D12.1: User Requirements

Author:

H.Wright, ADS

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

Version 1.0 - 8 July 2014

Authors	H. Wright (ADS)		
Version, last changes made, date	0.5	H. Wright, ADS	4-6-2014
	0.6	K. Fernie, PIN	12-6-2014
	0.7	H. Wright, ADS	25-6-2014
	0.8	H. Wright, ADS	27-6-2014
	0.9	K. Fernie, PIN, G. Geser, SRFG	30-6-2014

About this document

This document is a contractual deliverable of the ARIADNE research project (D12.1). 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.

Partner in charge of the deliverable:	ADS
Quality review:	PIN
Authors:	Holly Wright, ADS Julian Richards, ADS Kate Fernie, PIN Hannes Selhofer, SRFG Guntram Geser, SRFG Franco Niccolucci, PIN Carlo Meghini, CNR Paola Ronzino, PIN Christos Papatheodorou, ATHENA RC Hella Hollander, KNAW-DANS
Contributors:	Ruth Beusing, DAI Elizabeth Fentress, AIAC Costis Dallas, Athena-RC Cesar Gonzalez-Perez, CSIC Roberto Scopigno, Paolo Cignoni, ISTI-CNR Ulf Jakobsson, SND Emmanuelle Bryas, Amala Marx, Kai Salas-Rossenbach, Bernard Pinglier, INRAP with additional contributions from all partners

ARIADNE is a project funded by the European Commission under the Community's Seventh Framework Programme, contract no. FP7-INFRASTRUCTURES-2012-1-313193. 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

Executive summary

This document is a deliverable (D12.1) of the ARIADNE project ("Advanced Research Infrastructure for Archaeological Dataset Networking in Europe"), which is funded under the European Community's Seventh Framework Programme. D12.1 is associated with Task 12.1 within WP12, which is titled *Implementing Interoperability*, and falls within the larger ARIADNE conceptual framework for the ARIADNE e-Infrastructure.

WP12 is focussed on the design and implementation of the ARIADNE portal; the means through which the data standards and metadata gathered within the ARIADNE registry (WP3) will be integrated and accessed. As such, the purpose of Task 12.1 of WP12 has been to pull together the considerable information already gathered within several other ARIADNE tasks and their resulting deliverables, to understand the nature of the infrastructures provided for integration, including what data and metadata will be available within the registry, make it possible to identify gaps that may be present, and how they may be adapted for integration. The deliverables include:

- D3.1: *Initial Report on Standards and on the Registry*
- D3.2: *Report on Project Standards*
- D3.3. *Report on Data Sharing Policies*

This understanding was also informed by the recently completed D2.1 *First report on users' needs.* Task 12.1 is titled *Use Requirements*, and refers to the requirements for the design and specification of the subsequent tools and services necessary for integration (as opposed to User Requirements, which refers to the users within the research domain; the subject of D2.1). The design of the tools and services will be carried out within Task 12.2 *Design and Specifications* and Task 13.1 *Service Design and Specifications*. Task 12.2 includes assessment of the features of the contributed metadata itself, and the resulting needs in terms of interoperability resources and interfaces. Task 12.3 is the implementation, and Task 12.4 provides testing. In parallel to and informed by WP12, WP13 will design and deploy the service components for the integrated infrastructure.

This deliverable was structured so as to produce recommendations for Task 12.2 (and to a lesser degree, Task 13.1) for:

Datasets

- **Site and monuments databases:** Most European countries and/or regions have them, and combining may be useful for cross-border searching and geo-location
- **Intervention activity:** May have multiple activities associated with a geo-locatable site, which may allow linking of various activities to a single site or monument
- **Fieldwork databases:** Usually too diverse, so individual databases may not be useful for integration, but may be worth linking to intervention activities for bibliographic discovery
- Other categories are quite specific, but may be useful for integration:
  - **Scientific Databases**
  - **Artefact Databases**
  - **Burial Databases**

Balance data quality and quantity: specify requirements that datasets have to meet in order to be integrated, preferably using formal criteria.

The relationships between the types of data available from the content providing partners and the recommended integration activity to be designed within D12.2 are set out in the table below.

DATA **Balance data quality and quantity**	**ARIADNE datasets**					
Integration activity	**Sites and monuments databases**	**Intervention databases**	**Fieldwork databases**	**Artefacts**	**Burials**	**Scientific datasets**
Cross-border subject search	**X**	**X**	**X**			**?**
Cross-border period search	**X**	**X**	**X**			**?**
Map driven searching or visualisation	**X**	**X**	**X**		**?**	**?**
Bibliographic metadata from grey literature	**X**	**X**	**X**	**X**	**X**	**X**
Integration and interoperability from scientific databases						**X**
Integration of particular kinds of artefact data				**X**	**X**	
Dataset assessment required	**+**	**+**	**+**	**+**	**+**	**+**

Table showing types of data available from the ARIADNE content providing partners, categorised by the type of integration recommended for implementation within the ARIADNE infrastructure. The question mark "?" denotes cases in which the feasibility of the integration activity must be established case by case according to content type.

Metadata Standards, Schemas and Vocabularies

- The **use of international standards for the documentation of excavations and monuments** so as to render it transparent and comparable
- **Free access to tools,** particularly for data mapping, to make it easy to comply with these standards, and offering the means and guidance to archaeologists to deposit their digital records
- **The sustainability of digital datasets** must also be high on the agenda

The relationships between the wishes and concerns with regard to metadata and the recommended tools to be designated or designed within D12.2 are set out in the table below.

	Metadata schemas	**Vocabularies**		**Metadata mapping tools**	**Metadata input tool**	**Metadata description tool**	**SKOSifier tool**
Wishes							
Data transparency	+						
Data accessibility	++	+					
Metadata quality	+++	+++					
Data quality							
International dimension	++	+++					
Concerns							
Metadata quality (managers)						X	X
Effort for metadata creation (researchers)					X		
Anxiety about unfamiliar schemas (researchers)				X			

Table showing the wishes and concerns with regard to data standards, categorised by the type of schema or vocabulary which may address the wishes, and the tools which may address the concerns. The + signifies the level of impact.

Access and Sharing Policies

- **A common method of data citation should be established** for adoption by partners, and promoted by ARIADNE to the archaeological research community. Academic recognition is an important motivation for encouraging researchers to share access to their datasets
- **Allocation of DOIs or the equivalent to datasets ingested to the ARIADNE infrastructure** should be investigated. The system used should be capable of identifying sub-sets within collections. Persistent identification of datasets is important in underpinning data sharing and data citation
- **Content itself (databases, document archives, images, 3D models, etc.) should be provided to ARIADNE by content partners using the Creative Commons license suite** (version 4.0 is preferred) under license permissions agreed with the content owner. CC BY is recommended for open access. CC BY SA or CC BY SA NC licenses may also be applicable
- **A Collection description** (of the whole collection and sub-sets within the collection) should be published under a CC BY license for each dataset ingested to the ARIADNE infrastructure
- **Metadata records should be published under a CC0 license** – to enable integration of multiple datasets within the metadata repository, support resource discovery and enable linked open data

The creation of the ARIADNE infrastructure requires a wide variety of information, both to inform the design of the portal, and to understand the current situation with regard to archaeological data within the domain. To ensure the infrastructure is relevant, useful and represents a positive step towards meeting the needs of researchers, considerable work has been undertaken by all partners within the ARIADNE project to understand the wishes and expectations of our potential users, and gain an understanding of the current technologies, data structures and policies in use within the domain. The results of this work are spread across several deliverable reports produced by the partners, and have been synthesised in this report to inform the development of the ARIADNE infrastructure, Task 12.2 in particular.

1 Introduction

This document is a deliverable (D12.1) of the ARIADNE project (“Advanced Research Infrastructure for Archaeological Dataset Networking in Europe”), which is funded under the European Community's Seventh Framework Programme. D12.1 is associated with Task 12.1 within WP12, which is titled *Implementing Interoperability*, and falls within the larger ARIADNE conceptual framework for the ARIADNE e-Infrastructure.

1.1 The ARIADNE Conceptual Framework

The figure below shows the conceptual framework for the ARIADNE e-Infrastructure:

Level 1: Data created by research projects and groups

Level 2: The institutional archives and repositories where data may be stored initially

Level 3: The higher level data centres and repositories where data may be deposited for long-term preservation and access

Level 4: The ARIADNE e-infrastructure and integrated services, including the ARIADNE registry, portal and additional services

This deliverable concerns the area between Levels 3 and 4, *Data and Metadata Integration.*

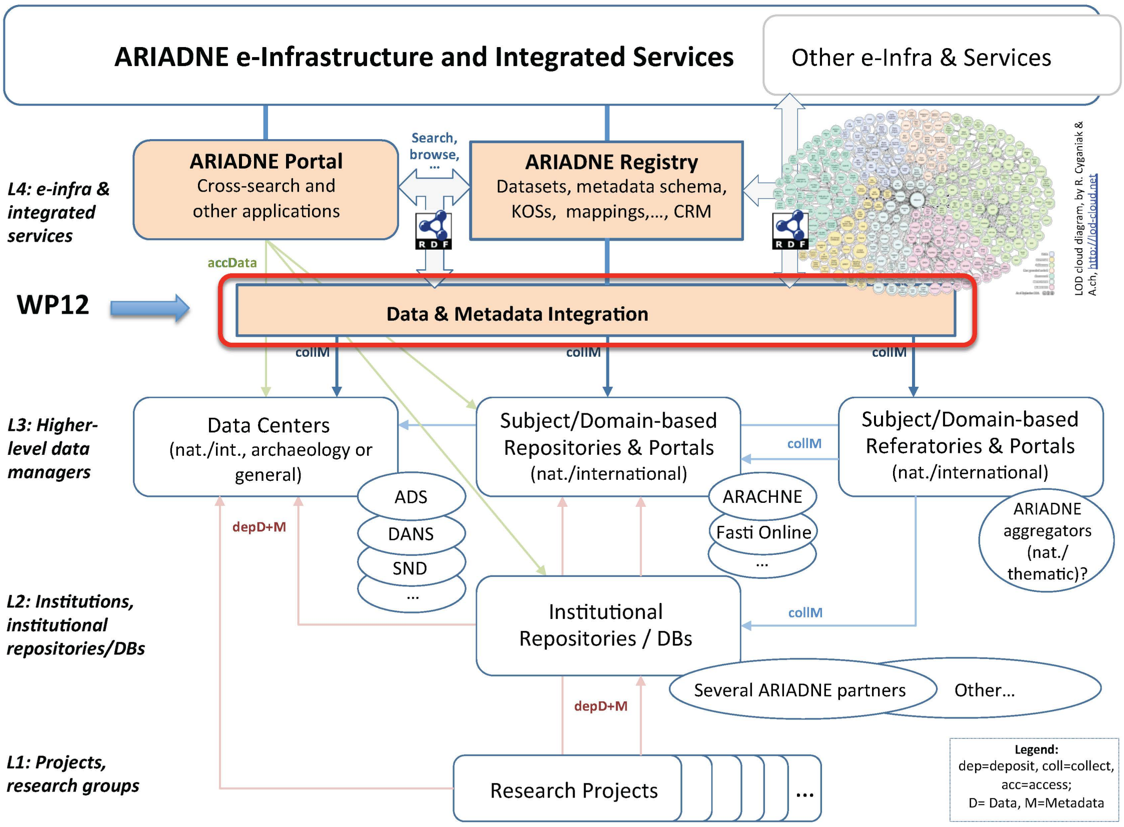

ARIADNE e-Infrastructure and Integrated Services
Other e-Infra & Services
L4: e-infra & integrated services
ARIADNE Portal
Cross-search and other applications
Search, browse, …
ARIADNE Registry
Datasets, metadata schema, KOSs, mappings,…, CRM
RDF
RDF
LOD cloud diagram, by R. Cyganiak & A.ch, http://lod-cloud.net
accData
WP12
Data & Metadata Integration
collM
collM
collM
L3: Higher-level data managers
Data Centers
(nat./int., archaeology or general)
ADS
DANS
SND
…
Subject/Domain-based Repositories & Portals
(nat./international)
ARACHNE
Fasti Online
…
collM
Subject/Domain-based Referatories & Portals
(nat./international)
ARIADNE aggregators (nat./ thematic)?
depD+M
L2: Institutions, institutional repositories/DBs
Institutional Repositories / DBs
collM
Several ARIADNE partners
Other…
depD+M
L1: Projects, research groups
Research Projects
…
Legend:
dep=deposit, coll=collect, acc=access;
D= Data, M=Metadata

Figure 1.1: The ARIADNE conceptual framework reproduced from D2.1, showing the four level scheme, with the addition of the role of WP12 falling between Level 3 and Level 4, as outlined by Achille Felicetti.

As stated in D2.1 [1] ARIADNE will not replace existing infrastructures, services or tools on Levels 1–3, but provide additional integration and services to make data sources and services more accessible for research.

The services or tools will include:

- Metadata held by repositories will be integrated into the ARIADNE Registry, either through direct upload, OAI-PMH targets, SPARQL endpoints, etc. which will allow resource discovery and access using controlled vocabularies, exchange formats, licenses etc.
- The ARIADNE Registry will also be a registry for metadata schemas and KOS, including metadata cross-walks, KOS mappings, etc.
- ARIADNE Portal with various search and browse features for researchers
- Additional services for aggregators and service providers (such as metadata deduplication, indexing, annotation)

Co-operation with developers and providers of other e-infrastructures and services were also deemed important in D2.1, to increase the potential for interoperability.

The ARIADNE Registry

At the heart of the ARIADNE e-infrastructure is the ARIADNE Registry. The Registry will house the metadata, schemas and vocabularies described in this deliverable, and partners are now using it to upload their data and metadata. It is the component upon which the ARIADNE Portal and associated services will be built, and is therefore introduced briefly here. The conceptual model used by the Registry is the ARIADNE Conceptual Data Model (ACDM). As stated in D3.1:

> …the ACDM is an extension of the Data Catalog Vocabulary (DCAT), a quasi-recommendation of the W3C Consortium [2] that *"is well-suited to representing government data catalogs such as Data.gov and data.gov.uk."* The reason for adopting the DCAT Vocabulary (apart from re-use) is that DCAT is proposed as a tool for publishing datasets as Open Data, therefore its adoption places ARIADNE in an ideal position for publishing datasets as Open Data as well.' To this end, ARIADNE will be following the recommendations made in the "DCAT Application Profile for data portals in Europe" using the DCAT ontology.'

As set out in D3.1, these are categorised within services, language resources, data resources and metadata schemas. As metadata is foundational to the ARIADNE registry, metadata is further categorized at collection or dataset level, record level and/or record level metadata that is part of a collection or dataset.

1.2 The Role of WP12 and Task 12.1 within ARIADNE

The overall objectives of WP12 are:

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

In short, WP12 is focussed on the design and implementation of the ARIADNE portal; the means through which the data standards and metadata gathered within the ARIADNE registry (WP3) will be integrated and accessed. As such, the purpose of Task 12.1 of WP12 has been to pull together the considerable information already gathered within several other ARIADNE tasks and their resulting deliverables, to understand the nature of the infrastructures provided for integration, including what data and metadata will be available within the registry, make it possible to identify what gaps may be present, and how they may be adapted for integration. The deliverables include:

- D3.1: *Initial Report on Standards and on the Registry*
- D3.2: *Report on Project Standards*
- D3.3. *Report on Data Sharing Policies*

This understanding will also be informed by the recently completed D2.1 *First report on users' needs.* Relevant sections from D2.1 and D3.1-3 are summarized or quoted in this deliverable for the convenience of the reader.

Task 12.1 is titled *Use Requirements*, and refers to the requirements for the design and specification of the subsequent tools and services necessary for integration (as opposed to User Requirements, which refers to the users within the research domain; the subject of D2.1). The design of the tools and services will be carried out within Task 12.2 *Design and Specifications* and Task 13.1 *Service Design and Specifications*. Task 12.2 includes assessment of the features of the contributed metadata itself, and the resulting needs in terms of interoperability resources and interfaces. Task 12.3 is the implementation, and Task 12.4 provides testing.

In parallel to and informed by WP12, WP13 will design and deploy the service components for the integrated infrastructure. The interactions between the two work packages and relevant tasks from other work packages can be seen in Figure 1.2 from D13.1 [3].

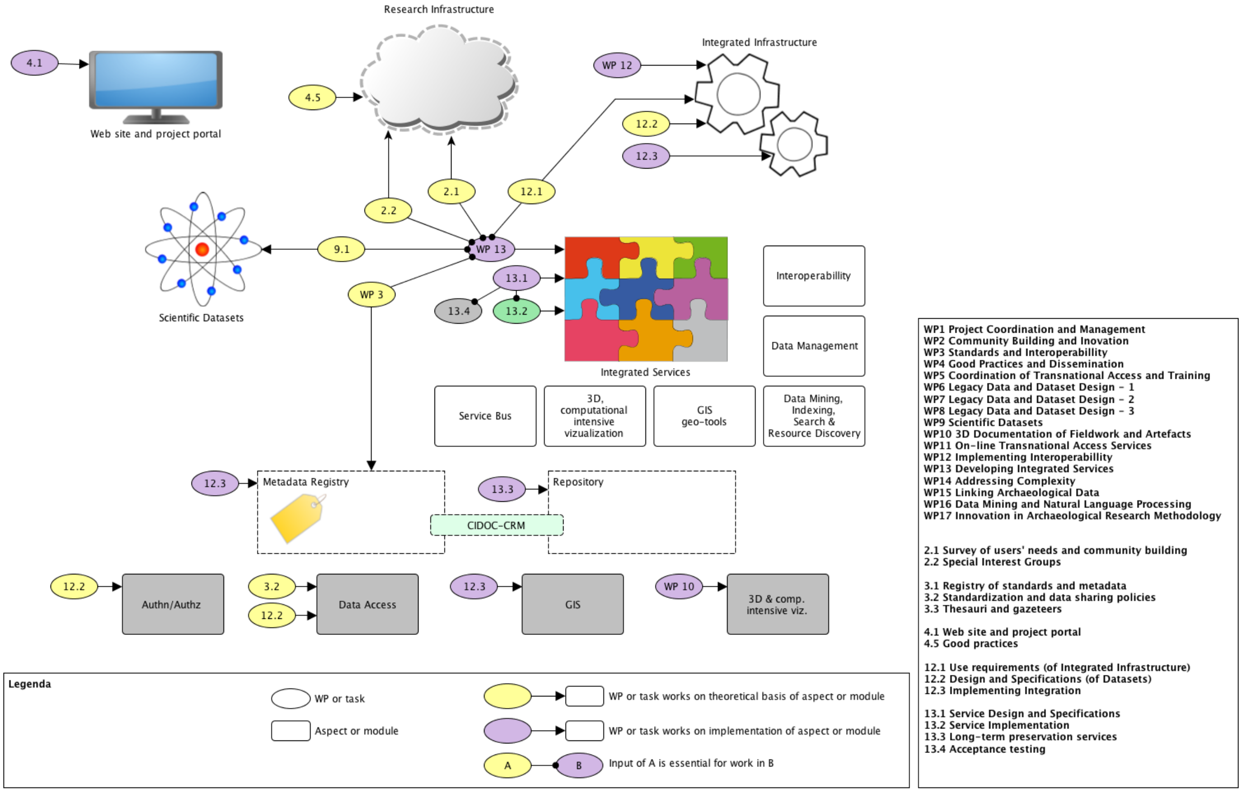

Research Infrastructure
4.1
Web site and project portal
4.5
Integrated Infrastructure
WP 12
12.2
12.3
2.2
2.1
12.1
9.1
WP 13
Scientific Datasets
WP 3
13.1
13.4
13.2
Integrated Services
Interoperabillity
Data Management
Service Bus
3D, computational intensive vizualization
GIS geo-tools
Data Mining, Indexing, Search & Resource Discovery
12.3
Metadata Registry
13.3
Repository
CIDOC-CRM
12.2
Authn/Authz
3.2
12.2
Data Access
12.3
GIS
WP 10
3D & comp. intensive viz.
Legenda
WP or task
Aspect or module
WP or task works on theoretical basis of aspect or module
WP or task works on implementation of aspect or module
A
B
Input of A is essential for work in B
WP1 Project Coordination and Management
WP2 Community Building and Inovation
WP3 Standards and Interoperabillity
WP4 Good Practices and Dissemination
WP5 Coordination of Transnational Access and Training
WP6 Legacy Data and Dataset Design – 1
WP7 Legacy Data and Dataset Design – 2
WP8 Legacy Data and Dataset Design – 3
WP9 Scientific Datasets
WP10 3D Documentation of Fieldwork and Artefacts
WP11 On-line Transnational Access Services
WP12 Implementing Interoperabillity
WP13 Developing Integrated Services
WP14 Addressing Complexity
WP15 Linking Archaeological Data
WP16 Data Mining and Natural Language Processing
WP17 Innovation in Archaeological Research Methodology
2.1 Survey of users' needs and community building
2.2 Special Interest Groups
3.1 Registry of standards and metadata
3.2 Standardization and data sharing policies
3.3 Thesauri and gazeteers
4.1 Web site and project portal
4.5 Good practices
12.1 Use requirements (of Integrated Infrastructure)
12.2 Design and Specifications (of Datasets)
12.3 Implementing Integration
13.1 Service Design and Specifications
13.2 Service Implementation
13.3 Long-term preservation services
13.4 Acceptance testing

Figure 1.2: Relationship of Task 12.1 within its related WPs and Tasks. Reproduced from D13.1.

1.3 Structure of D12.1

The main objective of D12.1 is to understand the current landscape from which the ARIADNE infrastructure can be created, in order to inform the development of the ARIADNE portal and services. This landscape includes the data, metadata, ontologies and vocabularies available for use, along with any associated issues of licensing and access, informed by users' needs.

D12.1 is structured to:

- Collate the information gathered from WP3 into the following areas:
    - Data
    - Metadata Standards, Schemas and Vocabularies
    - Access and Sharing Policies
- Investigate the collated information using the following criteria (as appropriate):
    - Resources and/or Policies in use and/or available
    - User Requirements as set out by D2.1
    - Issues for Consideration
    - Recommendations

Through the structure listed above, D12.1 attempts to synthesise information gathered about the domain across multiple deliverables listed in the previous section. These deliverables represent a comprehensive (though of course incomplete) picture of the way archaeological data is created and used across Europe, and how European practice fits within international practice. They further attempt to gain understanding of this information, incorporating user expectations, exploring the potential issues that are now known to exist, and creating recommendations for implementation of the ARIADNE infrastructure.

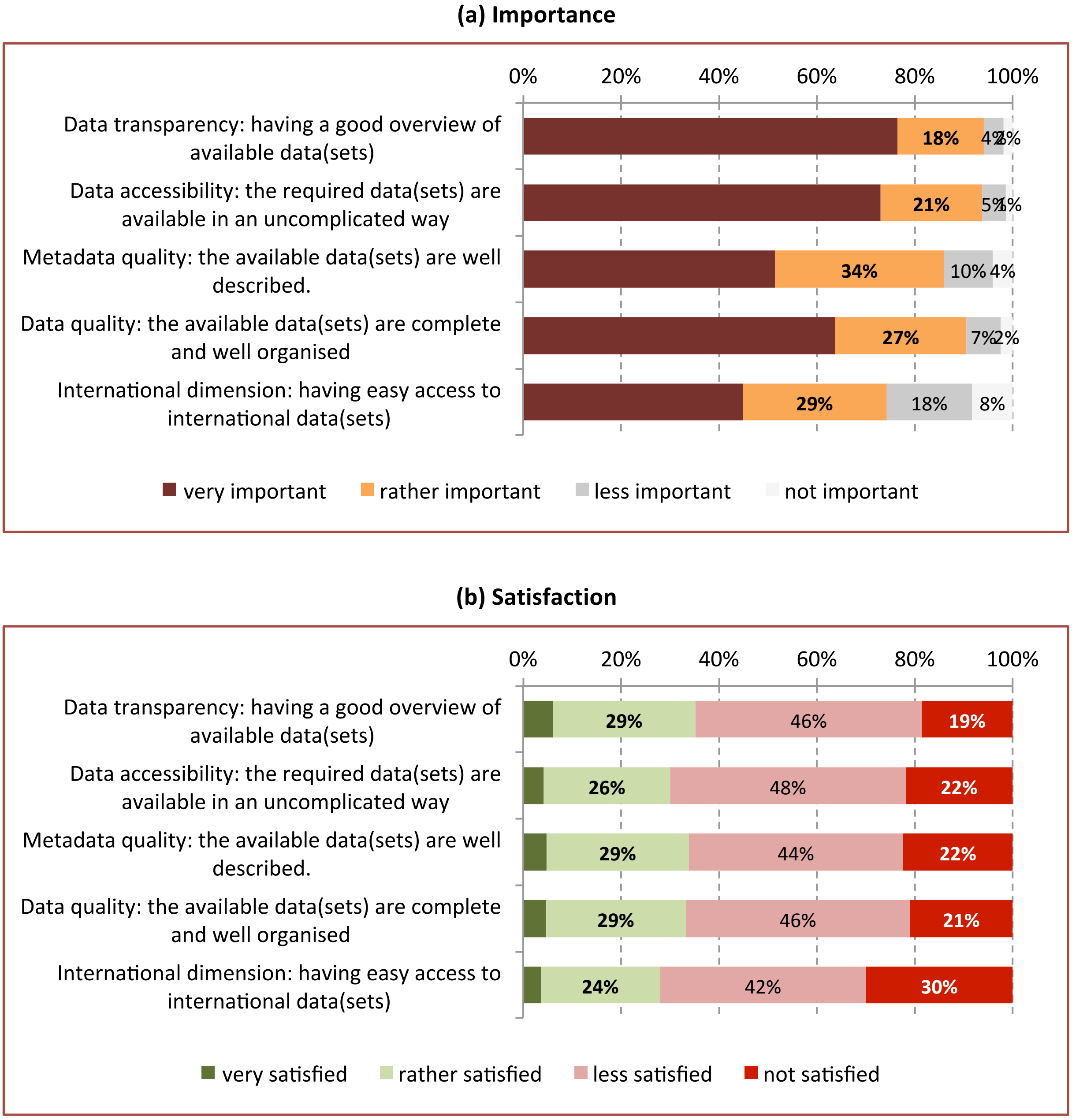
1.4 The Context of Users' Needs

In the recently completed *D2.1 First report on users' needs,* five areas of importance for archaeological research data were defined, along with corresponding levels of satisfaction for archaeological researchers with regard to these areas.

The five areas are:

- Data transparency needs: having a good overview of available data(sets)
- Data accessibility needs: the required data(sets) are available in an uncomplicated way
- Metadata quality needs: the available data(sets) are well described
- Data quality needs in general: the available data(sets) are complete and well organised
- The need for an international dimension: having easy access to international data(sets)

The graphs from D2.1 are shown below by percentage (Figure 1.3), first by level of importance, and then by level of satisfaction, followed by the figures based on the actual numbers of respondents (Figure 1.4).



(a) Importance

(b) Satisfaction

Figure 1.3: From question D.1 in D2.1 – "Please say how important the following aspects are for you in order to conduct your research, and how satisfied you are with the current situation in this regard."

N = 502-506 per item (depending on number of respondents without answer)

		++	+	-	--	
	D.1 – IMPORTANCE	very important	rather important	less important	not important	N
a	Data transparency: having a good overview of available data(sets)	381	88	20	9	497
b	Data accessibility: the required data(sets) are available in an uncomplicated way	365	104	24	7	499
c	Metadata quality: the available data(sets) are well described.	256	171	50	20	497
d	Data quality: the available data(sets) are complete and well organised	317	133	35	12	497
e	International dimension: having easy access to international data(sets)	223	146	87	41	498
	D.1 – SATISFACTION	very satisfied	rather satisfied	less satisfied	not satisfied	N
a	Data transparency: having a good overview of available data(sets)	30	145	230	92	497
b	Data accessibility: the required data(sets) are available in an uncomplicated way	21	129	240	109	499
c	Metadata quality: the available data(sets) are well described.	24	144	218	111	497
d	Data quality: the available data(sets) are complete and well organised	23	142	228	104	497
e	International dimension: having easy access to international data(sets)	18	121	210	149	498

Figure 1.4: From question D.1 in D2.1 – "Please say how important the following aspects are for you in order to conduct your research, and how satisfied you are with the current situation in this regard." Answers by actual number of respondents.

The results of question D.1 are interpreted in D2.1 as follows:

- All the items are highly relevant for researchers and they see important gaps between what they would ideally expect with regard to these aspects, and the actual situation.
- Two requirements in particular were attributed the highest importance: data transparency and data accessibility. In both cases, about three quarters said this was very important.

As the core work of ARIADNE will be to address data transparency and accessibility through the use of the Registry, Portal and additional services, this confirms both the relevance of the project rationale, and strong need to address the dissatisfaction within the domain.

The authors of D2.1 caution however, due to the data being highly aggregated, specific issues cannot be addressed. While it shows the broad range of opportunity with which the ARIADNE project can move forward with now, these areas will be broken down and explored in more detail within D2.2 *Second report on users' needs*.

To align the users' needs to what is currently available within the ARIADNE project and within the domain; the five user requirements have been incorporated into this report in the following structure:

- **Section 3: Data**
    - Data quality: the available data(sets) are complete and well organised
- **Section 4: Metadata**
    - Metadata quality: the available data(sets) are well described
- **Section 5: Access**
    - Data transparency: having a good overview of available data(sets)
    - Data accessibility: the required data(sets) are available in an uncomplicated way
    - International dimension: having easy access to international data(sets)

1.5 Summary

In order to understand the data landscape of the archaeological domain, the nature of the data and metadata to be collected within the ARIADNE Registry using the ACDM, and inform the relevant tasks within WP12 and WP13, Task 12.1 *Use Requirements* will explore the methodology and results of the relevant use requirements studies and their derived recommendations. Each area will investigate the collated information using the following criteria:

- Collate the information gathered from WP3 into the following areas:
    - Data
    - Metadata Standards, Schemas and Vocabularies
    - Access and Sharing Policies
- Investigate the collated information using the following criteria (as appropriate):
    - Resources and/or Policies in use and/or available
    - User Requirements as set out by D2.1
    - Issues for Consideration
    - Recommendations

The structure listed above will attempt to bring together the considerable information gathered about the domain across multiple deliverables, aligned with the users' needs survey, through which the ARIADNE partners can gain understanding of this information, incorporating user expectations, exploring the potential issues which are now known to exist, and creating recommendations for implementation of the ARIADNE infrastructure. Specifically this deliverable will inform Task 12.1 *Design and Specifications*, and Task 13.1 *Service Design and Specifications*.

2 Data

2.1 Background

Objective

To **assess the information collated about the datasets** held by the consortium with regard to users' needs and other issues. This assessment should result in recommendations for Task 12.2 *Design and Specifications* and Task 13.1 *Service Design and Specifications*.

The ARIADNE content providing partners gave detailed reporting on the datasets they hold and wish to include within the infrastructure, and this was reported within D3.2 *Report on Project Standards*. The metadata from partners' datasets will be recorded in the ARIADNE Registry with complete information on metadata structure, and size of the dataset. The survey revealed that the data held by the partners is highly diverse and includes databases, text, multimedia data in 2D and 3D formats; both raster and vector, and combinations of all of these held within collections. The survey also gathered the standards adopted by the project partners.

Data Collection Strategy

A questionnaire was sent to all 16 ARIADNE content providing partners (the details of which can be found in Appendix F in D3.2 [4]). ARIADNE partner CSIC has also subsequently chosen to become a content provider, and will be offering both site data and grey literature (as such, they are included in the following summary table for completeness, but their data was not part of the original survey).

All partners replied in detail, so the survey was a comprehensive snapshot of the datasets planned for integration taken at the date of compilation of D3.2, and provides the information necessary for proceeding with integration and services design. As more data (and therefore metadata) becomes available throughout the project, so will the Registry be updated. The questions were also formed to see how the Registry might be extended to include datasets outside of the project consortium, so other institutions who are not partners within the project might participate if they wished to. The summary table from D3.2 showing the data each partner plans to provide is included in the next section for the convenience of the reader, and the individual detail can be found in Section 4 of D3.2.

2.2 Datasets held by the consortium

The survey revealed the consortium holds a wide range of different types of datasets including:

- Archaeological databases relating to sites, settlements, burials, finds, objects specific to particular regions and time periods, and specialist artefacts
- Ethno-archaeological datasets
- Archaeological science databases including data relating to wood, charcoal, palaeo-environmental data, c14 and isotope data, and data relating to dendrochronology, zooarchaeology, petrology, geophysics, and ceramic analysis
- Collections of data relating to specific archaeological excavations
- Remote sensing datasets including lidar, aerial photography and geophysics
- Map-based data, including archaeological sites, monuments, field investigations within

regions, sites identified by aerial photography, and GIS data

- Grey literature, the majority of which is derived from contract-based archaeology
- Image datasets, video and audio files, and 3D models in a wide range of formats
- PDF and other text documents, Excel spreadsheets and CSV files

The following table from D3.2 shows a summary of the types and amounts of data held by the content providing partners, and is based on the initial details about the data provided in the survey. The data has been classified under a series of headings developed at the WP3-12-13 partners meeting in Pisa in November 2013, and the types of datasets may overlap and be found in multiple categories. Subsequently, it was decided these headings could also serve as a basic subject/type vocabulary list to be incorporated into the infrastructure. The data reported in the survey should be taken as a snapshot of what the partners were able to offer at the time, but may be expanded as the project proceeds, as with the case of CSIC previously mentioned. In addition, the heading Fieldwork Activity, which has been modified from the original heading Event/Intervention.

How best to incorporate the individual datasets will be explored within D12.2 Design and Specifications, but it is possible to see patterns in the data that show which directions may be most fruitful in terms of interoperability.

Sites and Monuments records

Many content providing partners have these inventories of records pertaining to the location and nature of particular sites. Some pertain to particular time periods (ie medieval) or site types (ie burials), but all are likely to contain similar types of basic metadata that can be made interoperable. A site may have had multiple excavations over many years, resulting in a variety of site reports and related databases.

Intervention Activity

Intervention activity can include excavation, survey, geophysics and a wide variety of other types of fieldwork. The vast majority of archaeological fieldwork is now carried out by contract-based units in advance of development (though whether the units are commercial, national or affiliated with an academic institution can vary), and therefore most fieldwork activity results in an unpublished report of some kind (grey literature). The ever-growing corpus of grey literature represents one of the most important sources of archaeological research data, but also one of the most problematic. Grey literature is often left undigitised, in the hands of local authorities and developers, and can only be accessed by consulting an archive in person. This is starting to change however, and one of the most promising aspects of the ARIADNE infrastructure is that many of the partners now hold digitised grey literature, which can potentially be linked to relevant sites and monuments, and made more accessible online. In addition, considerable work is now being undertaken with WP16 to explore both data mining and Natural Language Processing (NLP) for use with grey literature, to try to unlock much of the hidden potential within these reports.

Fieldwork Databases

While interoperability at record level will likely be beyond the scope of ARIADNE (though some case studies may be possible), fieldwork databases will be important at the collection level.

Scientific Data

The most promising area for interoperability within scientific data is dendrochronology. Several partners hold dendrochronology (or related) data. There is much interest within the consortium to work with dendrochronology data, and as part of WP4, a new Guide to Good Practice for dealing with this type of data is currently being produced.

Artefacts

Some partners hold data about specific types of artefacts, which may also have potential for interoperability. Virtually all types of artefacts from a wide variety of regions and time periods are represented, but there are particular concentrations data about stone tools and ceramics, so they may be the most promising. In addition, the experimentation with NLP has been focussing on finding references to stone tools in multiple languages.

Burials

Burials were singled out as their own data type, due to their specific characteristics. Only two partners reported having burial data, and both relate to very different regions and time periods.

Partner	Country	Sites and monuments	Intervention Activity	Fieldwork databases	Scientific	Artefacts	Burials
ZRC-SAZU	Slovenia	ZBIVA – 3000 early medieval sites in SE Alps					
OEAW	Austria	UK_Material-POOL – 442 LBA settlements		Thunau LBA and early medieval settlement			Franzhausen-Kokoron 3827 LBA graves
DISCOVERY	Ireland	NB Mapping Death – 174 early med cemeteries	SHARE-IT 50 site surveys		WODAN – wood and charcoal database for 533 sites	ISAP – 21000 Irish stone axes	Mapping Death – 174 early medieval cemeteries
ARHEO	Romania	BA and medieval sites survey		Geophysics data		Ceramics data	
INRAP	France		Arheozoom – 1100 excavation metadata records DOLIA – 500 pdf reports IDA – 3380 excavations				

Partner	Country	Sites and monuments	Intervention Activity	Fieldwork databases	Scientific	Artefacts	Burials
ARUP-CAS	Czech Republic		AMCR – all field interventions in Bohemia				
NIAM-BAS	Bulgaria	Archaeological map – 17000 sites					
SND	Sweden		361 reports for Ostergotland	361 excavation archives			
ADS	UK	ArchSearch: 1,300,000 site records for archaeology of UK	25,000 grey literature reports	c.400 excavation archives	Dendrochronology; multiple environmental databases; grey lit reports	CBA Stone axe database for England & Wales; Ceramics	
DANS	Netherlands		17,000 grey literature reports	c.3000 datasets	DCCD – 50,000 records		
MIBACTT-ICCU	Italy	SIGEC SITAR – UA dataset	SITAR – 11000 records	SITAR – PA dataset		CulturaItalia – 52000 objects	
MNM-NOK	Hungary		3000 sites; 1038 grey literature reports	geophysical survey of c. 50 sites	c. 550 reports	Ceramics, metal, glass, textile, human and animal bones, stone, archaeobotanical remains, other organic remains	

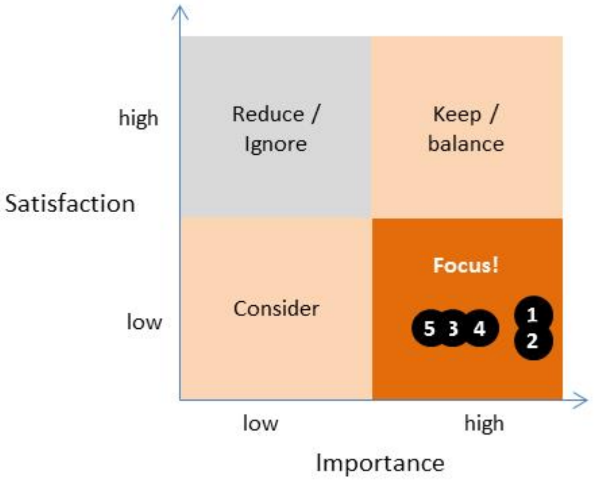
Partner	Country	Sites and monuments	Intervention Activity	Fieldwork databases	Scientific	Artefacts	Burials
ATHENA RTC- CETI	Greece					Ceramics – 1500 sherds	
AIAC	Italy +		5000 excavations				
Cyi-STARC	Cyprus					2000+ images plus 300+3D models	
DAI	Germany					ARACHNE – 200,000+ classical archaeology artefacts	
CSIC	Spain	5000 site records (plus associated finds and features)	Large number of grey literature reports in Spanish and Galician				

2.3 User Requirements

In D2.1: *First report on users' needs,* data quality was ranked as fourth in importance to the groups of archaeological researchers who participated in the user requirements survey, and has therefore been designated a priority area for the ARIADNE project for further strategic work within D2.2 *Second report on users' needs.* As such, it is important to understand the user expectations for the archaeological data slated for provision within the infrastructure.

Archaeological researchers and directors of research institutes were asked about the importance of different types of data, and their specific needs and expectations for ARIADNE with regard to data. The designers of the survey consulted with other members of the consortium to determine how to classify the data, and particular types of data emerged as being clearly more important than others.

User requirements of archaeological researchers according to their importance and the satisfaction with the existing situation in a strategy matrix



User needs as shown in the matrix:

(1) = Data transparency

(2) = Data accessibility

(3) = Metadata quality

(4) = Data quality

(5) = International dimension of available data

Importance of different types of data

D2.1 determined that the 'single most important type of data if measured by the number of researchers for whom they are important is excavation data. 75% of the respondents said that excavation data was "very important" for them to carry out their research projects. Also very important for a large group of researchers (about 50% each) was GIS data, data stemming from material or biological analysis, and data from field surveys. This is not to say other types of data are not relevant; quite the contrary, they all have their users; in most cases, at least 50% of the respondents said that they were at least "rather important" (Figures 2.1 and 2.2)'.

The survey not only revealed what types of available data were most important, but also showed that poor availability of some data potentially complicated the result. Some data was seen to have potential importance, but the difficulty in obtaining it meant it wasn't considered in the same way it might have been otherwise. This indicates that access is a major issue, which will be discussed further in Section 5.

The figures below show the full breakdown of the categories of data, and their level of importance to researchers. Fortunately, the ARIADNE content providing partners hold the types of data deemed most important, including large amounts of excavation data. The consortium must continue to keep the users' priorities in mind during the design of the portal and services, but we can move forward with confidence knowing the necessary data is available.

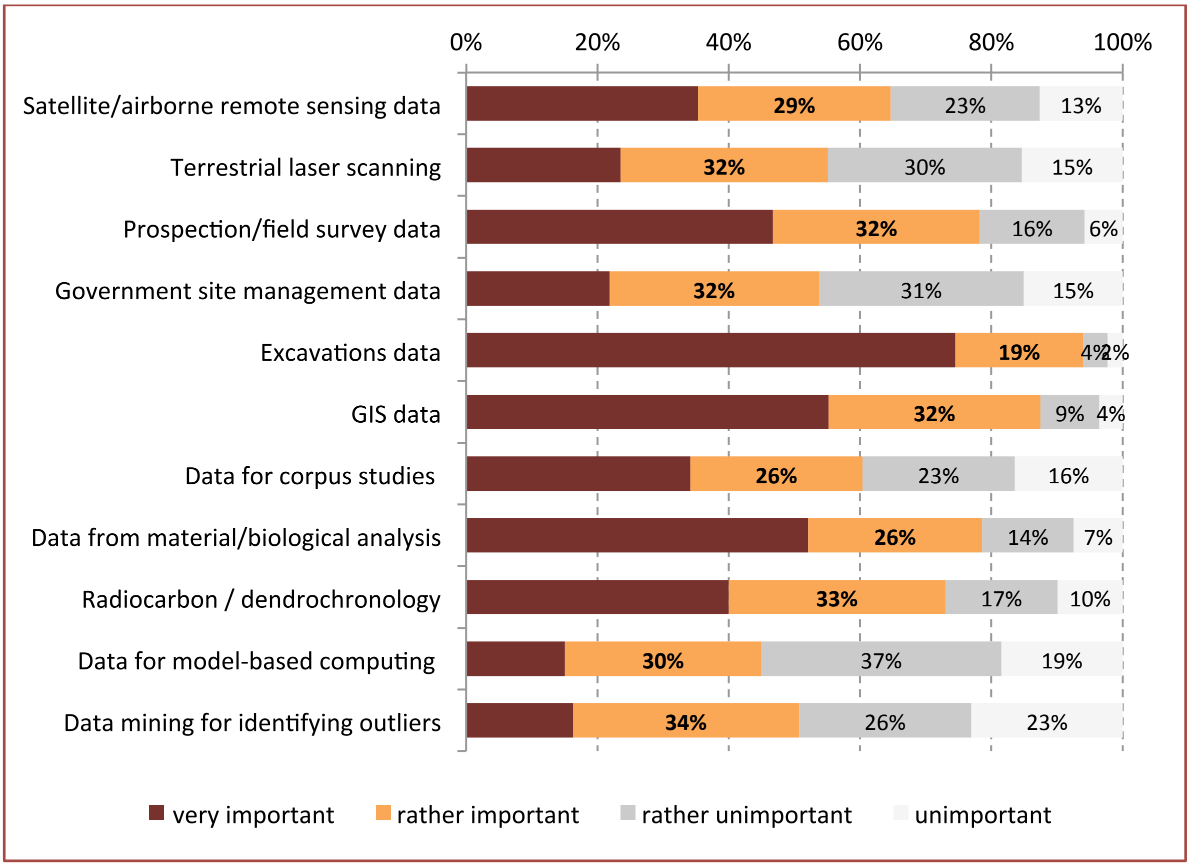


Figure 2.1: Question B.2 reproduced from D2.1 – "How important are the following types of data for you and your research group in preparing and carrying out your projects?"

		++	+	-	--	
	B.2	very important	rather important	rather unimport.	un-important	N
a	Satellite/airborne remote sensing data	209	174	134	75	592
b	Terrestrial laser scanning	138	185	173	90	586
c	Prospection/field survey data	273	184	93	34	584
d	Government site management data	126	184	180	87	577
e	Excavations data	445	116	22	14	597
f	GIS data	326	191	53	21	591
g	Data for corpus studies	199	153	135	96	583
h	Data from material/biological analysis	309	157	83	44	593
i	Radiocarbon / dendrochronology	237	196	101	59	593
j	Data for model-based computing	87	173	211	107	578
k	Data mining for identifying outliers	85	179	137	120	521

N = 521-592 per item (depending on number of respondents without answer)

Figure 2.2: Question B.2 reproduced from D2.1 – "How important are the following types of data for you and your research group in preparing and carrying out your projects?" Answers by actual number of respondents.

2.4 Issues for Consideration

The survey in D3.2 revealed a large quantity of data available for integration, but also the heterogeneous nature of much of this data. Several partners hold large collections of datasets of diverse structure, whilst other partners hold single datasets of identical structure. According to D2.1, **even when data was available online, it still failed to be useful because the data is structured in different ways, not up to date, is incomplete or lacking important details**. As the respondents stated: “the main problem is the variability and inhomogeneity of data content and structure”; “incomplete datasets; online databases that aren’t kept up to date (this is a big problem)”; “lack of details on how data was collected – it is difficult to assess the quality of data published online”. While simply making data available through the ARIADNE infrastructure will be an important first step, it will be equally important to remember it is only a first step. In order for the data to be useful, consideration will have to be made as to how to address the:

- Heterogeneity of structure (including use of differing file types and proprietary software): heterogeneity of the structure of different datasets is not necessarily a problem, what is important is how homogeneous clusters of datasets are (e.g. if the datasets of sites and monuments are homogeneous).
- Heterogeneity of content
- Datasets are invariably incomplete
- Datasets often do not hold the most recent data
- Datasets may be of poor or unknown quality (rarely subject to peer review): ARIADNE will not have control over the quality of the data submitted for inclusion, but can try to raise awareness about good practices for producing good quality data through our work in WP4. It may be useful to design a quality index and attach it to the datasets
- Much archaeological data is ‘legacy data’ which may have been gathered using very different methods, and may lack important contextual information (i.e. when a dataset was created) which should be disclosed in order to the data to be useful.

2.5 Recommendations on data

The following recommendations were reported in D3.2: Report on project standards, also corresponding to the main data subject headings, the recommendations are quoted here apart from the heading Intervention Activity, which has been modified from the original heading Event/Intervention.

Sites and Monuments Databases

Most European countries and regions hold inventories of known archaeological sites and monuments within their area. These are often developed for management purposes, but also provide an invaluable research resource. In addition, several partners are also offering datasets of sites for specific classes or periods. Although not comprehensive they could be usefully searched alongside the management-driven records. The inventories are generally characterised by field for place, period and multiple keywords for site or monument type. Spatial data may be precise enough to support a GIS function. Given that most archaeological research questions are cross-border (e.g. "Where are all the Bronze Age sites in southern Europe?") there is a value-added from combining inventories that ARIADNE is uniquely placed to provide.

Intervention Activity

This category of record is closely related to the previous grouping but provides instead a record of where archaeological fieldwork has been carried out, as opposed to where known sites are. In database terms there will frequently be a one-to-many relationship between monuments and events, as each site will have been subject to multiple investigations over time. It is the events, however, that provide most information about a site or monument. This special category of activity databases held by several partners are effectively bibliographic metadata records for unpublished fieldwork reports, or 'grey literature'. There is a widely recognized problem within Europe for researchers to gain access to this literature, which was highlighted in the description of work, and it has already been identified as a priority for online access and integration at the European level. As the records share several fields with the sites and monuments inventories the network will need to decide if it wishes to combine both categories of data in a single search interface.

Fieldwork Databases

Several partners have complete digital fieldwork archives available, including both excavation and geophysics data. Each archive may be a data collection, incorporating a range of file types with different record structures, and may range beyond simple text files to include GIS, databases, spreadsheets, images etc. In some cases the event databases (above) may provide an index record to the richer data collections. Given each collection will be site specific there will generally be little utility in integrating such datasets, although where WP16 is able to provide Linked Data at item level for finds records this may support research questions (e.g. "Where have Roman coins been deposited in later features?"

The remaining categories reflect a wide variety of quite specific digital resources, but there may be some utility in integration, dependent upon the outcomes of the user needs studies.

Scientific databases may include, for example, environmental or faunal datasets, dating records, or the results of scientific analyses. Where such datasets concern the same analyses or data types they may share a relatively similar structure which will aid integration.

Databases of **Artefacts** provide access to information about specific objects at item level. Several partners have dedicated artefact databases, but they may also be present in fieldwork data collections. Some databases may include images of finds, or thumbnail links to separate image files. If there is a critical mass in specific categories of find, such as stone axes or ceramics for example, there could be considerable value added from providing transnational and cross-border interoperability.

Finally, **Burials** represent a special class of database where the individual grave represents the item level. The attributes recorded may include finds as well as physical anthropological information. Integration may aid studies of population demographics and social analyses. Cemeteries may also appear as single records in sites and monuments and events databases, which could provide a pointer to the richer item level datasets.

According to the recommendations in D3.2, there is potential scope for creating:

- Cross-border searching of sites and monuments records (combining existing national inventories alongside those offered by the partners
- Map driven searching or visualisation where geo-spatial data exists for sites and monuments
- Event-based bibliographic metadata derived from the grey literature held by several partners, which may allow the association of events (i.e. excavations) with particular sites and monuments, making both more useful
- Greater integration and interoperability from scientific databases, which often share more similar structures than other kinds of data
- A higher level of interoperability through integration of particular kinds of artefact data held by partners that may have trans-national or cross-border significance.

An important recommendation was also reported within D2.1: First report on users' needs:

> **Balance data quality and quantity: specify requirements which datasets have to meet in order to be integrated**
>
> Data and metadata quality (completeness, degree of organisation) are key user requirements with regard to digital repositories. The ARIADNE project will have to carefully consider and specify the quality requirements for specific collections or data sets to be integrated in the e-infrastructure, so that the users regard the resulting services as valuable. In other words, the project needs to think about where and how "to draw the line". These criteria may be different for various types of data. In particular, ARIADNE will have to discuss how to deal with "legacy data".
>
> In order to be a truly valuable resource for archaeological researchers, once integrated into the ARIADNE Repository, it is suggested that criteria should be created to assess the datasets. These criteria should try to address the following questions:
>
> - Were the files created using a file type common to other datasets? Can the file types be made interoperable?
> - Is the content of the data common to other datasets? Will attempts to make the data interoperable be meaningful?
> - How complete is the data? What constitutes and acceptable level of completeness for the data to be considered useful?
> - Is the data sufficiently recent? Is it 'legacy data' which lacks contextual information or gathered using outdated methods? Will this affect the data's ability to be interoperable?

In order to meet the user expectations for balancing quality vs. quantity, each of these potential recommendations will have differing answers to the proposed assessment criteria, and will likely need to be addressed separately in order to know where to "draw the line".

To summarise the specific recommendations for ARIADNE with regard to data, the following table shows the areas where data is available, and what kinds of integration activity should be carried out within Task 12.2, taking into consideration the importance of data quality and assessment, and the most important types of data set out in D2.1.

DATA **Balance data quality and quantity**	**ARIADNE datasets**					
Integration activity	**Sites and monuments databases**	**Intervention databases**	**Fieldwork databases**	**Artefacts**	**Burials**	**Scientific datasets**
Cross-border subject search	**X**	**X**	**X**			**?**
Cross-border period search	**X**	**X**	**X**			**?**
Map driven searching or visualisation	**X**	**X**	**X**		**?**	**?**
Bibliographic metadata from grey literature	**X**	**X**	**X**	**X**	**X**	**X**
Integration and interoperability from scientific databases						**X**
Integration of particular kinds of artefact data				**X**	**X**	
Dataset assessment required	**+**	**+**	**+**	**+**	**+**	**+**

Figure 2.3: Table showing types of data available from the ARIADNE content providing partners, categorised by the type of integration activity which is recommended for implementation within the ARIADNE infrastructure. The question mark "?" denotes cases in which the feasibility of the integration activity must be established case by case according to content type.

3 Metadata Standards, Schemas and Vocabularies

3.1 Background

Objective

To assess the metadata standards, schemas and vocabularies currently available and **potentially of use to the archaeological domain,** and the metadata standards, schemas and vocabularies which are **actually in use by the ARIADNE content providing partners**, so as to inform the development of the ARIADNE infrastructure with regard to which standards may be adopted, and at what level interoperability may be possible.

Metadata standards, schemas and vocabularies currently available

The key to interoperability in any domain is a common set of standards that meet the needs of the sector, and the will to adopt them. D3.1 *Initial report on standards and on the project registry* was the synthesis of an extensive survey of the standards currently in use within archaeology, developed by other domains relevant to archaeology, and allowing interoperability with other domains useful to archaeology. The report also described the registry developed within the ARIADNE project, which will incorporate the standards best suited to building the infrastructure.

3.2 Archaeology related resources

D3.1 *Initial report on standards and on the project registry* reported on the metadata standards, schemas and vocabularies currently available and potentially of use to the archaeological domain. The available standards have been extensively described within D3.1, so they are merely listed here in their appropriate sections for the convenience of the reader. The sections include:

Archaeology - oriented metadata standards and conceptual schemas

- Reference models
  - CIDOC CRM
  - CHARM (Cultural Heritage Abstract Reference Model)
- Archaeological sites, monuments, landscape areas
  - International Core Data Standard for Archaeological Sites and Monuments
  - MIDAS Heritage
  - CARARE schema
  - ICCD Cataloguing Standards
  - EU-CHIC CHICEBERG
  - SIKB0102 - Excavation data
  - STARC schema
- Archaeological sciences
  - Tree Ring Data Standard - TRiDaS
- Museum objects
  - SPECTRUM

  - LIDO
  - Object ID
  - VRA Core
  - CDWA
- Dublin Core
  - Dublin Core Application Profiles
  - DCMI Metadata Terms
- Bibliographic materials
  - MARC
  - MODS
  - METS
- Archival standards
  - EAD
- Geospatial information
  - INSPIRE
  - CEN/TC 287 Geographic Information
  - ISO 19115:2003 - Geographic metadata
  - UK GEMINI
- Other standards
  - Data Documentation Initiative (DDI)
  - Digital Object Identifiers (DOI)

Archaeology related terminology resources

- International terminology resources
  - Art & Architecture Thesaurus (AAT)
  - Cultural Objects Name Authority (CONA)
  - Union List of Artist Names (ULAN)
  - European Language Social Science Thesaurus (ELSST)
  - PACTOLS Thesauri
- National terminology resources
  - British Museum Materials Thesaurus - UK
  - INSCRIPTION - England
  - RCAHMS Vocabularies - Scotland
  - RCAHMW Vocabularies - Wales
  - MiBACT - Italy
  - Referentienetwerk Erfgoed (ABR) - Netherlands
  - ZRC SAZU - Slovenia

  - Feldolg-R - Hungary
  - Finnish Ontology Library Service ONKI Vocabularies
  - Museums Vocabularies – Germany
  - Archaeological Dictionary of the DAI – Germany
- Geospatial resources
  - Getty Thesaurus of Geographic Names (TGN)
  - Geonames
  - Pleiades Ancient Places Gazetteer
  - Pleiades Ancient Places Time Periods Vocabulary
  - Place names database of Ireland

The metadata standards and conceptual schemas consist of the reference models (interlinked sets of defined concepts), the most important of which is the CIDOC-CRM and its extensions. The CIDOC-CRM is the cornerstone ontology for modeling data within the cultural heritage domain, and interoperability with the CRM is an important aspect of many of the other standards, but is by no means to only option. In addition to the reference models, the synthesis includes standards for archaeological sites, monuments and landscapes, archaeological sciences, museum objects, the Dublin Core properties for resource description, bibliographic materials, archival standards and geospatial information.

The terminology resources included in the synthesis are divided into national, international and geospatial categories. While the majority are available in English, irrespective of the country of origin, many have translations into a variety of European languages, which will form a vital basis for interoperability, but translation will not be enough. Terms used within the archaeological domain necessarily vary between languages, as archaeological resources vary from place to place. One of the great challenges to interoperability within archaeology is the difference in the development of human culture in different places at different times (so spatial data and temporal data in archaeology often has little meaning unless it can be linked).

3.3 Standards Currently in Use by the ARIADNE Partners

The ARIADNE content providing partners were surveyed about the actual standards they use with their data, and the result is reported in D3.2 *Report on project standards*. Due to the large number of standards currently in use, the ARIADNE Registry has also been designed as a metadata registry to bring these standards together, to explore their potential for interoperability. The registry follows the metadata registry standards, such as ISO 11179 and the framework defined by the DESIRE and ROADS projects. A synthesis of the results of the survey are listed below for the convenience of the reader, and show the wide variety of content being made available, and the equally wide range of approaches the partners are using with regard to the adoption of metadata standards.

Formal metadata standards
Standards in use: DDI, DataCite, MARC/UNIMARC, TriDAS, Dublin Core application profiles, INSPIRE, ISO 11915, CARARE, LIDO, CIDOC-CRM
Partners using formal standards: SND, KNAW-DANS, Discovery, MIBACT-ICCU, INRAP, ADS, CYI-STARC

Proprietary metadata schemas
Partners using proprietary schemas: ZRC SAZU, MIBACT-ICCU, ADS, AIAC, OAEW, MNM-NOK, CYI-STARC, ARUP-CAS, ATHENA RC, NIAM-BAS

Content derived metadata
Partners from whom metadata is not currently available, but could be derived from their content: DISCOVERY, INRAP, ARHEO

Controlled vocabularies (international standards)
Vocabularies in use: European Language Social Science Thesaurus, Irish place names database, Tree of Life project, Geological Survey of Ireland, GEMET, PACTOLS
Partners using international standard controlled vocabularies: SND, Discovery, INRAP

Controlled vocabularies (national standards)
Partners using national standard controlled vocabularies: SND (monument type), KNAW-DANS (archaeology + ABR), Discovery (monument type and artefact classification), MIBACT-ICCU (PICO, SITAR vocabularies)

Controlled vocabularies (proprietary)
Partners using proprietary controlled vocabularies: ZRC-SAZU, ADS, NUAM-BAS, AIC, MNM-NOK and CYI-STARC

Not currently using controlled vocabularies or wordlists
Partners not using controlled vocabularies or wordlists: OAEW, ARHEO and Athena RC-CETI

The adoption of metadata standards across the content providing partners spans the full range of use, and reflects both differing levels of awareness about standards and differing attitudes towards its necessity. Comparing the range of standards in use with the available standards surveyed in D3.1, there is significant overlap, but it illustrates there are many standards of which partners may not be aware that could be of use. At the same time, there were a few standards that are in use by partners that were not included in the D3.1 survey and were a surprise. These included DataCite (assigns persistent identifiers to datasets for archiving and resource discovery), The Tree of Life Web Project (taxonomies for living organisms) and GEMET (multilingual environmental thesaurus).

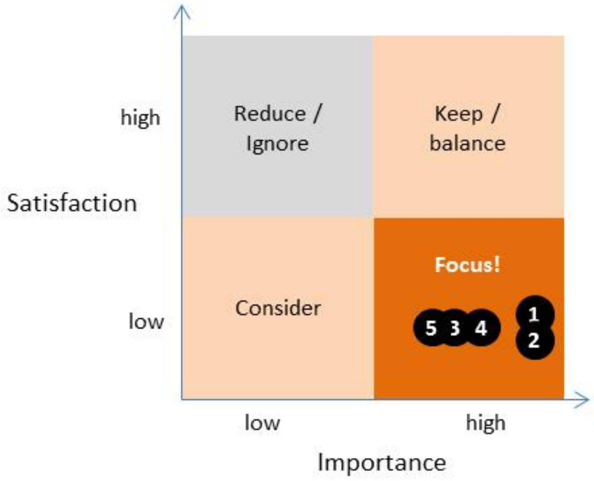
Pulling these standards resources together in the ARIADNE Registry will raise awareness of the standards in use and those that could be of use, among the partners and the ARIADNE stakeholders, and should more easily illustrate where gaps in standards may exist. It will also allow for comparisons and potential mappings, which will be explored in D12.2.

3.4 User Requirements

Metadata quality was ranked as third in importance to the groups of archaeological researchers who participated in the user requirements survey, and has therefore been designated a priority area for the ARIADNE project for further strategic work. As such, it is important to understand the user expectations for the archaeological metadata and any alignments to standards, schemas and vocabularies which may be incorporated into the ARIADNE infrastructure. The user needs survey did not include detail about particular standards, but rather the attitudes towards using them.

The users' needs literature review within D2.1 revealed anxiety about the adoption of unfamiliar (but predefined) data standards, schemata, vocabulary, and user interfaces, along with required markup of data to align it with more general Web or semantic standards (perceived as disconnected from immediate needs, and outside of practitioners' area of expertise).

User requirements of archaeological researchers according to their importance and the satisfaction with the existing situation in a strategy matrix



User needs as shown in the matrix:

(1) = Data transparency

(2) = Data accessibility

(3) = Metadata quality

(4) = Data quality

(5) = International dimension of available data

While they cite "data compatibility and interoperability are a concern", the will to adopt changes into their practice is low. Many do not see metadata as a priority and produce no metadata at all, as they don't see the use of the additional effort (as this effort is rarely factored into research project funding). At the same time, the number of digital datasets continues to increase, and the use of standards to ensure adequate resource discovery is becoming more and more pressing. With this increase, the major challenge data managers see themselves confronted with is ensuring metadata quality (by far the most important item on their list of six challenges), so there is a fundamental disconnection between those who are creating the metadata (or not) and those who are trying to manage their datasets.

This is the user context the ARIADNE consortium finds itself in with regard to standards. The importance of data is fairly well understood by both researchers and data managers, but this is not the case for metadata. It is going to take considerable work to bridge this gap over many years, but this is also an opportunity for the ARIADNE project to show good practice and raise awareness about the potential usefulness of data when it contains good metadata, and the available standards to allow further interoperability.

3.5 Issues for Consideration

Archaeology-oriented metadata standards and conceptual schemas

The archaeology-oriented metadata standards and conceptual schemas listed in section 3.2 show the broad range of areas where standards are being developed. Some are meant to be high-level standards for broad subject areas like cultural heritage (i.e. CIDOC CRM, CHARM or the CARARE schema), while others are meant to address particular types of data (i.e. geospatial data with INSPIRE or bibliographic data with MARC), still others attempt to create standards which are specific to archaeology (i.e. the International Core Data Standard for Archaeological Sites and Monuments, SIKB0102 - Excavation data, or the STARC schema).

Some have been foundational to broader interoperability initiatives for some time, and ARIADNE must certainly attempt to incorporate them into the building of the infrastructure. Cited in D3.2 as the most likely candidate, is the CIDOC CRM. As the ISO standard for cultural heritage, the CIDOC CRM has been core to many major data projects in the sector, and work to extend it into more domain specific areas useful to archaeology is currently being undertaken in association with WP14 *Addressing Complexity*, with the following extensions:

CRMdig: CRMdigital is an Extension of CIDOC-CRM to support provenance metadata. It is available in English as Resource Description Framework Schema (RDFS), and was created by FORTH.

CRMsci: CRMsci: the Scientific Observation Model is an Extension of CIDOC-CRM to support scientific observation. It is available in English as RDFS, and was created by FORTH.

CRMgeo: CRMgeo links the CIDOC CRM to GeoSPARQL through a Spatiotemporal Refinement. It is available in English as RDFS, and was created by FORTH.

CRMarchaeo: CRMarchaeo is an ontology and RDF Schema to encode metadata about the archaeological excavation process. The goal of this model is to provide the means to document excavations in such a way that the following functionality is supported:

- Maximize interpretation capability after excavation or to continue excavation
- Reason of excavation (goals). What is the archaeological question?
- Possibility of knowledge revision after excavation
- Comparing previous excavations on same site (space)
- All kinds of comprehensive statistical studies (“collective behavior”)

Most are already available in English as RDFS, from FORTH, and are freely available for download.

In addition, the **CRM-EH** is an existing extension to the CIDOC CRM that models the archaeological excavation and analysis process. Creation of the CRM-EH was prompted by a need to model the archaeological processes and concepts in use by the English Heritage, to inform future systems design and to aid in the potential integration of archaeological information in interoperable web based research initiatives. As such, it is modelled for use with the Single Context system of recording, which may not be suitable for data created using other types of systems. The possibility for expansion of the CRM-EH to accommodate other recording systems is a point of discussion in WP15 *Linking Archaeological Data.*

It is available in English as RDFS, was created by the University of South Wales in partnership with English Heritage and the Archaeology Data Service, and is freely available for download.

Even with more domain-specific extensions, the CIDOC CRM is still a conceptual model, and more work will be required to integrate it into the ARIADNE infrastructure. In addition to the work underway within WP14, the CRM Special Interest Group is currently defining a set of core concepts as a place to start.

Archaeology related terminology resources

Terminology resources are a critical component of any attempt to create an interoperable infrastructure for archaeology, but also one of the most challenging. Use of terminology within archaeology varies between languages, regions, nations, spatial and temporal research areas, research methodologies and ideologies and areas of expertise. Adoption of a terminology requires agreement by researchers that it is rich enough to describe their data, and a sufficient number of researchers must be willing to use it (or be required to use it) for the resource to be useful for interoperability.

No one currently available terminology resource will ever meet the needs of archaeological researchers across Europe, and it will be mappings between resources that will move interoperability forward. The fact that the ARIADNE infrastructure will be created using Resource Description Framework (RDF) means it will be built to make best use of mappings. Indeed, the ability to keep individual data terminologies and make them interoperable through mapping rather than forcing data to conform to a single terminology is one of the main reasons RDF, Linked Data and Semantic Web technologies were created. That said, while ARIADNE will experiment with using different types of terminology resources and attempt to create mappings between them, the partners must deposit their data into the ARIADNE Registry at the start of the project, so a few very basic terms (based on the natural categories which arose from trying to classify the data actually held by the content providing partners – see Section 2.2) have been accepted for use across the ARIADNE infrastructure. The category terms are:

- Site/monument
- Context
- Object
- Event
- Actor
- Document
- Time
- Period

3.6 Recommendations on standards

The concepts listed in the previous section will be the core terms partners should use to begin mapping the terms relevant for their datasets, from there partners can choose the concepts (classes and properties) which describe their datasets. These can then be mapped to the extended CRM (which is due later in the project). By creating the required concepts now, they can then be compared with the CRM extension, and new terms created. In addition, some partners are already working with archaeology domain-specific terminology mappings to the CRM, and these resources will help build the list of concepts.

Within D2.1, users cited two main needs with regard to standards: the need for international standards specifically for excavation and site/monument data, and access to tools to ease the mapping of their metadata to these standards.

To address these issues, and the five over-arching user requirements generally, the following table shows where schemas and vocabularies are best placed to address the over-arching users' needs categories, and what tools may address their concerns.

	Metadata schemas	Vocabularies		Metadata mapping tools	Metadata input tool	Metadata description tool	SKOSifier tool
Wishes							
Data transparency	+						
Data accessibility	++	+					
Metadata quality	+++	+++					
Data quality							
International dimension	++	+++					
Concerns							
Metadata quality (managers)						X	X
Effort for metadata creation (researchers)					X		
Anxiety about unfamiliar schemas (researchers)				X			

Figure 3.1: Table showing the wishes and concerns with regard to data standards, categorised by the type of schema or vocabulary which may address the wishes, and the tools which may address the concerns. The + signifies the level of impact.

The following tools should facilitate the creation of metadata and vocabularies. Some tools may need to be developed or adapted as part of the ARIADNE project, but there may also be existing tools which may be used. Within the D13.1, the use case “Discover tools & knowledge” describes how researchers will be able to use the ARIADNE Portal to discover resources to support their research/data management. Third party tools and any tools developed within the project which may be of use will be accessible there. Tools relevant to standards may include:

- Metadata mapping tool: A tool to facilitate mapping of data to appropriate metadata
- Metadata input tool: A tool to enhance existing metadata by adding further information (e.g. annotations), to improve metadata quality
- Metadata description tool: A system that forces and facilitates the description of the metadata structure (to enhance data quality)
- SKOS creation tool: A tool to facilitate the structuring vocabularies in SKOS

With these recommendations in place, the potential for creating more accurate and semantically rich data models is greatly increased, thus facilitating better integration and semantic search.

4 Access and Sharing Policies

4.1 Background

Objective

To assess data and metadata sharing and access policies, to inform the development of the ARIADNE infrastructure with regard to what policies should be adopted.

The ARIADNE project needs to consider the data access and sharing policies relevant to archaeological data and metadata, as one of the main goals of the project is to provide greater access. Within this, the project needs to consider the sharing policies already in place for the datasets ARIADNE partners plan to provide for integration and the broader context of the current move across all research domains towards open access for research publications and data.

The preservation and dissemination of archaeological data is of greater importance than that of data in many other research domains. Archaeology, in its most traditional form of excavation, is inherently destructive and the (increasingly digital) documentation gathered during this process becomes the primary data about that unique resource. At the same time archaeology as a discipline has a poor tradition of data sharing, and in the most extreme (though unfortunately quite common) circumstances, research can go unpublished for decades (or forever). This often occurs because funding tends to focus on fieldwork, which is often more exciting than the equally important post-excavation and publication phases of an archaeological project.

In much the same way that there is a disconnection between the priorities of researchers who provide metadata, and data managers who must ensure the quality of the metadata they hold (so they can in turn provide the resources researchers are looking for), there is a disconnection between researcher's willingness to share their data, and their desire to use the data of others. Equally, there is a disconnection between the sharing policies being put in place by research funders and institutions, and the attitudes of many researchers about those policies. Thus, the politics of sharing and access presents another challenging area for the creation of the ARIADNE infrastructure.

4.2 Rights and access policies implemented by partners

The content providing partners were surveyed about the rights and access policies they have in place for the data and metadata they plan to provide to the infrastructure. The results of this survey were reported in D3.3 *Report on data sharing policies* [5]. At the time the survey was carried out, the partners planned to make 28 collections available for ingestion. The rights and access details for these collections are summarised below:

Rights holders

- Multiple rights holders: 61%
- Single rights holders: 39%

Content copyright

- Subject to copyright: 83%
- Open: 8%
- Copyleft: 3%
- Restricted Access or temporary embargo: 6%

The percentages above reflect the fact that copyright agreement is often an individual agreement between the depositor and the repository, so researchers and institutions may fall into multiple categories.

Content Access

- Freely available online: 50%
- Freely available online to registered users: 39%
- Available offline: 5%
- Currently closed to users: 3%
- Freely available after click-through to accept license conditions: 3%

Content Licensing

- Non-standard: 47%
- Creative Commons: 42%
- CC BY NC SA: 22%
- CC BY NC ND: 14%
- CC BY SA: 3%
- CC0: 3%
- Open: 8%
- Gouv.fr Open: 3%

For the 47% of collections using non-standard licenses, users must typically apply for permission from the content holder to use the resource for publication etc.

Three national archives which hold archaeological data have developed their own licenses customised to their domain:

ADS – UK: All data is freely available for use with attribution for research, learning, and teaching; also for commercial archaeological projects with the provision that the outputs end up in the public domain

KNAW-DANS **–** The Netherlands: Depositors can choose between a license similar to CC BY, but can also choose to restrict some data, or impose an embargo period

SND – Sweden: Depositors can specify different levels of access for a whole collection, or for subsets within a collection

Metadata rights

As most partners will be supplying metadata to the ARIADNE infrastructure, while continuing to hold the data itself, it was particularly important to determine what kind of metadata rights were in use by the consortium. The rights under which the partners are prepared to make their metadata available will determine the level of resource discovery possible from the ARIADNE portal. It was hoped that most partners could make their metadata available in the least restrictive way (CC0 public domain), irrespective of the rights they hold for the data itself.

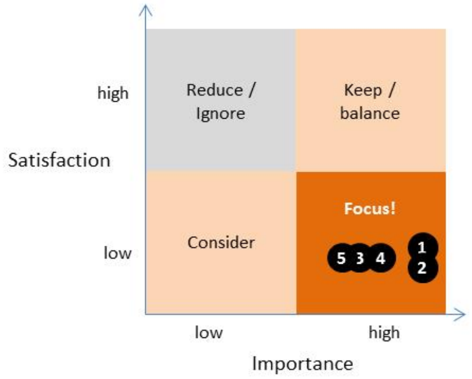
Partners able to make their metadata available under CC0: 12 = 60% of metadata

Other partners did not have firm policies in place with regard to metadata (or more complex policies), as they did with data. This is likely a further reflection of the less mature understanding of metadata versus data. Most were willing to try to come up with a solution however, and some were willing consider using CC0 if it was important to the success of the project.

4.3 User Requirements

Three out of five of the focus areas identified by the users' needs survey were related to access. Data transparency and data accessibility were ranked first and second, respectively, and the international dimension of available data was ranked fifth. The topic of access will therefore be one of great focus for further strategic work within D2.2 *Second report on users' needs.*

User requirements of archaeological researchers according to their importance and the satisfaction with the existing situation in a strategy matrix



User needs as shown in the matrix:

(1) = Data transparency

(2) = Data accessibility

(3) = Metadata quality

(4) = Data quality

(5) = International dimension of available data

Sharing Research Data

During the literature review on surveys of user needs and requirements in research for e-infrastructure, repositories and services, the authors of D2.1 found that sharing research data is the "hottest" topic in the sector. They cite that these services see sharing as not only best practice, but important to their continued existence.

This has resulted in policies being put in place that assume researchers "are willing and able to share their data in an open and trustful manner", but often have not taken into consideration whether this is actually the case. The literature review revealed that e-infrastructure surveys show that only 6-8% of researchers deposit their datasets in an external archive, which represents very shaky ground for the e-infrastructures sector. D2.1 states there are greater barriers to sharing for researchers than incentives. They include:

- Little academic reward for the development and curation of databases

- Priority of publications rather than data sharing
- Concerns that data could be scooped, misused or misinterpreted
- Issues of copyright, confidential and sensitive data
- Required additional effort for providing shareable data, including formatting, metadata creation, licensing

They go on to state: "Overall there are more barriers than incentives for open access sharing of reusable. Therefore data is primarily shared directly between trusted colleagues of the research community. Where open data recommendations and guidelines have been issued, actual provision is routinely ignored or under-performed. This makes the existing cases of open data sharing all the more valuable and exemplary."

4.4 Issues for Consideration

As set out in D3.3 *Report on data sharing policies,* access to data and metadata depends on a chain of data sharing activities, any link within which can influence whether data and/or metadata can ultimately be made freely available online. The analysis in D3.3 of the data gathered in D3.2 has shown the situation regarding data sharing policies and access is changing along with the general move towards open access, but is also quite variable. From D3.3:

> After consulting with partners it is clear that access and sharing policies are evolving. Management of IPR and licensing of content is well established and understood by some partners; others are still working through the process. There are national and institutional variations, and legacy datasets deposited under past frameworks to be taken into consideration. However it is clear that there is a common move towards the explicit licensing of content and metadata so that datasets can be made available for research, education and public use.

D3.3 goes on to set out the key components that need to be considered, puts them into the chain of data sharing activities (Figure 4.1).

Points when agreements must be made about the data:

Deposit agreements with content providers

When information about the provenance (research team, project) of the dataset and any underlying rights (objects, sites, data re-use) is collected and agreements reached for access permissions etc.

Agreements with ARIADNE

When agreements need to be reached about the licensing of resource description metadata and content (for research, education, public and/or commercial use), permissions for data re-use (making derivatives), and data citation (accreditation) etc.

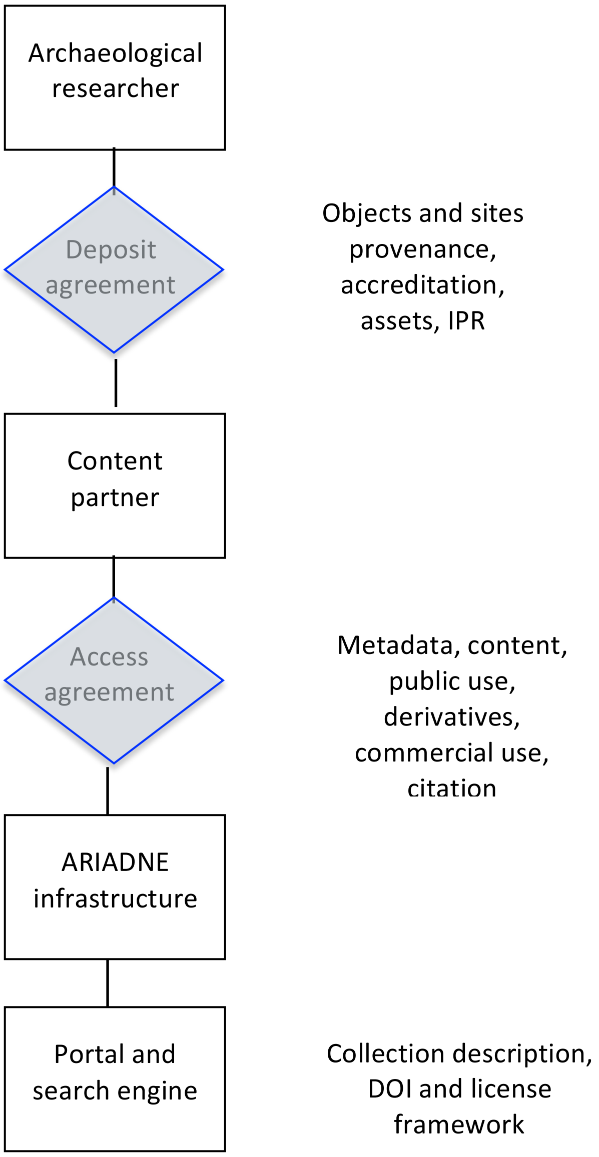

Archaeological researcher
Deposit agreement
Objects and sites provenance, accreditation, assets, IPR
Content partner
Access agreement
Metadata, content, public use, derivatives, commercial use, citation
ARIADNE infrastructure
Portal and search engine
Collection description, DOI and license framework

Figure 4.1: Data sharing activity chain

Decisions will have to be made about the frameworks under which ARIADNE will operate, and will need to take into consideration "policies for data citation, provision of unique persistent identifiers for datasets (and subsets), and licenses for resource description metadata and content." The ARIADNE consortium will also need to agree on licensing, though as most data and metadata being offered by the partners is subject to a form of Creative Commons license, this may be the logical solution, but D3.3 has set out several issues the ARIADNE partners will have to consider:

There is a general consensus amongst partners that open access should be provided for collection level metadata. The issue is whether to follow the Europeana model and adopt the CC0 (public domain) license or the CC BY license (to ensure attribution of the content provider).

Many partners are comfortable with Creative Commons licensing, but if the ARIADNE consortium decides to adopt the full suite of Creative Commons licenses rather than a specific license that suits the majority of partners, it will still become very problematic to implement the variants within the infrastructure (see D3.3 for details on the specific issues which may be encountered). As such, the consortium should consider whether it is better to introduce a single license and risk excluding some data/metadata from some partners, or negotiate the complexities to be more inclusive.

Another issue to consider is whether users should register to access the resources that will be held within the ARIADNE portal. Only 50% of the collections offered by the partners are currently freely available without users needing to register, the other 39% that are freely available require users to register. This will need to be thought through in task 12.2 and a decision made. As users may be accessing data held by multiple partners who require registration, it will be important that they only have to register once through ARIADNE, otherwise they will likely find using the portal a frustrating one.

4.5 Recommendations on ARIADNE's data sharing policy framework

Very specific recommendations were made in D3.3, all of which should be carefully considered for Task 12.2, therefore they are reproduced here in their entirety:

1. **A common method of data citation** be established for adoption by partners and promotion by ARIADNE to the archaeological research community. Academic recognition is an important motivation for encouraging researchers to share access to their datasets.

2. **Allocation of DOIs or the equivalent to datasets ingested to the ARIADNE infrastructure** be investigated. The system used should be capable of identifying sub-sets within collections. Persistent identification of datasets is important in underpinning data sharing and data citation.

3. **Content itself (databases, document archives, images, 3D models, etc.) be provided to ARIADNE by content partners using the Creative Commons license suite (version 4.0 is preferred) under license permissions agreed with the content owner**. CC BY is recommended for open access. CC BY SA or CC BY SA NC licenses may also be applicable.

4. It is recommended that together with the content itself, partners be requested to provide:

- **A Collection description (of the whole collection and sub-sets within the collection) be published under a CC BY license for each dataset integrated into the ARIADNE infrastructure.** Collection description is a useful way of capturing the provenance and contextual information about data collections, and can be used to underpin data citation.
- **Item level metadata records be published under a CC0 license** – to enable integration of multiple datasets within the metadata repository, support resource discovery and enable linked open data. As ARIADNE will be ingesting multiple datasets from different content providers under differing existing license conditions, it is recommended that ARIADNE follows the example of Europeana and defines a metadata element set that can be published under an open license (CC0 is the most open, CC BY if public domain licensing cannot be agreed).

6. Conclusion and Recommendations

The purpose of Task 12.1 of WP12 has been to pull together the considerable information already gathered within several other ARIADNE tasks and their resulting deliverables, to understand the nature of the infrastructures provided for integration, including what data and metadata will be available within the registry, make it possible to identify what gaps may be present, and how they may be adapted for integration. The deliverables included:

- D3.1: *Initial Report on Standards and on the Registry*
- D3.2: *Report on Project Standards*
- D3.3. *Report on Data Sharing Policies*

This understanding was also informed by the recently completed D2.1 *First report on users' needs.*

This deliverable was structured so as to produce recommendations for Task 12.2 (and to a lesser degree, Task 13.1) in the areas of:

- Data
- Metadata Standards, Schemas and Vocabularies
- Access and Sharing Policies

Below are the summary recommendations for each area:

Data

- **Site and monuments databases:** Most European countries and/or regions have them, and combining may be useful for cross-border searching and geolocation
- **Intervention activity:** May have multiple activities associated with a geolocatable site, which may allow linking of various activities to a single site or monument
- **Fieldwork databases:** Usually too diverse, so individual databases may not be useful for integration, but may be worth linking to intervention activities for bibliographic discovery
- Other categories are quite specific, but may be useful for integration:
    - **Scientific Databases**
    - **Artefact Databases**
    - **Burial Databases**
- **Balance data quality and quantity**: specify requirements that datasets have to meet in order to be integrated, preferably using a formal criteria

The relationships between the available types of data available from the content providing partners and the recommended integration activity to be designed within D12.2 are set out in the table below.

DATA **Balance data quality and quantity**	**ARIADNE datasets**					
Integration activity	**Sites and monuments databases**	**Intervention databases**	**Fieldwork databases**	**Artefacts**	**Burials**	**Scientific datasets**
Cross-border subject search	**X**	**X**	**X**			**?**
Cross-border period search	**X**	**X**	**X**			**?**
Map driven searching or visualisation	**X**	**X**	**X**		**?**	**?**
Bibliographic metadata from grey literature	**X**	**X**	**X**	**X**	**X**	**X**
Integration and interoperability from scientific databases						**X**
Integration of particular kinds of artefact data				**X**	**X**	
Dataset assessment required	**+**	**+**	**+**	**+**	**+**	**+**

Figure 6.1: Table showing types of data available from the ARIADNE content providing partners, categorised by the type of integration activity which is recommended for implementation within the ARIADNE infrastructure. The question mark "?" denotes cases in which the feasibility of the integration activity must be established case by case according to content type. Note that this table is the same as Figure 2.3: it is reproduced here for the reader's convenience.

Metadata Standards, Schemas and Vocabularies

- The **use of international standards for the documentation of excavations and monuments** so as to render it transparent and comparable
- **Free access to tools,** particularly for data mapping, to make it easy to comply with these standards, and offering the means and guidance to archaeologists to deposit their digital records
- **The sustainability of digital datasets** must also be high on the agenda

The relationships between the wishes and concerns with regard to metadata and the recommended tools to be designated or designed within D12.2 are set out in the table below.

	Metadata schemas	**Vocabularies**		**Metadata mapping tools**	**Metadata input tool**	**Metadata description tool**	**SKOSifier tool**
Wishes							
Data transparency	+						
Data accessibility	++	+					
Metadata quality	+++	+++					
Data quality							
International dimension	++	+++					
Concerns							
Metadata quality (managers)						X	X
Effort for metadata creation (researchers)					X		
Anxiety about unfamiliar schemas (researchers)				X			

Figure 6.2: Table showing the wishes and concerns with regard to data standards, categorised by the type of schema or vocabulary which may address the wishes, and the tools which may address the concerns. The + signifies the level of impact.

Access and Sharing Policies

- **A common method of data citation should be established** for adoption by partners, and promoted by ARIADNE to the archaeological research community. Academic recognition is an important motivation for encouraging researchers to share access to their datasets

- **Allocation of DOIs or the equivalent to datasets ingested to the ARIADNE infrastructure** should be investigated. The system used should be capable of identifying sub-sets within collections. Persistent identification of datasets is important in underpinning data sharing and data citation
- **Content itself (databases, document archives, images, 3D models, etc.) be provided to ARIADNE by content partners using the Creative Commons license suite** (version 4.0 is preferred) under license permissions agreed with the content owner. CC BY is recommended for open access. CC BY SA or CC BY SA NC licenses may also be applicable
- **A Collection description** (of the whole collection and sub-sets within the collection) should be published under a CC BY license for each dataset ingested to the ARIADNE infrastructure
- **Metadata records should be published under a CC0 license** – to enable integration of multiple datasets within the metadata repository, support resource discovery and enable linked open data

The creation of the ARIADNE infrastructure requires a wide variety of information, both to inform the design of the portal, and to understand the current situation with regard to archaeological data within the domain. To ensure the infrastructure is relevant, useful and represents a positive step towards meeting the needs of researchers, considerable work has been undertaken by all partners within the ARIADNE project. This work has been to understand the wishes and expectations of our potential users, and gain an understanding of the current technologies, data structures and policies in use within the domain. The results of this work are spread across several deliverable reports produced by the partners, and have been synthesised in this report to inform the development of the ARIADNE infrastructure, Task 12.2 in particular.

References

1. D2.1 *First Report on Users' Needs*
2. D3.1 *Initial Report on Standards and on the Registry*
3. D13.1 *Services Design*
4. D3.2: *Report on Project Standards*
5. D3.3. *Report on Data Sharing Policies*

All deliverables discussed in this report (both current and in progress) are/will be available at http://www.ariadne-infrastructure.eu/Resources.