Virtual Workshop on

Semantic mapping of archaeological excavation data

Version 1.0

01.09.2022

Grant Agreement number:	823914
Project acronym:	ARIADNEplus
Project title:	Advanced Research Infrastructure for Archaeological Dataset Networking in Europe - plus
Funding Scheme:	H2020-INFRAIA-2018-1
Project coordinator name, Title and Organisation:	Prof. Franco Niccolucci, PIN Scrl - Polo Universitario "Città di Prato"
Tel:	+39 0574 602578
E-mail:	franco.niccolucci@pin.unifi.it
Project website address:	www.ariadne-infrastructure.eu

The research leading to these results has received funding from the European Community's Horizon 2020 Programme (H2020-INFRAIA-2018-1) under grant agreement n° 823914.

Editors: **Markos Katsianis, PP**

Giorgos Styliaras, PP

Contributing partners: **George Bruseker, Takin.solutions Ltd.**

Paola Derudas, Lund University

Gerald Hiebel, University of Innsbruck

Florian Hivert, CNRS, MSH Val de Loire Tours

Markos Katsianis, PP

Vaggelis Kritsotakis, ICS-FORTH

Olivier Marlet, CNRS, University of Tours

Denitsa Nenova, Takin.solutions Ltd.

Federico Nurra, INHA

Giorgos Styliaras, PP

Christian-Emil Smith Ore, University of Oslo

Maria Theodoridou ICS-FORTH

Document History

- 7.7.2022 – Draft Version 0.1
- 25.7.2022 - Draft Version 0.8
- 1.9.2022 - Final version 1.0

Table of Contents

1 Executive Summary

On 15 June 2022 the "*Virtual Workshop on Semantic mapping of excavation data*" took place. The event was organised as an open forum to illustrate aspects of the work carried out by the *Archaeological Excavation Modelling Working Group*, a sub-group within WP 4.4.12. The presenters, both Partners and Associate Partners of the ARIADNEplus consortium, explored semantic modelling and the use of CIDOC CRM, as well as the tools developed to assist researchers with mapping their data. Five case studies on semantic mapping of excavation data were also presented. Each presentation was followed by a Q&A, while a discussion at the end of each session allowed participants to engage in conversation and contribute their experiences and ideas with a view to making excavation data FAIR (Findable, Accessible, Interoperable and Reusable).

The virtual workshop was executed online via Zoom video conferencing services. One hundred and four (104) people registered for the event with the following geographical distribution. Simultaneous participation peaked at 62 people.

Country	Initials	Participants
Argentina	AR	2
Austria	AT	13
Bulgaria	BG	2
Brazil	BR	1
Cyprus	CY	3
Czech republic	CZ	2
Germany	DE	4
Ethiopia	ET	1
Finland	FI	1
France	FR	13
Great Britain	GB	4
Greece	GR	18
Hungary	HU	4
Ireland	IE	1
Israel	IL	1
Italy	IT	12
Nederlands	NL	2
Norway	NO	3
Pakistan	PK	1
Portugal	PT	1
Sweden	SE	7
Slovenia	SI	5
Turkey	TR	3
		104

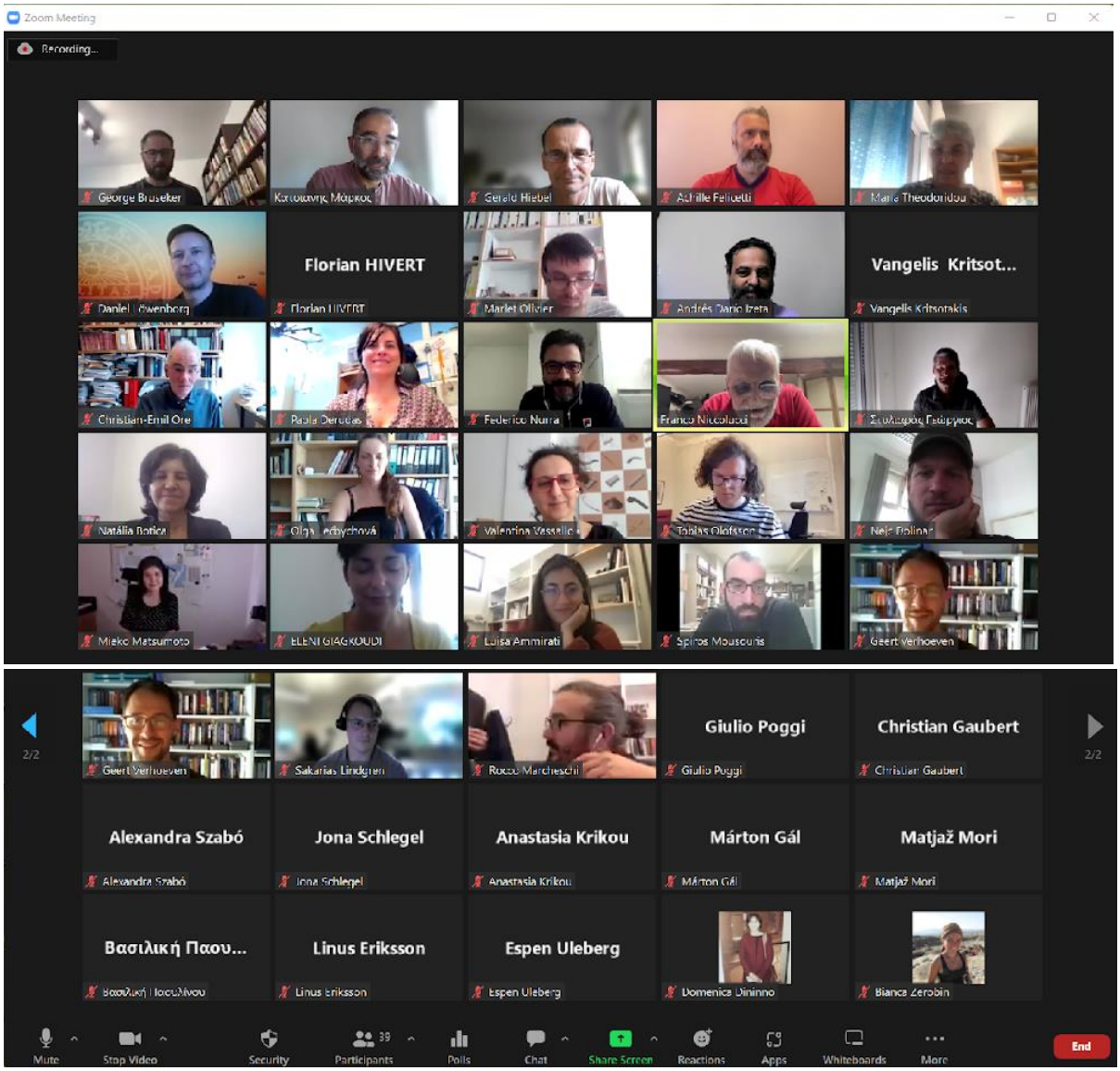

Figure 1. Snapshot from the event.

2 Introduction and Objectives

On 15 June 2022 the "*Virtual Workshop on Semantic mapping of excavation data*" took place. The event was organised as an open forum to illustrate aspects of the work carried out by the *Archaeological Excavation Modelling Working Group*, a sub-group within WP 4.4.12. The workshop explored semantic modelling and the use of CIDOC CRM, as well as the tools developed to assist researchers with mapping their archaeological excavation data. Five case studies were included to showcase the application of modelling workflows and tools on concrete examples.

The workshop's agenda:

Workshop Agenda

10.00 CET	Welcome & Workshop agenda *Franco Niccolucci, PIN-University of Florence* / Scientific Coordinator ARIADNEplus *Markos Katsianis, University of Patras*

10.10-12.00	**PART A: Introduction and tools for semantic mapping**
10.10	Semantic Data Modelling and Archaeological Research Data - Why, How and Where We are Now *George Bruseker, Takin.solutions Ltd.*
10.30	From modelling to mappings: how to appropriate the CIDOC CRM *Olivier Marlet, Centre national de la recherche scientifique (CNRS)*
10.50	Break 10’
11.00	The X3ML toolkit: How to map excavation data to CIDOC CRM *Maria Theodoridou & Vaggelis Kritsotakis, Foundation for Research and Technology - Hellas (FORTH)*
11.20	An approach to model archaeological data and create RDF from spreadsheets *Gerald Hiebel, University of Innsbruck*
11.40-12.00	Discussion 20’
12.00	Lunch Break 30’

12.30-14.20	**PART B: Case studies in excavation data semantic mapping**
12.30	OpenArchaeo: a semantic Web platform for archaeological data *Florian Hivert, Centre national de la recherche scientifique (CNRS)*
12.50	Modelling Archaeological Excavations. Theoretical Patterns and Practical Recipes *Denitsa Nenova, Takin.solutions Ltd.*
13.10	Reworking aged excavation mappings with new models and tools *Markos Katsianis & Giorgos Styliaras, University of Patras*
13.30	Break 10’

13.40	597 Norwegian excavation databases and CIDOC CRM - a practical exercise *Christian-Emil Smith Ore, University of Oslo*
14.00	Archaeological Interactive Report: a *trait d'union* between data management and semantic publication *Paola Derudas, Lund University* *Federico Nurra, Institute national d'histoire de l'art (INHA)*
14.20	Break 10'

14.30-15.00	**PART C: Discussion**
14.30	Discussion, wrap-up & next steps for the Excavation Modelling Group

3 Presentations

In this section, the content of each presentation is summarised and followed by the main themes discussed in the Q&A.

3.1 Semantic Data Modelling and Archaeological Research Data - Why, How and Where We are Now

In the event's introductory presentation **G. Bruseker** (Takin.solutions Ltd.) stressed that semantic data modelling and data sharing are not yet common place in archaeological field excavation practice. ARIADNEplus, as a European Research Initiative dedicated to the sharing of digital practices and competencies in the archaeological community, has as one of its key goals to work out common semantic data model profiles for different kinds of archaeological data. In the Archaeological Excavation Modelling Working Group, a sub-group within WP 4.4.12, the focus was directed on exploring how to create, use and disseminate the practical use of semantic modelling for archaeological excavation data. An overview of the state of the art of this work is provided as an introduction to the different projects and experiments that were taken up within the remit of this group's mandate.

In this regard, the primary question examined by G. Bruseker is why to semantically model archaeological excavation data in the first place and to what end could such an activity and practice be put. Two positive and one negative fact have been acknowledged, as to why there is good reason to consider semantic modelling and sharing archaeological data.

Positively speaking, the objects of archaeological investigation, even on the field, are understood as in a related continuum. An excavation cannot be rightly and fully understood on its own, but only as it relates to a broader picture of ancient, living culture and environmental situation. The sites excavated and represented by archaeological data did not stand in splendid isolation and can only be understood in their interconnections, including trade, exchange and other types of communication between communities. Consequently, given that the object of study is interconnected and only understood in this context, the data that describes and explains it should also manifest this interconnection, so that the past can be brought better into view in its broader interconnection and those connections' meanings.

Moreover, the subjects of scholarship who generate this data, i.e. archaeologists and archaeological institutions, exist themselves in a continuum and have an important bearing on our understanding and use of archaeological data through time. The author of data often disappears into the background, effaced by the data itself. But, to understand data is to have a critical relation, thereto, to understand its provenance, its relation to a research plan, an objective etc. Here too, the semantification of data and its explicit representation of the 'metadata' of data to contextualise and situate data pools is of crucial importance to building a new digital practice.

Conversely, whilst an aim of creating archaeological data is to create a record that fittingly mirrors the past, the present environment of un-standardized, in-explicit, under-documented data management practices may result in "dark" or "misleading" reflections of the

archaeological record. By and large and for the most part, archaeological data exist in data silos. Their structures may not necessarily follow an explicit standard but multiple, different and often incompatible standards. In addition, archaeological data is often only analogue and, therefore, inaccessible to digital scholarship in the first place.

In this respect, semantic data modelling offers a means to overcome these roadblocks towards a new digital scholarship over a unified set of archaeological excavation information, enabling the ability to ask and answer questions with more efficiency, while safeguarding the long-term readability and reusability of archaeological data. Semantic data modelling promises a new step and direction in digital scholarship, but the ways for reaching this new step is the challenge with which this working group was charged to contend. Elements of this challenge include working to create adequate models, methods and tools for the task at hand.

Models provide a means of expressing the world consistently, i.e. semantic models enable the consistent expression of diverse dataset according to agreed objects of reference. Objects of study here included: CIDOC CRM 7.1.1, AO Cat, CRMArchaeo 1.4.4, the ARIADNEplus "Application Profiles" and additional models by the CIDOC CRM family. Methods are the tactics for practically achieving data integration. Well expressed methods for data production, transformation and maintenance support the application of models in tools towards genuine integration. Here research is engaged in critical studies, application scenarios, the elaboration of semantic recipes and the investigation and elaboration of pedagogy around this topic. Tools consist of the software and code for facilitating data integration. Well maintained and supported software and code enables sustainable workflows for generating integrated data. Here some particular areas of focus include, but are not limited to: X3ML Toolset, CRMGame, OpenArchaeo and Spreadsheet Workflows.

Figure 2. The basic elements required towards a better-defined conceptual definition of the archaeological excavation domain.

The presentations of the Virtual Workshop showcase some of the results of the investigations of this working group in those directions. The working group critically took up the models of the ARIADNEplus project and subjected them to critique and stress testing through working on their

applicability in different application scenarios, outlining recipes for how to apply them, working on means to transfer this knowledge to a broader public and make them usable through various tooling for learning, producing, transforming and querying semantic archaeological research data.

3.2 From modelling to mappings: how to appropriate the CIDOC CRM

In the second presentation of the morning session, **Olivier Marlet** (CNRS, University of Tours) discussed the experience of the MASA consortium in France towards appropriating the entire family of CIDOC CRM models to model excavation reasoning. The MASA group has many partners working together to disseminate the FAIR principles and to find solutions to help archaeologists bring their data to the Semantic Web. For this purpose, MASA has created a digital ecosystem with several tools to help researchers process their data, from structuring to dissemination on the Semantic Web. Based on the data lifecycle, this digital ecosystem highlighted a workflow describing the different steps necessary for this processing.

Within this workflow, MASA worked on two aspects in particular. On the one hand, the implementation of a semantic Web triplestore for archaeological data; MASA has set up a SPARQL semantic web platform with a user-friendly interface (Sparnatural) allowing users to intuitively generate a query without having to write any SPARQL code. On the other hand, the training of archaeologists at CIDOC CRM using the table card game by G. Bruseker and A. Guillem, whose online release is managed by MASA. The online digital version was developed to provide a more optimal use of all the control and interaction ideas that the paper card game already offers. In addition to being fully customisable (ontology, instances, pedagogical progression), the digital version of the game also automates and systematises the pedagogical part of the game. Both infrastructures have allowed MASA to mobilise more and more archaeologists around the issue of Open Science.

Figure 3. Learning CIDOC CRM by playing the online card game.

3.2.1 Q&A with event participants

- *Are there any differences between the original CIDOC CRM table card game and the online digital version?* (M. Katsianis)
    - Screen space limitations in the online game means that the emphasis is on learning how to build expressions that link model entities, rather than building or structuring specific models. Real datasets are easier to be compiled using the table game. In both versions it is a one-player game.
- *Has anyone else used Sparnatural (https://sparnatural.eu/) querying mechanism in their work?* (K. May)
    - The French National Archives (Archives nationales de France) use Sparnatural with their own ontology, i.e. ICA RiC-O (Records in Contexts-Ontology) (https://www.ica.org/standards/RiC/RiC-O_v0-2.html).
- *Can a model that uses OpenArchaeo but also custom concepts or modified properties be considered as interoperable?* (C. Bouras)
    - OpenArchaeo can be used as a model basis and maintain a basic degree of interoperability. For mappings from databases, Protégé-Ontop (https://protege.stanford.edu/ & https://ontop-vkg.org/) are very useful allowing the direct connection to the database without the need to export data in another format (https://halshs.archives-ouvertes.fr/halshs-03561376v2/file/MappingOntop.pdf).

3.3 The X3ML toolkit: How to map excavation data to CIDOC CRM

In the third presentation **M. Theodoridou** (ICS-FORTH) and **Vaggelis Kritsotakis** (ICS-FORTH) presented the specifics of the X3ML toolkit to map excavation data. They argued that data aggregation and integration have the potential to create rich resources useful for a range of different purposes, from research and data modelling to education and engagement. CIDOC CRM and the family of compatible models provide a sufficient target schema for the integration of heterogeneous excavation data. The Centre for Cultural Informatics, ICS-FORTH, developed the X3ML toolkit, a set of small, open source, microservices that follow the SYNERGY Reference Model. They are designed with open interfaces, and they can be easily customised and adapted to complex environments.

The X3ML Toolkit consists of a set of software components that assist the data provisioning process for information integration. The key components of the toolkit are: (a) X3ML Mapping Definition Language, (b) 3M Mapping Memory Manager, (c) X3ML Engine and (d) RDF Visualiser.

• X3ML Mapping Definition Language

X3ML (https://github.com/isl/x3ml/blob/master/docs/x3ml-language.md) is an XML based language that describes schema mappings in such a way so they can be collaboratively created and discussed by experts. X3ML was designed on the basis of work that started in FORTH in 2006 and focuses on establishing a standardised mapping description, which lends itself to collaboration and the building of a mapping memory to accumulate knowledge and experience.

• 3M Mapping Memory Manager

3M is a web application suite containing several software sub-components and exploits several external services. It is available online and its main functionality is to assist users during the mapping definition process, using a human-friendly user interface and a set of subcomponents that either suggest or validate the user input. An online version of the system is available: https://demos.isl.ics.forth.gr/3m/Projectsf. For local installations, a Docker image is also available: https://gitlab.isl.ics.forth.gr/cci/3m-docker

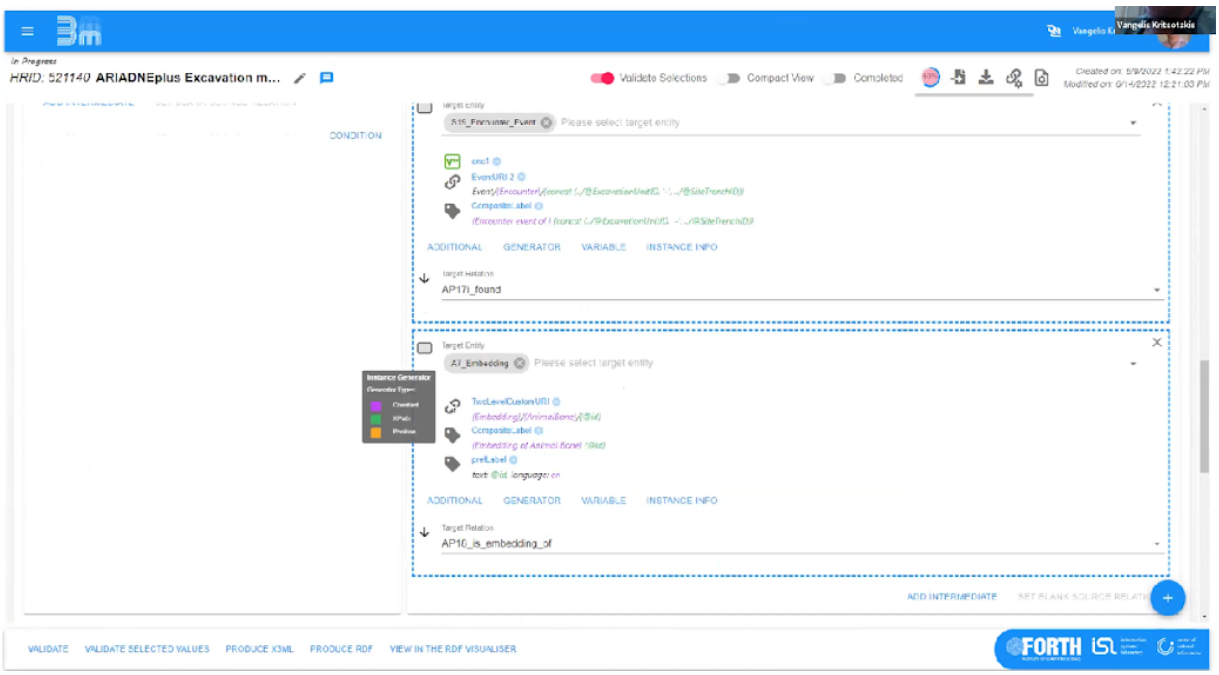

Figure 4. The new 3M interface.

• X3ML Engine

The X3ML Engine (https://github.com/isl/x3ml) realises the transformation of the source records to the target format. The engine takes as input the source data (currently in the form of an XML document), the description of the mappings in the X3ML mapping definition file and the URI generation policy file, and is responsible for transforming the source document into a valid RDF document, which corresponds to the input XML file with respect to the given mappings and policy. The source code is open source and is available on github.

• RDF Visualiser

RDF Visualizer (https://demos.isl.ics.forth.gr/RDFV-Demo/) is a generic browsing mechanism that gives the user a flexible, highly configurable, detailed overview of a dataset/database encoded in RDF. It has been integrated in the 3M interface of the X3ML Toolkit for data mapping and transformation. It enhances 3M with an important validation tool for transformed data. Domain experts can easily check and correct their mapping and transformations on the fly, enabling an iterative and collaborative evaluation of the resultant RDF.

A detailed presentation (manual) for the system is under development.

3.3.1 Q&A with event participants

- *Is this version available in the ARIADNEplus VREs in D4Science?* (D. Novak)
  - Not yet, but it is available online for usage: https://demos.isl.ics.forth.gr/3m/Projects.
- *The input for 3M is XML. What is the best way to transform from spreadsheet or a database to XML?* (G. Hiebel)
  - Spreadsheets and most relational database models have XML export capabilities. There is previous experience in potential conversion workflows and specific cases are welcome to be examined and provide concrete conversion examples.
- *How easy is it to add another extension to the 3M tool to create a graph RDF?* (M. Katsianis)
  - This can be hard. A certain degree of nested viewing of the graph is possible, but at present no viewing mechanism exists for the entire graph. Different types of tools are needed or should be used towards that end.
- *How easy is it to migrate mappings from the old tool version to the new?* (G. Bruseker)
  - Export/Import capability is supported between the two versions.
- *3M is a database for mappings. Can different instances for tool deployment be supported?* (G. Bruseker)
  - Yes, different deployments are supported. The new 3M can be deployed on your own server using the docker image: https://gitlab.isl.ics.forth.gr/cci/3m-docker.
- *Do you plan to update the running environment of 3M for the new version?* (G. Bruseker)
  - You can export a zip file of the old environment and install it locally.

3.4 An approach to model archaeological data and create RDF from spreadsheets

In the last presentation of the first session, **G. Hiebel** (University of Innsbruck) presented an approach to create RDF data that uses Excel spreadsheets for data entry, a Postgres database for data transformations and semantic tools, like Karma or OntoRefine, for RDF creation. For Thesaurus creation, mind mapping software has additionally been used to ease the creation of a hierarchical structure of terms. The entire approach was executed with data from the project "Prehistoric Copper Production in the Eastern and Central Alps" by the research centre HIMAT (History of Mining Activities in Tyrol and adjacent areas) of the Archaeological Department of the University of Innsbruck. The data were collected during several scientific research campaigns and are related to prehistoric mining activities in the eastern Alps of Austria. The documentations were done according to the guidelines of the Austrian Federal Monuments Office (BDA – Bundesdenkmalamt). As the data are of archaeological nature, the methodologies and guidelines of ARIADNE (Advanced Research Infrastructure for Archaeological Data Networking in Europe) were already used to process the data for making them FAIR.

For the creation of the metadata for all generated and deposited files and research documents, the CIDOC CRM ontology with its extension CRMsci and CRMarchaeo were used. CRMsci was used to model physical things and scientific observations and CRMarchaeo to model the documentation of archaeological excavations. Concepts specific to Mining Archaeology research are organised with the DARIAH Back Bone Thesaurus (https://www.backbonethesaurus.eu/), a model for sustainable interoperable thesauri maintenance, developed in the European Union Digital Research Infrastructure for the Arts and Humanities (DARIAH). SKOS (Simple Knowledge Organization System) was used to organise our

vocabularies, another semantic web standard for sharing and linking knowledge organisation systems such as thesauri, taxonomies, classification schemes and subject heading systems. In a further step to use the created data for aggregation in ARIADNE it was necessary to map the concepts used in the research data to the concepts of the Getty AAT (https://www.getty.edu/research/tools/vocabularies/aat/) and PeriodO (https://perio.do).

The modelling of the excavation data was based on the three different main classes of (1) the *material research objects*, like sites, trenches, stratigraphical units and finds, (2) the *activities*, like excavating or analysing a find and (3) the *actors*, like an archaeologist or a specialist. These were each entered in different excel sheets and related through identifiers, as it would have been done in a relational database. Because of such detailed modelled provenances, it is possible to explicitly specify the different research activities with their specific methodologies and linking them to the investigating person or institution.

The pipeline for RDF data creation consists of Excel spreadsheets for data entry, a Postgres database for data transformations and semantic tools like Karma (https://usc-isi-i2.github.io/karma/) or OntoRefine (https://disc-semantic.uibk.ac.at/ontorefine) for RDF creation. The RDF is ingested in a Triple store (Fig. 1).

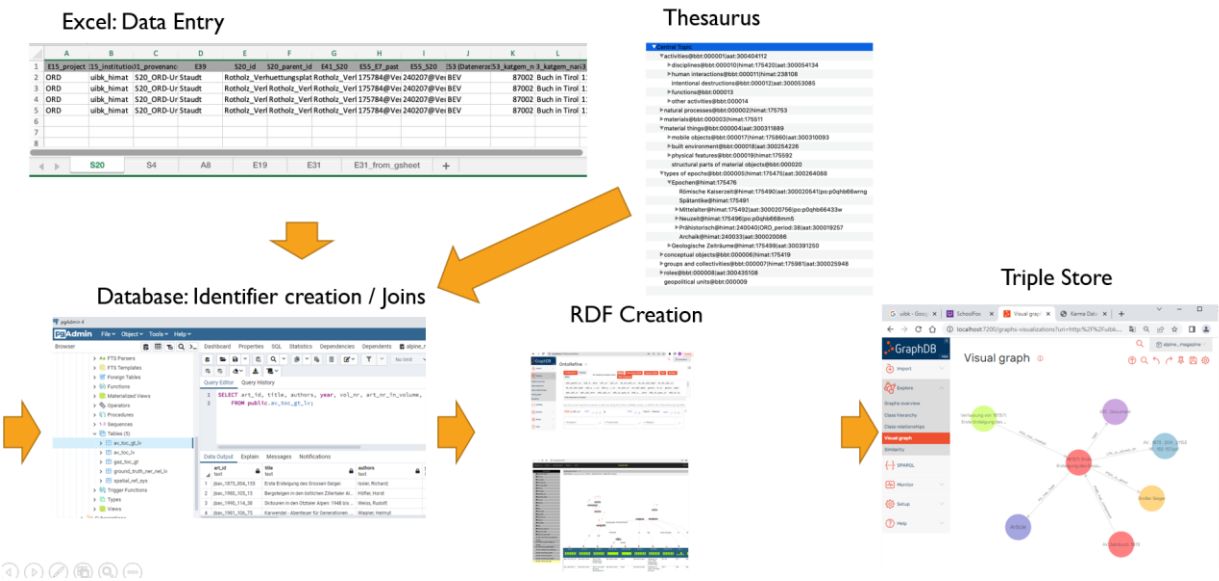

Figure 5. Pipeline for RDF data creation.

To create ARIADNE Aggregation Data, additional information specific to the ARIADNE Catalogue was entered in a “ADRIANE Metadata” Excel spreadsheet and a modelling to the ARIADNE catalogue had been performed to create the RDF necessary for the ARIADNE Portal. The URI identifiers from the RDF network created in the previous step were used in the “ADRIANE Metadata” Excel and had to be transformed in ARIADNE specific URIs for the ARIADNE repository in a repeatable way, again in Postgres. Thus, lots of the information provided in the original network can be accessed and used for the ARIADNE catalogue.

3.4.1 Q&A with event participants

- *If excavation data is already in a database can this process be also used, e.g. with just adding the URI generators in the tables?* (M. Katsianis)
    - Yes, the process can be initiated at the database level and start with the SQLs. Some sort of data restructuring may be required (e.g. removing elements that do not need to be mapped) to replicate the breaking-apart of the database into its constituent elements.
- *The spreadsheet approach to the data records looks much closer to an actual archaeological workflow that could/would be used or derived from archaeological fieldwork. There are many challenges to archaeological data archiving and many organisations, especially in the commercial sector, require workflows that allow revisiting old datasets. In terms of data archiving, what kind of excavation information should be included in such an approach, especially in datasets that are not structured with archiving in mind? The originals or the final RDF?* (K. May)
    - Attention should be directed to include long-lasting data structures at the original documentation. Then, depending on time, additional effort is usually required to align at least the essential information, especially if there are large scales/bulk data involved (i.e. multiple excavations). A balance should be sought between data volumes and available resources.
    - Perhaps, alternative workflows may be needed depending on the ways data are meant to be aggregated (file-based or direct data feed) (M. Katsianis).

3.5 OpenArchaeo: a semantic Web platform for archaeological data

In the first presentation of the afternoon session **F. Hivert** (CNRS, MSH Val de Loire, Tours) discussed the development of OpenArchaeo and its supported functionality. Since 2013, the MASA (Mémoire des Archéologues et des Sites Archéologiques) consortium of the TGIR Huma-Num has assisted archaeologists in digitising and making available their excavation archives, as well as disseminating the FAIR principles within the French archaeological community. The goal is to help archaeologists make their data interoperable and open up their datasets on the semantic web, by using the CIDOC CRM ontology as a shared structure layer for their heterogeneous data.

OpenArchaeo is a platform which attempts to achieve this goal by describing the data with a generic model using a subset of the CIDOC CRM and some extensions (CRMsci, CRMarchaeo and CRMba), as well as few gazetteers and standard vocabularies (PACTOLS, GeoNames, VIAF, ORCID). With this modelisation, the commonalities of the heterogeneous datasets are described with the same structure and metadata. The generic model describes some archeological main concepts and their relationship with each other, such as archeological site, artefact, documentation, etc.

The model in itself won't describe specific data or information, so that the researcher can cross-reference various data from various datasets easily. The production of knowledge graphs from the original dataset is assured by the usage of mapping tools like 3M and Protégé-Ontop. The

conformity of the knowledge graphs with the generic model is then verified with the tool SHACL (https://shacl-play.sparna.fr/play/).

The core functionalities of OpenArchaeo rely on Sparnatural (https://sparnatural.eu/), a Javascript based system allowing the user to query an RDF graph with a graphic interface and without any SPARQL to write. The generic model is interpreted and translated for the user by using an "editorial ontology". This editorial ontology allows the user to query the RDF graphs without a deep knowledge of the CIDOC_CRM ontology and facilitates the interpretation and translation of the OpenArchaeo generic model from Entities or Triplets to a more specific denomination: *E22_Man-Made_Object* becomes only an «Artefact» in Sparnatural, so that every instance of E22 described with this triplet is interpreted as an artefact. The result of a query provides a list of URIs which redirects to the source publication of the datasets.

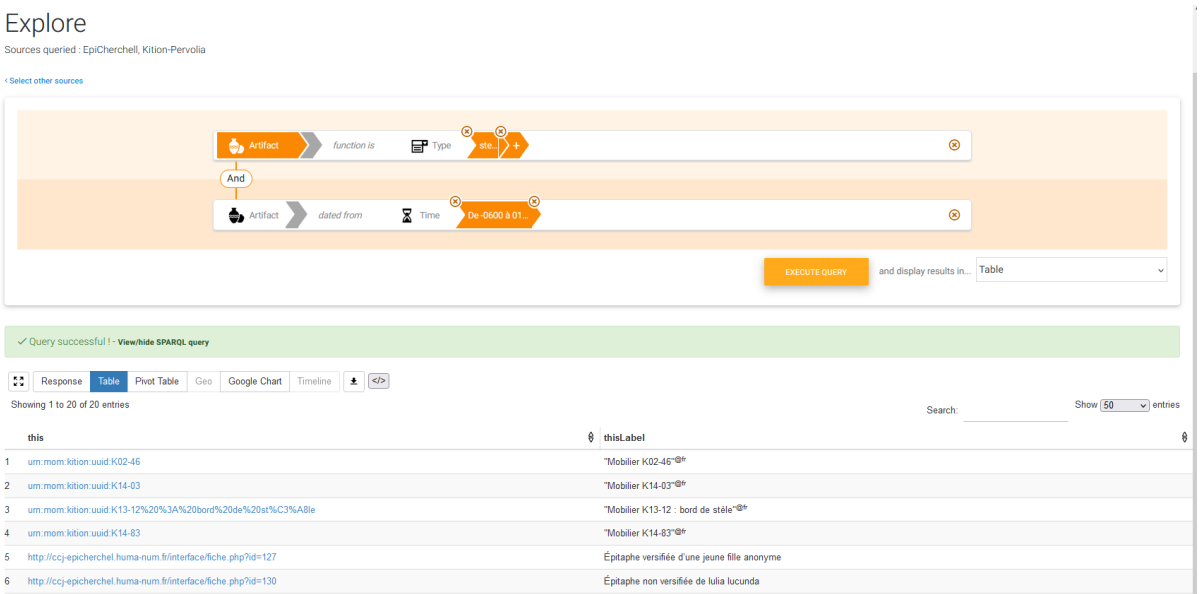

Figure 6. OpenArchaeo explorer querying and results.

A workflow has been set up from cleaning and enriching data to map the OpenArchaeo graphs to the AO_CAT ontology, so that our datasets become progressively more accessible and reusable through the workflow. There's still progress and improvement to make on OpenArchaeo with a deeper query system or with the possibility to retrieve what we would want from a much more complex RDF graph. But OpenArchaeo and Sparnatural are a new way to publish, visualise and query an RDF graph.

3.5.1 Q&A with event participants

- *Can Sparnatural be used to access any SPARQL end-point? To what depth can data be queried using this user-friendly version?* (G. Hiebel)
    - In the current version of OpenAchaeo, Sparnatural enables federated querying of both the MASA triplestore and any triplestore whose data respect the OpenArchaeo generic model (i.e. a subset of the CIDOC CRM). Currently the federation is not perfect because the response times on large volumes of remote data are not optimised.

However, Sparnatural can be adapted for any triplestore and access remote end-points, as long as there is semantic consistency between the different triplestores queried. Work towards customising or humanising the query syntax is currently underway.

- *Is Sparnatural developed as part of the MASA project?* (G. Hiebel)
  - Sparnatural (http://sparnatural.eu) is developed by Thomas Francart (Web of data and knowledge graph services - http://www.sparna.fr/) for the MASA Consortium. It has since been used by the French National Archives and adapted to the Records in Context ontology (RiC-O).
- *The graphs of the OpenArchaeo model are created manually or extracted from RDFs? Graphs can be a very good way to compare models and find commonalities* (G. Hiebel)
  - Yes, these are created manually using a graphic/flowchart editor.
  - CRITERIA (https://github.com/chin-rcip/CRITERIA) can be used for the automatic transform of RDF data to visualised graphs using Mermaid (https://mermaid-js.github.io/mermaid/#/), while diagrams.net Libraries (https://github.com/chin-rcip/diagrams.net_libraries) provides a handy library created by CHIN to facilitate standardised representation of graphs when making them manually using draw.io (https://app.diagrams.net/) (G. Bruseker).
- *Does OpenArchaeo include published or unpublished data?* (A. Katevaini)
  - Necessarily published, as excavation results have to be web resources.

3.6 Modelling Archaeological Excavations. Theoretical Patterns and Practical Recipes

In the second presentation of the afternoon session **D. Nenova** (Takin.solutions Ltd.) discussed the appreciation of modelling patterns while describing the archaeological excavation universe. As a part of the ARIADNEplus project, work has been directed on recognizing the immediate necessities and issues of archaeo-modelling and the potential moves forward. The development of standard models and patterns based on a wide range of archaeological experience and a variety of methodologies has the potential to cover the core entities and extensively outbranch within each individual model. Those models can then be documented and made available for use within the community, either as a whole or as individual fields or collections of fields. Common excavation practices have been targeted to identify core entities and problematic areas that need further development.

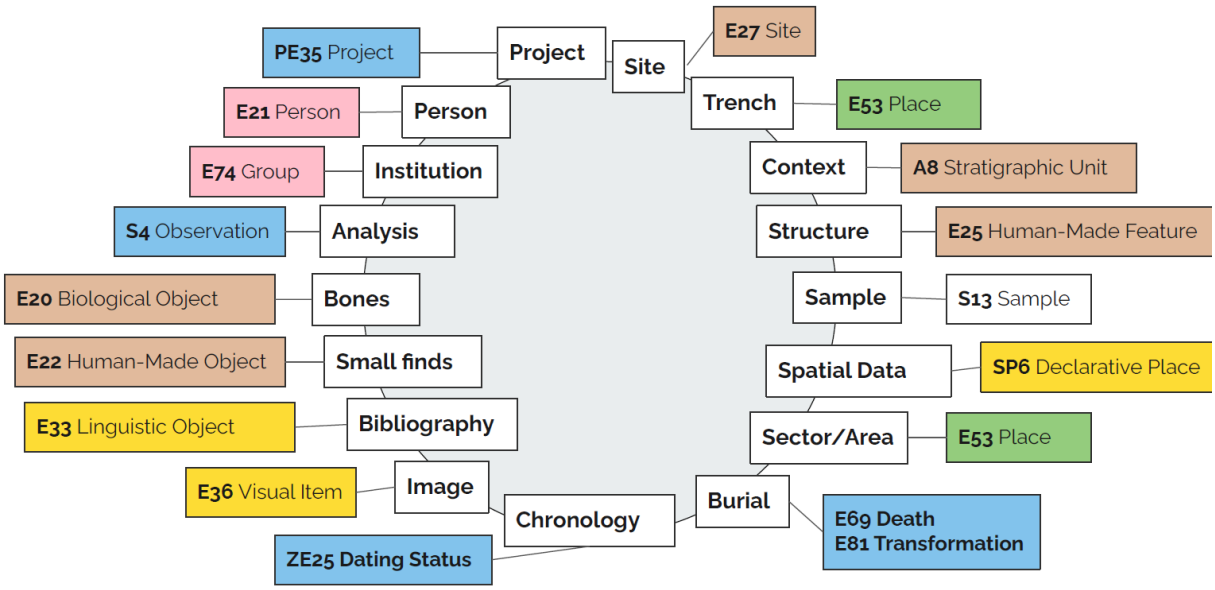

Figure 7. Recognition and assignment of base classes and models.

3.6.1 Q&A with event participants

- *Can this tool be applied to other projects?* (G. Hiebel)
  - There are two ways to go. An implementation has been developed using AIRtable (https://www.airtable.com/). Also, data shapes can be shared. In terms of applying the method, a manual has been created. However, in a sense this process entails a secondary form of ontological analysis. Developers can self-analyse their modelling work, identify certain semantic paths, and create modelling building blocks that can be used on a case-by-case basis. Further information can be accessed on https://docs.google.com/document/d/1rfBV-A8H_z2nWwjq2ShMY3fHkh35r3rebzaAIVPaVk4/edit?usp=sharing, which includes the basic semantic recipes developed in SARI (https://swissartresearch.net/). Note: *The modelling is a work in progress since it still needs validating with the community*.
- *To what extent does a domain expert is still needed to evaluate the way people take up these patterns?* (K. May)
  - There is plenty of room for expanding data modelling patterning research. Shacl shapes could also be included. Shacl is a validation tool (https://github.com/sparna-git/shacl-play) for RDFs. Ideally, as an example, a modeller could access these models, identify the corresponding pattern and get provided with a service for validating their data accordingly. If the query can get the exact data, then this would provide sufficient evidence for the consistency and case-specific applicability of the pattern. In this respect, there may be a point where the modelling expert may be implicated less in data alignment processes.
- *Can we ever dream of this time when people are using a query tool? There is a significant overhead in the retrospective data modelling of archaeological excavation data*. (K. May)
  - Hopefully, this is where this project targets, i.e showing the advantages of semantic data. So, the idea is that there is a central dig database to which individual specialists contribute based upon compatible models, and then ultimately, if there are a bunch of systems set up, then parallels from other sites can be retrieved and provide some clues as you work along.

3.7 Reworking aged excavation mappings with new models and tools

Markos Katsianis and **Giorgos Styliaras** (University of Patras) described previous work in building and implementing a conceptual model for 3D excavation research and discussed some of the challenges in archiving this dataset 10 plus years later, from the semantics point of view. The Paliambela Kolindros archaeological project in Greece, which became the testing ground for the advancement of a 3D documentation workflow between 2000-2010, provides the case-study.

Within the ARIADNEplus consortium, the dataset was used to explore item-level integration within its infrastructure. In this respect, the initial excavation data model, which was mapped using CIDOC CRM v. 4.4.3, has been reworked to update its compatibility and explicitness with respect to the current CIDOC CRM family of models and the ARIADNE Data Model, as well as considering FAIR data provisions.

Useful and restrictive aspects of semantic mapping procedures and tools for excavations were examined. As the excavation process implicates data descriptions from several domains, the meaning of several real-world entities or their documentation proxies can be mapped to different model concepts depending on the research context or stage. To overcome the problems, certain paths were explored:

- current tools to data mapping were tested for their functionality and potential usage,
- published research and online sources attempting to describe the excavation domain were examined, to detect how CIDOC CRM has evolved into an interrelated set of domain specific models,
- a spreadsheet approach to data mapping with basic data fields for describing, origin and target nodes as well as their linking paths were used, allowing the identification of re-usable paths,
- 3D content creation and usage were linked to archaeological reasoning processes.

This exercise provided certain realisations that have some wider implications to the current situation of excavation data modelling:

- a growing environment of models, implementations and ontological mapping tools may be difficult for the researcher to follow,
- data curation in the digital realm is constant, and that semantics mappings may also require updating, as standards can also evolve,
- data mapping implementations are context specific, i.e. depend on the focus of the researcher,
- the multiple instantiation of concepts can be useful in linking items that pose different conceptual meanings in different stages of the archaeological process,
- in the case of Paliambela Kolindros, the provision of layered content and the inclusion of the digital based recording and interpretation workflows provide the focus of reworking the original data model.

The presentation concluded by showing a possible description for digitally assisted excavation documentation and reasoning based on this dataset. Its main idea is based on the fact that later post-excavation study stages increasingly use digital datasets that substitute the material reality that was initially encountered in the field.

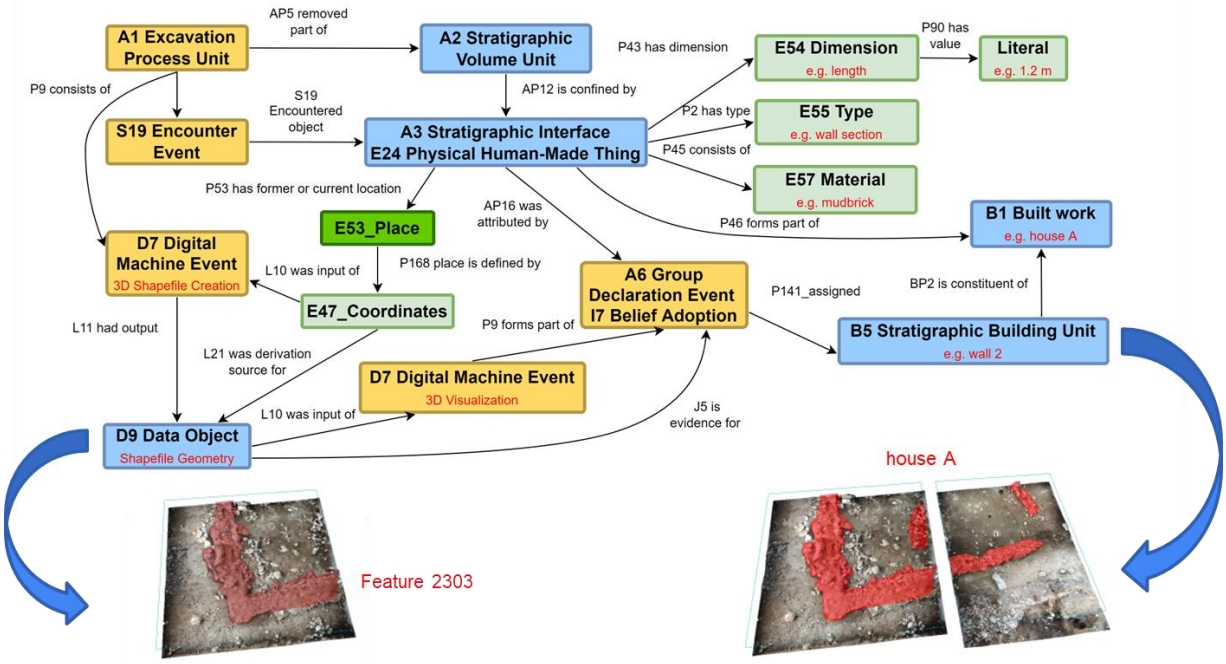

Figure 8. Reworking semantics to add digital processes and intermediate datasets.

3.7.1 Q&A with event participants

- *As suggested, going a step further from fieldwork recording and including the interpretation process, the more digital resources exist, the more the argumentation is based on these digital resources and not on the actual material recovered. This is something that might not have been targeted enough from the modelling side. The recording of archaeological evidence is somewhat fixed, but this further process of argumentation is not.* (G. Hiebel)
    - This was a necessity of the dataset itself. During post-excavation, data can be selected, regrouped and analysed, and then rejoined to the original dataset with added informational layers.
- *Perhaps the idea of multiple instantiations may foster such descriptions of layered datasets?* (M. Katsianis)
    - Multiple instantiations is a valid technique to be used in semantic mappings. Other ideas may include more complicated structures that, of course, have their own practical drawbacks. (A. Felicetti)
    - Multiple instantiation is a way to go, but I suppose that by splitting the dataset in smaller parts, it may be easier to make sense of data, identify patterns and re-use them. (M. Theodoridou)
- *How can proposed mappings be validated?* (M. Katsianis)
    - Make queries and get answers, that's the ideal way to validate. (M. Theodoridou)

  - Validation will happen at the end. You pose a question to the system and if the system replies in a satisfactory way and according to the criteria, then the modelling is validated (A. Felicetti)
- *Did you find any concept or information that you were not able to encode with any of the existing models or extensions in any way?* (A. Felicetti)
  - No, but I would be more satisfied with revising certain concept definitions or scopes. Whereas the logic of how stratigraphically things are created makes sense, I found more difficulty in descriptions from the point of view of the archaeologist that breaks apart the site into stratigraphic entities that need to be regrouped and reordered at a later stage. Notes will be made for informing the CRMarchaeo model.
- *If you go in the direction of Inference or Argumentation, the E13 Attribute Assignment may be a fitting way to go into this modelling domain. I would also suggest making a rough model and then maybe explore small parts of the dataset and work your way from the upper level down.* (G. Hiebel)
  - You are right, as I have found the *E13 Attribute Assignment* back in the modelling work for the CfA and realised that work from back then is still relevant. In order to retrieve patterns we also need to look at existing publications and see how things are described there.
  - CRMinf is great for describing argumentation processes, however the usage of *E13 Attribute Assignment* may be the correct level of representation in an archaeological context. CRMinf may be too far, as we are not building a syllogistic argument based on premises, inference logic etc. We basically apply an implicit logic procedure that requires to be linked to an individual, i.e. the archaeologist or the specialist. (G. Bruseker)

3.8 597 Norwegian excavation databases and CIDOC CRM - a practical exercise

C. E. Ore (University of Oslo) described the practical exercise of aligning 597 Norwegian excavation databases from INTRASIS using CIDOC CRM.

In Norway, INTRASIS has been the standard excavation documentation system since 2010: INTRASIS is extremely flexible and can be adapted to most excavation practices. This is done through a user defined metadata template, in which one can define object classes, subclasses, attributes and relations between the classes, as well as how these will be visualised on the maps. The flexibility also has a downside, since it can be hard to export data into a common database from a series of excavation databases based on adapted templates. As a result, a complete data integration requires some extra data cleaning.

As part of the ADED (Archaeological Digital Excavation Documentation) project (https://www.khm.uio.no/english/research/projects/aded/index.html) these datasets will be imported into a single searchable information system based on PostgreSQL/PostGIS. These data are not mapped to CIDOC CRM. A second track is the mapping to CIDOC CRM compliant data structures (XML intermediate format and RDF-triple store).

The first step in a mapping to CIDOC CRM is to create a common template. To do so it is necessary to get an overview of the variation in the templates and the actual usage in the databases, that is, which attributes, subclasses and relationships are actually used and create a conversion table. This conversion table will then be used for actual mapping of the data into the CIDOC CRM compliant XML-format.

This is done by creating a schema with the table structure of INTRASIS (https://www.intrasis.com/) with the following modifications: for each table we add two extra columns. The first is a new numeric primary key and the second is the name of the INTRASIS-instance the data come from. Then all the data in the 597 databases are copied into this new common database. On the basis of this common database it is easy to get an overview over the variation of names and identifiers in the templates and create statistics of the actual use of relations, classes and attributes in the various excavation databases. That is, the number of objects of a given (sub)class, which attributes are filled in and to which extent the relations are used.

Klassenavn	Class Name	CRM class	Type
Arkeologisk objekt	Archaeological Object	A8 Stratigraphic Unit	
Bergkunst_Motiv	Rock Art Motif	E73 Information Object	
Bergkunst_Område	Rock Art Area	S20 Rigid Physical Feature	
Bilde	Image	E36 Visual Item	
Båt	Boat/Ship	E22 Human-Made Object	
Dagbok	Diary	E73 Information Object	
Dokument	Document	E73 Information Object	
Funn	Find	E22 Human-Made Object	
Funnenhet	Find Unit	E22 Human-Made Object/ A8 Stratigraphic Unit	
Georeferanse	Geo reference	E73 Information Object	
Geo-objekt	Geo Object	E53 Place + E62 String (geojson)	
Graveenhet	Excavation Unit	S20 Rigid Physical Feature	
Hendelse	Event	E7 Activity	

Figure 9. Data mapping: INTRASIS class to CRM class

Based on the common database a flat table showing variations in the templates is created. Each row is a tuple (class id, class name, subclass id, subclass name, attribute id, attribute name). In addition, three extra columns are added: one with the list of databases in which these entities are defined, one with a list of where they are used and finally a list of the total number of instances of each entity. By adding columns for normalised values, the table serves as a tool for mapping the templates into a common template and systematically changing the values of the foreign keys in the data part. For the set of the 597 INTRASIS instances the table has 5,500 lines.

This is straightforward, but not very exciting, to go through each line and add normalised identifiers and names and took around two human weeks. The resulting table will be used as a conversion table in data cleaning and normalisation process. Although it was not difficult to

map the INTRASIS classes to CIDOC CRM, the mapping of the INTRASIS relations proved more challenging.

To link the excavation datasets, artefact information (from museums), site and monument information and excavation report archives one needs unique identifiers denoting the same item in all databases and the connection of catalogue information to common authorities. This is crucial but often neglected, due to lack of resources or ignorance. The slogan for linked data is data cleaning and digital discipline.

3.8.1 Q&A with event participants

- *We need to go through a data cleaning and data restructuring process in many cases, correct*? (M. Katsianis)
    - That's correct. We should instruct archaeologists on how to document or record their data, though this may be a risky experiment. Excavators don't consider events, but mainly documentation objects. It may be possible to employ semantic patterns towards that end.
- *What is the general impression of archaeologists? What do they expect to get out of this mapping exercise or data integration?* (G. Bruseker)
    - The project targets a PostgreSQL database with the datasets organised in a common semantic structure, so that archaeologists can try and find an element in that database. A general problem is that there is too little time and too little funding to analyse and align datasets. We are happy that it was possible to standardise a common INTRASIS template. The template may not be ideal, but it provides a common baseline and suits archaeologists.
- *What is the discussion in Norway in terms of modifying practice for excavation documentation?* (D. Lowerborg)
    - Archaeologists want to refine the current INTRASIS template to include information, such as who excavated what, which at present resides in free-text fields. In addition, all information about finds are stored in museum databases and the detailed catalogue of these finds are in another structure. These two need to be linked. Certain decisions need to be imposed from a managerial point of view.
    - To standardise free text expressions you could use a standard vocabulary and then use something like SKOS XL (https://www.w3.org/TR/2009/CR-skos-reference-20090317/skos-xl.html) to conserve the expression from the researcher (F. Hivert)
- *You raised a question about CRMgeo and GIS data. GIS data does not make sense in a triple-store. Postgres seems a much better option than in a GIS, because you can work much better with the relations and are not constricted by the INTRASIS system in terms of moving or ingesting the data. The mapping produced for the integration of the data makes sense, even if further effort will be required to add identifiers. Creating the identifiers and adding the concepts that relate the instances of objects of the same type to each other eventually will fulfil the added value of this work* (G. Hiebel)
    - You need to go through all the terms used in INTRASIS and different databases and map them onto some kind of vocabulary, such as the AAT. That's always a hard task to do. I remember, when I made the museum databases, that I was satisfied to end up with 25,000 different terms for artefacts from an original number of 80,0000.

3.9 Archaeological Excavations. Theoretical Patterns and Practical Recipes, Archaeological Interactive Report: a trait d'union between data

P. Derudas (Lund University) and **F. Nurra** (INHA) showcased the *Archaeological Interactive Report* (AIR), a blended online system for recording, collecting, managing the archaeological investigation data, and writing dynamic archaeological reports and other forms of editorializations. AIR provides a possible solution to overcome the limitations of scattered archaeological excavation data, by merging the 3D visualisation platform 3DHOP with the Content Management System (CMS) Omeka S (https://omeka.org/s/).

Design and development goals were:

- to advance a system for the documentation, management, and publication of data from archaeological excavations,
- to define a "working model" for an interactive archaeological report, meaning a solution that addresses the specific needs of field archaeologists.

The website of Västra Vång, illustrated within the presentation, is one of the websites published through AIR. It is the result of collaborative work between DARKLab at Lund University and the Digital Research Service (Service numérique de la recherche - SNR) at the National Institute of Art History in Paris (INHA).

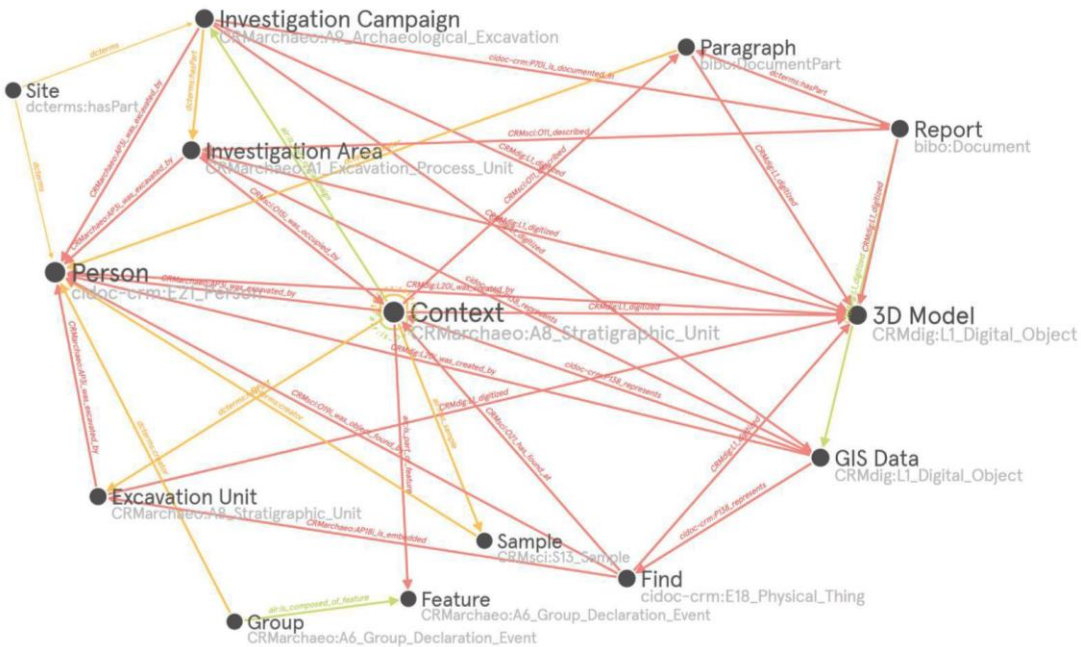

Figure 10. Data structuring in practice.

The presentation has described the steps taken towards the alignment of this model to the most recent archaeological semantic models, like the ones developed in the framework of ARIADNE plus, in order to store RDF data in a triple store, so that data can be queried from a

SPARQL endpoint. Provisions are also taken to map and align terms to the common reference thesauri as the AAT.

3.9.1 Q&A with event participants

- Is there a relevant data publication online? (K. May)
  - https://omeka.ht.lu.se/s/vastra-vang/page/home
- To what extent were you not covered by existing models or concepts that you decided to use auxiliary or custom models and concept definitions? (M. Katsianis)
  - We wanted to provide specialised descriptions that cover all the fields that are normally used within the INTRASIS documentation system. This was a requirement in order to standardise the documentation process in a detailed fashion. By adding this depth to description we wanted to increase the possibilities for the archaeologists to use the tool. We could also say that some of the custom concepts that we added are shortcuts to help archaeologists understand their value. Of course, having provided a data structure and since there is now an API that allows the export to JSON RDF, we can align and transform these nodes according to CIDOC CRM in an rdf format. The next step then would be to go through the SPARQL endpoint and triple store.
- *What was your experience with Omeka S? You can ingest vocabularies, but it can be difficult to ingest CIDOC CRM properties. Were the capabilities of Omeka S part of the modelling decisions made?* (G. Bruseker)-
  - In Omeka S in fact you can do it, but it increases model complexity, as an item needs to be created for each node developed. So, there is a sort of fractalization of the schema that becomes difficult for archaeologists to read or handle. The idea was to have a friendly tool for archaeological implementation and then put at the end of the chain a tool to align it in the form of a meta-model, so as to expose and store as an RDF format in a correct way. The use of the Dublin Core also facilitated web publication by creating the necessary meta-tag to the webpages for indexation purposes. In this respect, a balance was sought between achieving a data structure ready for the semantic web and the need to publish readable web resources.
  - There is a module that allows the addition of more semantics without sacrificing the readability of the model (https://forum.omeka.org/t/good-practice-cidoc-crm-data-properties-and-classes/13178/8) (G. Bruseker)

4 Discussion

The discussion took part in two separate periods during the virtual event. In the morning session, the discussion followed G. Hiebel's presentation and was related to the wider issue of aligning different excavation datasets as well as core excavation data to post-excavation data by-products from specialist accounts. Questions were discussed in the following order.

- *If you already have a well-structured relational database, what is the added value of creating RDF?* (M. Mori)
  - In RDFs the general scheme contains the semantic information, i.e. the way information is structured and understood. Also, the RDF allows extending a data structure, as well as joining data from multiple organisations. (G. Hiebel)

  - Using RDFs allows communication with other people and integration with other resources. (M. Theodoridou)
  - RDFs can also allow the detection of semantic patterns between mappings. (M. Katsianis)
- *Did the experience in France using aggregated data documentation schemas allow a better baseline for bringing data together?* (K. May)
  - Inrap is the biggest data collector in France, but other teams also exist that produce databases, meaning that working with everyone is still required. Even in Inrap there are different database structures that need to be aligned. (O. Marlet)
  - In Italy there is a national schema for documentation, but even there archaeologists provide their data in a minimum baseline, i.e. even if structure is imposed from above, that doesn't mean that everyone follows it. (F. Niccolucci)
  - Both, an idealist and a pragmatist argument could perhaps be made. The pragmatist argument is about efficiency. For example, Getty's solution was to use semantic structures from the start, so that it wouldn't be obliged to sustain hundreds of different databases and data structures. The latter is not practical. The idealist argument states that by committing to semantic compatibility, questions that are bigger than our individual data silos can be asked and their added value can be explored. Usually, it is suggested that data structures are the outcome of the unique or special nature of the data, but semantic work provides evidence that in general similar data exists and it should be compatible or linkable. (G. Bruseker)
- *Most archaeologists work with specialists. To what extent can semantic recipes be used to add specialists' data into wider excavation datasets? What is the current situation there?* (K. May)
  - ARIADNEplus tries to do that with the application profiles (APs) advanced for different domains. Not many examples are there, but we are working towards this end. (M. Theodoridou)
  - OpenArchaeo attempts to a certain degree to provide a minimum or basic data structure template (generic model) to bring together different types of data. More specific descriptions can follow different mappings of more specific implementations. (F. Hivert)
  - You can look for different parts of the overall information with the magnifier (e.g. a coin) or not (dataset). The system should be capable of incorporating all the details and the user should be able to employ only the part that is of interest. (F. Niccolucci)
  - If you link the notion of ARIADNEplus application profile with the work of the specialist, then applying a strategy of providing templates for data documentation can allow the integration of datasets within the same data pool. (G. Bruseker)

The discussion that followed the afternoon session took a wider take at several ontological issues. F. Niccolucci commented on the usefulness of the virtual discussion and suggested it be converted into a permanent point of reference for excavation data modelling and beyond. F. Niccolocci then went onto providing a key starting point for discussing excavation data modelling:

- *Can there be a "root" or "overarching" concept identified that acts as an entrance point to excavation data? Would this be a tangible (e.g. E77 Persistent item) or an intangible (E7 Activity) thing?* (F. Niccolucci)
  - Why do we need a root or overarching concept? (A. Felicetti)
  - Because otherwise we do not know what we are talking about. If we want to move onto a data-centric Archaeology, we need to model very basic concepts in archaeological research. (F. Niccolucci)

- If you put information coming from different sub-domains (i.e. a Republic of information), it may be politically incorrect to ask for a starting point. There is not a single starting point. (A. Felicetti)
- The answer may be given by ARIADNEplus and the concept of a "Resource", which may be many sorts of things. A resource might be a coin, a database, an entire site. (M. Theodoridou)
- On the one hand, Archaeology is a material science, and the starting point may be the materials encountered (tangible). But on the other hand, it is also a scientific activity in the present (intangible). Perhaps the starting point is what we do, while practising Archaeology. (M. Katsianis)
- I ask myself what are we talking about? Excavation produces and uses results that are relevant to other scientific processes. Can we separate the information created by excavation and how it is used in further deductions? If we manage to make an intelligent way of making deductions (e.g. an archaeological robot), what would be the information and what would be the deduction or the overlay? Understanding what the information is about may impact the way information is collected, organised and further used. So, how can we make sure that the links we choose are valid and others are not? (F. Niccolucci)
- There may be no single entry point. Data is defined by the way we are organising our questions and research. For example, there are siteless Archaeologies. There are theoretical and material aspects that can be very hard to get together. (G. Verhoven)
- So, if data organisation results from the research questions and if the same research questions are not met by others, then these data may not be used or needed? (F. Niccolucci)
- Data collection is steered by the research questions and scale of collection. For example, 3D models are also difficult to integrate or exchange because different models are built/used for different purposes or usages. (G. Verhoven)
- This may undermine the idea of data sharing, as data may not be relevant or even biassed and so their use may be misleading. (F. Niccolucci)
- I don't think it undermines it, but you have to ensure relevance of the scope and methods of data collection. This is a very interesting aspect of Archaeology, that we try to find the common schema. If you ask several people to map the same thing with CIDOC CRM you may come up with very different results. A priori knowledge and the cultural environment structure the way we understand our domains and how archaeological data are defined, experienced and collected. (G. Verhoven)
- When data is integrated new data questions can be defined. The idea would be to create agnostic graphs that analyse these integrated data. The reason we use ontologies is that they provide useful names for concepts for machine data processing. (A. Felicetti)
- "But why should I care for data collected by someone else for a different research question?" This was a criticism received by J. Richards and myself in a past article discussing the scope of ARIADNEplus by a reviewer. Possibly from the discussion today he was right. If there is a way of bypassing these obstacles, this virtual event provides the ground for further developments. (F. Niccolucci)
- It is useful to find common descriptions. There is always a choice on what to use and how to use it. At least we should be able to find potential data candidates and then decide upon their usage. We may be required to make these choices, even if a data analysis robot someday exists. (M. Katsianis)

5 Conclusions – Next steps

In conclusion, the workshop attempted to map the current environment with respect to archaeological excavation data modelling and highlighted the relevance of further conceptual modelling work in the context of item-level information integration at the ARIADNEplus portal. Many of the points raised and discussed are indicative of the way research is going to be structured in the next few years, as well as to the activities of the respective Archaeological Excavation Modelling Working Group. Key points include:

- An excavation cannot be rightly and fully understood unless it is connected or combined with information from other excavations and/or wider scale archaeological research data bodies.
- The provenance of excavation data is extremely important for contextualising and situating their production and construction. This realisation especially applies to digital documentation and analytic workflows creating complex digital outputs that often need to be traced back to fieldwork documentation.
- Despite the ongoing effort towards the standardisation of archaeological documentation practices, archaeological excavation research still often operates in an environment of non-standardized, inexplicit and under-documented data management practices. As a result, archaeological data are compartmentalised in several data silos, both in analogue and digital form, that prevent the full benefits of digital scholarship.
- Semantic data modelling may be a way to overcome methodological differences in excavation practice, speak a common language (or at least trace our thoughts to a common vocabulary pool) and create a backbone structure to integrate data created by different professionals, in different times, with different documentation means and targets.
- To advance semantic modelling in the archaeological excavation we need to focus on:
  - **Models**, by developing those that have the capacity to describe the application domain consistently (e.g. CRMarchaeo etc),
  - **Questions**, by identifying meaningful queries that can be posed on integrated archaeological excavation datasets,
  - **Methods**, by comparing existing excavation modelling examples, establishing modelling patterns and basic application scenarios, encouraging digital pedagogy and ontological thinking,
  - **Workflows & Tools**, by employing the major software tools that can be used in ontological modelling and data mapping and addressing their potential expansion or combination.
  - **Learning and training**, by providing educational material, digital facilities and teaching opportunities for semantic modelling.

  With respect to **Models**, the presentations showcased the extensive use of several models from the CIDOC CRM family (e.g. CRMarchaeo, CRMba, CRMsc, CRMinf, CRMdig, CRMgeo) and beyond. The evolution of CIDOC CRM has allowed domain specific definitions that are more successful in capturing domain-specific meaning. However, the fact that an archaeological excavation is a multi-level process, happening on multiple fields, using different methodologies and documentation media or tools, illustrates the difficulties still encountered in its complete description. In most presentations the main concern has been to

achieve baseline descriptions or a minimum level of general descriptions covering main aspects of the archaeological process that are compatible with CIDOC CRM, using a combination of concepts and models from its respective extensions. However, given the specificity of the CIDOC CRM extensions and their gradual evolution, there are occasions where different modelling versions of the same excavation process or concept may be equally valid and act in parallel due to the domain specific intended use. The possibility for the multiple instantiation of concepts was acknowledged as a useful approach for mitigating these problems. In addition, it was acknowledged that the more specific and all-encompassing a description of the excavation process becomes, the more difficult it becomes to stick to baseline descriptions. In this respect, there are many occasions where datasets are partly compatible with CIDOC CRM descriptions and employ custom ontologies or complementary semantic standards (e.g. Dublin Core) to better fit existing excavation methodologies or dissemination needs. This is not necessarily a drawback, as data interoperability is required to those information fields that make sense to be connected.

To understand which excavation information makes more sense to investigate in aggregated datasets, certain effort must be directed to formulating meaningful **Questions** and examining both their semantic syntax, as well as their performance with respect to the results returned. At a first level, the exploration of the kind of questions archaeologists want to pose to aggregated excavation datasets may help provide a basic abstraction level that can facilitate data linkages. At a second level, such questions can be explored as to their performance with respect to practically returning query results. After all, the ability of an integrated data system to correctly return results from multiple datasets, was recognised as the most decisive way to validate selected semantic descriptions. Attempts to harvest this possibility in practice were effectively presented in the case of OpenArchaeo, where a Javascript based- systems (SPARNATURAL) allows the user to query an RDF graph by assembling the data query parts using a graphic interface (i.e. keeping actual the SPARQL syntax in hide).

With respect to **Methods**, calls and well-established cases were also made for the purposeful bringing together of data mapping examples to explore similarities and differences, compare their scope and meaning and decide upon standardised semantic descriptions or semantic data “patterns”. Equally, the analysis of existing modelling examples and implementations has the potential to identify such modelling scenarios and provide a closer understanding of the evolution of the respective CIDOC CRM extensions. Such a strategy can obviously start from core or generic entities involved in the excavation process and subsequently be outbrached to cover more specific meanings. In all cases, these semantic patterns can then be documented and made available for use within the community as standardised data modelling *recipes* in the form of a modular data description building process. The gain of the approach may be two-fold. From one hand, it can enable a less challenging familiarisation with semantic modelling processes for domain experts, increasing the possibilities for a greater number of compatible implementations. From the other hand, it can provide critical studies of model definitions in an applied form and identify the problematic areas that require further development with respect to their ontological integrity or their compatibility with different archaeological excavation methods and interpretive procedures. A final comment can be

made with respect to archaeological terminology, as the use of knowledge organisation systems, such as thesauri, taxonomies, classification schemes and subject heading systems (e.g. ARIADNEplus requires mapping thematic concepts to the Getty's Art and Architecture Thesaurus - AAT and chronological descriptions to PeriodO - https://perio.do/) expedites vocabulary homogenization processes and data discipline, the same time opening up further engagement with the Linked Open Data community.

With respect to **Workflows and Tools**, several main pathways were identified with respect to semantic mapping.

1. Within the ARIADNEplus project a data mapping workflow is based on the X3ML toolkit, which comprises a set of small, open-source software components for information integration. The X3ML Mapping Definition Language provides a standardised mapping description for producing RDF in various serialisations, streamlines the URI Generating process and attempts to bridge the gap between human author and machine executor. The 3M Mapping Memory Manager is a web app providing a human-friendly interface for data mapping and a set of user input suggesting or validation facilities. It includes the RDF Visualiser, which provides validation facilities for transformed data and implements a configurable and detailed view of the RDF transformation. Finally, the X3ML Engine executes the transformation of source records to the target format by combining the X3ML mapping definition and the URI generation policy into an RDF document. Together these tools can be used within a complete workflow for transforming XML exports of datasets/databases into CIDOC CRM compatible RDFs. This suite of tools can be combined with the Vocabulary Matching Tool (VMT) by the University of Wales (accessed from within the ARIADNEplus VRE services and https://heritagedata.org/vocabularyMatchingTool/). The workflow, although specific for the purposes of data integration within the ARIADNE infrastructure, can be deployed online or locally.
2. Under the MASA consortium a workflow that covers the data-lifecycle of archaeological excavation data includes several steps and multiple tools, such as dataset cleansing (OpenRefine), ontology structuration (via a generic CIDOC CRM backbone model) and standard vocabulary alignment (AAT, Pactols, Geonames, PeriodO, VIAF) into an interoperable dataset that is mapped (Protégé-Ontop) and validated (SHACL) as an RDF TripleStore with a SPARQL endpoint. The generic backbone allows the linkage of several excavation datasets at the expense of the specificities of each dataset. The workflow manifests its efficiency in the gradual addition of external datasets into a running data pool that maintains a minimum level of accordance in its content.
3. Another option, within the *Archaeological Interactive Report* (AIR) attempts to standardise the archaeological publication process by providing an extensive alignment of archaeological information fields with multiple semantic models. This solution standardises the entire informational potential of the publication platform data schema, to allow further data integration with data aggregators using RDfs.
4. A different approach attempts the direct creation of RDFs from spreadsheets or databases. Datasets are analysed and rearranged into sets of spreadsheets or data tables that correspond with a combination of semantic standards. Tables are then integrated

within a database (Postgres) that allows further data structuring (e.g. URI addition). Afterwards, semantic tools (e.g. Karma, OntoRefine) are used for RDF creation. In this workflow, rather than the serial mapping or subsequent concept mapping of custom database XML exports (like in the X3ML workflow), semantic information required for creating the RDF is encoded within the data structures and subsequently transformed in an iterative manner that allows further corrections and structure re-assessments. This approach may be useful for aligning finalised or closed datasets. In many cases, though, similar processes that involve the alignment of several datasets with respect to both their transformation to a conceptual reference model and among themselves, may need to be complemented by decisions on what to include to the final deposition files, as well as significant manual data cleaning, conversion and alignment (e.g. the ADED project).

Several other data mapping solutions exist in between or even outside these four main workflows and from the presentations made, it seems that the current ecosystem of data modelling and knowledge organisation tools is far from standardised and in a state of rapid evolution through sets of interrelated open-source micro-tools that are usually selected based on the familiarity in their employment or the knowledge level of the data modeller. Pluralization is obviously beneficial especially in the domain of archaeological excavation that lacks the level of formality evidenced in other disciplinary domains. However, certain steps towards the further standardisation of certain workflows, the purposeful further development of selected tools by the archaeological community and the explanation of the benefits or overlaps of specific toolsets with actual examples is required.

Accordingly, the area of ontological modelling digital pedagogy received repeated attention during the event. Domain experts can easily get discouraged by the steep learning curve involved in all aspects of knowledge organisation systems (understanding the models, selecting workflows, using the tools). **Learning and training** opportunities should be targeted, if further audiences are to be attracted for building new or sharing existing semantically compatible archaeological excavation datasets. At the rookie level, the *CIDOC CRM Game*, either in the table or its online edition, can provide an entry point for domain experts who want to understand the mechanics and benefits of semantic structures. At a more intermediate level, the identification of excavation data modelling patterns may be compiled into the form of a *cookbook* for data modellers, allowing the creation of a kind of an archaeological marketplace. At an even more practical level, a guidebook with data modelling examples and workflows, complemented by structured tutorials in using existing tools can be seen as a major aid to potential data modellers. The creation of educational material may also increase teaching opportunities either in the form of research workshops (such as the ones organised in the context of ARIADNEplus) or even in the integration of the theme in archaeological curricula. As discussed by many experts during the event, training in archaeological knowledge organisation may bring together different views of the excavation universe and even impact archaeological theory and data management methods.

To conclude, the virtual event provided an opportunity to map the current research environment with respect to archaeological excavation data modelling and detect the main elements required for a potential roadmap to the item-level integration of relevant datasets.

A generic roadmap description could consist of a workflow attracting new implementation examples that starts from learning ontological modelling, moves to developing semantically informed excavation data models, continues to their transformation or mapping to existing semantic schemas and validating their applicability through query mechanisms.

At the end of the day, and following the more philosophical twist of the second discussion, it was generally appreciated that there is still much more to be done towards the integration of archaeological data. This provides additional incentives to further collaboration between domain experts and data modellers, as well as to bringing together existing and new modelling examples and workflows.

In terms of the next steps of the group, a series of working documents have been included in the ARIADNEplus portal aiming to: a) compile a reference list and a tool registry for excavation data modelling, b) collect data modelling examples, c) explore archaeological dataset questions and d) indicate problematic data encodings to inform relevant reference models.

A presentation by the group members entitled “Bringing excavation data together. Are we there yet and where is that?” that includes many of the points raised from the virtual event has been scheduled for the 28th EAA Annual Meeting in Budapest, Hungary, 31 August - 3 September 2022 at Session 273: *FAIRly Front-loading the Archive: Moving beyond Findable, Accessible and Interoperable to Reuse of Archaeological Data*.

Following the presentation, a final meeting of the group will be scheduled in Autumn 2022 to integrate feedback from all participants. The overall conclusions from the event, the presentation and the group’s meetings will inform the final report for ARIADNEplus.

6 Participants

George Bruseker is co-CEO of Takin.solutions, a semantic data management consulting firm. He is active in the semantic data modelling community in various roles including as Vice-Chair of CIDOC CRM SIG and member of Linked Art, Arches Resource Modelling Working Group and the Canadian Heritage Information Network Semantic Committee amongst others. Through Takin.solutions, he is the consulting semantic architect for Getty Digital. His focus is in the application, adoption and use of semantics to further cutting edge research in the humanities and sciences.

Paola Derudas is a PhD candidate in Archaeology (digital archaeology) @LU, a member of the DARK Lab and her dissertation project focuses on the definition of new forms of documentation, management and publication of excavation data.

Gerald Hiebel is a senior scientist at the University of Innsbruck and affiliated with both the Department of Archaeologies and the Digital Science Center (DiSC). He is also a member of the Research Center Digital Humanities. His academic work centres on methodologies that can be used to represent the complex knowledge arising from a (pre-)historic reality and the scientific research about this reality.

Florian Hivert is a research engineer in semantic web at the Maison des Sciences de l'Homme Val de Loire, in Tours. His work focuses on the mapping of data sets to the platform OpenArchaeo, on reflections around the data modelling of humanities with standards like the CIDOC CRM, and on how to query the RDF graphs optimally.

Markos Katsianis is Assistant Professor in Antiquity and Digital Culture at the University of Patras. His research centres on the application of digital technology in Archaeology with a special focus on GIS, 3D modelling, data preservation and re-use.

Vangelis Kritsotakis is a Research & Development engineer in the Information System Laboratory of the Institute of Computer Science, FORTH. His main interests include information management systems, conceptual modelling, knowledge representation and cultural informatics.

Olivier Marlet is a French researcher engineer at the Laboratoire Archéologie et Territoires, in Tours (CNRS, University of Tours). He leads the working group on the Semantic Web within the MASA consortium in France. This consortium supports archaeologists who want to bring their data to the Semantic Web.

Denitsa Nenova is a co-founder and a co-CEO of Takin.solutions Ltd. Her academic and professional background is in archaeology, database solutions for Cultural Heritage (CH), and Geographic Information Systems (GIS). Her main research interests are in the fields of digital archaeological documentation, semantic reference data models and their archaeological application, landscape archaeology, aerial photography and airborne-scanning technology, photogrammetry, Mediterranean and Balkan Prehistory, and archaeological ceramics.

Franco Niccolucci is the coordinator of ARIADNEplus, the European research consortium that targets the creation of a research infrastructure on archaeological data and the supporting force behind the present event. He is the Director of VAST-LAB research laboratory at PIN in Prato, Italy. He has coordinated several EU-funded projects on the applications of Information Technology to Archaeology. His main research interests concern knowledge organisation of archaeological documentation and the communication of cultural heritage.

Federico Nurra is Head of the Digital Research Service at the French National Institute of Art History. Between 2015 and 2018 he worked at the French National Institute for Preventive Archaeological Research (Inrap), within the ARIADNE project. His main research topics are digital development and data management applied to the protection and valorization of cultural heritage.

Christian-Emil Ore is Associate Professor and Head of the Unit for Digital Documentation (EDD) at the University of Oslo. He has worked with digital methods in the Humanities for 25 years. He is currently participating in the Norwegian infrastructure ADED project (Archaeological Digital Excavation Documentation) with the objective to create a common open repository for archeological data from excavations and from museums and archives.

Giorgos Styliaras is Associate Professor at the University of Patras. His research focuses on the application of multimedia systems, cultural technology and spatial hypermedia in the field of Cultural Heritage.

Maria Theodoridou is a Research & Development engineer at the Foundation for Research and Technology - Hellas, Institute of Computer Science. She is the Vice Chair of the Centre for Cultural Informatics and coordinates the mapping technology activities of CCI. Her research interests include semantic web technologies and semantic interoperability.