INFRAIA-2018-1-823914

The views and opinions expressed in this presentation are the sole responsibility of the author and do not necessarily reflect the views of the European Commission

Links

- ARIADNEplus project <https://ariadne-infrastructure.eu/>
- ARIADNE portal <https://portal.ariadne-infrastructure.eu/>
- Vocabulary Matching Tool <https://vmt.ariadne.d4science.org/vmt/>
- USW Hypermedia Research Group <https://hypermedia.research.southwales.ac.uk/>

Contact

- ceri.binding@southwales.ac.uk
- ORCID: 0000-0002-6376-9613

