FASTI ONLINE DOCUMENTS & RESEARCH

The Journal of Fasti Online (ISSN 1828-3179) ● Published by the Associazione Internazionale di Archeologia Classica ● Palazzo Altemps, Via Sant'Appolinare 8 – 00186 Roma ● Tel. / Fax: ++39.06.67.98.798 ● http://www.aiac.org; http://www.fastionline.org

FAIR survey: improving documentation and archiving practices in archaeological field survey through CIDOC CRM

Tymon de Haas∗ – *Martijn van Leusen*∗∗

∗ Faculty of Archaeology, Leiden University
∗∗ Groningen Institute of Archaeology, University of Groningen

This paper aims to contribute to the improvement of documentation and archiving standards (conforming to the FAIR principles) for systematic Mediterranean archaeological field survey. It reports on the initial stages of work by the authors to build an extension to the CIDOC CRM ontology to accommodate concepts underlying the description of archaeological field survey data. We first constructed, based on our own experience as survey directors, a general process model for archaeological field survey; we then defined the concepts central to such survey practices in consultation with other domain experts; and we produced a draft conversion of these concepts into CIDOC CRM 'classes' with the help of members of the CRM Special Interest Group. While this work has resulted in a fairly robust conceptual model of field survey as practiced in the Mediterranean, we also identify and discuss several issues relating to the tension between the desire to enable comparative analysis of survey databases by improving documentation standards, and the apparent inability of the survey domain to achieve standardization of field procedures. Although the process of formally agreeing a CIDOC CRM extension for field survey is a slow one, we believe a global solution to the problem of comparability is worth pursuing over a local, temporary one. We lay out the steps needed to resolve the remaining conceptual issues, to formalise the CRM extension, and to implement it in the form of a 'mapping' tool.

1. Introduction

Archaeological field-walking surveys have, especially since the 1970s, been conducted in great numbers all over the world. They generate archaeological data about the landscape and about the surface archaeological record that is typically kept in relational (geo-) databases. Although of a relatively simple structure, and therefore in theory easy to standardise and integrate, little or no progress has been made in the past two decades in the Mediterranean area in combining and comparing survey data generated by different surveyors. Among the causes for this unfortunate situation are the fact that survey databases are still not regularly published, archived, or even sufficiently documented; and that field survey directors tend to adopt documentation procedures optimized for the aims and context of their specific study, without consideration of the comparability of their data with those of others. This situation is all the more undesirable if we consider that field survey results can often not be reproduced because of the rapid and ongoing destruction of the landscape archaeological record by mechanised agriculture and construction work. Existing and 'legacy' field survey data are therefore irreplaceable, and at the same time of major societal relevance because they form our main source of documentation of the archaeological heritage at regional and supraregional scales. Being unable to access, share and combine such data is therefore a major impediment to both research and effective heritage management.

The authors have been involved in field walking surveys in central and southern Italy since the late 1990s. For central Italy, these surveys are embedded within the Pontine Region Project (PRP), a major and on-going landscape archaeological research program that has since the mid-1980s gathered a wealth of information to reconstruct long-term developments in settlement and land use in the Pontine Region. We were faced with the question of proper archiving ever since the first of these surveys were published[1] and, later on, deposited in the Dutch national archaeological data repository, DANS-Easy.[2] Lately, in the course of wrapping up two substantial field survey programs that were part of the PRP,[3] we felt an increasing responsibility to improve our own standards for data documentation and, more generally, to contribute to the improvement of survey data documentation and archiving practices. This article reports on the resulting work, which began in 2016 with a pilot study supported by the Netherlands Royal Academy of Sciences KNAW, and is now continuing under the name FAIR Surveys Project.[4] The pilot study aimed to improve the quality of archived archaeological field survey datasets by applying the so-called FAIR principles for archiving research data and, more specifically, by extending the CIDOC CRM ontological approach to this specific subdomain of archaeology. The FAIR Surveys Project is pursuing the long-term aim of improving the potential for re-use and integration of such field survey datasets.[5]

Thinking about archiving worldwide has crystallized around the acronym FAIR: archives must be Findable, Accessible, Interoperable and Reusable.[6] The first two of these only ensure that data can be found and accessed by those who have the right and the desire to do so; the latter two ensure that those data will make sense to others and that there are no technological obstacles preventing them from being re-used. The FAIR data principles act as an international guideline for high quality data stewardship. As we shall see below, however, even current best practice in archaeological archiving (as practised by repositories such as DANS-Easy and, in the UK, ADS) does not yet achieve the crucial R criterion, which ensures reusability.

Within the broader field of Heritage Studies, similar concerns with data interoperability and re-use have, from about 2000 onwards, led to the development of CIDOC CRM, a conceptual reference model (or ontology) that identifies and defines concepts central to specific domains (e.g., archaeology, museum studies, …), and links these to the various terminologies in use within these domains.[7] Anyone can use this to 'map' their own data and procedures to this common reference model, and then convert their data to an internationally accepted exchange format such as RDF (Resource Description Format). This allows digital data, although physically stored on different servers, to be brought together and analysed in a meaningful way, irrespective of differences in terminology, database platform, and layout of the underlying databases. If enough researchers in a given field do so, this allows us to bring a critical mass of heterogeneous data together using a single framework.

The potential of CIDOC CRM to stimulate the exchange and integration of archaeological datasets is increasingly being recognised in the academic research community. This has so far led to the development of CRMarchaeo, an extension of CIDOC CRM for describing stratigraphic excavation records, and related initiatives developed by the ARIADNE and ArcheoInf programs.[8] This work focuses, however, on the documentation of archaeological objects in collections and related excavation data; it does not yet cover archaeological field survey data.

This paper has two main aims. Firstly, to raise awareness of the crucial role that improving current documentation and archiving standards has in enabling re-use of archaeological field survey data; and secondly, to present the work we undertook to develop a survey extension to CIDOC CRM whilst highlighting relevant issues and future work. For now, we are focusing strictly on *systematic* field-walking as currently practised in the Mediterranean in our efforts to extend CIDOC CRM, excluding other traditions such as topographic and architectural surveys, and closely related geo-archaeological, geophysical, and remote sensing approaches (fig.

[1] VAN LEUSEN 1998; BURGERS *et al.* 1998; ATTEMA *et al.* 2001; VAN LEUSEN & ATTEMA 2001/2002.
[2] VAN LEUSEN & DE HAAS 2005.
[3] These are the Hidden Landscapes Project (FEIKEN 2014) and The Minor Centres Project (TOL *et al.* 2014; TOL & DE HAAS 2017).
[4] https://www.rug.nl/research/groningen-institute-of-archaeology/research/major-gia-projects. The project archives: DE HAAS 2016; VAN LEUSEN 2017.
[5] The results of this pilot were presented at EAA Maastricht 2017 and at a meeting of the CRM Special Interest Group in early 2018. We here include the results of further extensive tests on colleagues' datasets and a roadmap towards our ultimate goal: high quality Linked Open Access to survey data.
[6] WILKINSON *et al.* 2016; https://www.force11.org/fairprinciples.
[7] OLDMAN & KURTZ 2014; CROFTS *et al.* 2011.
[8] CRIPPS *et al.* 2014; Doerr *et al*. 2016. ARIADNE, a European Integrating Activity project, ran between 2013 and 2017 and was recently extended (as ARIADNE plus) to 2022. It intends to "[…] bring together and integrate existing archaeological research data infrastructures so that researchers can use the various distributed datasets […]". It runs a portal for archaeological datasets (http://www.ariadne-infrastructure.eu/About). In addition, various institutions and projects, such as the ArcheoInf project (LANG *et al.* s.d.) apply the CIDOC CRM ontology to describe and make accessible specific archaeological datasets.

1). However, the CIDOC CRM offers excellent opportunities to include these related activities, as much of the relevant groundwork has already been laid down in CRM extensions such as CRMgeo, CRMsci and CRMinf.[9]

In section 2 we first discuss current practices in the documentation and archiving of Mediterranean field survey data, outlining in what respects these practices (fail to) comply with the FAIR principles and arguing why and how this situation should be improved. In section 3 we explain why CIDOC CRM may be an appropriate solution. In sections 4 to 5 we then present our work on extending the CRM ontology to enable the mapping of survey datasets, and evaluate this work and outline the steps we need to take in order to reach our longer-term aim of enabling re-use and integration for scientific purposes in section 6.

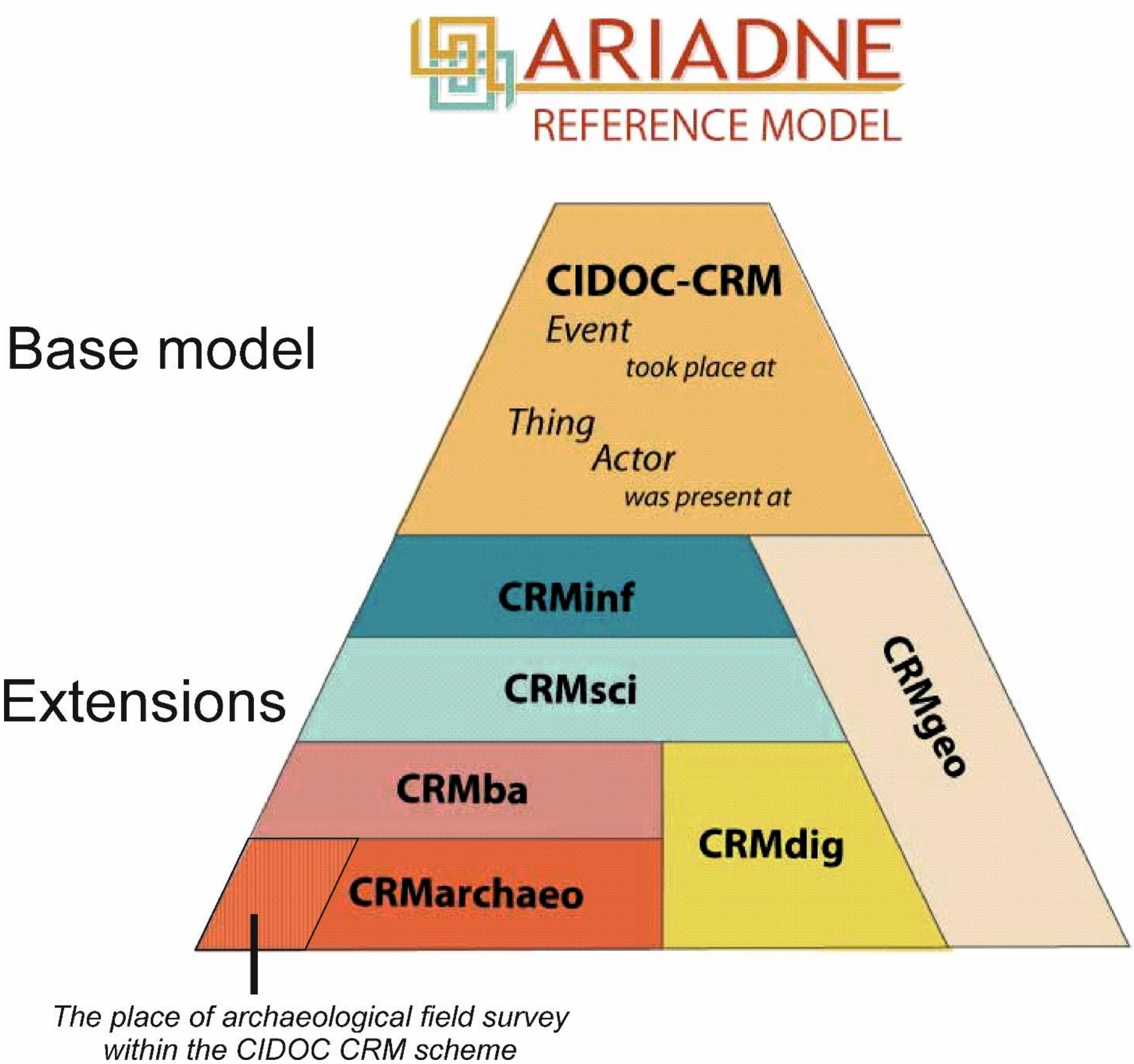

Fig. 1. The ARIADNE reference model (modified from https://www.ics.forth.gr/isl/index_main.php?l=e&c=229).

2. Mediterranean Field Survey: current practices in data documentation and archiving

The best current overview of field walking surveys taking place in the Mediterranean theatre is provided by Fasti Online Survey, a web-based platform and open-access archive recently developed by scholars from the Royal Netherlands Institute in Rome with the support of the International Association for Classical Archaeology

[9] HIEBEL *et al.* 2015; DOERR *et al.* 2018; STEAD *et al.* 2018.

(AIAC) and L-P Archaeology (UK).[10] This attests to the vast body of data that has been generated by hundreds of surveys conducted since the 1950s, with dozens of field walking surveys still taking place every year across the Mediterranean. Fasti Online Survey contains basic metadata on circa 120 survey projects in the Italian peninsula, and it would only seem logical that those data should be made available for large-scale analysis. It turns out, however, that many of these projects have only published their datasets partially (e.g., providing site gazetteers), or not at all. Few of them have published their data fully, and even fewer have supplied a digital data archive to a repository.[11] In other words, and despite repeated international calls to work towards survey data compatibility,[12] Mediterranean survey practice is still a long way from adhering to the FAIR data management principles: only a small proportion of survey data are Findable and Accessible, and as we shall see below, an even smaller proportion is Interoperable and Re-usable.

This point is also illustrated by a review of current documentation and archiving practices in Dutch and British archaeology, which may be taken as typical or even best practice within Europe.[13] Repository services for Dutch archaeology are supplied by Data Archiving and Networked Services (DANS), an institute residing under the Royal Dutch Academy of Sciences KNAW and the Dutch Foundation for Scientific Research NWO. In 2007 DANS created the E-Depot for Dutch Archaeology, EDNA, specifically for archiving archaeological datasets.[14] This archive now contains documentation on a vast number of mainly development-led archaeological projects in the Netherlands, but there is also a small component of university-led projects, including survey projects in the Mediterranean, that have found their way to EDNA. A similar development can be seen at other repositories such as ADS in the UK (fig. 2).

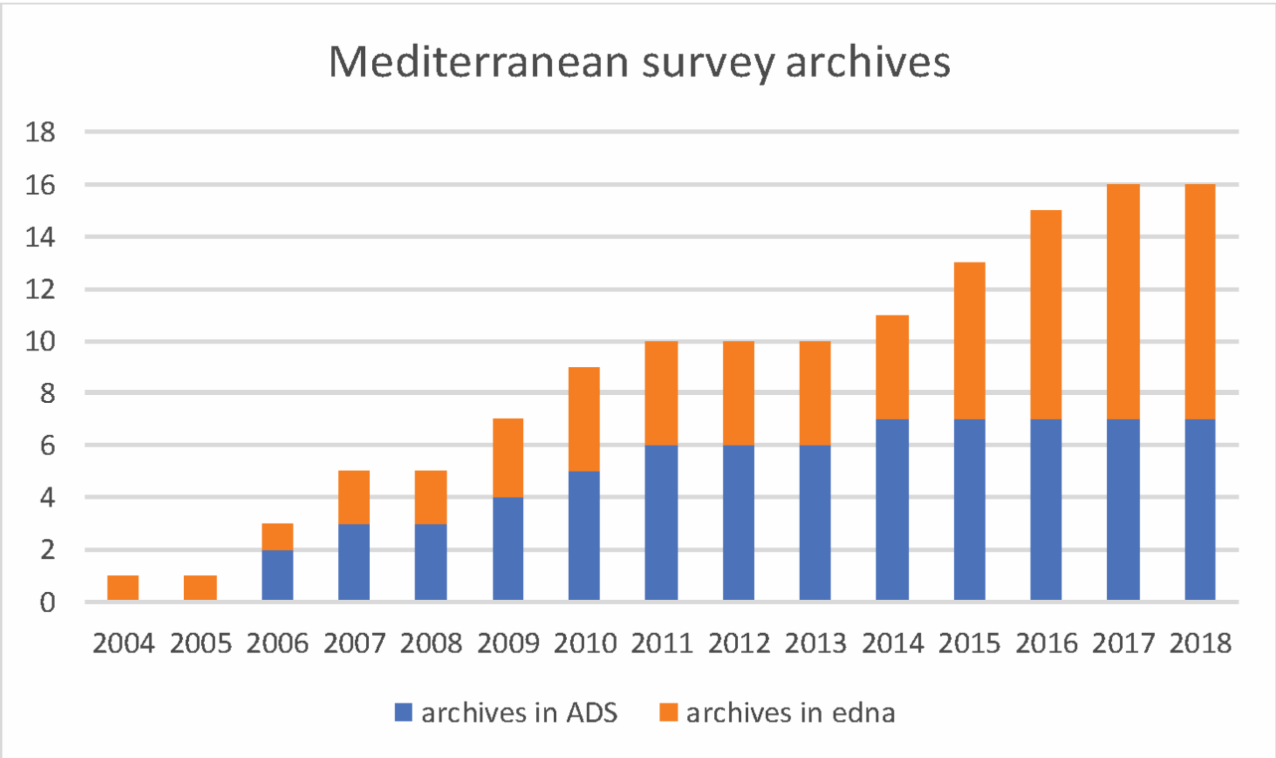

Fig. 2. Cumulative numbers of deposited Mediterranean survey archives at ADS and DANS-EDNA, 2004-2018.

These few archaeological field survey datasets are now Findable (structured metadata about them is available online), Accessible (the archives themselves can be consulted under a system of rights management),

[10] Fasti Online Survey (http://www.fastionline.org/survey/about.php?view=aboutfastisur) originates in Pedar Foss' MAGIS database at DePauw University, http://cgma.depauw.edu/MAGIS/.

[11] Examples include the Ager Tarraconensis survey (KEAY & MILLETT 2003), deposited with ADS; the Agro Pontino Survey Project (HOLSTROM *et al.* 1989), a subset of the Pontine Region Project (ATTEMA 2004), the Dzarylgac Survey Project (Groninger Instituut voor Archeologie 2010), and the Zakynthos Project (VAN WIJNGAARDEN 2015), all deposited with DANS; and a subset of the Metaponto Survey, available through the project website at http://ica.tacc.utexas.edu/metaponto.

[12] BARKER & MATTINGLY 1999; BLANTON 2001; ALCOCK & CHERRY 2004.

[13] RICHARDS & ROBINSON 2000; PERRIN *et al.* s.d.

[14] https://doi.org/10.17026/dans-xk5-n3me

and Interoperable (provided in formats that can be read and interpreted platform-independently, with the repository guaranteeing any future migration needs).

So what about the fourth criterion: Reusability? We have serious doubts that current documentation standards at these repositories allow, let alone guarantee, the possibility of re-use. When we (as trained survey archaeologists) studied other Mediterranean survey archives from DANS-EDNA and ADS, using *only* the information supplied in those archives, we were not able to understand fully the meaning, scope, and limitations of these datasets. Even though datasets are accompanied by a metadata file explaining the general aims of the project, the structure of the archive, and the codes used in the data files, they still do not provide essential information on (for example) how 'sites' were defined, what is meant by terms such as 'tract', or which are the precise protocols used in the collection of ceramics. We were left to wonder whether a tract might be the area inspected by a single walker within a single survey unit, or by a group of walkers within that unit, or that it might even mean a series of such units linked together spatially. Likewise, it was unclear whether the standard pottery sampling protocol was 'total' (that is, an attempt to collect all visible fragments from the defined tract), 'grab' (unsystematic collections that may be part random part selective), or 'representative' (systematic but incomplete). If such terms are left undefined, and procedures undescribed, we may well not be comparing like with like between two surveys.

We want to emphasise that this is not intended to criticise those colleagues who have made their data available via a repository (our own deposited survey archives suffer from the same issues), but rather to point out that *de facto* reusability can only be established by having your dataset tested by independent reviewers, and can only be ensured by applying more stringent meta-documentation standards to deposited datasets. In any case, such examples show that current best practice can be described as fulfilling the FAI principles but *not the R principle*.

The fact that we are not succeeding in publishing and archiving our data in a reusable manner puts severe limits on their scientific value: many research questions regarding large-scale societal processes cannot be studied unless we can put together multiple regional datasets.[15] It is no surprise that the desire to conduct comparative analyses across Mediterranean survey datasets was already expressed in the POPULUS conferences of the mid-1990s.[16] However, this desire then and now has been frustrated not only by the limited accessibility of datasets, but more fundamentally by the incompatibility of datasets based on different and often poorly documented survey designs and recording methodologies. Such designs and procedures depend, naturally, on specific research contexts (geographic conditions, research questions, financial and logistical considerations, legal limitations in terms of collection and storage of artefacts, et cetera) and consequently vary considerably between projects.[17] Little progress has so far been made in the standardization of these procedures, and this will not likely change in the near future given that many consider such standardization undesirable.[18]

The limited availability of survey datasets, and the (often undocumented) differences between these datasets, force scholars to rely on a 'highest common denominator' approach to comparative analysis of survey data. Launaro (2011), for example, studied Roman demographic change and social organization on the basis of 27 Italian field surveys but was forced to make a series of simplifying assumptions about site types and dating. In his assessment of 19 survey projects in Tunisia, Stone (2004) discusses many problems in comparing the resulting datasets – several of them insurmountable – before deciding that only the *ratio* of pre-Roman to Roman sites provides a relatively reliable indicator of long-term socio-economic developments. Discussing the well-known Tiber Valley Project and several other surveys in Etruria, Witcher too noted that differences in field and artefact processing procedures between the different component surveys mean that it is impossible to compare chronological and spatial settlement patterns and trends directly, and was forced instead to rely on site *interpretations* to come to comparisons.[19] Attema *et al.* (2010) took a similar approach when comparing the development of settlement systems and land use in central and southern Italy.

In sum, we are currently in a situation in which effective comparison and synthesis of Mediterranean field survey data is extremely difficult, as the bulk of the data is not even Findable and Accessible, and the data that is available are hardly suited for Re-use because of the differences between the datasets and the fact that they

[15] BLANTON 2004; LAWRENCE *et al.* 2016; KINTIGH *et al.* 2014.
[16] BARKER & MATTINGLY 1999; FRANCOVICH & PATTERSON 2000. Even the subsequent criticism of 'Mediterranean myopia' could be construed to indicate that Mediterranean survey archaeologists were constructing datasets without regard for comparability (BLANTON 2001).
[17] Cf. MATTINGLY 2000; MILLETT 2000.
[18] TERRENATO 1996; MATTINGLY 2000; ALCOCK & CHERRY 2004.
[19] WITCHER 2006.

are often not sufficiently documented. As things stand, each research team is only capable of interpreting the survey datasets they themselves have produced. This is an intolerable situation: the discipline has an obligation to make primary and often irreplaceable data available in a manner that can be verified and re-used by others, for their own purposes and without any misunderstandings arising. To do that, we must be able to effectively share and merge our datasets, and provide tools for extracting information across them.[20] How can we best do this? We present here the approach that, in our opinion, offers the best perspective of a long-term and high-quality solution.

3. Moving forward: why describe archaeological field survey using CIDOC CRM

So far we have argued that a substantially greater effort is needed to properly document our field-survey datasets, to enable both re-use of single datasets and meaningful comparison between (and analysis of) multiple datasets. One obvious way of producing comparable datasets would be standardisation of the way they are conducted and documented in the field. However, this would only affect future datasets and therefore not help exploit the wealth of legacy data available for the Mediterranean.[21] Moreover, past attempts to discuss standardisation of field protocols have always met with fierce resistance from practitioners unwilling to 'give up their freedom of research'. Whilst we disagree with this particular argument, which amounts to a denial that a best (or even good) survey practice can be defined, a less confrontational (and more feasible) approach would be to pursue the standardisation of survey *documentation* practices instead. In the absence of a professional body to discuss, promulgate and implement such a standard, however, this would probably be a long-drawn-out process.

A second approach, currently being explored by an international research group for the immediate hinterland of Rome, is to attempt a 'hard' merger of existing survey datasets from the Tiber Valley Project, the Suburbium Project and the Pontine Region Project.[22] It is already clear that this is a major and time-consuming undertaking, requiring a detailed analysis of the participating databases, close collaboration between the data owners to resolve queries and explicate what is normally left implicit, and the setting up of a complex system of concordances between the type lists used in each dataset. Moreover, it is unlikely to lead to a durable solution, because the complexity of the merged database rises with the number of participating datasets.

In view of the above, we believe that the only feasible and sustainable approach to allow the technical merger and meaningful re-use (if certain problems with vocabularies and type lists can be solved; see our discussion of these below) of large numbers of existing field survey datasets, owned by many different institutions, is a decentral one in which each dataset owner has an incentive to 'map' their data to a central reference model. CIDOC CRM is a useful tool to do so: as a conceptual reference model, CIDOC CRM provides an agreed set of concepts that can be used to describe data and the ways in which these data have been generated. The CIDOC CRM uses two main types of concepts: *classes* that define real world things, and *properties* that define relationships between these classes. Classes can be subdivided into *persistent* ones (people, things, ideas, concepts) and *temporal* ones (events and activities), and they are hierarchical: things can be subdivided into more specific things.[23]

When mapping data generation processes such as occur during archaeological field survey, activities and events are central. Data are generated through activities (field walking, artefact processing) carried out by actors (we as survey archaeologists) in specific places (in fields, on sites), and linked to objects like the artefacts we collect (fig.3). Thus, any dataset (typically residing in a relational database) can be described or 'mapped' in terms of CRM classes and properties. Subsequently, it can be made available online as Linked Open Data (LOD), and searched and integrated with other datasets through data mining techniques.[24]

---

[20] Current digital archiving services do not yet allow the possibility for Linked Open Data (DANS) or have only few datasets available in this format (ADS).

[21] WITCHER 2008; CASAROTTO 2018.

[22] The Roman Hinterland Database project, http://comparativesurveyarchaeology.org/. The Pontine Region Project (PRP) itself has generated data during multiple 4-5-year research funding cycles, with attendant changes in field protocols and project management, resulting in slightly different data structures and type lists (see DE HAAS & TOL forthcoming for the phasing of the project and related changes in field strategies, sampling procedures and artefact analysis protocols).

[23] E.g., things may be subdivided into man-made things and natural things; the former can in turn be subdivided into physical man-made things and conceptual man-made things.

[24] This requires a conversion to a so-called 'triple' format - a 'triple' is a pair of classes linked by a property - using RDF (resource description framework); to do so efficiently, the STELLAR project has already developed templates for archaeological excavation data (BINDING *et al*. 2015; STELLAR s.d.). SPARQL provides data mining capabilities (OLDMAN & KURTZ 2014).

In light of the above, the benefits of mapping our data to CRM concepts seem considerable. By describing in a standardized way how they are generated we will more fully understand the similarities and differences between datasets. By mapping our individually owned datasets to a common set of concepts (e.g., the CIDOC Conceptual Reference Model), we will allow these datasets to be compared in a meaningful way. And by transforming them to LOD, we allow them to be mined and thus used for comparative purposes efficiently. Finally, the CIDOC CRM ontology has already been used within archaeology to document excavation data, for which it seems fairly successful in describing the process of data generation as well as in enabling (automated) extraction of standardised data from multiple datasets.[25] A complementary benefit of using CIDOC CRM is that we can connect to an internationally established standard and build on the expertise of the CIDOC CRM community. Finally, it can be applied to both legacy and new data, and it reduces the total effort required for the field as a whole: each data owner only has to map his or her own data structure to the reference model.

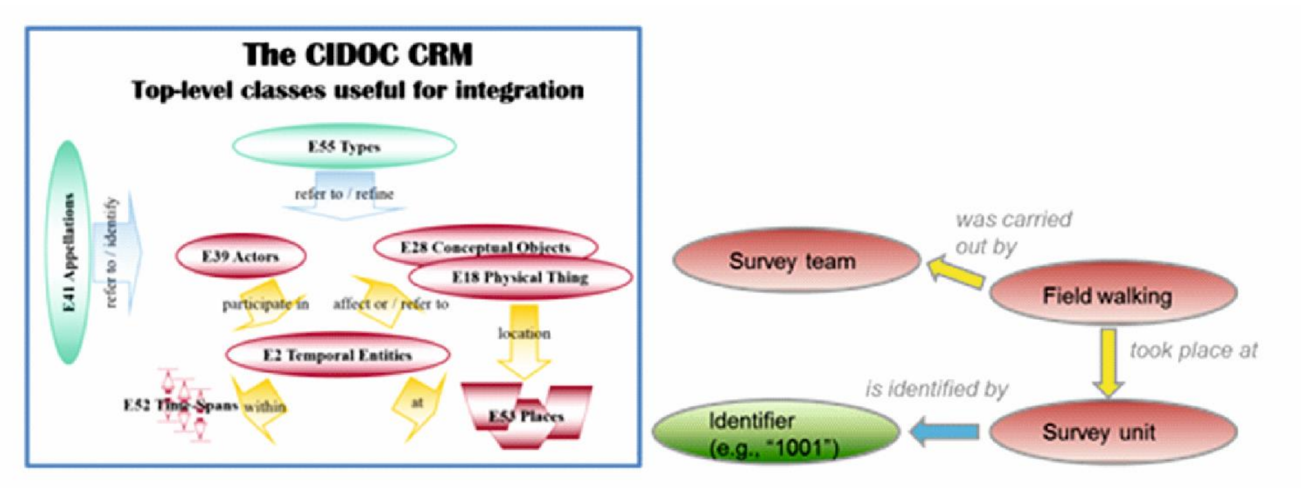

Fig. 3 Main classes and properties of the CIDOC CRM, with application to field survey (left part of image from Oldman & Kurtz 2014, 8).

4. Process model: field survey as a chain of activities

As a first step to applying CIDOC CRM to field survey data we reformulated the process of conducting an archaeological field survey in terms of activities and related actors, places, procedures and objects. We based our initial process model on the databases of our own projects, which represent considerable methodological differences between intensive off-site surveys in high visibility arable farmland and more site-oriented surveys in lower visibility uplands. However, because the process description must be applicable to archaeological field surveys employing quite diverse methods, we then tested it against three of our older surveys that employed different field and ceramic classification procedures, as well as against survey datasets described in the literature (Mattingly 2000), before discussing it with colleagues working in different geographic settings and research traditions. Despite the variations in the procedures followed in these different surveys, we managed to extract a sequence of activities, summarised in fig. 4, that are central in all surveys.[26]

Conducting a field survey means conducting several related activities: *making observations* about the area being field walked (land use, visibility conditions, archaeologically relevant features), *collecting* artefacts, *defining* (parts of) the area as a site, and of course documenting all of this through *record taking*. These activities may take place concurrently or consecutively, for example with an initial non-site-oriented survey followed by one or more stages of site-directed survey including finds collection; so the sequence of and relations between these activities are flexible (as visualised in the arrows in the diagram).

Here it is relevant that, though practitioners, including the authors, often define sites in slightly (or even very) different ways, there is a common if often implicit understanding of a site as a locus of human activities identified by associated material remains. In our model, we therefore conceptualise sites very broadly as any such location identified through field-walking survey, thus including both settlements *and* non-settlement loci. We

[25] Doerr *et al.* 2016; Binding *et al*. 2015.

[26] Note that the process model follows current common practice in focusing on the definition of 'sites'. Non-site survey units are of course included but, since none of the test datasets contain any explicit interpretations of off-site materials, our process model does not include such interpretations either.

also acknowledge that such sites may be defined at different moments in the chain of activities that constitutes a survey – immediately during the field survey, or later during data analysis. The precise position of the concept of ‘site definition’ in our process model can therefore be disputed, but the concept itself is still needed.

After the actual field work, artefact collections are classified and described by finds specialists. The depth of classification varies depending on the ‘diagnosticity’ of the finds and on the needs and goals of the survey project, but in general we can distinguish the classification of groups of objects (classification) from the more detailed documentation and classification of individual objects (often including typological assignment and dating). A final activity in many surveys concerns the interpretation of sites, usually on the basis of characteristics as observed in the field but also later, during artefact studies and spatial data processing. This involves the assignment of periods, functions and ‘types’ to a site.

We acknowledge that figure 4 simplifies matters: for example, a single survey team’s fieldwalking activity can be deconstructed into many separate individual activities. But although the process could be described in much more detail (e.g., how were field walkers instructed before the survey? how were records being taken? with what type of equipment are measurements of sherds taken?), the goal of CIDOC CRM is simply to *model the concepts that are actually used* in the datasets being modelled. The concepts as defined here cover the survey process as documented in our test survey databases, and allow us to efficiently map these. Moreover, the CRM allows us to define ‘Design or Procedure’ documents in which field methodologies, artefact processing procedures, and site classifications can – and indeed should - be documented in much more detail.

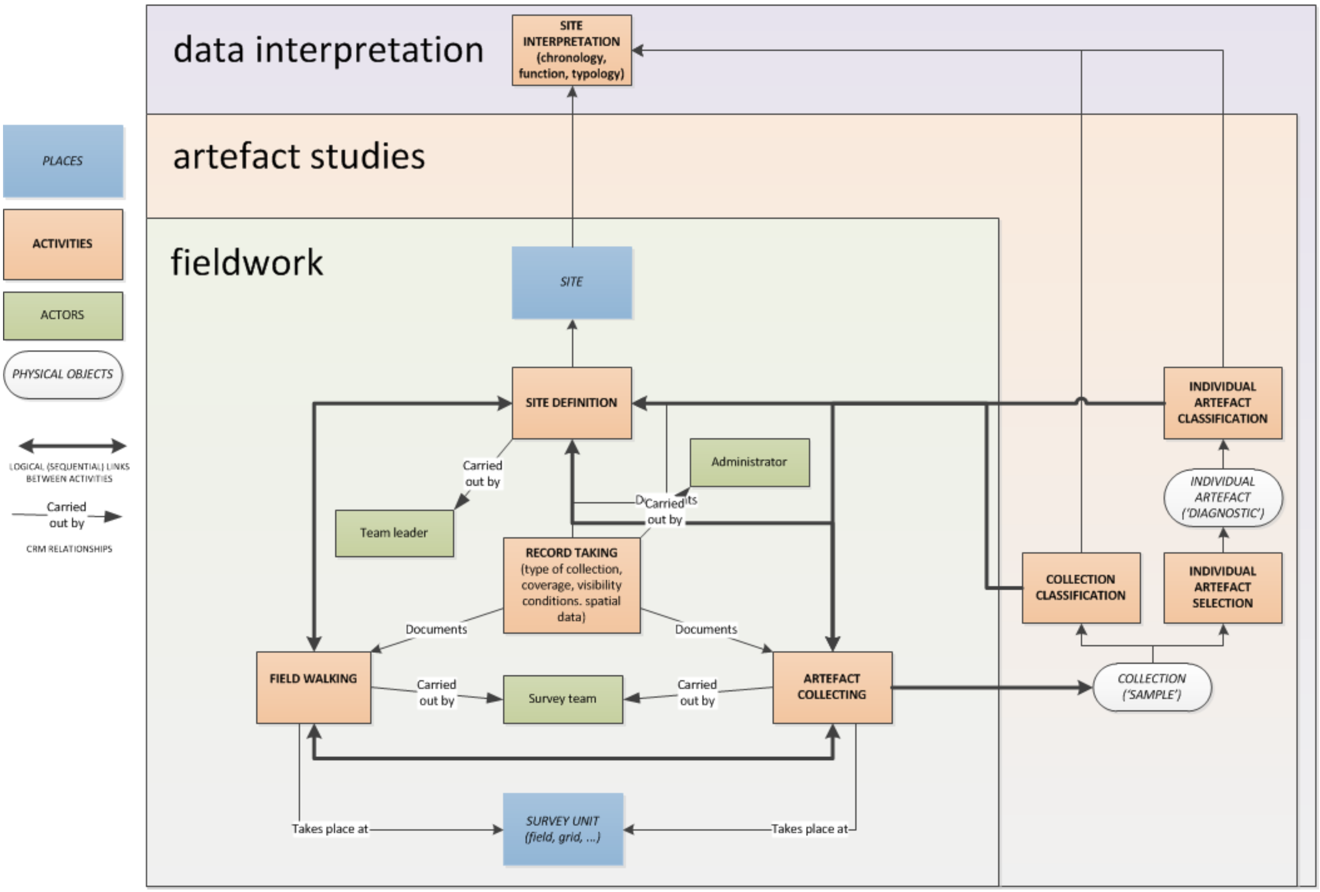

Fig. 4. Schematic rendering of the process of archaeological field survey.

5. *Extending CRMarchaeo with field survey concepts*

The chain of activities involved in a survey as described in section 4 formed the basis for compiling an exhaustive list of domain-specific concepts needed to describe the process of field survey. This was modified and extended after a test mapping of GIA’s Minor Centres Project database and a round of consultations with domain experts, in which (perhaps surprisingly) we were able to agree on the set of basic concepts (including site, visibility, coverage and collections) and their definitions.[27] We provide a concise overview of all concepts and their definitions in Table 1, but highlight only some concepts and issues here.

As already mentioned, the archaeological community has contributed to the development of CIDOC CRM through the development of the CRMarchaeo extension.[28] This provides concepts needed to document archaeological excavations, defining amongst others the stratigraphic relationships between archaeological deposits, their characteristics, genesis and any later modifications, and the objects contained in them. However, it does not provide us with the concepts needed to document archaeological field survey taking place at the land surface. We therefore aimed to link our concepts to corresponding ones already defined in CIDOC CRM and its extensions, and only propose new concepts to be added to CRMarchaeo when unable to find corresponding ones.[29] The outcomes of this work have been documented in the format normally used for CRM extensions.[30]

Table 1. Selected concepts needed to describe an archaeological field survey.

Entity type	**Concept**	***Description***
Activities/events	Field walking	Systematically traversing an area with the objective of making observations (which result in taking records and/or collecting objects)
	Collecting	The collection of physical objects (artefacts) from an area
	Taking records	Documenting properties of surveyed areas and/or objects encountered in these and/or any activities undertaken as part of a survey
	Site definition	Assigning site status to a specific area
	Visit	A specific instance at which one or more activities (e.g., field walking, record taking and/or collecting) take place at a Survey Unit or at a Survey Site
	Collection Classification	The grouping of collections (of artefacts) on the basis of one or more physical characteristics
	Object selection	Selecting of individual (´diagnostic´) artefacts to be described in more detail
	Object classification	The interpretation of individual artefacts ('diagnostics') according to set chronologies and typologies
	Site interpretation	The interpretation (classification) of sites according to set typologies and/or chronologies
Actors	Survey team	Group of people that as a team carries out any of these activities
	Individual	Single person (in role of team leader, artefact specialist, field technician, …) carrying out any of these activities
Other Classes	Survey unit	A uniquely defined piece of geographic space in which field walking and/or collecting were undertaken
	Survey site	A spatially coherent concentration of artefacts and/or other archaeologically relevant features recorded during a survey
	Collection	An artefact or group of artefacts collected in a specified way from a survey unit or site
	Finds class	A group of objects within a Collection defined during artefact classification as having similar characteristics
	Artefact	A uniquely identified physical object (part of a collection) that is studied because of its perceived analytical value
	Coverage	The notional proportion of the survey unit that has actually been investigated through field walking and/or collecting (usually expressed as a %)

[27] This mapping was discussed with members of the CIDOC CRM Special Interest Group and colleagues involved in the International Mediterranean Survey Network.
[28] CRIPPS *et al.* 2014.
[29] This course of action was taken after initial consultation with Steve Stead (Paveprime Ltd., London) and after presenting our plans to the CRM Special Interest Group at a conference late February 2016 in Prato (Italy).
[30] DE HAAS & VAN LEUSEN 2016.

	Total Visibility	The proportion of artefacts on the surface estimated to have been observed within the covered area during field walking and/or collecting (usually expressed as a score or a %)

Many of these concepts proved to map easily onto existing CRM classes. For example, the concept of field walking was mapped as a specification (type) of the class 'Observation' as defined within the CRMSci extension;[31] the concepts of finds collection and finds class as Physical Objects defined in the core CIDOC CRM;[32] the concept of coverage as a type of 'Dimension';[33] and the concept of Visibility as a type of E3 Condition State. Furthermore, the use of 'types' conveniently allows us to distinguish different instances of such classes. For example, the Collecting activity can have types 'systematic' (collection of a given proportion of all artefacts), 'diagnostic' (collection of a subset of artefacts according to its perceived diagnosticity), and unsystematic or 'grab' (collection not following any set strategies).

Other concepts, however, proved less easy to map directly onto existing CIDOC ontology; again, 'collecting' provides a good example. Many survey archaeologists consider finds collecting to be a form of sampling (in the statistical sense) from a population, which is the entire observable surface archaeological record. In the CRMSci extension the concept of Sample is indeed defined in those terms.[34] Upon closer inspection, however, we find that samples are defined there as 'amounts of matter' taken from a 'material substantial', which is not commensurate with our own conception of a sample as consisting of a group of objects. In other words, either the CRMSci definition of 'Sample' must be changed, or we need to define an all-new concept for survey samples.

Other concepts proved to be not so much incompatible with CIDOC CRM, as too broad. For example, our concept of 'record taking' encompasses the CIDOC CRM classes of ´taking measurements´, ´attribute assignment´, ´type assignment´ and, when including visual records (drawing and/or photographing), types of 'production'.[35] Likewise, our concepts of 'collection classification', 'object classification' and 'site interpretation' all encompass multiple CIDOC CRM attribute and type assignment activities.

Some queries brought up by our work even highlighted fundamental ontological issues, and these are subject of continued discussion between ourselves and CRM SIG members. We here present two examples. One of the more intractable problems concerns the difference between the concepts of Survey Site and Survey Unit. We have followed suggestions from CRM SIG members to consider the Survey Site, defined as a spatially coherent concentration of physical objects and/or features, as a new subclass of the existing class 'Site'.[36] By contrast, Survey Units are conceptualised as unique pieces of geographic space, not necessarily having a material component, and therefore as so-called SP6 'Declarative places'. While this seems fine in principle, it becomes problematic if we consider that in practice (during a survey), both are pieces of space where identical types of activities (especially collecting objects) are carried out. This would force us to model activities such as Collecting in two different ways, depending on whether they are carried out within a Survey Site (where collection can be modelled as 'Part Removal' of the material component of the Site) or within a Survey Unit that, as a Declarative Place, does not have a physical component and can therefore not support Part Removal. The issue becomes even more complicated if we consider that a Survey Unit may turn into a Survey Site *after* such collecting takes place (we have defined site definition as 'the act of assigning site status to a specific area/location').

A second unresolved issue concerns our conceptualisation of – again - collecting. There are instances in which we *intend* to collect artefacts (e.g., if we conduct a systematic survey of a Survey Unit), but in practice the survey does not result in a collection of physical objects because these were simply not present. Some argue that in such a case we *do* have a collection (an empty bag), and record this in their database. An actual bag filled with finds can be satisfactorily modelled by the CIDOC CRM concept E78 Curated Holding, but if there are *no* finds in the bag then there is nothing to curate and therefore this concept cannot be used. It turns out that the modelling of such negatives – the absence of something - is a broader and still unresolved issue with CIDOC CRM.

---

31 DOERR *et al.* 2018, class S4.
32 CROFTS *et al.* 2011, class E19.
33 CROFTS *et al.* 2011, class E54.
34 DOERR *et al.* 2018, class S13.
35 CROFTS *et al*. 2011, classes E16, E13, E17 and E12.
36 CROFTS *et al*. 2011, class E27.

6. Concluding discussion

In the preceding sections we have described the unfortunate lack of re-use of field-walking survey data, argued that turning to CIDOC CRM provides the best long-term solution for this, and shown how far we have come along this path in recent years. It now remains to consider whether, as a subdiscipline, we should not also work towards greater harmonisation of field and documentation protocols.

Our review of current practices has clearly highlighted that we as survey archaeologists do not yet comply with the FAIR principles in archiving our data. Currently archived and accessible survey datasets in DANS and ADS only form a small minority, and although these are Findable, Accessible and largely Interoperable, they do not comply with the Re-usability criterion, which requires higher standards of documentation. Only the data owners have the knowledge and expertise to produce such documentation, e.g. by adding extensive metadata to their archives, after the fact. Perhaps more importantly, substantial data re-use could be achieved in future more easily if field survey practitioners could agree to adhere to sufficiently high data documentation standards in a spirit of enlightened self-interest.

Another, possibly more feasible way to achieve re-use of survey data, would be to enforce a quality standard. This is already done in some areas, for example in the case of *documentation* practices in commercial archaeology, enforced by national heritage management bodies. Extending such standards to non-commercial research including field surveys, would go a long way towards achieving this goal, but would also require an international authoritative body (perhaps an EAA working group?) to define and maintain such standards. Such internationally agreed standards could then be enforced by the separate national heritage management bodies, repositories, and research funding bodies. We have already seen this process at work with the enforcing of Open Access standards in publication. Even if successful, however, this solution would only apply to datasets produced in future surveys, and would not have any effect on the reusability of the sea of legacy data produced in the decades since WWII.

In the absence of an enforced documentation standard, we have suggested that the next best solution is to adopt CIDOC CRM as a metadata standard. This internationally acknowledged documentation standard, our pilot study has shown, is well-suited to incorporating survey data. The pilot also showed that the process model of systematic archaeological fieldwalking is not overly complex, and that consensus on the main concepts exists. This means that many different survey data structures, including both new and legacy datasets, can be efficiently mapped to a single reference model, enabling re-use and comparative analysis after conversion to RDF and serving as LOD. It also means that the total effort required for the field as a whole (with hundreds of practitioners) can be distributed: each data owner only has to map his or her data structure to a single CRM for the entire dataset to become available for re-use. [37]

While CIDOC CRM thus has substantial benefits, we must also acknowledge its limitations. It offers no magical solution to issues of standardization and comparability that Mediterranean survey archaeologists have been struggling with for decades. For example, it is established good practice to record the visibility and coverage at which a survey takes place, but whilst there is broad agreement on terminology and formatting – 'coverage', for example, being recorded as the percentage of the surface actually observed during field survey – there is no such agreement yet on methods to estimate the values of such parameters. Thus visibility scores remain incomparable across survey projects, and coverage percentages are based on varying assumptions about the width of the strip of land a surveyor observes during fieldwalking. If we leave this kind of thing unstandardized, or even undocumented, CRM will not help us make it into comparable and reusable data.

Another limitation concerns the difference between CRM concepts and type lists. Survey projects commonly employ classifications, for example for recording visibility conditions, find types, site types, and periodizations. Such classifications are easily implemented in the CIDOC CRM in the form of type lists. In some cases, such as broad ceramic types, standardisation can easily be reached by adhering to widely accepted multilingual type lists (e.g., SKOS vocabularies) developed by bodies such as English Heritage. However, in many other cases – not just 'technical' types such as coarse ware pottery fabrics but also fundamental ones such as site types – a large number of different and partially incompatible type lists are in use. Work already done for

[37] We reiterate that the scope of our model for now is limited to systematic field-walking survey, but should at some point be extended to include data collected by other non-invasive methods, such as non-systematic field survey and various types of remote sensing.

the Rome Hinterland Project suggests that such lists, if they are sufficiently documented, and if we impose a superstructure of broad 'container' classes, can be harmonized to a substantial degree.[38]

A further practical limitation lies in the time-consuming process of developing, gaining support for, and adopting a CIDOC CRM extension. This process will require several more rounds of discussion with colleagues in the survey domain and in the CRM community. Whilst we have extensively discussed both our definition of concepts and our initial database mappings with members of the CRM Special Interest Group in meetings and via e-mail, several issues remain unresolved and even seem to require adjustment to existing CRM concepts. Discussions with both the CRM SIG and the Mediterranean survey communities must be continued in order to resolve these. We still aim to have our new concepts included into the official CIDOC CRM standard, continuing work on resolving the issues in the model extension and subsequent testing in collaboration with both the CRM SIG and field survey practitioners. Our involvement in the ARIADNEplus program will hopefully facilitate this.

Once we have established the CRM extension as an official standard, we can take up the challenge of mapping well-documented survey databases to the CRM. This is technically still quite challenging, as it requires exporting the databases to RDF triple format using dedicated software. Within the STELLAR project, tools have already been developed to map archaeological datasets to the CRM and convert them to RDF format, but templates tailored to field survey data are still lacking.[39] Also, in order to be harvested as LOD, the data triples must be served on a dedicated server that allows the data to be searched using an RDF query language. We are hopeful that the ARIADNEplus program will provide the opportunities to do so. Data repositories such as DANS and ADS, already essential in providing FAI versions of the datasets generated by researchers, could take on the task of converting to triple format and serving as LOD.

In this paper we highlighted the partial and fragmentary nature of current documentation and archiving practices in Mediterranean field survey, which clearly do not yet conform to FAIR principles. Although more survey databases are coming online bit by bit, current documentation practices are still insufficient to allow others to understand these data. We realise that it will be challenging to remedy this, as it requires that researchers spend more of their already scarce resources on documenting and archiving their data, but we expect that more strict 'good practice' requirements from funding bodies and repositories will stimulate researchers in this direction. The fundamental point remains, that survey archaeologists cannot be satisfied with a status quo that prevents meaningful re-use and comparison of regional datasets.

Tymon de Haas: t.c.a.de.haas@arch.leidenuniv.nl
Martijn van Leusen: p.m.van.leusen@rug.nl

[38] In our experience, meaningful integration and re-use of site typologies is more likely to be achieved, for example, than that of the omnipresent Common Ware types. Legacy datasets, which are typically poorly documented, will present additional problems for such harmonization efforts.

[39] Besides the Stellar tools, software such as 3M (www.ics.forth.gr/isl/3M/) can also be used to map data to the CRM, to generate LOD triples and create URIs (OLDMAN *et al.* s.d.; DOERR *et al.* 2016, 445-6; OLDMAN & KURTZ 2014, 9-12).

Acknowledgements

This research is based on two survey projects funded by the Netherlands Organisation for Scientific Research under grant numbers 276-61-002 (the Hidden Landscapes Project) and 360-61-030 (The Minor Centres Project), and a Small Data Project funded by the Royal Dutch Academy of Sciences under grant no. 20160226_007. We thank the following colleagues for discussing the structure and concepts of their survey databases with us: Gert-Jan van Wijngaarden and Jitte Waagen (University of Amsterdam, Zakynthos Survey), Franziska Lang and Jana Skundric Rummel (Darmstadt Technical University, Olympia Survey), Rogier Kalkers (Leiden University, Landscapes of Early Roman Colonization project, Molise survey), Michael Galaty (University of Michigan, Regional Archaeology in the Peja and Istog Districts of Kosova project), and Patrick Willett (University of Leuven, Sagalassos Mountain Survey). We further wish to thank the members of the CIDOC CRM special interest group, especially Steven Stead, George Bruseker, Achille Felicetti and Martin Doerr, for on-going discussions regarding our concepts and their position in the CIDOC CRM 'family tree' and to two reviewers for their helpful comments on this paper.

References

ALCOCK, S., CHERRY, J., 2004, 'Introduction', in ALCOCK, S., CHERRY, J. (eds), *Side by side survey. Comparative regional studies in the Mediterranean world*, Oxford, 1-9.

ATTEMA, P., 2004, *Pontine Region Project (It.) fase 2*. DANS. https://doi.org/10.17026/dans-zrf-f8kx

ATTEMA, P., VAN JOOLEN, E., LEUSEN, M. VAN 2001, 'A Marginal Landscape: Field work on the beach ridge complex near Fogliano (South Lazio)', in *Palaeohistoria* 41/42 (1999/2000): 149-162.

ATTEMA, P., BURGERS, G.-J., LEUSEN, M. VAN 2010, *Regional Pathways to Complexity: settlement and land-use dynamics from the Bronze Age to the Republican period* (Amsterdam Archaeological Studies 15). Amsterdam: Amsterdam University Press.

BARKER, G., MATTINGLY, D., 1999, 'General editors' introduction: the POPULUS project', in BINTLIFF, J., SBONIAS, K. (eds), *Reconstructing past population trends in Mediterranean Europe*, Oxford: iii-v.

BINDING, C., CHARNO, M., JEFFREY, S., MAY, K., TUDHOPE, D., 2015, 'Template Based Semantic Integration: From Legacy Archaeological Datasets to Linked Data', in *International Journal on Semantic Web & Information Systems* 11-1: 1-29

BLANTON, R., 2001, 'Mediterranean myopia', in *Antiquity* 75: 627-629.

BLANTON, R. 2004, 'Settlement Pattern and Population Change in Mesoamerican and Mediterranean Civilizations: A Comparative Perspective', in ALCOCK, S., CHERRY, J. (eds.) *Side-by-Side Survey: Comparative Regional Studies in the Mediterranean World*, Oxford: 206-240.

BURGERS, G.-J., ATTEMA, P., LEUSEN, M. VAN, 1998, 'Walking the Murge: Interim report of the Ostuni field survey (Apulia, southern Italy)', in *Studi di Antichità* 11: 257-282

CASAROTTO, A., 2018, *Spatial patterns in landscape archaeology: a GIS procedure to study settlement organization in early Roman colonial territories*, ASLU 43, Leiden University Press.

CRIPPS, P., DOERR, M., HERMON S., HIEBEL, G., KRITSOTAKI A., MASUR, A., MAY, K., SCHMIDLE, W., THEODORIDOU, M., TSIAFAI, D., 2014, *CRMarchaeo: the excavation model. An extension of CIDOC-CRM to support archaeological excavations*, FORTH [version 1.2, march 2014]

CROFTS, N., DOERR, M., GILL, T., STEAD, S., STIFF, M. (eds.), 2011, *Definition of the CIDOC Conceptual Reference Model Version 5.0.4,* available from: http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_5.0.4.pdf

DOERR, M., THEODORIDOU M., ASPÖCK, E., MASUR, A., 2016, 'Mapping archaeological databases to CIDOC CRM', in CAMPANA, S., SCOPIGNO, R., CARPENTIERO, G., CIRILLO, M. (eds), *CAA 2015. Keep the revolution going*: 255-260.

DOERR, M., KRITSOTAKI, A., ROUSAKIS, Y., HIEBEL, G., THEODORIDOU, M., 2018, 'Definition of the CRMsci An Extension of CIDOC-CRM to support scientific observation', Unpublished Report (Version 1.2.5), FORTH.

FEIKEN, H., 2014, *Dealing with biases. Three geo-archaeological approaches to the hidden Landscapes of Italy*, Phd thesis, University of Groningen.

FRANCOVICH, R., PATTERSON, H. (eds), 2000, *Extracting Meaning from Ploughsoil Assemblages*, Oxford.

Groninger Instituut voor Archeologie (Rijksuniversiteit Groningen), 2010, *The Dzarylgac Survey Project*, DANS, https://doi.org/10.17026/dans-zwa-fma6

HAAS, T.C.A. DE, 2016, *Fora, stationes and sanctuaries: the role of minor centres in the economy of Roman central Italy*, DANS, https://doi.org/10.17026/dans-xtq-kvav

HAAS, T. DE, LEUSEN, M. VAN, 2016, *Extending CRMarchaeo for archaeological field survey - a pilot*, Unpublished report, Groningen Institute of Archaeology.

HAAS, T.C.A. DE, TOL, G.W., forthcoming, 'The analytical potential of intensive field survey data: developments in the collection, analysis and interpretation of surface ceramics within the Pontine Region Project', in MEENS, A., NAZOU, M., VAN DER PUT, W. (eds.), *Fields, Sherds and Scholars*.

HIEBEL, G., DOERR, M., EIDE, Ø.,THEODORIDOU, M., 2015. *CRMgeo: a Spatiotemporal Model An Extension of CIDOC-CRM to link the CIDOC CRM to GeoSPARQL through a Spatiotemporal Refinement,* Unpublished Report (Version 1.2), FORTH.

HOLSTROM, S, VOORRIPS, A., KAMERMANS, H., 1989, *Agro Pontino Archeological Survey*, DANS, https://doi.org/10.17026/dans-2bq-pjrf

KEAY, S.J., MILLETT, M., 2003, *Ager Tarraconensis Field Survey Project [data-set]*. York, Archaeology Data Service, https://doi.org/10.5284/1000351

KINTIGH, K., ALTSCHUL, J., BEAUDRY, M., DRENNAN, R., KINZIG, A., KOHLER, Y., LIMP, F., MASCHNER, H., MICHENER, W., PAUKETAT, T., PEREGRINE, P., SABLOFF, J., WILKINSON, T., WRIGHT, H., ZEDER, M., 2014, 'Grand challenges for archaeology', in *Proceedings of the National Academy of Sciences* 111-3: 879-880. https://doi.org/10.7183/0002-7316.79.1.5

LANG, M., CARVER, G., PRINZ, S., s.d., 'Standardised Vocabulary in Archaeological Databases', Available from: https://www.academia.edu/9704806/Standardised_Vocabulary_in_Archaeological_Databases [accessed 3-11-2016]

LAWRENCE, D., PHILIP, G., HUNT, H., SNAPE-KENNEDY, L., WILKINSON, T., 2016, 'Long Term Population, City Size and Climate Trends in the Fertile Crescent: A First Approximation', in *PLoS ONE* 11(3): e0152563. doi:10.1371/journal.pone.0152563

LAUNARO, A., 2011, *Peasants and slaves. The rural population of Roman Italy (200 BC to AD 100)*, Cambridge.

LEUSEN, M. VAN, 1998, 'Archaic Settlement and Early Roman Colonisation of the Lepine Foothills', in *Assemblage* 4: http://www.shef.ac.uk/~assem/4/

LEUSEN, P.M. VAN, 2017, *Monti Lepini Survey 2006-2008*, DANS, https://doi.org/10.17026/dans-z2d-45zz

LEUSEN, M. VAN, ATTEMA, P., 2001/2002, 'Regional archaeological patterns in the Sibaritide; preliminary results of the RPC field survey campaign 2000', in *Palaeohistoria* 43/44 (2001/2002): 397-416.

LEUSEN, M. VAN, HAAS. T. DE, 2005, 'eDNA: naar een e-Depot voor de Nederlandse Archeologie', in *PaleoAktueel* 16: 3-7.

MATTINGLY, D., 2000, 'Methods of collection, recording and quantification', in FRANCOVICH, R., PATTERSON, H. (eds), *Extracting meaning form ploughsoil assemblages*, Oxford: 5-15.

MILLETT, M., 2000, Dating, quantifying and utilizing pottery assemblages from surface survey, in FRANCOVICH, R. & PATTERSON, H. (eds), *Extracting Meaning from Ploughsoil Assemblages*, Oxford: 53-59.

OLDMAN, D., KURTZ, D., 2014, *The CIDOC Conceptual Reference Model (CIDOC-CRM): PRIMER*. FORTH [version 1.2, august 2014]

OLDMAN, D., THEODORIDOU, M., SAMARITAKIS, G., s.d., *Using Mapping Memory Manager (3M) with CIDOC CRM,* Unpublished paper, available from: http://83.212.168.219/DariahCrete/sites/default/files/mapping_manual_version_4g.pdf

PERRIN, K., BROWN, D., LANGE, G. *et al.*, s.d., *A Standard and Guide to Best Practice For Archaeological Archiving in Europe*, EAC Guidelines 1, Namur.

RICHARDS, J., ROBINSON, D. (eds), 2000, *Digital Archives from Excavation and Fieldwork: A guide to good practice*, Oxford.

STEAD, S., DOERR, M., ORE, C.-E., 2018, *CRMinf: the Argumentation Model. An Extension of CIDOC-CRM to support argumentation*, Unpublished Report (Version 0.9), Paveprime Ltd.

STELLAR, s.d., *STELLAR mapping and extraction guidelines*, Available from: http://hypermedia.research.southwales.ac.uk/resources/STELLAR-applications/ [consulted 3-11-2016]

STONE, D.L., 2004, 'Problems and possibilities in comparative survey: a North African perspective', in ALCOCK, S., CHERRY, J. (eds.), *Side-by-Side Survey: Comparative Regional Studies in the Mediterranean World*, Oxford: 132-143.

TERRENATO, N., 1996, 'Field Survey Methods in Central Italy (Etruria and Umbria). Between Local Knowledge and Regional Traditions', in *Archaeological Dialogues* 1996-2: 216-230.

TOL, G., DE HAAS, T., ARMSTRONG, K., ATTEMA, P., 2014, 'Minor Centres in the Pontine plain: the cases of Forum Appii and Ad Medias', in *Papers of the British School at Rome* 82: 109-134

TOL, G., DE HAAS, T., 2017, 'The role of minor centres in local economies: new insights from recent archaeological fieldwork in the lower Pontine plain', in *The Melbourne Historical Journal* 44-2: 33-61.

WIJNGAARDEN, G.J.M. van, 2015, *The Zakynthos Archaeology Project*, DANS, https://doi.org/10.17026/dans-27f-zd2u

WILKINSON, M. D. *et al.,* 2016, 'The FAIR Guiding Principles for scientific data management and stewardship', in *Scientific Data* 3: 160018. doi:10.1038/sdata.2016.18

WITCHER, R.E., 2006, 'Settlement and Society in Early Imperial Etruria', in *Journal of Roman Studies* 96: 88-123. http://dx.doi.org/10.3815/000000006784016161

WITCHER, R.E., 2008, '(Re)surveying Mediterranean Rural Landscapes: GIS and Legacy Survey Data', in *Internet Archaeology* 24: https://intarch.ac.uk/journal/issue24/witcher_toc.html