Open Research Data of the Finnish Academic Institutes Abroad: Pilot projects

Harri Kiiskinen Laura Nissin Manna Satama

2020

Researchers in pilot projects:
Suvi Kivimaa
(*Segregated or Integrated? Living and Dying in the Harbour City of Ostia, 300 BCE – 700 CE*)
Sirpa Ollila
(*I bolli doliari romani dell'Italia centro-occidentale*)
Kira Pihlflykt
(*Digitizing the Hilma Granqvist Archive*)
Tommi Turmo
(*Archaeological Sites in Thesprotia*)

Suomen Ateenan-instituutin säätiö sr – Suomen Japanin-instituutin säätiö sr
Suomen Lähi-idän instituutin säätiö sr – Säätiö Institutum Romanum Finlandiae sr

Pilot Projects of the Finnish Academic Institutes Abroad Open Resaerch Data
2020
Harri Kiiskinen (ORCID: 0000-0003-4187-5551)
Laura Nissin (ORCID: 0000-0001-8050-3916)
Manna Satama (ORCID 0000-0003-3775-9363)
Tiedeinstituuttien avoin tutkimusdata -hanke
Institutum Romanum Finlandiae sr
Suomen Ateenan-instituutin säätiö sr
Suomen Lähi-idän instituutin säätiö sr
Suomen Japanin-instituutin säätiö sr
DOI: 10.5281/zenodo.4118489
Keywords: Research Data Management, RDM, Report

Contents

Chapter 1

Pilot projects in the context

The project *Finnish Academic Institutes Abroad Open Research Data*, funded by the Finnish ministry of culture and education, 2018–2020, had as its aim to develop a set of policies and best practices for digital research management in the Finnish scientific institute abroad. The institutes participating in the project were the *Institutum Romanum Finlandiae*, the Finnish Institute in Athens, the Finnish Institute in the Middle East, and the Finnish Institute in Japan. *Institutum Romanum Finlandiae* was the co-ordinating organization.

At the outset of the project, it was deemed important to focus on practical issues of digital research data management. The institutes are very small organizations, and a set of documented abstract policies would have been of little use for these organizations; instead, a set of documented experiments with real data from the institutes was deemed to be useful for gaining understanding on the practical issues of digital research data management in all its stages.

For this purpose, four pilot projects were selected.

The projects had different aims, and all aimed to produce usefuls experience in managing research data. The particularities of each pilot project are described in connection of each project below.

Chapter 2

Segregated or Integrated? Living and Dying in the Harbour City of Ostia, 300 BCE – 700 CE

The pilot project *Segregated or Integrated? Living and Dying in the Harbour City of Ostia, 300 BCE – 700 CE* in integrally connected with the research project lead by professor Arja Karivieri, the director of the *Institutum Romanum Finlandiae* (2017–2021).

2.1 Purpose

The purpose of the project was to figure out a solution for long term storage adn publication of the digital photographic collection of prof. Karivieris project. The collection contained mostly photographs taken by prof. Karivieri herself, but is also contains digital versions of the collection of old photographs relating to Ostia owned by the institute. These photographs were digitized as part of the pilot project.

The role of this pilot project was to investigate the possibility of using an institutional solution for managing research data.

2.2 Premises

The dataset contained a collection of originally digital photographs. In addition to these, digitized versions of old photographs from the institutes own archives were appended to the dataset.

The photographs were described using a set of descriptive metadata, including the main object depicted in the photograph as well as a geographical location of the object described. The locations were marked using the standard Ostia Antica addressing scheme, similar to the one used also in Pompeii. In this scheme, the town is divided into parts composed of blocks. The houses are identified with a numbering scheme where each door and opening that opens to the streets surrounding the block is given a number. In addition to these, certain houses have room numbers, which usually are not standardized but follow a scheme adopted in a particular publication.

Specific buildings usually have names, and both English and Italian nomenclature is well known.

Table 2.1: Metadata fields describing the photographs.

Metadata field name	Description
PhotoID	Unique ID for the photo; also used for labelling also the digital file.
Photographer	Name of the photographer.
Source	Possible external source for the digital object.
Photo date	
Subject	Name of the house, god on statue, etc.
Type of object	If the photo depicts a specific object, this field contains the type of the object, using a Taxonomy.
Inventory number SBAO	
Location	Specific reference to the location where the photo was taken.
Findplace	Specific reference to the place the depicted object was found at.
Room	Room in a spceific house, using a scheme from the publication in Plan reference.
Area	A specific area. Used e.g. in Isola Sacra to refer to tombs.
Plan reference	The plan used for detailed spatial description of the structure.
View	Direction of the photographic view.
CIL	Reference to CIL.
AE	Reference to AE.
EDCS	Reference to EDCS.
PHI	Reference to PHI.
Bibliography	Bibliographic references.
Approximate dating	Dating, using approximative expressions.
Description	Free-form description of the image.

In addition to architectural views, the collection also contains images of items, statues, and other non-structural elements. These have a set of own metadata, such as Object Type. In many cases, there are references to publications and external references.

The photographs are in the form of JPEG files, except for the digitized old photographs that in TIFF.

The metadata describing the photos is created in spreadsheet files, where the columns are the data fields. The metadata fields are described in Table 2.1.

2.3 Plan

The project plan was to have a trainee documenting the photographs and then to find a solution for storing the photographs in an online repository or store where they could be further used.

The solution chosen for institutional data storage was Islandora, a data management module designed over Drupal, a widely used digital content management system (CMS).

Islandora features a basic set of metadata that caters to a wide variety of needs. A perfect match, however, between the Ostia photographic data and the Islandora default repository object content model was not possible to make, so a set of new fields and in some cases, related Taxonomies, were created. The set of Islandora fields and custom fields that was used for importing the photograph data is shown in Table 2.2

Table 2.2: A mapping between the field used in Islandora and the fields of the metadata. The field "Custom?" marks the fields that were added to the original Islandora configuration. The field Taxonomy describes the taxonomies used for the specific fields.

Islandora field	Photo descriptions	Custom?	Taxonomy (default with *)
id	PhotoID		
parent_id	(collection)		
field_weight	(serial)		
field_identifier	PhotoID		
title	PhotoID		
file	(Path to file)		
field_model	(islandora model)		Islandora Models
field_member_of	(empty)		
field_description	Description		
field_linked_agent	Photographer and Editor		Corporate Body, Family, Person*
field_edtf_date	Photo date		
field_subject	Subject		Corporate Body, Family, Geographic Location, Person, Subject*
field_object_type	Type of object	x	Object Types
field_sbao_invno	Inventory number SBAO	x	
field_geographic_subject	Location + Room/Area		Geographic Location
field_find_location	Findplace	x	Geographic Location
field_plan_reference	Plan reference	x	
field_view_direction	View	x	View Direction
field_reference_cil	CIL	x	
field_reference_ae	AE	x	
field_reference_edcs	EDCS	x	
field_reference_phi	PHI	x	
field_reference_bibliographic	Bibliography	x	
field_temporal_subject	Approximate dating		Temporal
field_source	Source	x	

The use of taxonomies is especially recommended becaus it allows for easier searching and grouping of the data.

Islandora was also the suggested solution for the image storage for the *Bolli doliari* pilot. In that case, the individual photographs have so little metadata that the basic Islandora metadata scheme was deemed enough.

In addition to the online publication of the data on the Islandora platform, the plan was to store the complete dataset including the metadata in a long term research repository, of which Zenodo.org would be the best option. However, before doing this stage, the rights related to the photographic material should be well defined.

2.4 Implementation

The work itself consisted of two parts. First, a trainee / researcher was hired to process the images and to document the metadata. Then, the project researcher set up the Islandora on a cloud server, and created the necessary data import system for reading the metadata and image files to Islandora, and processing the metadata appropriately.

The setup process includes severeal stages:

1. Create the server on the cloud service
2. Set up necessary user accounts and permissions
3. Set up a customized Islandora playbook for installing in an online environment
4. Run the playbook using ansible, a simle IT automation engine.
5. Set up post-installation settings for server, like mail server and SSL
6. Create the required set of metadata
7. Import the data using Islandora Workbench tools

2.5 Successes and failures

The main successes of the project are:

- The photographs with their metadata was successfully imported to the server.
- The Islandora server was set up in suvh a way, that searching the data is possible.

The main failures of the project are:

- The geographical linking of the photographs was never fully done.

2.6 Achievements

The pilot managed to show, how an institutional photo archive could be created also using existing data. The resulting storage is suitable for items other than photographs as well. This was not tested yet.

2.7 Current State

Currently, the photographic collection is stored on an Islandora installation running at `https://irfimages.tatd.ovh/`. The installation runs on a cloud-based VPS, on Ubuntu 18.04 LTS. The collection contains all photographs and metadata provided for the pilot. Several of the metadata fields were turned into vocabularies during the import process, meaning, that the set of terms describing the photographs can be browsed and edited..

2.8 Future work and proposed follow-ups

In case the Islandora system is deemed suitable for the purpose, the next logical step would be to integrate this as part of the Institute's IT architecture. There are two options for doing this, depending on the possibilities of the institute's own service provider:

1. To take control of the current virtual server, and add it to the services controlled and serviced by the IT service provider. The ownership of the cloud server can be transferred to the institute, and control access to the IT service provider's personnel.

2. To create a similar Islandora system on the service provider's infrastructure. This could work in the case the service provider already runs an Islandora system that could be used for this purpose.

At the time when the use rights of the photographic dataset are agreed upon with the local authority, the long term storage option has to be negotiated, especially, whether it is the institute or the local authority who is responsible for long term storage.

The agreements must include stipulations regarding the long term status of the photographic material. Is it closed or publicly accessible now? If the access has restrictions, who has the authority to grant access and on what terms? How long will the access be limited, and at what point can the dataset be opened for public use?

Chapter 3

I bolli doliari romani dell'Italia centro-occidentale

3.1 Purpose

The purpose of the pilot was to find a solution for long-term preservation and publication of Professor Eva Margatera Steinby's database *I bolli doliari romani dell'Italia centro-occidentale*. The database consisted of several thousand stamps that were pressed during production on bricks and tiles, *dolia* and other large pottery items.

The original data was in the form of several Microsoft Word documents that were used by prof. Steinby for recording the stamp data. The stamp records were structured in a formal manner, and were in cleartext, meaning, that stylistic information was not used for semantic markings.

The stamp database already has one digital publication in the form of the site http://www.bollidoliari.org/. This site is and important part of the original dataset, for that is the place where the photographs of the stamps, and the information connecting each stamp to photographs and other graphical representations is stored. The MS Word documents by professor Steinby do not contain this information, and therefore, the original data exists in two places.

The current form of data publication is as follows:

1. Prof. Steinby stores the information in her private Word files and records the changes made to the file
2. The file is sent periodically to an assistant who transfers the changes to the files on the web site

3.2 Premises

A set of premises was defined that must be taken into account while developing any solutions for the pilot:

1. This was an ongoing process that had been running for years and would not be completed before the end of the pilot. Therefore, the dataset could not be seen as a set of static files that must be stored but a set of dynamic, changing files.
2. Any solutions offered must be dynamic and allow for continual updating of the data.

3. The main researcher, prof. Steinby, would not and could not be expected to adopt any new technologies at this stage. Her only aim was to document the stamps as fully and completely as possible while working in a way that has been well defined over the decades of her work.

3.3 Plan

While looking for suitable solutions for this pilot, the guiding principles were the FAIR-principles. Any proposed solution should address the issues of Findability, Accessibility, Interoperability and Reuseability. As is common to all pilots in this project, long term preservation and stability are not easily combined with accessibility and useability. Therefore, the solution planned for this pilot includes two parts: a server implementation focusing on the accessiblity and useability of the data, and a long term storage solution focusing on the preservation of data. In this case, this resulted in various factors that were used in considering options:

The current format of the data was already rather good. Even though the original data files were stored as Microsoft Word documents, they were so well structured that a convertion to a text-only format did not lose any information. The stamp descriptions were structured in a well known structure inherited from the first stamp publications. As such, it was very accessible for the researchers familiar with the data. The documentation was written in Italian, which is very relevant in the context of Italian pottery studies. From the viewpoint of the traditional researcher, the data was very useable.

However, in the context of digital research tools, there were certain features that were lacking.

The earlier stamp publications, like Steinby's own *LSO*, had included extensive indices which allowed to search for stamps in the database based on fragmentary information. Compilation of these indices was time consuming and difficult, and even then the results were unable to answer questions such as how to find all possible stamps when the beginning of a word was missing, i.e., when the first letters of a word were not known. Also, information regarding the physical form and extent of the stamps was well recorded, but currently, there was no way to systematically search for this kind of information.

There existed and earlier database schema that had been created for storing the stamp data. This schema was obviously based on a traditional relational database and on the concept of fields where the data was stored for each stamp. As is often the case, the real data is more complicated than a simple schema, and this route did not seem the best one to follow. Additionally, it was made clear by prof. Steinby that the stamps should not be considered similar to, for example, funerary inscriptions, because for each stamp, its position in the corpus as a series of stamps was as important as its independent data content. This made the stamp data very different from other inscriptions that are individual documents. In this case, it became important not only to record the data content of the individual stamps, but also the ordering of the stamps into corpora.

It soon became evident that a very suitable format for describing and documenting the brick stamps could be Epidoc. Epidoc is a TEI specialization that is developed for describing and documenting inscriptions and other epigraphical material. It has been originally developed for the digital editions of ancient inscriptions, and is used for publications such as the Vindolanda Tablets, and most recently, the *Inscriptiones Siciliae* -database[1] (https://isicily.org).

EpiDoc files are a starting point for digital publication. They allow for the transcription of original texts and for adding the documentation and commentaries regarding the texts, but the resulting files are just XML-files. These can be used as they are by anyone with experience in XML but for a typical

[1] Prag 2020.

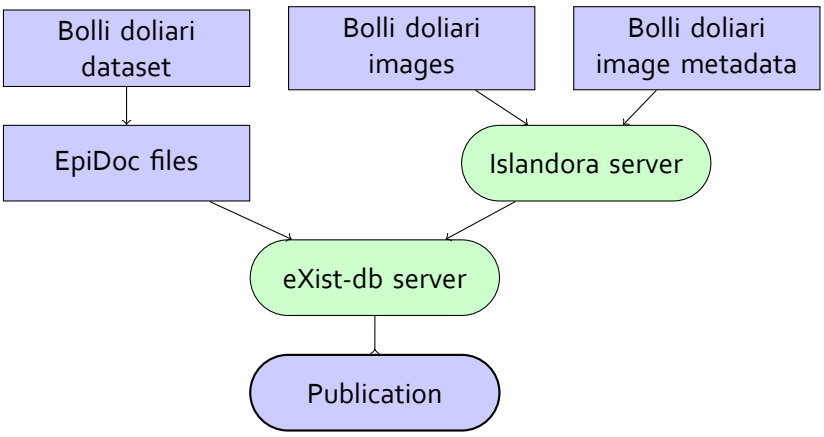

Figure 3.1: Structure of the Bolli doliari data process

humanities researcher, they are not yet easily accesible. In addition, the resulting XML files do not constitute what is commonly understood by publication.

The EpiDoc work group provides a tool for viewing and displaying EpiDoc XML documents called EFES but that tool is not suitable for production use in a public environment. A valid option was found in eXist-db that has been used for other TEI-based publications.

The Bolli doliari -database currently in use also includes ca. 2700 photographs and other graphical representations of the stamps. Originally, it was thought that these could be handled with the same publication system but in light of the FAIR requirements, it was realized that for properly storing these images, another option would have to be selected. Intuitively, it would feel simple to just store the image files in a directory on the same server running the publication platform and just to publish the images from there. At this point, it was realized that as with all research data, we were dealing with more than just image files; instead, we were working with photographs and drawings. A photograph, or a drawing, always has a maker with a set of Intellectual Property Rights for the graphical representation. According to the Finnish law, these cannot be waived. In addition, several of the graphical representations used in the Bolli doliari database were from earlier publications, being used with permission.

The "image files" thus could not be treated just as files. They needed accompanying metadata to document their origins.

Since this part of the Bolli doliari dataset was very similar to the data that was used in the Ostia pilot (see Chapter 2), it was decided that the same platform could also be used for storing the Bolli doliari images.

Thus the structure of the process was defined as can be seen in Figure 3.1.

So the intention was to do the work in following stages:

1. Convert the stamp test databases to EpiDoc/TEI XML
2. Import the EpiDoc XML files to the eXist-db server and configure the server appropriately for publishing the stamps
3. Import the Bolli doliari images to the Islandora server with added metadata
4. Link the images to the stamp files from the eXist-db -server

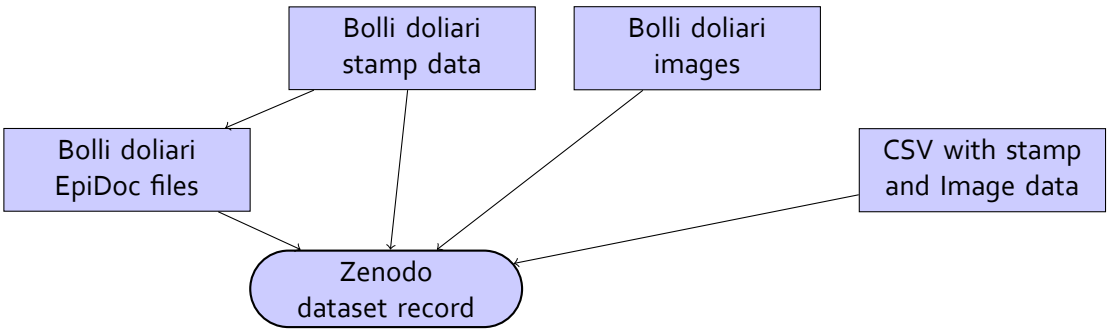

Figure 3.2: Bolli doliari Zenodo data storage

5. Publish the database as one online resource

Long-term preservation at Zenodo

The above described solution anwers to the needs of useability of the data by other researchers in various aspects. An eXist-db -server running on a Cloud VPS does not answer the requirements for long term preservation, though, and for this purpose, a storage of the main data files at zenodo.org was envisaged.

3.4 Implementation

When the pilot project was planned, it had not become clear that the dataset was dynamic. The original concept was based on the idea of static files that should and could be permamently converted to EpiDoc and this format could be used as a basis for further use. While EpiDoc would be a perfect format for working with epigraphic editions, it requires a very different set of skills and experience, and while it is the recommended way to write text editions for researchers aiming for long-term presence in academia, in the case of prof. Steinby it was not possible to even consider moving to a wholly new set of skills and techonolgies. This required a re-evaluation of the plans.

The original data must be kept in the form of text files. Also, it was soon realised, that the only place where the information connecting the individual stamp data to its graphical presentation was on the current web site, in the form of hyperlinks to the image files. In addition, the current publication process was streamlined and functional, and filled the basic needs for stamp data publication, its replacement with something that was not yet functional nor reliable and perhaps not even implementable could not be considered as an option.

The solution therefore had to be found in a process that uses the current stamp data publication process and the data it produces as the starting point.

In the first round of work in the Spring 2019, the trainee worked with the stamp data by manually converting stamps to EpiDoc XML, producing important and interesting results.

In one month, the trainee was able to convert ca. 100 stamps to EpiDoc. Based on her experience, after getting well versed in editing XML and learning to transcribe the stamps, it was estimated that the transcription of a typical stamp could perhaps be done in ca. 30 minutes, perhaps allowing for 15 stamps / day. There is no decent figure of the size of the whole corpus. There are ca. 2700 graphical representations, but not all stamps have images. On the other hand, some stamps have two images. A conservative estimate could be that there are ca. 5000 stamps altogether in the dataset, meaning,

that their conversion to EpiDoc would take $5000/15 = 333$ work days, i.e. ca. a year and a half full time work. Manual conversion was thus deemed impossible to achieve without extensive external funding.

However, the structure of the stamp entries was very formal. The information appeared in exactly the same order and was (or should be) written in exactly the same form in every stamp. This made the implementation of a parser for reading the stamp database possible. Once this parser was implemented, it could be used to read the data, which in turn could be used for producing EpiDoc XML files. The conversion script was implemented in Python. For the parser, the library Parsimonious was used, and in some parts of the XML production, Lxml.

Since the linking information between the stamps and their graphical representations was present only on the current web site, the conversion scripts were written so that they could be used for reading the current www-pages, parse the data on the pages and convert it to EpiDoc XML. However, since the parser is very strict about the correctness of the stamp entries, the work process was defined to include an intermediate step where the www-pages were downloaded to the local computer and converted to plain text files. The various errors could then be corrected in these files, which could be used for producing the resulting Epidoc XML files.

In addition to the Epidoc XML files, a script was written that parses the whole current Bolli doliari site for images. It reads each image link and stores information regarding that image to a CSV file that can be used in importing the images to the Islandora server. The corresponding migration definitions for Drupal (the CMS system on which Islandora is implemented) are included in the same Github repository.

In the second phase, from February to April 2020, the researcher hired for the pilot collected all sources for the graphical representations of the stamps so that informations regarding the image sources could be joined with the images themselves in the image database. In addition to this, her work consisted of running the stamp data files through the parser–converted in order to find typographical and formal errors in the data.

3.5 Successes and failures

The main successes of the project are as follows:

Automatic conversion of stamp data The parser / converted was not particularly simple to write, but the results are good, and the parser is able to read a variety of data. There are very few absolute requirements for the original data, and most of the information can be considered optional. The parser / converter was also modified to read stamps that consisted of more than one separate individual stamps.

Use of EpiDoc as stamp documentation format EpiDoc seems well suited for documenting this kind of data. It has several strengths in relation to stamp data:

1. It has very strong support for text critical annotations
2. It supports "tagging" individual words and concepts, both in the stamp text as well as in the commentaries. This tagging could be done either manually or using a NER (Named Entity Recognition) system for person names, place names, etc.
3. XML-based tagging and semantic structuring allows for context-aware indexing and faceted searchs.

Main failures of the pilot project are as follows:

Failure to process all data The amount of data in the stamp collection was too large in relation to the resources reserved for the pilot, and thus, it was not possible to process all stamp data

Conversion not perfect The conversion of the diplomatic text editions of the stamps was not perfect. In an optimal situation, the resulting EpiDoc files would contain one diplomatic text that is used for producing both the diplomatic edition as well as the interpretation. In the original data, these two are separate and an automated conversion that merged these two data is at least very difficult to produce, and may be very close to impossible. The current solution was to include both from the original data and to store the interpretation as "translation".

Communication and final product The pilot project failed to communicate satisfactorily with the material owner and the proposed solution failed to fill the owner's expectations regarding the visual presentation of the material.

3.6 Achievements

The main achievement is the creation of the automatic conversion process that turns the old text-based brick stamp database to TEI/Epidoc XML documents with little manual work involved. Also, the importing of the whole stamp photography collection to the Islandora setup described in the Ostia pilot can be considered an achievement..

3.7 Current state

Current state of the data is as follows:

Processing tools for the data exist in the github repository. These can be used the convert the stamp pages on the current site to EpiDoc XML files.

There was a test server running eXist-db at address https://bollidoliari.tatd.ovh/ with a rudimentary application for Bolli doliari database installed. The XSL styles used for rendering stamp XML are modified to use the image data that has been uploaded to the Islandora system for Ostia pilot (see Chapter 2), but at the request of the material owner, this server was closed.

Currently, the data exists only in the old system. Tools for conversation to Epidoc XML are completed but the environment for publishing the resulting files requires further work. It includes complex searching tools and automatic bibliographic links and citation indices, but the presentation of invidual stamp data is not satisfactory yet.

3.8 Future work and proposed follow-ups

The main future work, if permission from the data owner is gained, should concentrate on polishing the presentation on the eXist-db server. This is mostly a question of tweaking the XSLT files and should not required more than a few weeks. Further work, requiring more time, would be to create tools for editing and updating the stamp data directly on the server, either directly editing the XML files (easy) or using a non-technical user interface (more difficult). This work could integrate semantic annotation of the data (persons, geographical locations, etc.), and in the long run, should integrate the current and future work done with the Epigraphic Ontology.

Also, independent of whether Epidoc is chosen as working format for the data in the future, the whole current dataset should be stored somewhere secure, like Zenodo.org. The dataset produced by the current conversion tools is almost ready to be stored as such.

Chapter 4

Archaeological Sites in Thesprotia

4.1 Purpose

The purpose of the pilot project of the Finnish Institut at Athens was to investigate the preservation and publication of archaeological research data in a form that would preserv the internal complexitiy of archaeological excavation documentation.

The pilot aims to publish the archaeological documentation from the Gouriza excavation site. Gouriza is located in Thesprotia, in the valley of the Kokytos-rivera, and the recent excavations at the site brought to light one of the biggest owens ever discovered in Greece. The owen had been used in production of roof tile during the 4th c. BCE.

4.2 Premises

The material consisted of digital photographies and drawings, as well as a publication using this documentation.

4.3 Plan

The plan was to use the Arches system for storing the documentation. Arches is a web-based, geospatial information system for cultural heritage inventory and management. It is especially interesting for its strong inclusion of CIDOC-CRM cultural heritage ontologies, which would provide a good system for documenting archaeological excavations.

4.4 Implementation

On the whole, this pilot was very complicated. The chosen solution, Arches, was found only later on in the project.

It was discovered early on in the project that CIDOC-CRM would provide a very good structure for the documentation model and that using that ontology the documentation would really achieve another level of semantic interoperability and re-usability. CIDOC-CRM integrates the logic of cultural heritage management in a particular way. It does not focus on a static description of the object but provides schematics for describing the processes affecting the described object. In archaeological context, this means that CIDOC-CRM is suitable for describing the actual process of archaeological

excavation. In the logic of CIDOC-CRM, for example, a description of an object is not one and definite; the description is the result of an action of describing by an actor, and consequently, another act of description by another actor could well result in another description which may be similar or different, depending on the particularities of the second actor and the act of description. The point is to document these acts of description.

This sounds rather complicated but is actually something that is often done during an archaeological project: the identity of the person who does the description is often implicit in the material, perhaps in the form of the organization where all descriptions of finds are done by the same person, and therefore, this information is not explicitly documented anywhere in connection of the actual descriptions.

The challenge with CIDOC-CRM is that while it is conceptually sound and intellectually invigorating, it is rather complicated to use in practice. While the scheme itself can be understood and the concepts defined in it can be used to link pieces of information together, the tools that could be used to achieve are not really that common.

A solution is offered by the Arches Project (`https://www.archesproject.org/`) developed by the J. Paul Getty Trust and the World Monuments Fund. Altghough not strictly built around CIDOC-CRM, the Arches system has very strong support for managing data using the CIDOC-CRM ontologies. Therefore, it was decided to try to import the Gouriza dataset to an Arches instance. This was also considered an useful excercise because currently some other institutes in Athens are testing their own Arches-based solutions.

An Arches server was set up on a Cloud server at `https://arches.tatd.ovh/`. A set of Resource Types was created to reflect the data in the Gouriza dataset; see Table 4.1 for a description of the Resource models, their data fields, and the CIDOC-CRM classes used.

Once the Resource models were created, the data could be uploades according to the documentation of the Arches system. (`https://arches.readthedocs.io/en/latest/import-export/`)

Geographical data and map data

Arches is able to load geographical datasets in the form of ArcGIS shapefiles. The shapefiles need to be prepared suitably and the geographical data in the file must be in the correct projection (EPSG:4326 to be exact). This process was used for importing the Finds data, and the result can be seen as both individual Find resources in the database, as well as an "overlay" map that shows all Finds on the map window. The process is quite straightforward, and import of digital field measurements should be easy; sadly, the Gouriza dataset did not include much of GIS data that could be used in this way. Ready-made plans are much more complicated.

When presenting excavation data, the ready made plans, be they drawn by hand or prepared with a GIS, are very useful, for they show an interpretation of the site that is much more accessible to everyone who does not yet know the site well. This is not even a question of professions vs. amateur, because even for a GIS professional, it is much easier to gain an overall understanding of a site when looking at a plan that has already been prepared. Therefore, the publication of raw digital measurement data is never enough and any decent publication must also include real plans.

This proves to be much more complicated using Arches. The system by itself is not able to use raster data for plans even if they are georeferenced. Arches relies on an external map server for this type of data. In addition, the use of an external map server also allows for a wider variety and more complex vector based data used for basemaps and overlays.

The Gouriza data included one raster drawing, the so-called Figure 13, and for testing purposes, it was deemed worth the while to install a map server so that the use of these maps could be tested.

Table 4.1: The Arches Resource Models used in describing the Gouriza data set.

Resource model	Data Field	Comment
Excavation (A9)	Excavation Project Name (E41)	Gives name to resource (P1)
	Excavation Site (E27)	Links to Site (P7)
Site (E27)	Site Name (E44)	Gives name to resource (P87)
	Geographical Location (E53)	Link to map location (P156)
Organization (E40)	Name (E82)	Gives name to resource (P131)
	Member (E21)	Links to Person (P107)
	Institutional Role (E55)	Role of member, links to vocabulary (P2)
	Period (E4)	Period for this role (P132)
	Start date (E63)	Beginning date of period (P116i)
	End date (E64)	End date of period (P115i)
Person (E21)	Project Public Name (E82)	Gives name to resource, for public use (P131)
	Address (E45)	Contact details (not used) (P76)
	Date of Birth (E67)	(not used) (P98i)
Image (E38)	Kuvan nimi (E75)	Gives name to resource (P149)
	Find Category (E55)	Gouriza vocabulary (P2)
	Illustration Type (E55)	Gouriza vocabulary (P2)
	Find Context (E55)	Gouriza vocabulary (P2)
	Object Type (E55)	Gouriza vocabulary (P2)
	Burial Contex (E55)	Gouriza vocabulary (P2)
	Year (E65)	Creation date (P92i)
	Value (E61)	Value of the creation date (L31)
	Owner (E39)	Links to Organization (E40) or Person (E21) (P105)
	Location on Image (E53)	Link to map location (P129)
	Part of Project (A9)	Link to Excavation (A9) (P19i)
	Depicts (E1)	Link to Site, Find, Image, Person (P138)
	Image File (D1)	Image file (P138)
Find (E22)	Name / ID (E41)	Gives name to resource (P1)
	Find Category (E55)	Gouriza vocabulary (P2)
	Found at location (E53)	Link the map location (P53)
	Revocered in Project (A9)	Link to Excavation (A9) (P12i)

The codes within parentheses show the CIDOC-CRM classes (A, E, D) and properties linking the resources (P,L).
A CIDOC-CRM ontology setup is offered by the Arches project at `https://github.com/archesproject/cidoc-crm-ontology`, and this is the ontology setup that was also used for Gouriza data. The different class letters refer to different data classes. The Data classes used in this setup come from the following ontologies:

E, P CIDOC Conceptual Reference Model 6.2 (`http://www.cidoc-crm.org/Version/version-6.2`)

A CRMarcheo excavation model 1.4.1 (`http://www.cidoc-crm.org/crmarchaeo/ModelVersion/version-1.4.1`)

D, L CRMdig Model for provenance data 3.2.1 (`http://www.cidoc-crm.org/crmdig/ModelVersion/version-3.2.1`)

The solution chosen was Geoserver, a Java-based map server, that has a rather intuitive graphical user interface for configuring the map layers. Geoserver was installed on the same server as the Arches system.

The raster maps "Figure 13" was georeferenced using the known find locations on the plan, using the QGIS system, and ther resulting raster was converted to GeoTIFF and installed on the Geoserver.

4.5 Successes and failures

- Arches (together with Geoserver) can be used for storing and publishing archaeological documentation, GIS data, and plans.
- CIDOC-CRM can be used with Arches to define the description basis for archaeological data.
- Data can be imported using the command line interface and relatively simple mapping definitions

Failures:

- The data is very difficult to export from Arches in the current state of the system.
- CIDOC-CRM is so complicated, that the definition of the resource models is not possible for just anyone. In this pilot, the only person really able to work with it was the project researcher, who could not, however, spend enough time with this pilot.

4.6 Achievements

The dataset that was made available for the Gouriza project has been imorted to the Arches system and is available for registered users. Data is described using the CIDOC-CRM ontology and some related, compatible, discipline-specific ontologies.

4.7 Current State

The Gouriza archaeological data set is stored in the Arches server of the pilot (`https://arches.tatd.ovh/`). The dataset included mostly photographs and one set of Find documentation with coordinate points. Also, one of the plans produces by the excavation projecs has been georeferences and can be used as "basemap" under the Arches system.

4.8 Future work and proposed follow-ups

The challenge of using Arches is that it really required three different sorts of expertise to be useful.

1. There must be an archaeologist involved who is capable of understanding the needs of the documentations that is to be stored on the server.
2. There must be an ontologist who understands how ontological modelling works, and is able to understand the CIDOC-CRM on the level, that they can assist the archeologist in formulating the ontological structure of the resource models.

3. There must be an IT-person who is able to install and configure the Arches system. In its current state, the system looks very nice and graphical but there are several things that must be taken care of when the system is installed on an open, production server which cannot be done graphically.

Arches is not suitable for long-term digital data preservation and its support for digital object storage seems limited. Based on the pilot, it seems that the actual digital files are stored just as files on the server disk by default. On the other hand, the files could just as well stay in an external storage in which case Arches would be the storage for information about the files.

An integration with the solution from the Ostia pilot might also be suitable. In this case, the actual digital files (photographs, measurement data, etc.) would be stored in a Digital Object Storage, using S3 protocol or similar. There are several commercial options available, some of which are low-cost.

Chapter 5

Digitizing the Hilma Granqvist Archive

5.1 Purpose

The purpose of the pilot of the Finnish Institute in the Middle East was to examine the archiving and publishing of digital research material in co-operation with external actors. In this case, a suitable actor was the *Svenska Litteratursällskapet i Finland* (*SLS*), who had expressed an interest in such co-operation.

The work in the pilot was focused on the ethnographic field archives of Hilma Granqvist from the Palestine field work in 1926–32. Granqvist was an internationally recognized scholar of the Middle East. Her Palestine collections contain unique and still relevant documentation on the everyday life of the local communities. Most of the archive has never been availabe before, not even for researchers.

5.2 Premises

The Hilma Granqvist Palestine field material has been deposited at the Palestine Exploration Fund archives in London. It consists of 95 archival storage boxes. Of interest to the pilot from these were 32 boxed of field notes, 8 boxes of diaries, 6 boxes of letters and post cards, and 14 boxes of photographs and negatives.

5.3 Plan

The platform used for storing and publishing the data is the archival storage platform used by SLS. A researcher was sent to London to digitize the documents at the PEF archives, and the digitized documents were further processed and entered into the archival system at the SLS.

5.4 Implementation

The work in this pilot was done mostly by the trainee/researcher who was responsible for the digitizing of the archival material, and the support personnel at SLS, who entered the data into the digital archive system.

5.5 Successes and failures

THis pilot was very successful in managing to get the intended data digitized and published. From the institutes point of view, this is an excellent solution for old archival material that could and should be published in a form that is accessbile both for other researchers and the large public.

However, the resulting online publication is at the time of writing this report still lacking in several features.

- There does not seem to be any machine-readable metadata about the archival items available.
- There is very little metadata in general on any of the items in the archive. This is especially bad in the case of the photographs.
- It is not possible to link to individual items within image collections, or individual pages in diaries.
- There does not seem to exist any possibility for any kind of semantic annotation using internal or external tools.
- Search possibilities are very limited.

5.6 Achievements

This pilot demonstrates well how much more is possible to achieve when there is an external partner that already has resources and knowledge about digital archives available. A large amount of the material has been digitized and made available in online storage.

5.7 Current State

The digitized data has been published on the digital archive platform at `http://granqvist.sls.fi/`.

5.8 Future work and proposed follow-ups

The publication could be improved in some respects.

- Linking to the documents is only possible on the level of an individual diary; it is not possible to get a permanent link to a page in the diary, which would be important for external referencing of the material.
- They keywords (Pesons, Places, Subjects) link correctly to individual pages within the diary, but the page is not shown, so for a researcher, it is difficult to tell the links from each other.

Chapter 6

Conclusion

There are some overall conclusions that can be drawn from the pilot projects.

Most pilots had problems related to the nature of the dataset chosen for the pilots.

In experimenting with storing and publishing research data, it is very important to recognize the state of the data: is it final, provisional or under work.

A final dataset is something that only should be archived. The project that produced it is no longer using it, and the RDM challenges are related to digital object and metadata formats. The data has to be converted to formats that are suitable for long-term preservation and the metadata has to be converted to reuseable and possibly semantically annotated. After this work has been done, the whole dataset can described, documented and stored somewhere safe.

The only pilot in this project that had such data was the Hilma Granqvist archive where the dataset consisted of already archived documents from the 1920's. Consequently, the process from original data to digital storage and publication was straightforward, as Ms. Granqvist no longer was contributing new material nor correcting earlier documents.

A provisional dataset is something that has been left over from a project but which has not been systematically organized nor documented. The data may contain intermediate data that is not necessary nor useful to preserve; the data might contain sensitive information and the ordering of the data might be inconsistent. However, the project is already over, so no new contributions are coming to the data, and the process of cleaning, ordering, and documenting the data is still rather straightforward. None of the pilots had this kind of datasets.

A dataset under work is practically a dataset belonging to an ongoing project. There is no stability in the data, new files could be coming in, old files can be changed and documentation is changing. In this case, the problem is actually of research data management during a project and the resulting set of challenges are very different. The data has to be stored in such a way that it can be altered and changed. All other three pilots except the Hilma Granqvist archive were of this kind.

In hindsight, it would have been much better to run all the pilots using final or provisional data. The aims of the main project were always related to long-term preservation and management of research data. In the beginning, the project researchers also assumed this to be the case, and the original planning of the pilots was aimed towards finding solutions for managing static, completed datasets; at some point, it became apparent that the datasets could not just be taken in and turned into something different, since the original datasets were still under use.

This problem was most apparent in the Bolli Doliari pilot. A simple conversion of the data to another format and the consequent treatment of the data in the resulting format would have been much simpler to achieve than the dynamic system. Similar problems were present in the Ostia and

the Gouriza datasets.

Perhaps the main issue to be learned is that it is very different to find solutions for research data management for an ongoing research project and for long term preservation of the data. Also, these operations should not be joined. A well done research data management make long term preservation of the data much easier, but the the tools available for long term preservation are not useable for research data management in an ongoing project; conversely, tools for research time data management do not fill the criteria for long term preservation. At best, a well-managed set of research data can be moved to long term preservation as it is.

Another question is the publication of data. Publication is not a netural term for one has to consider the audience: to whom is a publication intended? Different users may have different needs.

An example of this is the Hilma Granqvist archive which admittedly is the only pilot that has actually published any data. In the digital publication, the data is available for anyone to see and read but as already noted earlier, the search capabilities are limited and references to individual pages within the dataset are not possible. The data is therefore available in a form that responds to the needs of the open public but already for a typical humanities researcher, the useability is limited. For digital humanities research, the data is almost non-existent. There is no machine access to the data and any big data -type access to the data is impossible.

In this sense, the best solution is the Ostia image archive that offers REST endpoints of the data and metadata within the collection. It is also possible to get a stable, PID-like reference to individual digital objects, and it is possible to get the metadata using that reference. The Epidoc-based Bolli Doliari collection could also be used in this way, and the TEI/Epidoc scheme used to annotate the stamps is semantic, so any user with a suitable parser is able to extract information about the entries. In a further state of the data, also person and place names would be semantically annotated and linked to external vocabularies which would further enhance the semantic capabilities of the system. The Arches-based solution for Gouriza data has potential, but in the current state, the Arches system does not yet seem to offer good API's for machine useability. The data itself is semantically very well structured, but it is not easy to read.

Consequently, the publication of data has many possibilities, and one solution may not be able to cover them all.

During the main project, the researchers were following the international development of this field. At the outset of the project, there seems to have been an assumption that the tools and systems available for this kind of work were much more advanced than what they actually were. Of the solutions envisaged during the planning stages, most were not yet actually possible to achieve, because there were no systems available.

As a conclusion, we can state that in many ways, the institutes and this project were in the forefront of the development.