



D12.2: Infrastructure Design

Authors:

Dimitris Gavrilis, ATHENA RC

Christos Papatheodorou, ATHENA RC

Panos Constantopoulos, ATHENA RC



Ariadne is funded by the European Commission's 7th Framework Programme.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7-INFRASTRUCTURES-2012-1) under grant agreement n° 313193.'

Version: 10 (*final*)

28th February 2015

Authors:

Dimitris Gavrilis, ATHENA RC

Christos Papatheodorou, ATHENA RC

Panos Constantopoulos, ATHENA RC

Contributing partners:

Julian Richards, ADS

Holly Wright, ADS

Guntram Geser, SRFG

Ceri Binding, UoSW

Douglas Tudhope, UoSW

Achille Felicetti, PIN

Martin Doerr, FORTH

Carlo Meghini, CNR

Nicola Aloia, CNR

Benjamin Štular, ZRC SAZU

Dorel Micle, ARHEO

Elizabeth Fentress, AIAC/Fasti

Nadezhda Kecheva, NIAM-BAS

Emmanuelle Bryas, INARP



ARIADNE is a project funded by the European Commission under the Community's Seventh Framework Programme, contract no. FP7

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

Table of Contents

1	Executive Summary	7
2	Introduction	8
3	Overall Architecture	9
4	Content analysis	15
4.2.	Integration strategy	16
4.2.1.	Metadata Integration.....	16
4.2.2.	Data Integration.....	18
5	Functional requirements.....	21
5.1	ARIADNE Catalogue & Data Access Layer	21
5.2	Digital Assets Management	21
5.3	Catalogue Preservation Service	22
5.4	Catalogue Record Quality Service.....	22
5.5	Vocabulary Directory Management.....	23
5.6	Resource Discovery.....	23
5.7	Previewing Service.....	24
5.8	Catalogue Record Enhancement	25
5.9	Deposit Service	25
5.10	Configuration & Management	26
5.11	Integration & Interoperability	26
6	Information Organization.....	27
6.1	Metadata Integration within ARIADNE Catalogue.....	27
6.2	Integration using CIDOC-CRM.....	28
6.3	Integrating data from native repositories	28
7	The ARIADNE Portal	31
8	Access conditions and Interfaces	33
8.1	Authentication and Authorisation	34
8.2	Logging and Audit	34
8.3	Machine Interface Specification	35
8.4	Prototype Data Layer Machine Interface Specification	36
9	Implementation technologies	38
10	Implementation roadmap.....	40
11	References	42

Annex I – Sample ARIADNE Catalogue Records.....	43
Annex II – ACDM Catalogue.....	47
1. ArchaeologicalResource.....	47
1.1 DataResource.....	48
1.2 Language Resource	49
1.3 Service.....	51
2. Other Main Classes	52
2.1 DataFormat.....	52
2.2 DBSchema.....	53
2.3 MetadataSchema.....	53
2.4 MetadataRecord	54
2.5 Distribution	54
2.6 TemporalRegion.....	54
2.7 SpatialRegion	54

Acronyms and abbreviations

ACDM	ARIADNE Catalogue Data Model (see Annex)
API	Application Programming Interface
BC/AD	Before Christ / Anno Domini, i.e. dates after Jesus's birth
CIDOC	International Committee for Documentation of the International Council of Museums
CIDOC-CRM	CIDOC - Conceptual Reference Model; ISO 21127:2006 standard: A reference ontology for the interchange of cultural heritage information
CRUD	Create, Read, Update, Delete operations conducted through a REST API (see below)
CSS3	Cascading Style Sheets
DC	Dublin Core (metadata standard)
DCAT	Data Catalogue Vocabulary, an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web
GIS	Geographical Information Systems / Services
HTML	HyperText Markup Language
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
ISO/IEC 11179	International standard for representing metadata for an organization in a metadata registry
Java / Javascript	Programming language
JDBC	Java Database Connectivity
JSON	JavaScript Object Notation
lat/lon	Latitude and longitude coordinates
LDAP	Lightweight Directory Access Protocol
LOD	Linked Open Data
Log4J	A Java-based logging package

METS	Metadata Encoding and Transmission Standard
MySQL	An open source relational database management system
OAI	Open Archives Initiative
OAI-PMH	OAI - Protocol for Metadata Harvesting
OAI-ORE	OAI - Object Reuse and Exchange Specification
OAuth	Authentication protocol
ODBC	Open Database Connectivity
PHP	PHP Hypertext Preprocessor (server-side scripting language for making dynamic and interactive web pages)
RDF	Resource Description Framework
REST	Representational State Transfer
REST API	RESTful Application Programming Interface
REST services	RESTful web services
SAML	Security Assertion Markup Language
SKOS	Simple Knowledge Organization System
SKOSify	Transforming a thesaurus, classification system or other knowledge organization system into the SKOS format
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
Syslog	Standard for computer message logging
UI	User Interface
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
WGS84	World Geodetic System 1984
XML	Extensible Markup Language
XSD	XML Schema Definition Language

1 Executive Summary

This document is a deliverable (D12.2) of the ARIADNE project (“Advanced Research Infrastructure for Archaeological Dataset Networking in Europe”), which is funded under the European Community's Seventh Framework Programme. This deliverable reports the results of Task 12.2. It comprises an analysis of the contributed datasets to be integrated into the project infrastructure, a needs assessment in terms of interoperability, resources and interfaces, the design of the interoperability architecture, and specifications of the integration tools to be developed within Task 12.3 to follow.

The main goal of Task 12.2 is to specify a resource integration and discovery mechanism for use in ARIADNE. The resources to be integrated are datasets and collections, GIS data, metadata schemas, ontologies and vocabularies available from the project partners, as well as institutions outside the ARIADNE consortium.

This deliverable provides an overview of the ARIADNE architecture, including a summary of the conformance of the architecture to the data and standards requirements set out in D12.1 [1], as well as to the specifications of the Services of the ARIADNE Infrastructure, presented in D13.1 [6]. This is followed by a content analysis of the main content types defined in D12.1 [1], and the integration strategy for the two levels of content: the metadata integration, and the data integration. This will attempt to integrate selected resources (datasets and/or metadata) from particular partners/data providers and provide cross-search and access mechanisms to integrated resources, using a faceted search on “what”, “where”, “when” and “resource type”.

The functional requirements of the components identified in the overall architecture are discussed for the ARIADNE catalogue and data access layer, digital assets management, the catalogue preservation service, the catalogue record quality service, management of the vocabulary directory, the resource service, the previewing service, record enhancement within the catalogue, and the deposit service. Requirements are also set out for configuration, management, integration and interoperability.

The information and components required for developing the ARIADNE interoperability framework are then set out as four main groups of information objects: the ACDM, vocabularies and thesauri, the CIDOC-CRM core ontology, and the users. This will be accomplished within the ARIADNE Catalogue by using thematic, spatial and temporal metadata, and then integrated into the ARIADNE portal, which aims to integrate all of the major outcomes of the project. It will be centred on three axes: resource discovery, searching integrated data and services.

The various access conditions and interfaces are then discussed, including authentication and authorisation, logging and audit, the machine interface specification, and the prototype data layer machine interface specification. This is followed by a listing of the implementation technologies to be used, and an implementation roadmap, which includes a progress table.

2 Introduction

This document is a deliverable (D12.2) of the ARIADNE project (“Advanced Research Infrastructure for Archaeological Dataset Networking in Europe”), which is funded under the European Community's Seventh Framework Programme. It presents the results of work carried out in Task 12.2 “Design and Specifications”.

The main objectives of WP12 are:

- To adapt infrastructures provided to ARIADNE for integration.
- To design, implement and set up the necessary mechanisms (crosswalks, mappings) and resources for interoperability.
- To set up the internal (APIs) and external (human) interfaces to access the integrated resources.

WP12 comprises the following tasks:

- Task 12.1 Assessment of Use Requirements
- Task 12.2 Design and Specifications
- Task 12.3 Implementing Integration
- Task 12.4 Testing

The main goal of Task 12.2 is to specify a resource integration and discovery mechanism for use in ARIADNE. The resources to be integrated are datasets and collections, GIS data, metadata schemas, ontologies and vocabularies available from the project partners as well as by institutions outside the ARIADNE consortium.

This deliverable reports the results of Task 12.2. It comprises an analysis of the contributed datasets to be integrated into the project infrastructure, an assessment of the needs in terms of interoperability, resources and interfaces, the design of the interoperability architecture, and specifications of the integration tools to be developed in Task 12.3. The report focuses on the following interoperability aspects:

- Content: Data to be integrated (datasets, GIS, collections, metadata, schemas and vocabularies)
- Information structure and metadata
- Integration technologies: mapping, GIS integration, etc.
- Access conditions and access interfaces
- Implementation technologies

3 Overall Architecture

According to the deliverable D12.1 “User Requirements” [1] which outlines the user needs for the interoperability framework of the ARIADNE Infrastructure, a general view of the integration functionalities, organized in four successive levels, is presented in Figure 1 [1]. At Level 1, data is created by research projects and groups, and subsequently stored in institutional repositories (Level 2). Then, at Level 3, this data is aggregated by higher level data managers such as data centres, portals and thematic information gates. At Level 4 the descriptions of the data are aggregated into a Catalogue (termed the registry in Figure 1); the Catalogue data model is ACDM [2,3], (presented in Annex II). Moreover the ARIADNE infrastructure provides data integration services, and updates the Catalogue with the descriptions of new, integrated datasets. Finally the ARIADNE portal will provide added-value information services to the archaeology community.

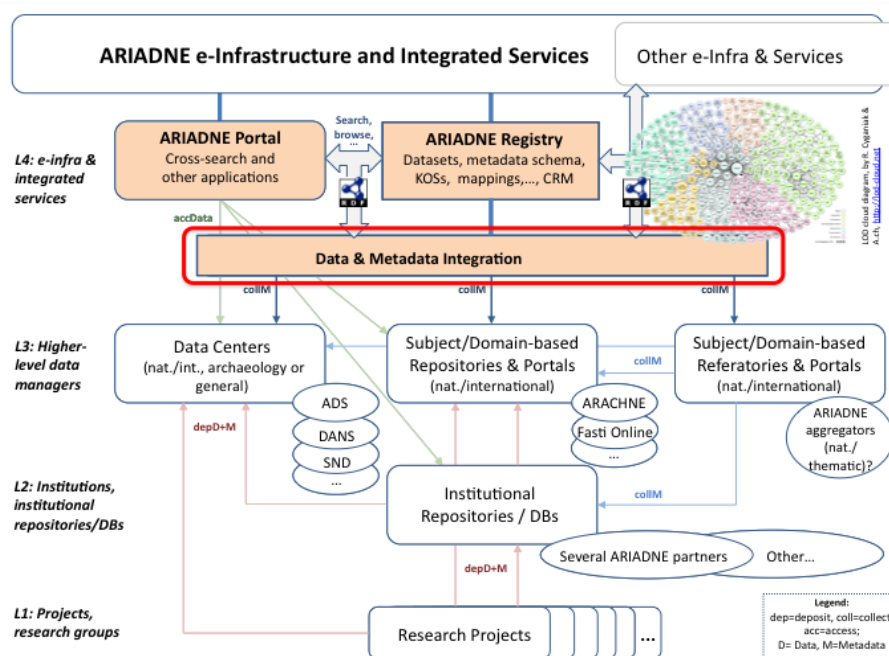


Figure 1. The ARIADNE data integration architecture.

Thus, ARIADNE aims to provide an interoperability framework consisting of:

1. A Catalogue (Registry) that aggregates descriptions of data and resources (data sets, databases, metadata schemas, vocabularies, etc.) from archaeological infrastructures like data centers, repositories etc.
2. Tools enabling the integration of data (such as databases, geographical datasets) and

metadata (such as collection-level descriptions, catalogues and finding aids).

3. Information access services to integrated data/metadata. These services are intended to be available not only to researchers and related stakeholders, but also to a wider audience requiring access to collections and datasets.

Figure 2 presents the services to be built within the framework of WP12, and identifies the interoperability framework. These services make use of data contained in the ARIADNE Catalogue, accessed through a Data Access Layer. Therefore there will be services for the ingestion and harvesting of metadata, which will be available to the ARIADNE partners and stakeholders that will contribute to the content enrichment. In particular the deposit service allows registered users to deposit data and metadata descriptions following the ACDM schema. The provided metadata is managed through a Digital Assets Management service, and is presented to the public through the ARIADNE portal. The Resource Discovery Services (mainly indexing and retrieval) will enable access to data resources and integrated viewing of data resource descriptions, through the ARIADNE Portal. The vocabulary management service is responsible for maintaining a list of SKOSified vocabularies and thesauri. The metadata enhancement service allows for automatic enhancement of metadata found in ACDM records. These enhancements include the mining of relations, and automatic linking with thesauri and vocabularies etc.

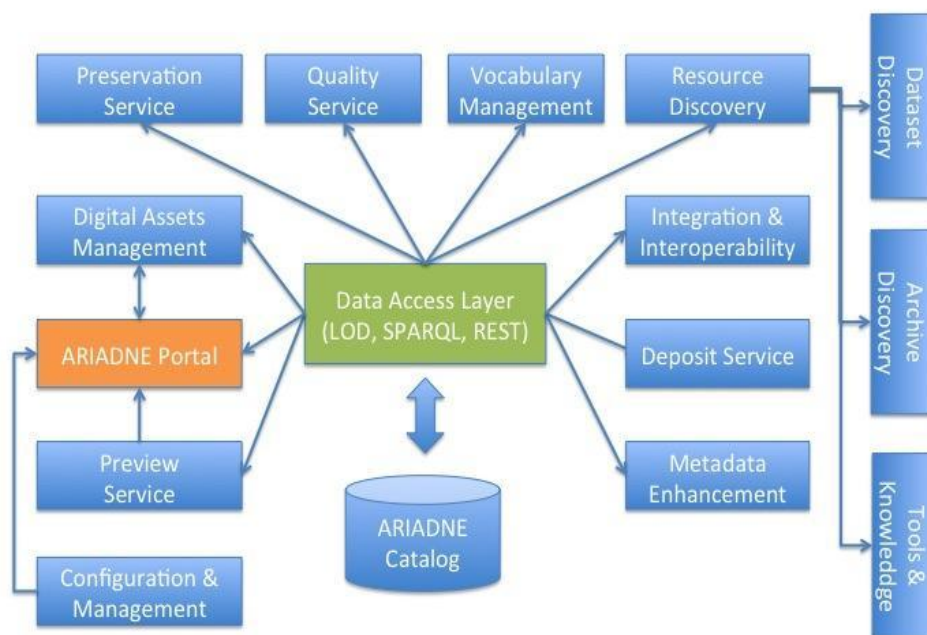


Figure 2. The ARIADNE data integration architecture.

The integration and interoperability service will enable:

- Integration of the registered and gathered metadata into the ARIADNE catalogue so as to support cross-search of the ARIADNE Catalogue according to the following facets: (i) What (such as Event Types (e.g. excavation, survey etc.), Topic/theme (e.g. Monument Type, Collections/Objects), (ii) Where, enabling resource discovery according to spatial criteria, (iii) When, to enable resource discovery according to temporal criteria, (iv) Resource type. The service will utilize mappings of registered vocabularies to a core (spine) vocabulary, such as the Arts and Architecture Thesaurus (AAT) [5] for term equivalence (see Section 4.2.1).
- Integration of data (e.g. datasets) and metadata (e.g. finding aids, collection-level metadata, etc.) of archaeological resources. These resources will be registered in the ARIADNE Catalogue. The service will utilize mapping tools based on a core ontology, like CIDOC-CRM [4] (see Section 4.2.2).

Furthermore, there will be services for the preservation of the infrastructure content (the metadata). The Quality service will be responsible for measuring the quality of the metadata and information provided (such as completeness of the metadata records, etc.), while the Preservation service will be capable of storing the full lifecycle of each entity of the ACDM model. Finally the Configuration & Management service will allow administrators and content owners to define certain parameters to do with the operation of the various services, as well as the access conditions on certain records/collections found within the ARIADNE Catalogue.

In summary the architecture provides:

1. the functionalities of the ARIADNE Catalogue;
2. interoperability and integration functionalities for archaeological resources;
3. the platform upon which the ARIADNE services [6] will run.

Table 1 summarizes the conformance of the architecture to the data and standards requirements analyzed in D12.1 [1], as well as to the specifications of the Services of the ARIADNE Infrastructure, presented in D13.1 [6], which focus mainly on resource discovery and preview, and are dependent on the implementation of the infrastructure. The number of “+” indicate the degree of relevance.

Table 1. Conformance of the ARIADNE Infrastructure to User requirements

	Data Access Layer	Integration & Interoperability	Deposit Service	Metadata Enrichment	Vocabulary Management	Quality Service	Preservation Service	Digital Assets Management	ARIADNE Portal	Preview Service	Resource Discovery
D12.1 Requirements											
Data Transparency	+								+	++	
Data Accessibility	+	++	+++						+	+	+
Metadata Quality	+	+		+		+++					
Data Quality	+					+++					
International Dimension	+	++			++				+	++	
Cross & border period search	+	++							+	+	+
Cross & border subject search	+	++							+	+	+
Map driven searching or visualisation	+								++	+++	
Bibliographic metadata from grey literature	+								+	+	+
Integration and interoperability from scientific databases		+++	+								+
Integration of particular kinds of artefact data		+++									
Dataset assessment required						+++					

	Data Access Layer	Integration & Interoperability	Deposit Service	Metadata Enrichment	Vocabulary Management	Quality Service	Preservation Service	Digital Assets Management	ARIADNE Portal	Preview Service	Resource Discovery
D13.1 Requirements											
Navigate to search page									+++		+
Enter Search Parameters									+++		+
Free text search	+								+++		+
Multilingual search	+	+++							+		+
Geo-integrated search	+	+++							+		+
Collection search	+	+++							+		+
Timeline search	+	+++							+		+
Navigate to tools & best practices									+++		+
Provide mapping tools								+++			
Change switch between the list, map and timeline view		++							+	+	
Multiple views at the same time (map and timeline view)		+							+	++	
API-access: The result-set must be machine-readable, preferably in standard XML or RDF	+++										

	Data Access Layer	Integration & Interoperability	Deposit Service	Metadata Enrichment	Vocabulary Management	Quality Service	Preservation Service	Digital Assets Management	ARIADNE Portal	Preview Service	Resource Discovery
Preview - The entries in the result-set can contain previews of the data.	+	+								+++	
Change								+	+		
Download data	+++								+	++	
Authenticate								+			
Preserve							+++				
Enhance metadata				++	+	+					

4 Content analysis

4.1. Content overview

Most of the innovative and significant services to be provided by the proposed infrastructure are tightly connected with the types of content within the ARIADNE Catalogue. According to deliverable D12.1 [1] the main content types are:

- Archaeological databases and spreadsheets
- Ethno-archaeological datasets
- Archaeological science databases
- Collections with a variety of formats
- Remote sensing data
- Map-based data
- Grey Literature
- Multimedia
- Vocabularies

Both deliverables D3.2 [7] and D12.1 [1] emphasise that the data are heterogeneous, and consist of:

- collections of data with diverse structures, and individual datasets with the same structure, while the content regarding the types of metadata schemas are classified within:
 - reference models,
 - archaeological sites, monuments, landscape areas,
 - museum objects,
 - bibliographic materials,
 - archival material, and
 - geospatial information.

Finally, the vocabulary types include international terminology resources and national terminology resources.

A live overview of the datasets/collections uploaded by the ARIADNE partners is available at: http://schemas.cloud.dcu.gr/ariadne-portal/index.php?op=provider_data.

4.2. Integration strategy

The Integration strategy refers to two levels: (i) the metadata integration, which will be realised inside the ARIADNE Catalogue, and (ii) the data integration, which will attempt to integrate selected resources (datasets and/or metadata) from particular partners/data providers and provide cross-search and access mechanisms to integrated resources.

4.2.1. Metadata Integration

At the first level, ARIADNE integrates metadata about resources that come in three different types: Data Resources, Language Resources, and Services. In particular:

- *DataResource*: whose instances represent the various types of data containers used by the ARIADNE partners, and provided to the project for integration. This class is created for the sole purpose of defining the domain and range of a number of associations. It is therefore an abstract class, whose instances are inherited from sub-classes.
- *LanguageResource*: whose instances include vocabularies, metadata schemas, gazetteers and mappings (between language resources). As new linguistic resources are added to the catalogue (such as subject heading systems and thesauri) the corresponding classes will be added to the model as a sub-class of this class. To describe language resources we have used ISO/IEC 11179 'Specification and Standardization of Data Elements'.
- *Services*: whose instances represent the services used by the ARIADNE partners and provided to the project for integration.

The ACDM provides two properties for the thematic description of the ARIADNE Catalogue resources:

- `ariadne:subject`, which associates the resource with one or more values from the following controlled list:
 - Fieldwork databases
 - Event/intervention databases
 - Sites and monuments databases
 - Scientific databases
 - Artefacts
 - Burials
- `dct:subject`, which associates the resource with one or more items from an existing controlled vocabulary the content providers may use.

It is estimated the ARIADNE Catalogue will be contain hundreds of thousands of resources, and therefore the main challenge is to develop a service that enables their integration. The purpose of this service is to provide semantic discovery [3] and to allow users to identify resources that relate to a specific topic, event, or spatio-temporal region.

In particular it is planned that the Integration and Interoperability service will support the cross-search of the ARIADNE Catalogue according to the following facets:

- **What:** to enable resource discovery according to (i) Event Types (such as excavation, survey etc.), (ii) Topic/theme (such as Monument Type) and (iii) Collections/Objects. For this integration faceted mappings to thesauri and vocabularies of archaeological object types will be developed; the vocabularies will be SKOSified and available through the Vocabulary management service.
- **Where:** to enable resource discovery according to spatial criteria. For this integration facet, latitude/longitude conversions will be developed.
- **When:** to enable resource discovery according to temporal criteria. For this integration facet vocabularies of local/national period terms are needed, along with their mappings to an absolute date range.
- **Resource type:** to enable resource discovery based on the classification of the resource types [2,3,7]: Fieldwork databases, Event/intervention databases, Sites and monuments databases, Scientific databases, Artefacts (and Collections of Artefacts), and Burials. This classification is already encoded by the attribute `ariadne:subject`, of the class `ArchaeologicalResource` within the ACDM schema.

Major requirements identified with respect to content are [1]:

1. Consistency of presentation: The content should be presented uniformly, according to the ACDM schema, and should be accessed via the interfaces of the ARIADNE portal.
2. Data and metadata quality: The data and metadata provided should be reliable, and follow particular requirements concerning their correctness and completeness. As mentioned, the infrastructure will provide services for ensuring data quality.

Although the ARIADNE Catalogue resources are described using subjects originating from controlled vocabularies, cross-search remains a problem, as there are multiple local vocabularies in use with no formal semantic links or mappings between them. Some indicative examples of existing local vocabulary resources identified for use in ARIADNE are listed below [5]:

- Data Archiving and Networked Services (DANS) provides a list of monument types (Archeologische complextypen) which have been made available online (see <http://rce.rnviewer.net/nl/structures>). The data is currently available in an XML format, having embedded SKOS concepts with unique identifiers.
- FASTI Online (FASTI) uses a flat list of monument types in the “advanced” search interface (see http://www.fastionline.org/data_view.php).
- The Italian Ministry for Heritage and Cultural Activities: Central Institute for Cataloguing and Documentation (Istituto Centrale per il Catalogo e la Documentazione, ICCD) publishes terminology for types of archaeological sites. The ICCD terminology will be made openly available in SKOS format for use within the ARIADNE project. (<http://www.iccd.beniculturali.it/getFile.php?id=182>).
- The English Heritage Monument Types Thesaurus is available online for download as SKOS (see <http://www.heritagedata.org/blog/vocabularies-provided/>) through the University of South Wales SENESCHAL project.

- Deutsches Archäologisches Institut (DAI) have produced a multilingual archaeological dictionary (<http://archwort.dainst.org/thesaurus/en/>).

Mappings between vocabularies are needed in order to implement a mediation platform that facilitates cross-search. Given the number of vocabularies, a ‘hub’ architecture is recommended for scalability and efficiency [5]. The hub architecture is based on an intermediate vocabulary or semantic structure, onto which the concepts from each local vocabulary may be mapped.

A query on a subject (concept) originating from one vocabulary can then utilise the intermediate vocabulary as a route through to concepts originating in other vocabularies, obtaining integration of the ARIADNE Catalogue resources. The Arts and Architecture Thesaurus (AAT) is one candidate to consider as a “hub” upon which local vocabularies can be mapped. Moreover the partners at the University of South Wales [5] provide a scenario for the implementation these mappings, as well as the semantic expansion of queries using the AAT SPARQL endpoint. The vocabulary mappings, as well as the alternative solutions for obtaining interoperability between sub-domain thesauri, including multilingual aspects, will be addressed in the framework of Task 3.3 “Thesauri and gazetteers” within WP3 “Standards and Interoperability”.

4.2.2. Data Integration

Data integration in ARIADNE is going to be realised in several phases. For this purpose the infrastructure will provide a set of tools for mapping and integrating archaeological resources. Primarily, the ARIADNE partners from PIN, CNR and FORTH will implement the tools. Furthermore, they will be exploited in an experimental process within the framework of WP14, aiming to provide enhanced access to the integrated resources.

In the first phase, called the preparation phase, access functionality will be designed and implemented in the registry, which would allow users to obtain enough information about which datasets are “good” candidates for integration. The suitability is based on the similarity of the following three fields:

- ariadne subject
- temporal
- spatial

A simple ranking function will be defined, so that given the identifier of a data resource, a ranking of the other data resources will be returned in order of decreasing of usefulness. In addition, the ARIADNE Catalogue will check other necessary conditions, such as the existence of mappings between schemas and the existence of appropriate rights. Also this information is possible to obtain from the ARIADNE Catalogue, because the required properties are part of the ACDM.

In the second phase, termed the decision phase, sophisticated access functionality will be used to select a number of candidates for integration testing, working with the appropriate data providing partners.

In the third phase, termed the execution phase, data integration will be carried out using:

- the open source schema mapping and data transformation tools being developed by FORTH, within the framework of ARIADNE (the “X3ML” schema mapping format, the “3M” mapping

editor and mapping memory, and X3ML data transformation tool) in order to make the transition to a CIDOC CRM compatible form (expressed in RDF).

- We will further upload transformed datasets to an open source triplestore-based platform called “Metadata Repository”, together with an intuitive user search interface, a hybrid search, and complete version management of uploaded RDF data sets. It will represent an ARIADNE adaptation of background technology developed by FORTH. It will allow for seamless searching and querying, and their explicit and implicit relationships contained in the ingested data - the ultimate goal of information integration.

In the last phase, named demonstration phase, we put the resulting integrated datasets on-line and build a number of demonstrators.

Within the execution phase, the metadata mapped and converted to CIDOC CRM format needs a solid platform to handle the complexity of semantic information, in order for it to be used in an efficient way. Thus an ideal ARIADNE integration platform would be conceived as a complex, modular system providing advanced interfaces and functions, and an architecture able to interact with the distributed repositories in a transparent way. The system architecture would likely include:

- the ARIADNE Catalogue, which returns information about the availability of digital collections, datasets, and their providers. It provides the first level of integration and completeness of access;
- one or more *curated* triplestores based on *thematic* aggregation services, which are scalable and able to handle a large number of RDF triples efficiently. Rather than aids for resource discovery, they provide deep integration of related data across resources, providers (and sub-disciplines) and LOD publishing;
- depending on the amount of data collected at a later stage, a simplified (deduced) form of the entire ARIADNE semantic graph could be automatically extracted from the triplestores, and could be served using a Solr system and/or exported to Europeana. This sort of component could handle a high data load, and answer simple questions effectively, at a global level;
- a set of modules to define mappings between legacy data formats, metadata schemas and CIDOC CRM compatible ontologies (and other cross-walks and migration paths), in order to enable long-term, sustainable data conversion and migration; in particular feeding thematic aggregation services and LOD publishing;
- a set of modules for communication with legacy archives, to implement the creation of semantic information on-the-fly and on-demand, for the population of the ARIADNE platform in a quick and effective way. This includes the data transformation tools using mapping definitions and crosswalks;
- a set of modules for controlling the internal consistency of the semantic graph, and the management of co-references. It may include advanced consistency and quality control mechanisms of data between providers, aggregators and researchers;
- a terminological service capable of taking advantage of the various multilingual terminology

resources available in ARIADNE, to be used throughout the various data integration and query operations;

- a hybrid query management system providing the necessary features to query the semantic graph, including a semantic query layer, text/image-based information retrieval, faceted browsing facilities, geographic and temporal visualization services;
- the ARIADNE Portal, the unique access point to the whole system, which includes a configuration and management module for user administration and data workflow control; the portal also provides all the required interfaces for interacting with the query management layer;
- modules for interaction with the newly developed or existing services, associated with the legacy data, such as annotation, geographic and 3D reasoning and statistical evaluation, etc.

Such a system would make able to query and extract information in many different ways (especially combinations of semantic and CBIR-based search), to integrate the results into a unique semantic graph and to present them to the user in a coherent manner by providing all the tools to analyse and use them as part of the user's research. The updates of the ARIADNE Catalogue, according to the modifications developed for use with legacy archives, would also be provided through advanced features, which always return the most updated version of the data to be queried.

In particular, the user would be able to query, in a semantic fashion, using advanced query mechanisms and interfaces, all the information held within the legacy archives, shared and unified by the architecture, and return relevant results in different views. Information concerning objects, places, events, actors and types can be retrieved and displayed in different ways, for instance on a timeline or a map if they contain temporal or spatial relationships, or browsed and refined with facet views, issued based on the most common fields. Knowledge provenance will be available and controllable for all data. Personalized annotations and scholarly arguments about interpretation and credibility of information can be created, shared and search collaboratively and act as scholarly publications.

This series of operations involves constant interaction with the ARIADNE Catalogue, which holds all the information relating to the distributed archives. The descriptive information stored in the ARIADNE Catalogue is able to drive the queries towards the most relevant archives, presumably containing information of interest to the user. Interaction with the terminological data and aggregation services is also very important for getting support at query and retrieval time.

References (i.e. URLs) to the legacy archives are always provided, to aid users navigating the original information, should they require custom searches tailored to specific needs.

The ARIADNE Portal, which represents the highest layer of the system architecture, will constitute the entry point for the users to the entire query mechanism. Through it, users can extract, analyse and use all the available information, as well as access it through the various services provided by the system itself.

5 Functional requirements

The functional specifications of the components identified in the overall architecture (Section 3, Fig. 2) are detailed below.

5.1 ARIADNE Catalogue & Data Access Layer

Component:	<u>ARIADNE Catalogue & Data Access Layer</u>
Definition:	The ARIADNE Catalogue contains the database where information about the archaeological resources is stored. This information follows the ACDM model [2,3] and is accessible through a Data Access Layer.
Input:	REST based calls for CRUD operations. GET/POST/PUT/DELETE/RDF/XML parameters containing ACDM related information and operation requests.
Output:	RDF/XML objects containing ACDM records Status & return codes
Function:	Model and store all ACDM information Provide CRUD (Create, Read, Update, Delete) operations through a REST API, as well as through an ODBC/JDCB driver and through SPARQL. The Data Access Layer requires authentication in order to perform certain tasks. The authentication options are presented in a separate section below.

5.2 Digital Assets Management

Component:	<u>Digital assets management</u>
Definition:	The digital assets management service provides all the necessary tools for content owners to manage their digital assets within the ARIADNE Catalogue.
Input:	An operation request An ACDM record in XML/RDF
Output:	The updated record and status/return code
Function:	Create new records in the ARIADNE catalogue Manage existing records in the ARIADNE catalogue This service allows creating and managing digital assets of any entity type

	defined within ACDM. Management is performed by registered users taking ownership of the data as they create it.
--	------------------------------------------------------------------------------------------------------------------

5.3 Catalogue Preservation Service

Component:	<u>Catalogue Preservation service</u>
Definition:	The service is responsible for preserving information entered in the ARIADNE Catalogue, and possibly information found within native repositories such as records, vocabularies etc.
Input:	<p>A record in XML/RDF format accompanied by:</p> <ul style="list-style-type: none"> - a CREATE/UPDATE/DELETE operation type - a log message used for auditing - user information <p>The record can be of the following types:</p> <ul style="list-style-type: none"> - ACDM - User - Service - Model - Vocabulary
Output:	A response/status return code
Function:	<p>To preserve information related to the various information objects within the ARIADNE infrastructure.</p> <p>The preservation service can store the full lifecycle of each entity of the ACDM model including vocabularies.</p>

5.4 Catalogue Record Quality Service

Component:	<u>Catalogue Record Quality service</u>
Definition:	<p>The Quality service is responsible for measuring the quality of metadata and provides information for two main classes of information:</p> <ul style="list-style-type: none"> - ACDM entities - Metadata located in native repositories
Input:	<p>An ACDM record in XML/RDF format</p> <p>An list of potential uses for the record</p>

Output:	An XML record containing quality measurement about the record
Function:	<p>To measure the quality of a given record with respect to a given use.</p> <p>With regard to ACDM entities, the metadata quality service evaluates the provided information about an entity (e.g. a dataset, collection or distribution) and returns a quality measure. Known quality criteria are metadata record completeness, correctness, etc. [8]. As the main purpose of the catalogue is information, this quality measure should reflect it.</p> <p>With regard to metadata in native repositories, metadata quality is evaluated for information outside the ARIADNE Catalogue. This information is accessible through endpoints that are described in ariadne:distribution entities and contain descriptions about: a) their metadata formats and b) metadata schemas. This is the minimum amount of information required to harvest and evaluate the quality of the records.</p>

5.5 Vocabulary Directory Management

Component:	<u>Vocabulary directory management</u>
Definition:	The vocabulary management service is responsible for maintaining a list of SKOSified vocabularies and thesauri, and make them accessible within the ARIADNE infrastructure.
Input:	A vocabulary in RDF/XML/Excel format
Output:	An SKOS representation of the vocabulary
Function:	<p>Ingest and manage vocabularies</p> <p>SKOSify vocabularies</p> <p><u>Manage</u> SKOS concepts, editing and interlinking with each other</p> <p>Identify owner information (marked as Agents) of various vocabularies so they can be linked to other entities found in the ARIADNE Catalogue.</p>

5.6 Resource Discovery

Component:	<u>Resource discovery</u>
Definition:	The resource discovery services direct data consumers to resources of interest based on various criteria.
Input:	A search request containing a list of search criteria and terms

Output:	A list of records that match the search request
Function:	<p>The resource discovery service searches within the ARIADNE Catalogue and allows data consumers to search using the following criteria:</p> <ul style="list-style-type: none"> - <u>Languages</u>: in which language(s) the resources are catalogued - <u>Geographical</u>: to which geographical location the resources refer - <u>Temporal</u>: to which period the resources refer - <u>Thematic</u>: which thematic areas the resources cover - <u>Keyword</u>: various keyword-based search on keywords, descriptions, etc. <p>The resource discovery service performs a search, based on the above criteria (or a combination of them). These search results can be grouped into templates and are displayed to the user through the Previewing service.</p>

5.7 Previewing Service

Component:	<u>Previewing service</u>
Definition:	The previewing service is responsible for previewing: a) search results, b) individual records.
Input:	<p>A list of ACDM record entries in XML</p> <p>An ACDM record in XML</p> <p>A display type (web, mobile, map, etc.)</p>
Output:	An HTML snippet or page displaying the results
Function:	<p>To display ACDM records using a number of outputs such as web, mobile, map, etc.</p> <p>For previewing search results, a group of records are displayed to the user in a variety of ways:</p> <ul style="list-style-type: none"> • As a list • On a map • On a timeline • Using a combination of map, timeline and other criteria (e.g. thematic) <p>For previewing individual records, the service displays a full record. The preview is adapted based on three criteria: a) the type of the entity (e.g. dataset, agent, distribution) displaying each record's individual properties, b) the detail level ranging from brief, normal and detailed and c) the access conditions (e.g display only specific properties or collections).</p>

5.8 Catalogue Record Enhancement

Component:	Catalogue record enhancement
Definition:	The metadata enhancement service allows for automatic enhancement of metadata found in ACDM records.
Input:	An ACDM record in XML/RDF format
Output:	The enhanced ACDM record plus an XML record with information related to the changes.
Function:	To automatically enhance ACDM related metadata records. The metadata enhancement service makes use of other resources such as ontologies and vocabularies. Enhancements include mining of relations, automatic linking with thesauri and vocabularies, etc. The metadata enhancement service relies on the data access layer to access the catalogue records, and on the vocabulary service to access vocabulary information.

5.9 Deposit Service

Component:	Deposit service
Definition:	The deposit service allows registered users to deposit metadata following the ACDM schema.
Input:	An ACDM record in XML/RDF
Output:	A status/response result in JSON/XML
Function:	<p>To ingest ACDM related information and check for mandatory fields, integrity checks, perform auditing on newly ingested records, etc.</p> <p>The deposit service has to follow the constraints laid out by both the ACDM schema (with its mandatory requirements) and the quality service. The deposits can be made through a variety of ways such as:</p> <ul style="list-style-type: none"> • Through a web based set of tools • Through a REST based ingest service that received XML instances that are compliant with an ACDM XSD

5.10 Configuration & Management

Component:	<u>Configuration & management</u>
Definition:	<p>The configuration & management service allows administrators and content owners to define certain parameters that have to do with:</p> <ul style="list-style-type: none"> - The operation of the various services - The access conditions on certain records/collections found in the ARIADNE catalogue
Input:	An service identifier/name plus a list of parameters (service specific) encoded in JSON/XML
Output:	A status/response result
Function:	<p>To configure the various services and the way they operate.</p> <p>Due to the large number and distributed nature of the ARIADNE services, this service should reside within the ARIADNE portal so that administrators and users can have an overview of all the configuration parameters in one place.</p>

5.11 Integration & Interoperability

Component:	<u>Integration & interoperability</u>
Definition:	The integration and interoperability service ensures the integration of the various elements of the ARIADNE Catalogue, and develops the tools and components to enable the integration of archaeological resources.
Input:	A list of integration parameters encoded in XML/JSON format
Output:	A list of ACDM records
Function:	Provides integration and interoperability functionalities for the ARIADNE project, according to the integration strategy described in the section 4.2. This service will have direct access to both the Data Access Layer and the Vocabulary Service.

6 Information Organization

The information and components required for developing the ARIADNE interoperability framework consist of four main groups of information objects:

ACDM: This comprises 10 major classes (ArchaeologicalResource, DataResource, DataFormat, DBSchema, MetadataSchema, MetadataRecord, Distribution, TemporalRegion, SpatialRegion, Service) [2,3], as well as a number of other entities.

Vocabularies and thesauri: These include a number of vocabularies that are either provided as linked open data through SKOS or are to be provided in some other format and SKOSified.

CIDOC-CRM core ontology: The ontology will facilitate integration and interoperability through semantic mapping/linking of resources.

Users: Information object that describes the various user access requirements on the ARIADNE infrastructure (use of the various services, access parts of the ARIADNE Catalogue, etc.).

The main purpose of the ARIADNE Catalogue is to register existing datasets and metadata, and to facilitate data integration. Integration can take place at two levels: (a) integration of the metadata for the registered resources and (b) integration of data (e.g. datasets) and metadata (e.g. resource discovery) using CIDOC CRM.

6.1 Metadata Integration within ARIADNE Catalogue

This can be accomplished within the ARIADNE Catalogue by using thematic, spatial and temporal metadata according to the integration strategy described in section 4.2.1. The registered heterogeneous datasets can be integrated on three levels:

Thematically, based on a generic thematic type. This will group information into six parts (see section 4.2.1): Fieldwork databases, Event/intervention databases, Sites and monuments databases, Scientific databases, Artefacts (and Collections of Artefacts) and Burials. For a further categorization, the subject information can be used when provided.

Spatially, based on the provided spatial information, which provides three levels of granularity: place name descriptions, address information and lat/lon coordinates. In most cases, spatial information is provided in lat/lon coordinates in WGS84 format. In cases where providers use the Web based UI, they can have direct access to services like Geonames [9], which provide accurate spatial information.

Temporally, based on the temporal information provided by the partners, which accommodates various levels of granularity such as: period name, BC/AD, rough year periods, and date ranges. For this purpose, precise formats for date ranges will be specified. For this integration, faceted vocabularies for local/national period terms are needed, mapped to absolute date ranges.

6.2 Integration using CIDOC-CRM

One of the most ambitious goals within ARIADNE is to provide ontology-based information integration services. This approach allows semantic integration of ACDM-related information using a core ontology. Due to the nature of ARIADNE, the CIDOC CRM (ISO 21127:2006) [4] is the most appropriate, and has been adopted by the most users/providers.

There is much textual data, which contains scientific or historical facts that could be usefully integrated from structured data. Extraction of such data into a knowledge representation format such as the CIDOC CRM is still in an experimental phase, typically as part of a semiautomatic process. Nevertheless, ARIADNE will allow for managing knowledge extracted in CRM-compatible form from texts in aggregation services. The ARIADNE Catalogue will register that such resources also exist as CRM extracted knowledge in the respective aggregation service.

Information in raster-based images is not accessible using knowledge extraction, and therefore in a CRM-based form. However, content-based information retrieval methods (CBIR) can provide similar functionality, which can be filtered by query results from a knowledge base, using a so-called “hybrid search”. This is a potential future application for the aggregation services.

3D-Models of immobile objects and sites can be regarded as “geometries” in the real world, and query results from a knowledge based aggregation service can also be constrained to provide information as part of a 3D Model.

6.3 Integrating data from native repositories

One of the main advantages of the ARIADNE Catalogue is that it sets the foundations for performing item level integration directly through the native repositories that are described. This can be made possible when the following conditions apply:

- the native repository information is provided as an ARIADNE distribution (through a machine readable interface such as OAI-PMH [10]),
- the metadata schema of the native repository is provided (in any kind of schema that is described in the respective DataFormat instance),
- the item structure of the native repository is provided (in OAI-ORE [11] or METS [12]).

It is possible to harvest information from the remote repository, return the results and map them in a desired format. Information such as the structure of the records (including the binary data streams) in the remote repository can be retrieved using protocols like the OAI-ORE [11] (Figure 3).

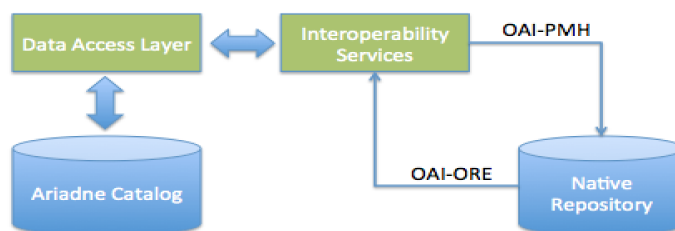


Figure 3. Integration with Native Repositories at record level.

The major advantage of holding the data structure of remote repositories is that it allows the use of mechanisms already in place for retrieving digital object structures in a formal way. Such mechanisms (e.g. OAI-ORE, METS) have existed for many years, and provide all the necessary information for item-level integration. Figure 4 presents a hypothetical scenario for a federated attempt at integration with native repositories at the item level, without having to harvest the remote repository.

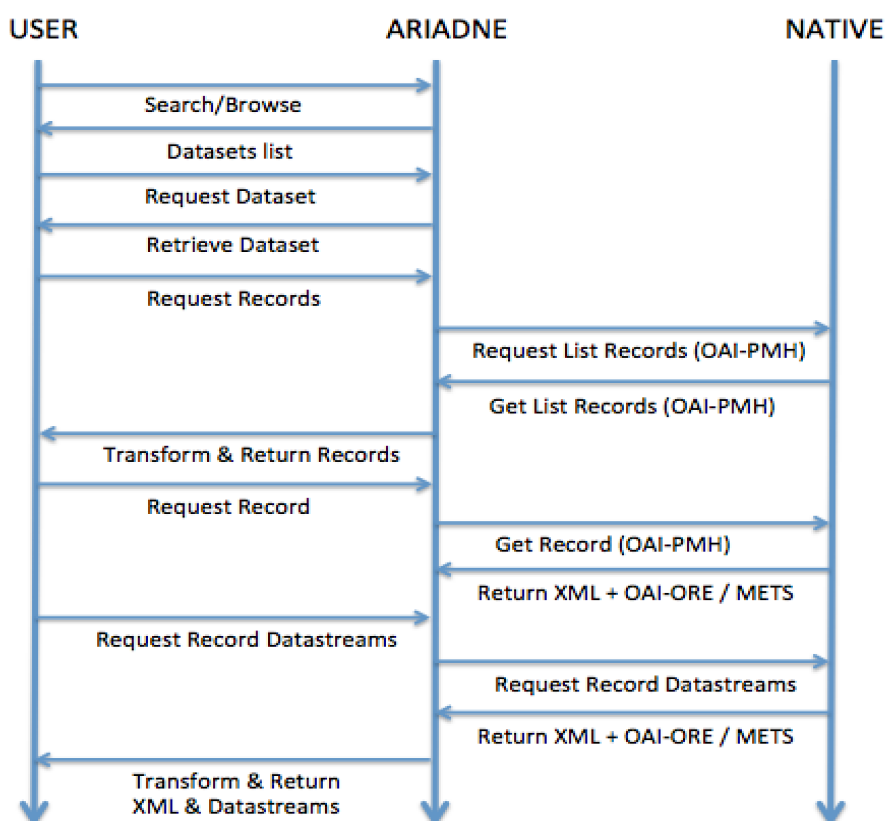


Figure 4. Integration with Native Repositories at record level.

Additionally, according to the Section 4.2.2, the following are needed for item level integration:

- a target schema for the integration, which is CIDOC CRM,
- the mapping from the original schemas to the target schema.

Furthermore for each property in the integrated item, it is needed:

- to agree on the domain of the integrated property,
- to agree on the mapping from the original domains to the target domains.

7 The ARIADNE Portal

The ARIADNE portal (Figure 5) aims to integrate all of the major outcomes of the project. These will be centred on three axes:

1. Resource discovery

This will be based on the ARIADNE Catalogue (in ACDM format) and will be provided through (simple and advanced) search mechanisms. As described, the resource discovery service allows users to search using the following criteria:

- Geographical: to which geographical location the resources refer,
- Temporal: to which period the resources refer,
- Thematic: which thematic areas the resources cover,
- Keyword: various keyword-based searches on keywords, descriptions, etc.

Resource discovery will be provided through the ACDM triplestore, plus an indexing component (e.g. Lucene based).

2. Searching Integrated data

These will include advanced search mechanisms based on the item level integration experiments that will be carried out within the project.

The searching of integrated data will include specialised applications that will perform item level integration based on: a) collections and b) databases/GIS based on technologies such as:

- specialised mappings based on the ARIADNE CRM reference model,
- NLP technologies that make use of vocabularies and LOD.

3. Services

These will include the various services that will be: a) created and provided as part of the project, b) provided by external parties, but are relevant to the project.

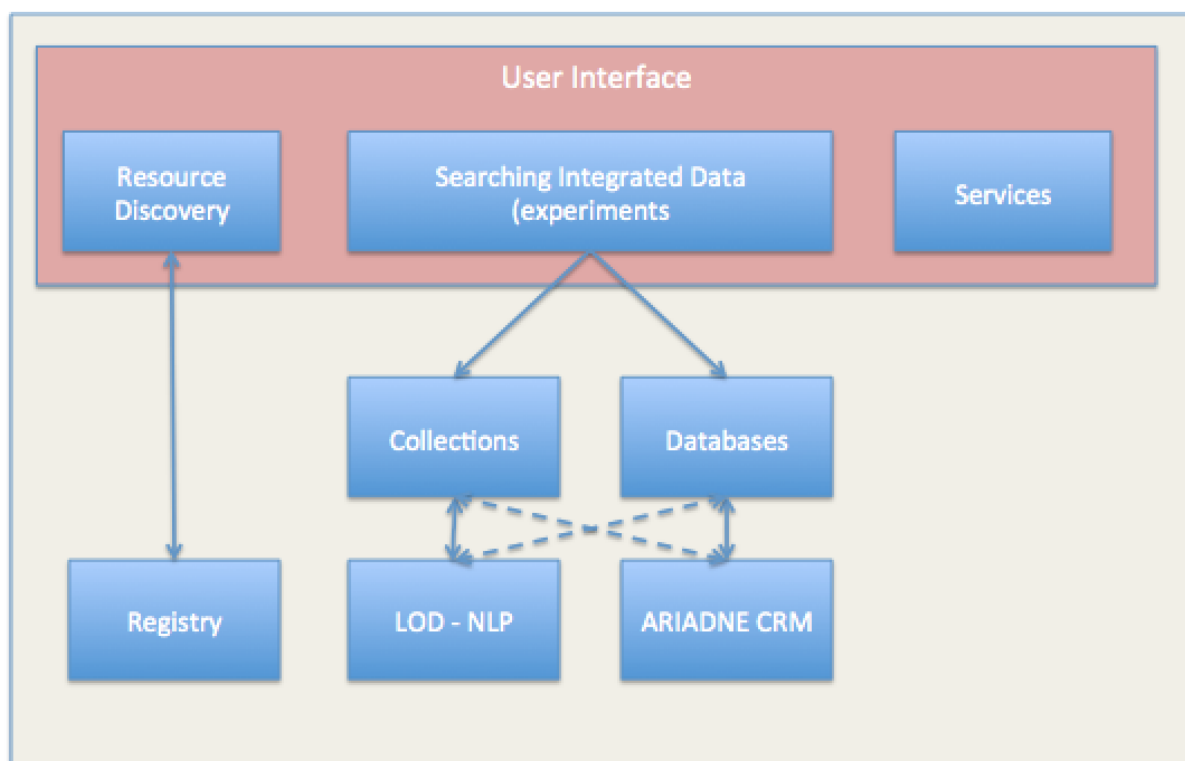


Figure 5. A general view of the ARIADNE portal architecture.

8 Access conditions and Interfaces

The proposed architecture presents a number of different services and technologies, along with a Catalogue, which may carry sensitive information. Therefore, there is a need for a robust and flexible authentication and authorization service that can be utilized throughout the infrastructure. The service should be able to authenticate the following entities:

1. ARIADNE Catalogue core services
 1. Web UI
 2. REST Services
 3. SPARQL Interface
1. Integration services
 1. REST Service
2. Enrichment services
 1. REST Service
3. Preservation services
 1. Web UI
 2. REST Service
4. Vocabulary services
 1. Web UI
 2. REST Service
5. Portal
 1. Access to the Web UI of the portal

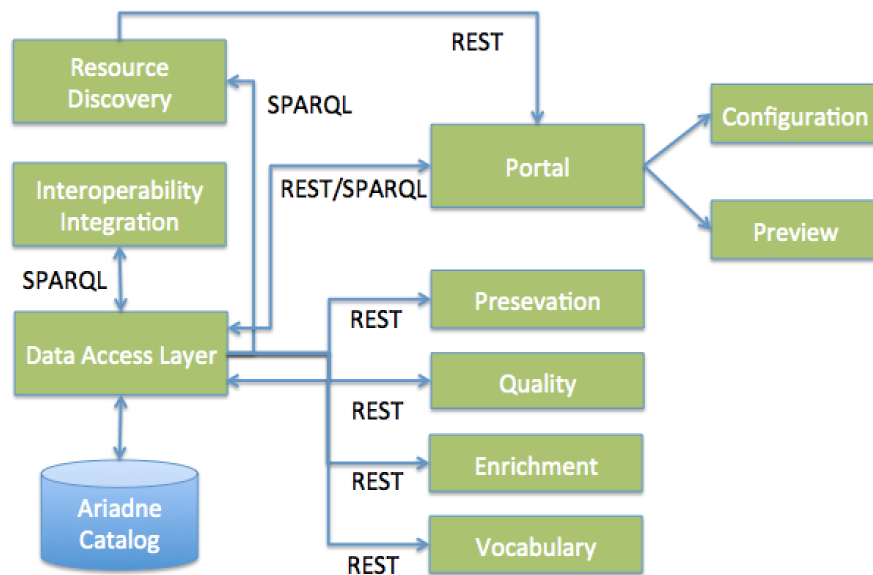


Figure 6. Inter-Service interfaces with Data Access Layer

The distributed nature of the architecture, along with the heterogeneous services that are specified,

require them to be handled using different protocols and data types. For instance, ideally a vocabulary service would serve a set of thesauri concepts in RDF format to a resource discovery service, and in JSON to the portal.

Similarly, major machine interface components can provide either a REST or a SPARQL interface (or both, according to their function). A proposed inter-component (or service) interaction based on the exposed machine interfaces is presented in Figure 5.

8.1 Authentication and Authorisation

The main technologies that can be used for Authentication & Authorisation among the various services include:

1. LDAP for maintaining a user directory with various types of flexible access and authorization structures,
2. SAML for defining and exchanging authentication and authorization information,
3. OAuth for authenticating services using a secure delegated access model.

An Authentication and authorization service will be setup to hold all related information, and provide an access framework throughout the infrastructure. This approach enables the easy integration with other infrastructures and provides better control both from the administrator and user point of view. From the above options, OAuth is proposed as it can facilitate the authentication of distributed services, which fits the model of the ARIADNE infrastructure architecture.

There are two main entities that require authentication:

1. Human interfaces: these typically include Web-based applications,
2. Machine interfaces: these typically include REST services.

In the first case, a user/password challenge is employed, and when combined with protocols such as OAuth 2.0, the distributed architecture of services is facilitated. In the second case, an API key is used because it is widely accepted and hence this approach is familiar to most developers.

8.2 Logging and Audit

All services should present a logging mechanism for holding a number of events such as:

1. Exceptions
2. Usage information

Especially for security reasons, a centralized log of all authentication attempts (both successful and not) should exist. Widely adopted technologies with a successful application record include:

1. Syslog and remote syslog,
2. Log4J.

8.3 Machine Interface Specification

All machine interfaces should specify in detail their API Calls. Especially in the case of REST services, the following template is proposed where for each REST API, the base URL is provided, along with a detailed description of all its requests. For each request, the parameters are specified with a method, data type and description. Similarly, the possible responses (along with Status Codes) and examples are also specified.

[API BASE URL]

Request

Method	URL
<i>[POST or GET or PUT or DELETE]</i>	

Parameter	Datatype	Description
api_key	String	The API Key provided by the infrastructure
format	String	The format for the returned information (preferably xml / rdf / json)
...

Response

Status	Response
200	[Example of response]
404	
500	

8.4 Prototype Data Layer Machine Interface Specification

Table 3 illustrates a prototype data layer machine interface for accessing the ARIADNE Catalogue and performing basic operations. This API will facilitate the implementation of services on top of the ARIADNE Catalogue.

Table 3. Data Layer Interface Specification

API Method	Description	Request parameters	Response parameters
/ListProviders	Get a list of providers	-	An XML/JSON response with the provider name and identifiers
/ListClasses	Get a list of all available classes	-	An XML/JSON response with the available classes (id, name)
/ListClass	Get information on a specific class	Class name	An XML/JSON response with the available class information
/Search	Search the catalogue for instances of specific class having specific property values	Class Term Properties	An XML/JSON response with the matched results
/GetRecord	Get information on a		The XML/RDF

API Method	Description	Request parameters	Response parameters
	specific record		representation of the record
/ListRecords	List all records for a specific class or provider. Return views with varying detail. Also possible to return specific properties for each record.	Class Provider View Properties list	An XML/RDF response with the records
/UpdateRecord	Update a specific record	An XML record with the updated record	A status code
/CreateRecord	Create a new record	A provider identifier The XML record	A status code
/DeleteRecord	Delete an existing record	The record identifier A reason for deletion	A status code
/Auth	Authenticate user	Username Password	An XML/JSON object containing session token plus the user details.
/UpdateProfile	Update the user's profile	A list of POST variables or JSON object with the user's data	A status code
/Profile	Get the user's profile information	The user identifier plus the session token	An XML/JSON object containing the user profile information

9 Implementation technologies

Data integration will be facilitated through technologies such as RDF, SPARQL, and Persistent Identifiers. The ARIADNE Catalogue data is currently stored within an SQL database following the schema described in the previous sections. In order to facilitate integration the SQL database it will be synced to an RDF triplestore which will provide an open SPARQL interface. The RDF encoded information will make use of unique and persistent identifiers.

The components of the proposed architecture are quite diverse, requiring the use of various technologies such as: storage technologies, programming languages, and existing frameworks. Furthermore, different partners with experience in different technologies will implement many of the components. Thus, the infrastructure should focus on common technologies that will act as glue among all heterogeneous components. HTTP REST and RDF/XML/JSON will be the primary building blocks of the infrastructure.

The reasons for following this approach have to do with the distributed nature of the infrastructure where multiple partners develop a number of services. A lightweight protocol such as HTTP REST can ease the development process, while also meeting all the infrastructure demands (in terms of complexity and performance). There are no complex data structures and protocols for exchanging information. Furthermore, HTTP REST is consistent with Linked Open Data (LOD) implementations, which are primary goals for this infrastructure.

The design of the human interfaces must fulfill the following requirements:

1. they should be accessed through the web,
2. they must be capable of being accessed by mobile devices and a variety of (web based) clients,
3. they must be able to make use of HTTP REST protocols and directly consume web services that produce / respond in XML/JSON,
4. they must provide an enhanced user experience.

The current state-of-the-art approach for the above includes HTML5/CSS3 plus a number of frameworks to meet these requirements.

- The Twitter bootstrap approach is proposed for the portal implementation due to its highly adaptive nature and wide acceptance worldwide.

The encoding of information is proposed to follow an RDF/XML approach mainly because of the ACDM model, which is encoded in both these formats.

- The implementation technologies include:
 - Overall architecture
 - HTTP REST
 - Programming languages
 - Java
 - PHP

- Web Interfaces
 - Javascript
 - HTML5/CSS3
 - Bootstrap
- Data encoding
 - XML
 - RDF
 - JSON
- Software architectural patterns such as MVC
 - Laravel php framework
 - Spring Web MVC framework

10 Implementation roadmap

The implementation roadmap of the ARIADNE infrastructure consists of a number of steps that can be seen in Table 4 below. It starts with the ARIADNE Catalogue and continues with the data access (layer three) interfaces. The roadmap ensures the proper ingestion and management of data followed by the various services that operate on those data. Most services contain two interfaces: a) a human and b) a machine interface. Both are necessary, but the latter allows the construction of services.

The main dependencies are:

1. the deposit service depends on the Human and REST interface of the data access layer,
2. the resource discovery service depends on the REST and SPARQL interfaces of the data access layer,
3. the preview service depends on the ARIADNE portal,
4. the digital assets management depends on the REST interface of the data access layer,
5. the preservation service depends on the REST interface of the data access layer,
6. the metadata quality service depends on the REST interface of the data access layer,
7. the vocabulary service depends on the REST interface of the data access layer.

Table 4. Implementation Progress

Service	Status
1. ARIADNE catalogue	100%
2. ARIADNE data access layer	
2.1 RDBMS interface	100%
2.2 REST interface	50%
2.3 SPARQL interface	0%
3. Deposit service	
3.1 Human interface	100%
3.2 Machine interface	100%

Service	Status
4. Resource discovery	
4.1 Human interface	50%
4.2 REST	0%
4.3 SPARQL	0%
5. ARIADNE Portal	50%
6. Preview service	
6.1 ACDM preview	50%
6.2 Record level preview	0%
7. Digital assets management	
7.1 Human interface	100%
7.2 Machine interface	0%
8. Preservation service	50%
9. Quality service	
9.1 Metadata quality	50%
9.2 Data quality	0%
10. Vocabulary management	0%
11. Integration & Interoperability	
11.1 Metadata Integration (within ARIADNE Catalogue)	0%
11.2 Data Integration (based on CIDOC CRM)	0%

11 References

1. H. Wright, ARIADNE D12.1. Use Requirements, July 2014.
2. C. Papatheodorou, D. Gavrilis, K. Fernie, H. Wright, J. Richards, P. Ronzino, C. Meghini, ARIADNE D3.1 Initial report on standards and on the project registry, November 2013.
3. N. Aloia, C. Papatheodorou, D. Gavrilis, F. Debole, C. Meghini, Describing Research Data: A Case Study for Archaeology, 13th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2014), Amantea, Italy, October 2014, in R. Meersman et al. (Eds.): "On the Move to Meaningful Internet Systems: OTM 2014 Conferences", Lecture Notes in Computer Science (LNCS) No. 8841: Springer-Verlag, 2014, pp. 768-775.
4. CIDOC-CRM (Conceptual Reference Model), <http://www.cidoc-crm.org>.
5. Ceri Binding, "Cross search of ARIADNE via subject", ARIADNE Report.
6. H. Hollander, M. Hoogerwerf, ARIADNE D13.1 Service Design, July 2014.
7. P. Ronzino, K. Fernie, C. Papatheodorou, H. Wright, J. Richards, ARIADNE D3.2 – Report on project standards, November 2013.
8. M. A. Gonçalves, B. L. Moreira, E.A. Fox, L. T. Watson, "What is a good digital library?" - A quality model for digital libraries", Inf. Process. Manage. 43(5): 1416-1437 (2007)
9. Geonames, <http://www.geonames.org>.
10. OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting, <http://www.openarchives.org/pmh/>.
11. OAI-ORE - Open Archives Initiative, Object Reuse and Exchange Specification, <http://www.openarchives.org/ore/>.
12. METS - Metadata Encoding and Transmission Standard, <http://www.loc.gov/standards/mets/>.
13. DCAT - Data Catalogue Vocabulary, <http://www.w3.org/TR/vocab-dcat/>.
14. European Union Open Data Portal, <http://open-data.europa.eu>.

Annex I – Sample ARIADNE Catalogue Records

This annex presents a few XML samples of ACDM provided records.

EASY Sample

```
<acdm:dataResource xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:acdm="http://ariadne-registry.dcu.gr/schema-definition"
  xmlns:dc="http://www.w3.org/ns/dcat#"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:eas="http://easy.dans.knaw.nl/easy/easymetadata/eas/"
  xmlns:emd="http://easy.dans.knaw.nl/easy/easymetadata/"
  xmlns:abr="abr.lookup"
  xsi:schemaLocation="http://ariadne-registry.dcu.gr/schema-definition http://ariadne-
registry.dcu.gr/schema-definition/sample_ariadne_xml.xsd">

  <dcterms:title>Bureauonderzoek ten behoeve van het leidingtracé tussen de NAM-locaties Opende
Oost 1 en Marumerlage 1, gemeente Grootegast en Marum (Gr.)</dcterms:title>

  <dcterms:description>In april 2009 is in opdracht van de NAM door ingenieursbureau Oranjewoud BV
een bureauonderzoek uitgevoerd voor een leidingtracé tussen de NAM-locaties Opende Oost-1
(Gemeente Grootegast) en Marumerlage 1 (Gemeente Marum). Het plangebied ligt in een relatief
laaggelegen dekzandgebied dat vanaf het laat-Neolithicum bedekt was met veen. De lage
dekzandruggen in de omgeving werden vanaf de midden-Bronstijd vrijwel ontoegankelijk. Mogelijk
waren de hoogste dekzandruggen in de bredere omgeving wel sporadisch bewoond: op de dekzandrug
waar het dorp Marum ligt zijn resten uit de IJzertijd aangetroffen. In het gebied komen veel pingoruïnes
voor, deze liggen als dobben in het landschap. Niet elke dobbe is echter een pingoruïne: het kan ook
een veendepressie zijn of uitblazingsvlakte. Op basis van reeds bekende archeologische waarnemingen
blijkt dat het gebied een lange bewoningsgeschiedenis kent. De dekzandruggen waren reeds in de
prehistorie bewoond. Er zijn echter in de omgeving van het plangebied geen samenhangende
vindplaatsen aangetroffen. Het betreft uitsluitend losse vondsten.</dcterms:description>
```

```

<dcterms:issued>2009-03-25</dcterms:issued>

<dcterms:modified>2009-11-19</dcterms:modified>

<acdm:originalId preferred="false">AIP_ID twips.dans.knaw.nl-4920366299359862317-
1258618514487</acdm:originalId>

<acdm:originalId preferred="false">eDNA-project a11267</acdm:originalId>

<acdm:originalId preferred="false"> 2009/43 (rapportnr)</acdm:originalId>

<acdm:originalId preferred="false"> 197685 (projectnr)</acdm:originalId>

<acdm:originalId preferred="true">urn:nbn:nl:ui:13-h73-jys</acdm:originalId>

<acdm:originalId preferred="false">Archis_onderzoek_m_nr 34302</acdm:originalId>

<acdm:originalId preferred="false">DMO_ID easy-dataset:11244</acdm:originalId>

<dcterms:language>nl</dcterms:language>

<dcterms:landingPage>http://www.persistent-identifier.nl/urn:nbn:nl:ui:13-h73-
jys</dcterms:landingPage>

<dcterms:accessRights>Access restricted to registered members of a group.</dcterms:accessRights>

<dcterms:isPartOf>DANS</dcterms:isPartOf>

<dcterms:creator>Oranjewoud BV</dcterms:creator>

<dcterms:creator>Spoelstra, A.</dcterms:creator>

<dcterms:creator>Kaptein, I.</dcterms:creator>

<acdm:ariadneSubject>Fieldwork databases</acdm:ariadneSubject>

<acdm:SpatialRegion>

  <acdm:lat>53.15612478</acdm:lat>

  <acdm:lon>6.23994777</acdm:lon>

  <acdm:coordinateSystem>http://www.opengis.net/def/crs/EPSSG/0/4326</acdm:coordinateSystem>

</acdm:SpatialRegion>

```

```

<dc:keyword>06H</dc:keyword>

<dc:keyword>Marum</dc:keyword>

<dc:keyword>Groningen</dc:keyword>

</acdm:dataResource>

```

DENDRO Sample

```

<acdm:dataResource xsi:schemaLocation="http://purl.org/dc/dcmitype/
http://dublincore.org/schemas/xmls/qdc/dcmitype.xsd          http://purl.org/dc/terms/
http://dublincore.org/schemas/xmls/qdc/dcterms.xsd          http://purl.org/dc/elements/1.1/
http://dublincore.org/schemas/xmls/qdc/dc.xsd                http://www.w3.org/ns/dcat# http://ariadne-
registry.dcu.gr/schema-definition/dcat.xsd                    http://ariadne-registry.dcu.gr/schema-definition
http://ariadne-registry.dcu.gr/schema-definition/sample_ariadne_xml.xsd"
xmlns:acdm="http://ariadne-registry.dcu.gr/schema-definition"
xmlns:dcmitype="http://purl.org/dc/dcmitype/" xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:dcat="http://www.w3.org/ns/dcat#">

<dcterms:title>Eindhoven, Stratumseind 5</dcterms:title>

<dcterms:issued>2012-04-16T11:37:38.383Z</dcterms:issued>

<dcterms:modified>2012-04-16T11:37:38.383Z</dcterms:modified>

<acdm:originalId preferred="true">dccd:5843</acdm:originalId>

<acdm:originalId preferred="false">99.001</acdm:originalId>

<dcat:landingPage>http://dendro.dans.knaw.nl/project/dccd:5843</dcat:landingPage>

<acdm:owner>BAAC bv</acdm:owner>

<dcterms:language>nl</dcterms:language>

<dcterms:isPartOf>DCCD</dcterms:isPartOf>

<acdm:ariadneSubject>Scientific Databases</acdm:ariadneSubject>

```

```
<dcterms:rights>Restricted access for levels more detailed than: series</dcterms:rights>
<dcterms:subject>gebouwd erfgoed</dcterms:subject>
<dcat:keyword>datering</dcat:keyword>
<dcat:keyword>gebouw</dcat:keyword>
<dcat:keyword>balk</dcat:keyword>
<dcat:keyword>dakspoor</dcat:keyword>
<dcat:keyword>Quercus</dcat:keyword>
<acdm:SpatialRegion>
<acdm:lat>51.4364121007867</acdm:lat>
<acdm:lon>5.48176339586946</acdm:lon>
<acdm:coordinateSystem>http://www.opengis.net/def/crs/EPSG/0/4326</acdm:coordinateSystem>
</acdm:SpatialRegion>
<acdm:temporalRegion>
<acdm:from>1336</acdm:from>
<acdm:until>1337</acdm:until>
</acdm:temporalRegion>
</acdm:dataResource>
```

Annex II – ACDM Catalogue

The model for the Integration data and metadata within the ARIADNE Infrastructure is called the *ARIADNE Catalogue Data Model* (ACDM), which extends the Data Catalogue Vocabulary (DCAT), a quasi-recommendation of the W3C Consortium [13], which has been used in several approaches for the development of dataset registries, such as the European Union Open Data Portal, [14]. ACDM is presented in Figure 7. This annex presents the last updated version of the model at the time of writing. The most recent version is available from the project web site.

ARIADNE integrates resources that come in three different types: Data Resources, Language Resources, and Services. The central notion of the model is the class *ArchaeologicalResource*, specialised in:

- *DataResource*, whose instances represent the various types of data containers owned by the ARIADNE partners and contributed to the project for integration. This class is created for the sole purpose of defining the domain and the range of a number of associations. It is therefore an abstract class, whose instances are inherited from sub-classes.
- *LanguageResource*, having as instances vocabularies, metadata schemas, gazetteers and mappings (between language resources). As new resources of a linguistic nature are added to the catalogue (such as subject heading systems and thesauri) the corresponding classes will be added to the model as a sub-class of this class. To describe language resources we have used ISO/IEC 11179 ‘Specification and Standardization of Data Elements’.
- *Services*, whose instances represent the services owned by the ARIADNE partners and contributed to the project for integration.

1. *ArchaeologicalResource*

The *ArcheologicalResource* class defines the properties common to its subclasses, mostly using the terms of the DCAT vocabulary, to which it adds properties for specifying: the access policy and the original identifier of the resource. The main associations having this class as a domain are (cardinality constraints are omitted for brevity, whereas XML notation is used for property names):

dct:isPartOf associates any archaeological resource in a catalogue with that catalogue.

dct:publisher: associates any archaeological resource with an agent responsible for making the resource publicly available.

dct:creator: associates any archaeological resource with an agent primarily responsible for creating the resource.

owner: associates any archaeological resource with an agent that is the legal owner of the resource.

legalResponsible: associates any archaeological resource with a person holding the legal responsibility of the resource.

scientificResponsible: associates any archaeological resource with a person holding the scientific responsibility of the resource.

technicalResponsible: associates any archaeological resource with a person holding the technical responsibility of the resource and contact person.

ariadneSubject associates any archaeological resource with one or more archaeological subjects defined by ARIADNE, namely:

- Fieldwork databases
- Event/intervention databases
- Sites and monuments databases
- Scientific databases
- Artefacts
- Burials

dct:subject associates any archaeological resource with a subject drawn from an existing vocabulary.

1.1 DataResource

This class specializes the class *ArchaeologicalResource*, and has as instances the archaeological resources that are data containers such as *databases*, *GIS*, *collections* or *datasets*. Two important attributes of this class are **dct:temporal** and **dct:spatial**, giving the spatial and temporal coverage of each instance data resource. The attributes will be used for establishing the degree to which two data resources are worth integrating.

The main associations having this class as domain are:

dct:isPartOf associates a data resource with the collections of which the data resource is part.

dcat:distribution: associates a data resource with the distributions, i.e. the accessible forms of the resource.

hasItemMetadataStructure: associates a data resource with the format of the metadata of the members

(or items) of the data resource (e.g. metadata of each record in a dataset, or of each item in a collection).

hasMetadataRecord: associates a data resource with the metadata of the resource as created by the organization holding the resource (for instance, the record describing a dataset in the organization holding the dataset).

1.1.1 Collection

This class is a specialization of the class `DataResource`, and has as instances collections in the archaeological domain. The items in a collection are data resources themselves; for instance, a collection may include a textual document, a set of images, one or more datasets and other collections. For interoperability, `Collection` is a sub-class of `dcmitype:Collection`. The main association having this class as domain is `dct:hasParts`, which associates a collection with the data resources that are in the collection. This association is used in the ARIADNE Catalogue only for stating membership of data resources in collections, since the Catalogue does not store information on individual objects.

1.1.2 Database

This class is a specialization of the class `DataResource`, and has as instances databases, defined as a set of homogeneously structured records managed through a Database Management System (such as MySQL), recorded as an attribute of the class. The main association having this class as domain is `hasSchema`, which associates a database with the schema defining the structure of the data in the database. Such schema is an instance of the class `DBSchema`.

1.1.3 Dataset

This class is a specialization of the classes `DataResource` and `dcat:Dataset`, and it has archaeological datasets as instances. An archaeological dataset is defined as a set of homogeneously structured records that are not managed through a Database Management System. The main association having this class as domain is `hasRecordStructure`, which associates a dataset with a data format defining the structure of its records. Such format is an instance of the class `DataFormat`.

1.1.4 GIS

This class is a specialization of the class `DataResource`, and has as instances Geographical Information Systems (GISs). The GIS technology used for each instance is modelled as an attribute of the class.

1.2 Language Resource

A language resource is a resource of a linguistic nature, whether in natural language (such as a gazetteer) or in a formal language (such as a vocabulary or a metadata schema). It also includes

mappings, understood as associations between expressions of two language resources that may be of a formal (e.g., sub-class or sub-property links) or an informal (e.g., natural language rules) nature. The LanguageResource have as instances vocabularies, metadata schemas, gazetteers and mappings (between language resources). The most significant subclasses of the class are:

1.2.1 MetadataSchema

This subclass has as instances metadata schemas used in the archaeological domain. The main attributes of this class are:

- standardUsed: the standard that the schema is based on, if any
- dct:description: a description of the format, recommended for proprietary formats
- foaf:homepage: an HTTP URI pointing to the web page describing the schema

The main associations having this class as domain are:

- isRealizedBy associates a metadata schema with a data format that realizes it in some specific encoding language.
- hasElements: associates a metadata schema with its elements.
- hasVersion: associates a metadata schema with a version of its.
- usedby: associates a metadata schema with the organizations using it.

1.2.2 Vocabulary

This is a subclass has as instances vocabularies used in the archaeological domain. The main attributes of this class are:

- dct:identifier: a URI, which identifies the original name or location of the vocabulary
- dct:title: the name of the vocabulary
- dct:description: a description of the vocabulary
- dct:language: the vocabulary language
- status: the current status of the vocabulary

The main associations having this class as domain are:

- hasConcepts: associates a vocabulary with its concepts.

- **hasVersion**: associates a vocabulary with a version of it.
- **usedby**: associates a vocabulary with the organizations using it.

1.2.3 Gazetteers

Since Gazetteers are geographical vocabularies, the class *Gazetteer* has a similar structure.

1.2.4 Mapping

An instance of this class represents a mapping between two language resources (e.g. metadata schemas). The main attributes of this class are:

- **name**: The name of the mapping.
- **source**: The id of the source language resource.
- **target**: The id of the target language resource.
- **xsltURI**: The URI to the XSLT describing the mapping (if any).

1.3 Service

The modelling of the services to be integrated by ARIADNE is at a preliminary stage of development. The goal is to provide the primitives for describing the services developed by the project partners for which integration or reuse can be envisaged. A preliminary survey has brought about the following categories:

- Services that make use of GIS software;
- Services that make use of databases management systems;
- Ad hoc systems developed in-house that do not use any of the previous technologies;
- Composite services that use a combination of the previous categories.

An ARIADNE service is therefore understood as an instance of one of the 4 categories of software listed above. Another important feature of services to be considered in the context of ARIADNE is how they can be accessed. From our preliminary investigations, the following types can be distinguished:

- Services to be used locally, which require installation on a specific hardware/software platform;
- Services to be used locally independent from any specific hardware/software platform;
- Services to be used on a website (web applications);
- Web services based on a standard protocol.

For the first three categories, it is important to know whether they provide Application Programming Interfaces (API) and whether they are Open Source.

A third feature, relevant to the service description in the ARIADNE context is the kind of functionality offered by the services (e.g. map viewer, data entry system, etc.)

We did not find a shared ontology to express the characteristics of services as discussed above. The best approximation that we found is the ontology for describing software adopted in DBpedia (<http://dbpedia.org/ontology/Software>), so we defined the ARIADNEService class as a specialization of the dbpedia.org/ontology/Software class.

The main associations having this class as domain are:

applyTo: the DataResource to which the service can be applied.

isInRepository: if the source code is available in a repository URI and other information like credential to access the repository are supplied.

hasAttachedDocuments: the documents that are attached to a service for illustration purposes.

hasTechnicalSupport: the person responsible for the technical support

hasAPI: if the service provide an API, a description must be supplied.

hasComponents: a service may include some other components (e.g. <http://dbpedia.org/page/Lucene>).

2. Other Main Classes

2.1 DataFormat

An instance of this class describes the structure of a Dataset, or of a MetadataSchema, or of metadata record. The XSD document describing the format may be stored in the Catalogue as value of an apposite attribute of this class.

The main association having the class as domain are:

usesVocabulary: associates a data format with the vocabularies that are used in the instances of the data format.

hasAttachedObject: associates a data format with the types of digital objects that are attached to its

instances. This association applies to datasets with a complex record structure.

hasSimpleDigitalType: associates a data format with the type of the object that make up an instance of the format, in case the data format is simple media object (text, image, video, audio, and so on).

expressedIn: a data format may be expressed in one or more encoding languages and one language can encode one or more DataFormats. For instance a Dublin Core metadata schema may be encoded in XML or in XML:RDF.

2.2 DBSchema

An instance of this class describes the structure of a Database.

The main attributes of this class are:

type: describes if the database is simple or hybrid

description: describes in an informal way the entities (table) and the associations of the Database.

2.3 MetadataSchema

This is a subclass of LanguageResource having as instances metadata schemas used in the archaeological domain. This is the main class in the metadata registry of ARIADNE, which is a part of the ARIADNE Catalogue.

The main attributes of this class are:

standardUsed: the standard that the schema is based on, if any

dct:description: a description of the format, recommended for proprietary formats

foaf:homepage: an HTTP URI pointing to the web page describing the schema

The main association having this class as domain are

isRealizedBy associates a metadata schema with a data format that realizes it in some specific encoding language.

hasElements: associates a metadata schema with its elements.

hasVersion: associates a metadata schema with a version of its.

usedby: associates a metadata schema with the organizations using it.

2.4 MetadataRecord

An instance of this class is a metadata record, typically associated with a data resource. Since the Catalogue only stores information about data containers, the metadata records in this class will be collection- or dataset-level metadata records. These are stored in the Catalogue in order to support discovery of similar resources that are natural candidates for integration. An XML version of the metadata record can be stored in the catalogue as value of an apposite attribute of this class. The main associations having this class as domain are:

- conformsTo: associates a Metadata Record with the (only) DataFormat to which the record conforms,
- usesVocabulary: associates a Metadata Record with a vocabulary used in the MetadataRecord.

2.5 Distribution

This class represents an accessible form of an ARIADNE resource as for example a downloadable file, an RSS feed or a web service. It extends the class dcat:Distribution by adding three attributes: the number of records in the distribution, the URI of the PMH Server, and a textual description of the platform supporting the distribution. The main associations having this class as domain are:

- hasLicense: associates a distribution with the class of licenses.
- dct:publisher: associates a distribution with an agent responsible for making the resource available.

2.6 TemporalRegion

An instance of this class is a temporal region of one of two forms: a temporal interval (e.g., from 155 bC. to 243 aC.) or a named period (e.g., neolithic). In the former case, the extremes of the interval are given as values of the from and to attributes. In the latter case, the named period is given as value of the periodName attribute.

2.7 SpatialRegion

An instance of this class is a spatial region of one of four forms:

- a region identified by latitude and longitude expressed via two apposite attributes, respectively;
- a bounding box identified by four vertices (expressed via four apposite attributes);
- a postal address (expressed via the :address, :numberInRoad, :postcode and :country

- attributes);
- a named place (expressed via the :placeLabel attribute).

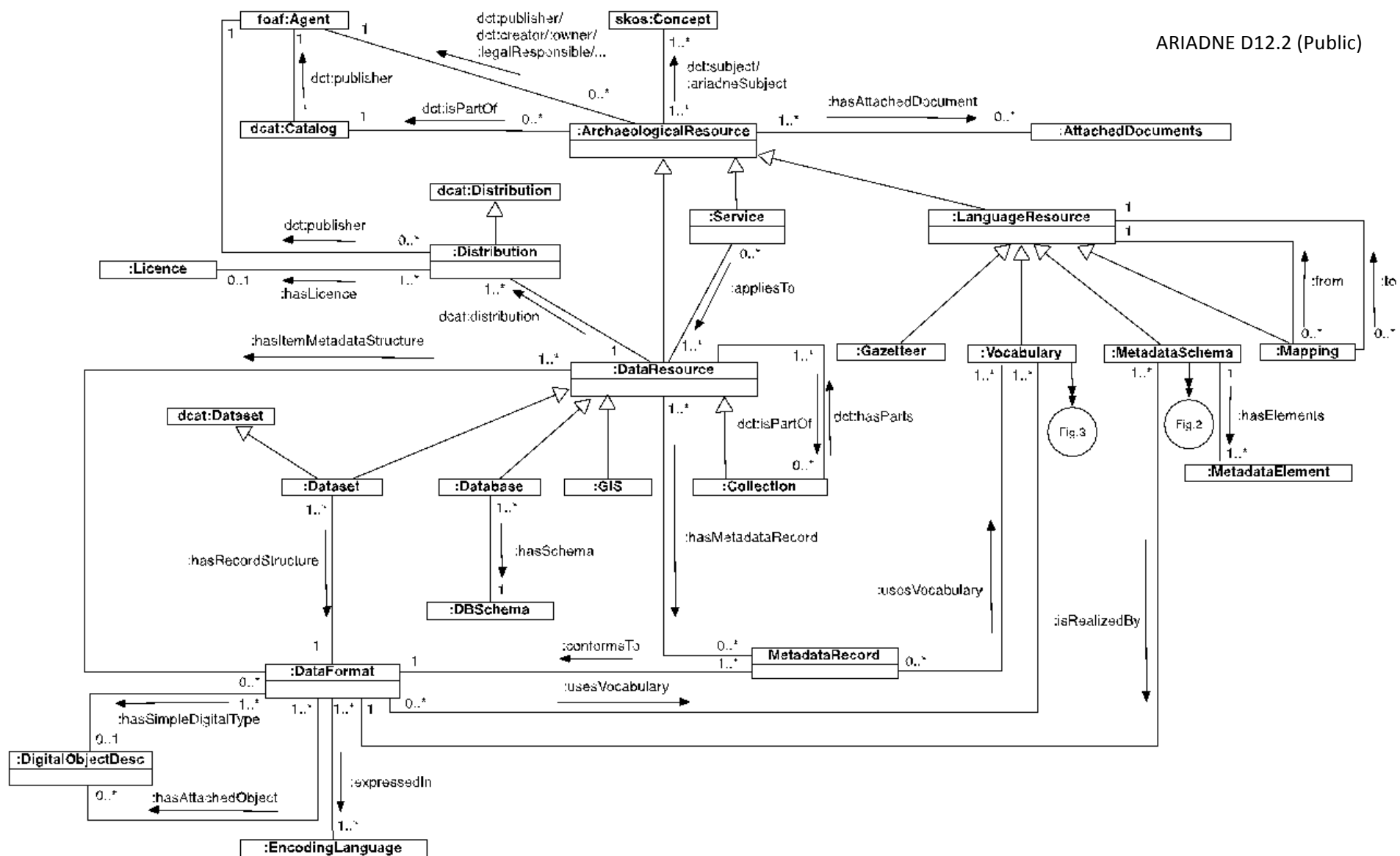


Figure 7. The Ariadne Catalogue Data Model.