KULA KNOWLEDGE CREATION, DISSEMINATION, AND PRESERVATION STUDIES

Börjesson, Lisa, Olle Sköld, Zanna Friberg, Daniel Löwenborg, Gísli Pálsson, and Isto Huvila. 2022. Re-purposing Excavation Database Content as Paradata: An Explorative Analysis of Paradata Identification Challenges and Opportunities. *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 6(3). https://doi.org/10.18357/kula.221

RESEARCH ARTICLE

Re-purposing Excavation Database Content as Paradata: An Explorative Analysis of Paradata Identification Challenges and Opportunities

Lisa Börjesson
Department of ALM, Uppsala University

Olle Sköld
Department of ALM, Uppsala University

Zanna Friberg
Department of ALM, Uppsala University

Daniel Löwenborg
Department of Archaeology and Ancient History, Uppsala University

Gísli Pálsson
Department of Archaeology and Ancient History, Uppsala University

Isto Huvila
Department of ALM, Uppsala University

Although data reusers request information about how research data was created and curated, this information is often non-existent or only briefly covered in data descriptions. The need for such contextual information is particularly critical in fields like archaeology, where old legacy data created during different time periods and through varying methodological framings and fieldwork documentation practices retains its value as an important information source. This article explores the presence of contextual information in archaeological data with a specific focus on data provenance and processing information, i.e., paradata. The purpose of the article is to identify and explicate types of paradata in field observation documentation. The method used is an explorative close reading of field data from an archaeological excavation enriched with geographical metadata. The analysis covers technical and epistemological challenges and opportunities in paradata identification, and discusses the possibility of using identified paradata in data descriptions and for data reliability assessments. Results show that it is possible to identify both knowledge organisation paradata (KOP) relating to data structuring and knowledge-making paradata (KMP) relating to fieldwork methods and interpretative processes. However, while the data contains many traces of the research process, there is an uneven and, in some categories, low level of structure and systematicity that complicates automated metadata and paradata identification and extraction. The results show a need to broaden the understanding of how structure and systematicity are used and how they impact research data in archaeology and comparable field sciences. The insight into how a dataset's KOP and KMP can be read is also a methodological contribution to data literacy research and practice development. On a repository level, the results underline the need to include paradata about dataset creation, purpose, terminology, dataset internal and external relations, and eventual data colloquialisms that require explanation to reusers.

Keywords: metadata; paradata; metadata extraction; data reuse; research data; unstructured data; archaeological data

Introduction

Improving the efficiency and effectiveness of scholarly work through data sharing and reuse is a central theme in the contemporary research policy discourse (see e.g., Wilkinson et al. 2016 on the FAIR principles[1]; the EU's open science policy, "Open Science" n.d.). To this end, collections of aggregated legacy data play a key role by enabling reuse and allowing new research questions to be asked and addressed. Aggregation of data can, for example, underpin the creation of new knowledge by facilitating comparisons and cross-analysis of results from investigations carried out in different locations and contexts during different periods of time. These new opportunities do, however, come paired with challenges: new research questions often instigate a need to know more about how the data was created than what was recorded in the original metadata.

Many of the challenges associated with data reuse are common to all research disciplines working with legacy data. The challenges are especially prominent in field sciences like archaeology that are hallmarked by large methodological and epistemological variation, a diversity of study contexts, and long-term usefulness of legacy data. For example, in archaeology, knowing undocumented details about the excavation techniques and tools that led to particular observations and conclusions can be crucial in new studies that use aggregated legacy data. Such information might not have been deemed important enough to be included in the metadata during the original investigation, but may at a later date prove to be crucial for estimating the usability and reliability of the original findings for, for instance, cross-site comparisons (Ullah 2015).

A key step towards facilitating extended and more purposeful data reuse is to better understand and assess the process of data creation. In this article, we focus on paradata, a subset of contextual information that describes data creation and manipulation processes and their underpinnings, which is often left undocumented in structured dataset descriptions but commonly is implicitly present and, at least to a competent reader, to varying degrees identifiable in the data itself (Huvila 2020a; Huvila, Sköld, and Börjesson 2021). Paradata, like descriptions of methods and tools used to produce data, is particularly interesting because it is often central to making and communicating assessments of data reliability and an important facilitator of productive and efficient data reuse (Faniel et al. 2013).

The aim of this article is to elucidate the challenges and opportunities of identifying paradata in fieldwork data. We meet this aim by reporting the findings from an explorative exercise of extracting paradata from a dataset exported from a fieldwork database created using Intrasis, a Swedish geographic information system (GIS)-based information management system developed for archaeological field documentation. The article addresses the following research questions:

RQ1. Which paradata categories are possible to identify in the fieldwork database?
RQ2. What are the technical challenges and opportunities of identifying paradata?
RQ3. What are the epistemological challenges and opportunities in assuming identified paradata as evidence of data creation processes?
RQ4. What are the implications of the technical and epistemological challenges and opportunities for using identified paradata as a basis for data reliability assessments?

The article provides insights into the desiderata and challenges of data reuse in field sciences like archaeology and explores the possibilities and obstacles inherent in repurposing information originally created as in-field documentation for data description and evaluation purposes. In this article, we assume a functional perspective on data reliability assessments as activities guided by the objective to evaluate if a certain dataset is suitable for a specific reuse purpose. From this perspective, reliability assessments can evaluate different types of data reliability such as internal reliability (is data commonly acceptable), relative reliability (is data acceptable to the user), and absolute reliability (does data resemble reality) (Agmon and Ahituv 1987). Each type of reliability assessment draws on different variables internal or external to the dataset and sets different limits for acceptable versus non-acceptable levels of reliability. The paper does not go as far as elaborating on how each type of paradata serves the different types of data reliability assessments. Still, the paper contributes towards the advancement of research data description, provides conceptual foundation for explorative studies of how paradata could be used in different types of data reliability assessments, and furthers data literacy by explicating principles and ad-hoc solutions underpinning dataset structuring and content.

Literature Review

Needs and Challenges in Data Reuse

Data reuse literature shares a broad consensus that a reuser needs high-quality data and high-quality contextual information about the data in question. However, the literature also acknowledges the difficulty

[1] The FAIR principles for data management describe making data Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016).

of knowing, as a data creator or indeed as repository staff, what high quality is for a potential reuser and how to go about providing good or even good enough data. Quality is not a universally defined feature of data. Data quality can be about accuracy or completeness but also relate to suitability, relevance, and availability and depend on consistency, verifiability, or believability (Koesten et al. 2020; Huvila 2020b).

In order to understand and judge data quality, data reusers need basic or intrinsic metadata (e.g., the identity of data creator, date of creation, and terms of use), information on the context and methods of data creation (Kim, Yakel, and Faniel 2019), and processual information (i.e., paradata)—for example, declaration of selection and exclusion decisions made during the research process (Allison 2008)—about the data creation. Koesten et al. (2020) identify common features associated with the reuse potential of datasets in an extensive review of data reuse literature across several disciplines. The difficulties and importance of understanding the context of data creation are mentioned in multiple studies. There is a consensus on the critical significance of methodological information about data creation for informed reuse. However, Kim, Yakel, and Faniel (2019) found in a study of data description requirements that repositories rarely require detailed contextual information on how data is constructed or manipulated. Only a few of the repositories surveyed in their study asked for methodology-related information on, for example, data cleaning and edits (Kim, Yakel, and Faniel 2019). The findings point to a gap between users' needs and what descriptions are provided by data creators or mandated by repositories, especially in relation to process information. An obvious remedy is to reconsider the requirements and ask for more comprehensive descriptions. Another, ostensibly less labour-intensive and parallel approach investigated closer in the present article is to see to what extent relevant process information (i.e., paradata) about, for example, methodology, versioning, and provenance, can be found and extracted from the data itself.

Kim, Yakel, and Faniel (2019) note also that existing data descriptions are often incomplete, inaccurate, and inconsistent in format and terminology. A partial explanation can be found in lacking standards and repository requirements, but there are also several additional barriers to sharing data for reuse. As Faniel et al. (2021) show, there are considerable differences in data practices even within research fields with strong documentation standards, and there is great variation on an institutional, or even personal, level in what is deemed important to document (Faniel et al. 2021; cf. Börjesson 2016). When surveying data reuse in the Earth system sciences, Yan et al. (2020) found that a lack of incentives for investing in facilitating data reuse was regarded as a major challenge by the study participants. Their responses indicated that institutions and funding agencies do not encourage the time-consuming work of producing well-documented datasets. Respondents stated also that the contemporary rapid publishing culture prioritises multiple and novel results above fewer and more meticulously documented studies (Yan et al. 2020).

Legacy Data Use in Archaeology

Even if it would be possible to find resources and establish satisfactory procedures for documenting data and data-related procedures, there will always be a lot of potentially useful data that was not created with reuse in mind. This applies not least to legacy data. The store of archaeological legacy data is huge, and the amount of existing data is also continuously growing as a result of ongoing fieldwork arising from the historical environment legislation that regulates urban and rural development in many countries (for an overview, see the 2021 *Internet Archaeology* special issue on digital archiving edited by Richards et al.). Legacy data of varying ages is increasingly used for secondary research in archaeology (Wylie 2017; Secci et al. 2019; Brown, Goodchild, and Sindbæk 2014; Ellis 2008; Boozer 2014), a trend which can be expected to continue as the sharing and availability of data grows.

As data infrastructure projects—like the Digital Archaeological Record (tDAR) and Open Context in the United States, Urdar in Sweden, Archaeological Digital Excavation Documentation (ADED) in Norway, and the ARIADNE and ARIADNEplus projects funded by the European Union—succeed in making data increasingly findable and accessible, questions about how to make data interoperable and reusable come to the fore. The Chaco Research Archive (CRA), an online database with over one hundred thousand searchable records from archival sources on architecture and artefacts, surveys, and excavations (Heitman, Martin, and Plog 2017), provides an illustrative example of the data description challenges curators face when attempting to increase the reusability of accessible resources. Reporting on the establishment of the CRA, Heitman, Martin, and Plog (2017) discuss the intellectual and technical work needed to provide access to archaeological data accumulated over many years during numerous excavations. They point out the importance of capturing both original observations without intrusion and providing metadata in normalised fields for searchability to preserve continuity. Montoya and Morrison (2019) report on comparable challenges of data continuity that they observed in the curation history of the Angel Mounds (AM) collection at the Glen A. Black Laboratory of Archaeology, Indiana. They remark that the practices of archaeologists, archivists, and museum curators differ and that this presents a challenge for maintaining continuity of the contextual information tied to the data (Montoya and Morrison 2019).

While there is some research into the challenges and opportunities of data reuse, relatively few studies engage in a detailed exploration of the technical or epistemological issues of repurposing legacy data. Challenges in data reuse are occasionally described in the reporting of projects using legacy data. For instance, Sobotkova (2018) expands on the difficulties encountered in her aggregation of data for a large-scale regional study on burial mounds. Sobotkova suggests that archaeology is one of several fields where data production is slow and labour-intensive, which in turn decreases the pace of archaeological data sharing. Data collection may require years of work in the field and in archives to derive data from multiple sources. As a result, datasets in archaeology are rarely consistent in form or structure. Even though data creation in archaeology is labour-intensive, Sobotkova (2018) reminds us that reusing other people's data often requires as much or even more work. Considerable efforts must be made to understand, test, and reconstruct a dataset, and to delve into what Allison (2008) describes as the complex layers of selection and exclusion present in legacy data. Additionally, finding the right sources and extracting associated data from, for example, PDF documents can require years of experience, implicit knowledge, and informal contacts with people with the needed expertise (Sobotkova 2018). Some of the problems of archaeological data reuse are primarily technical. For instance, Ullah (2015) notes that errors in spatial survey data are hard to discover and correct without proper contextual information. The findings of Atici et al. (2013) demonstrate in parallel that technical challenges are not necessarily solvable by technical means but require interpretation. In their study, three experts examined a zooarchaeological legacy dataset. Even if the experts all chose similar approaches to study the data and asked similar questions when cleaning and preparing the data, their non-identical interpretations and decisions on how to fill the lacunae in the contextual information resulted in three substantially different datasets (Atici et al. 2013).

Sobotkova (2018) laments the perceived unattractiveness of data reuse as compared to archaeological fieldwork and calls for the archaeological community to put more value on using and citing secondary data. The sentiment that archaeology is too preoccupied with fieldwork at the expense of enabling future archaeological research is echoed in the literature on a curation crisis in archaeology (e.g., Voss 2012; Kersel 2015). A part of the problem lies in the funding structures that, as both Voss (2012) and Sobotkova (2018) note, are geared toward financing fieldwork and data generation rather than reuse.

Even a brief overview of data reuse in archaeology illustrates the central message of Kansa and Kansa (2021): data use, and the accompanying need to understand the data, is as intellectually demanding as any other research approach and a successful data (re)user needs to be adequately literate for the task. Literacy in this context does not refer to solely technical skills but rather to the competency to understand the "underlying principles and challenges of data" (Bhargava et al. 2015; cf. Kansa and Kansa 2021)—that is, to analyze and interpret the structuring of data and patterns in data file content as demonstrated in the forthcoming analysis. A paramount reason that data (re)use takes analytical and interpretative literacy is that data seldom is *prêt à utiliser*—ready to use as such (e.g., Ullah 2015; Atici et al. 2013; Sobotkova 2018). A reuser needs to read data in its disciplinary context to bridge gaps in contextual information (Atici et al. 2013) and might need supplementary information from sources external to the dataset, knowledge of where to find such sources, and the know-how to extract information therefrom (Sobotkova 2018).

Material and Methods

Metadata and Paradata Generation by Extraction

In the present study, we explore the challenges and opportunities of identifying paradata in fieldwork data by extracting paradata from a dataset exported from an archaeological fieldwork database. In the following, we first review issues related to metadata and paradata extraction and then describe the excavation database analysed in the study.

Research data, including archaeological fieldwork data, is commonly described on a dataset level. Metadata can be assigned by a researcher involved in the data creation or, for example, by a project manager, data curator, or librarian. Dataset descriptions follow repository-specific metadata schemes that are in turn based on general or discipline-specific standards. While the traditional role of metadata is to support resource discovery, more recent community- and discipline-specific standards for data description often aim to facilitate (re)use beyond mere retrieval. Such schemes incorporate descriptors requested in many of the previously discussed studies (e.g., Faniel, Frank, and Yakel 2019; Sobotkova 2018) on, for instance, provenance, collection methods, and quality assessment categories (Börjesson, Sköld, and Huvila 2021). For example, geographic information standards from the earlier-used Content Standard for Digital Geospatial Metadata (CSDGM) (n.d.) to the current ISO 19115-1:2014 (2014) contain categories explicitly intended to facilitate reuse.

However, archaeological fieldwork data has most commonly been produced to enable the primary analysis and reporting of a single site rather than to facilitate data aggregation and reuse. Datasets have traditionally not been considered standalone products or genres, and consequently they have not been defined and prepared as fieldwork output (Huvila 2016). Data provenance and process information may exist in the documentation produced during, for example, an archaeological excavation, but it is likely to be scattered across fieldwork diaries, context sheets, notes on maps and drawings, site photos, and the field reports (Huvila, Sköld, and Börjesson 2021; Huvila 2006). Metadata and process information in attached read-me-files, table definition files, or in the dataset itself (e.g., supplementary information in column headers) may be sparse or non-existent.

Considering the spread of information across the documentation, an ideal approach to extract the maximum amount of paradata would be to harvest all existing documentation for provenance and process information. When such an approach is not feasible due to the lack of documentation or access to it, or lack of funding for manual harvesting, it is relevant to consider—as we do in the present study—if sufficient paradata can be extracted from the relatively structured fieldwork data file alone. First, data documented during fieldwork or analysis has the advantage of being intrinsic to the research activity as opposed to being "manufactured" for descriptive purposes (cf. Silverman's notion of "manufactured" qualitative data [2013, 32]). In a sense, assigned dataset-level metadata distills dataset content and adds an interpretative layer. Second, a fully developed paradata extraction approach could enable meta- and paradata extraction ranging from single data units, through selections of units, to a full dataset. Such scaling would enable, for example, paradata such as "X percent of the units were excavated by machine, Y percent of the units were excavated manually," giving a richer insight into the range of methods used to produce the data presented. Although such approaches should also be treated carefully in terms of the evidentiary quality of the output, metadata extraction based on file content would provide a basis to test and nuance manually assigned descriptors in, for instance, metadata or reports.

In addition to data proximity achieved by extraction approaches, a third potential gain for fieldwork data description practices would be the potential to enable adaptable meta- and paradata (i.e., extraction based on research interests) following situated information needs of data reusers. Suppose a researcher aims to aggregate data on sifted soil samples from several locations to learn more about past aquatic environments. High-precision information about sieving mesh size is crucial for understanding which types of fishbones have been found and recorded (and which ones potentially slipped through the sieving mesh). Thus, knowing about sieving methods and equipment used in the field can be critical in deciding how to combine and compare fishbone datasets to understand aquatic environments (Olson and Walther 2007), but is likely insignificant for a researcher with an interest in, for example, settlement patterns. Since fixed metadata standards build on explicit and implicit assumptions about what things and processes the scheme should describe and for what purpose, a set standard inevitably has its limitations (Börjesson, Sköld, and Huvila 2021). Therefore, the need to find ways to generate intentionally purposed meta- and paradata based on topical user needs grows along with the ambition to make use of legacy data and use data across disciplinary boundaries.

Established approaches for metadata extraction rely on various methods to identify where to find certain informative content like a volume number in a journal article PDF (Tkaczyk et al. 2015) or types of cell contents in a spreadsheet (Roy et al. 2016). A common denominator is that the extraction approaches use defined metadata categories to identify areas and segments with potential informative value. Thus, the first step in performing paradata extraction on fieldwork datasets is to tease out what content could provide information about research processes—for example, indications of excavation technique—and where to find it in the dataset.

The next section of the article outlines our case study by first introducing the research data infrastructure project and the data migration process necessary for making this type of archaeological fieldwork legacy data accessible and interoperable. We continue with a descriptive analysis of the structure of, and types of contents in, the dataset before proceeding to the explorative analysis of how the table structure and the table content could be identified as knowledge organization paradata (KOP) and knowledge making paradata (KMP).

The Excavation Database Case

Urdar,[2] the research infrastructure project enabling access to the analyzed data, aims to make digital excavation data openly accessible and possible to integrate with other data sources with the intent of

[2] The Urdar project is funded by the Swedish Riksbankens Jubileumsfond and is a collaboration between Uppsala University, the Swedish National Heritage Board, Umeå University, and Lund University. Read more at: https://www.arkeologi.uu.se/Research/Projects/urdar-en/.

enabling data-driven research on human prehistory and history (Larsson and Löwenborg 2020). Urdar's main goal is to ensure that existing data will be secured in open and safe formats (CSV and Geopackage) so that it can be archived in the Swedish National Heritage Board e-archive and linked to the national excavations register. The long term goal of Urdar is that this information should be included in a permanent research infrastructure so that researchers, heritage management professionals, and other stakeholders can access and analyse the information in an aggregated form. The project is running from 2020 to 2023 and is at the time of this writing in a phase of exploring available information, re-creating missing information, and investigating ways of making the data useful outside of its original context of creation.

The bulk of digital archaeological data produced in Sweden is managed using Intrasis, a Swedish GIS-based information management system developed for archaeological field documentation. A major challenge of working with Intrasis data is the complexity of the relational database at the core of the system. Making the information usable for archaeologists and others with varying computational skills, while at the same time preserving as much of the data's inherent potential as possible, is difficult. As data in Intrasis is only semi-structured and the design of the database varies to some extent from project to project, there is no straightforward way of exporting information that ensures its aggregability for reuse. Intrasis has also been designed to not enforce controlled vocabularies but to allow archaeologists flexibility in what terms to use. Hence, archaeological information in Intrasis is heterogeneous. In Urdar, that diversity will initially be preserved in an archival copy which is retained as a representation of the original decisions made during the fieldwork. However, to analyse the material as a whole, a dedicated database ("Intrasis archive") with all information, where data can be rearranged and repurposed to suit different research questions, will be maintained for research purposes.

Similar to the experiences described in the literature (Faniel, Frank, and Yakel 2019; Yan et al. 2020), making Intrasis data usable beyond its context of creation requires adequate paradata—that is, documentation of data collection and management procedures. This includes, for instance, information on georeferencing methods used (see also Ullah 2015), the granularity of survey (e.g., how many of the identified post holes were fully excavated), and the perceived level of certainty of interpretation (is the structure "a wall" or "a wall?") to evaluate and potentially correct or enrich the data before reuse (see also Sobotkova 2018). This information is only partially organised into separate fields in the Intrasis data, which means that paradata must be identified across data categories. In the next section, we introduce a subset of the excavation database used in this paper as a testbed to explore how content with the capability to function as paradata could be identified and its reliability assessed.

Excavation Data as Paradata Identification Testbed

The dataset analysed in this study comes from an excavation in the locality of Örja in the municipality of Landskrona in Scania (Swed. Skåne) in southernmost Sweden. The excavation was conducted in 2010 by the southern Swedish local office (UV Syd) of the Contract Archaeology Service of the Swedish National Heritage Board (since 2015 "The Archaeologists," a part of the National Historical Museums) and reported in 2013. The investigation was initiated before a major land development project on the site and resulted in the identification and investigation of a large number of remains from the early Neolithic period to the modern age. The project report (Sabo et al. 2013) is an archaeological field report with 358 pages, including references and an appendix. The documentation data was managed using the Intrasis documentation and data management software.

The spreadsheet with 24,424 rows of data (here referred to as "data units") analysed in this article is a prototype export with a selection of data connected to the excavated objects from the Örja project in the Intrasis database. The data units were first exported from Intrasis to PostgreSQL and enriched (Table 1) with a running index (Column A) and spatial descriptors (Columns C, D, E), an array of all associated attributes (L), information about table relations (M, N), and finally exported to an XSLX-file for analysis. The purpose of enriching the content retrieved from the Intrasis database with spatial descriptors was a) to situate the information geographically, and b) to add the information required for linking the data to the national excavations register. Certain information included in the Intrasis database was omitted in the export, including log and event registries with information on individual users' interactions with the data. The file thus makes up a prototype of a flat version of the data, which demonstrates how the available archaeological information can be aggregated without relational information to make it easy to use in conventional GIS software designed for the analysis of flat-format tabular data. Besides the analysed spreadsheet, Urdar is also developing an export version of the data that maintains the full relational structure of the database. This version will be incorporated in the final delivery package to the e-archive that will be linked to the national excavations register.

Table 1: The dataset with excavation data from Örja 1.9., compiled from data in the Intrasis excavation database together with spatial metadata produced by the Urdar project

Column	Column title (from first row)	Data type	Column content
A	pk	Integer	Running index from 1 to 24,424
B	site_name	String	Site name: "Örja 1:9 SU" for all posts
C	landskap	String	Province where site is located: "Skåne" for all posts
D	lan_name	String	County where the site is located: "Skåne län" for all posts
E	parish_name	String	Parish of the site: "Örja" for all posts
F	intrasis_archive	String (ID)	ID of the Intrasis archive file: "s2009045" for all posts
G	object_id	Integer (ID)	ID for individual documented features: series (e.g., 4XXX, 1XXXX, 4XXXX)
H	object_name	String (code)	Codes for object types: e.g., "G 16 Filling trench" (data unit 61), "JÄÅ B Filling posthole" (data unit 217), "G 20 Burnt layer" (data unit 615), numerous empty cells
I	class	String (code)	Code for feature class: e.g., "Stratigraphic object," "Find"
J	subclass	String (code)	Code for a subclass: e.g., "Dug hole, posthole" (data unit 621), "Lead" (data unit 6099), "Copper alloy" (data unit 6098), "Flintstone UV Syd" (UV Syd refers to the contractor, data unit 7190), "Stone and brick construct" (data unit 7294)
K	spatial_type	String (code)	GIS object type: Polygon, Point, Multipoint, or Polyline
L	attr_array	Attribute array	Common attributes include: e.g., "Interpretation_temp," "Interpretation," "Find retrieval technique," "Composition" (of feature), "Consistency" (e.g., "wet"), "Contamination"
M	parent_relations	Array	Stratigraphic relations to other features: e.g., "Above 58449," "Contains 56027, Above 4610"
N	child_relations	Array	Stratigraphic child relations of the feature: e.g., "Belongs to 1312, Below 5525"
O	description	String	Contains fields—e.g., "Interpretation" and "Grouping discussion" (e.g., to which building the feature belongs, data unit 13,002)—with free-text descriptions

The flat and the relational database formats serve different purposes. The flattened structure is suitable for statistical analysis and machine learning, while also being more accessible to users who do not have experience of relational data structures. The relational format closely adheres to the way in which archaeological knowledge is produced in the field by modeling the knowledge organization of the widely used single-context recording system (see Roskams 2001).

The structure of the dataset is described in Table 1 with column titles, the data type in the column, and a description of the column content with examples from the data. All translations of dataset content from Swedish to English were made for the purpose of this article. Names of individuals are replaced by "NN."

Identifying Paradata in the Excavation Dataset

The analysis of the data is based on an iterative close reading of the dataset, including its structure and the contents of the data units, with the purpose of identifying structural and content features that can function as paradata. The initial round of analysis was conducted by co-authors Börjesson and Huvila. The findings were cross-checked and the analysis was elaborated by Börjesson and Huvila together with the Urdar project staff who compiled the dataset, co-authors Löwenborg and Pálsson.

In future studies, this procedure could be—if possible and feasible—complemented by an additional feedback round with the original creators of the data, the archaeologists who performed and documented the fieldwork. Doing this would also address any ethical concerns of working with data which is legally public but was not originally intended to be published for reuse. In the current analysis, data ethics were taken into consideration by being mindful of interpreting intentionality in the data creation, respectful of data creators' choices and work, and careful of not referring to personal names. The research procedure followed relevant international and national guidelines for research ethics and current legislation in Sweden (Swedish Research Council 2017). A formal ethics review was not applicable for the study.

Unsurprisingly, the project file does not contain a specific paradata column. Only a few metadata standards document paradata explicitly in specific constructs (exceptions include CARARE; see D'Andrea and Fernie 2013). However, the dataset contains two types of information with the capability to function as paradata: 1) the structure of the dataset and the columns provide *knowledge organization paradata* (KOP) of how the site and its physical features have been transformed into data units (KOP:DU) and into archaeological information (KOP:AI); 2) the dataset contains traces of *knowledge-making paradata* (KMP)—data that describes fieldwork and analysis processes and transformations (Table 2). Here follows a descriptive account of the project file structure and content. In the section "Using Paradata for Reliability Assessments," we explain how KOP and KMP could be analysed for data evaluation purposes.

KOP:DU contains evidence and traces of how an archaeological site is transformed into information as documentation, representation, or substitute of the site (cf. categories of documentation in, for example, Furner 2021) by categorising and classifying the site in information objects (data units in the dataset). It is created by *relating* data units to each other either hierarchically through child-parent relationships or horizontally across the dataset and to a specific Intrasis archive file; by *assigning* ids, names, descriptions, or attributes to them; and by *declaring* the type of a related resource (the type of the corresponding spatial object—e.g., polygon). It is notable, however, that some categories only inform about the aggregation of disparate data units into a single file; in this case, the index was introduced at the compilation of the analysed dataset and does not reflect on any order or chronology of in-field observations.

In comparable terms, KOP:AI contains evidence and traces of how a physical archaeological site is transformed into archaeological information. This is done by *relating* archaeological features to each other. In addition, archaeological features represented by data units are *contextualised* (i.e., assigned to a specific context, that of a named archaeological site), *situated* in geographical space by relating them to a geographical point of reference (e.g., parish, region), and *assigned* a sequential id and a name (literally, naming). Archaeological features are also *documented* using information types (an Intrasis archive, spatial shapes such as polygon or point) and *classified*, and their aspects are *described and interpreted* using attributes and textual descriptions.

In addition, the dataset contains KMP, although with a wide variety of systematicity. Whereas archaeological documentation is expected to systematise both observations (i.e., site and archaeological features) as archaeological information and describe the investigation (transformation) process (Hodder 1989), the latter tends to be subordinate and instrumental to the aims and understanding of the results of the first one (Gustafsson and Magnusson Staaf 2001). The immediately legible KMP elucidates interpretations and their change in time and investigation methods as well as provides occasional indications of who did what (Columns L, O) and to whom certain tasks were assigned (Column J). Moreover, the dataset contains sedimented traces of knowledge-making processes in its KOP:AI that would require explanation to be legible as sources of paradata. As an example, the assigned "object_id" (Column G) can reflect several different types of sequences. A range of numbers (e.g., 10–19) out of a chronological sequence (e.g., 1–100) can be reserved for a certain group of finds or a certain area of the site without the rationale behind this breakout sequence being explicated.

Mostly, the data is not explicit about who wields description and selection power (Warner 2010) to decide what is a site and how it is delimited (Column B)—whether it is an archaeologist with a research question, the priorities of the land development that led to the excavation, or the premises of singling out features (i.e., data units) in the dataset. The principal exceptions are Column J, which contains references to organisations where specific find types were planned to be submitted for analysis, and Column O, which contains questions directed to and comments signed by individual archaeologists who participated in the investigation, such as "NN: Several trenches under this trench. In this trench a lead ball F 2150" (data unit 61) or "Comment during post-analysis by 11.03.18/NN" (data unit 284).

Table 2: Paradata categories identifiable in the analysed dataset. Columns marked with * were added to the data after it was exported from Intrasis

	Knowledge organisation paradata (KOP)		
Column	**Transformation to data units (DU)**	**Transformation to archaeological information (AI)**	**Knowledge-making paradata (KMP)**
A*			
B	Relating to site	Contextualising within site defined by the project	Reflects conceptualisation of "site"
C*	Relating to province	Situating in province	Reflects foregrounded geographical entities
D*	Relating to county	Situating in county	Reflects foregrounded geographical entities
E*	Relating to parish	Situating in parish	Reflects foregrounded geographical entities
F	Relating to origin in Intrasis (spatial database) archive	Documenting in Intrasis archive	
G	Assigning object id (numeric)	Assigning object id (numeric) in a certain sequence	Reflects several different sequences (chronology, but also series of numbers reserved for specific purposes interrupting the chronology)
H	Assigning object name (literal)	Naming (literal)	When assigned, some combination of elements: an alphanumeric code (e.g., "G 16"), context reference (e.g., "house 12"), literal name (e.g., "coin"), condition (e.g., "sharpened"), dating (e.g., "1812"), indication of level of certainty (e.g., "?")
I	Assigning class (ontology)	Classifying general (ontology)	
J	Assigning subclass (approximate ontology)	Classifying subclass (approximate ontology)	Assigned with higher or lower degree of ontological systematicity (e.g., variations like "CU-.leg" and "CU-leg" appear); several ontologies applied (e.g., naming of object OR material); communicates the division of responsibilities/involvement between contributors to knowledge-making (e.g., "Flintstone UV Syd," "Ceramic NN")
K	Explicating spatial type in Intrasis archive	Documenting as a particular spatial type in Intrasis archive	
L	Assigning attributes	Describing and interpreting	Attributes "Interpretation," "Interpretation_temp" (preliminary/temporary interpretation—i.e., trace of self-reported reliability), "Find retrieval technique/method," "Note"
M	Relating to other data units in the dataset	Relating to other archaeological features	
N	Relating to other data units in the dataset (child-parent relations)	Relating to other archaeological features	
O	Assigning descriptions	Describing and interpreting	E.g., "Interpretation," "First interpreted as . . . ," "One section was excavated . . . " (data unit 16), "Grouping discussion: see 4227" (data unit 21)

Challenges and Opportunities in Identifying Paradata

The analysis reveals several types of technical challenges in identifying paradata and epistemological challenges in analysing identified paradata as evidence. Simultaneously, the analysis points to several opportunities to use the diverse aspects of the dataset to derive information about processes. Some of the challenges and opportunities are easily discernible, while others remain relatively invisible barring closer scrutiny of the data and its structure.

A challenge that pertains to both the technical and epistemological realms is that the project dataset contains data at different levels of completeness that is communicated by different means. Many features of the data—for instance, personal references to colleagues, diverse free-form utterances of uncertainty, and colloquial styles of writing—demonstrate that the dataset is a working document. There are also certain discrepancies between the published report and the dataset that likely stem from the dataset not having been updated with the latest information. As a whole, it is apparent that its creators did not foresee that it would eventually be made available for others to reuse. The same applies to many other legacy datasets (e.g., Ullah 2015). In practice, datasets often move from one context to another—with different minor or major implications for who gets access to the data—even when they are produced by major public actors. For example, the analysed dataset was passed from the Swedish National Heritage Board to National Historical Museums in 2015 when the state-run contract archaeology operations were separated from the archaeological heritage administrative organisation (see Löwenborg et al. 2021).

Technical Challenges and Opportunities

The work-in-progress nature of the data means that it is not proofread and not fully normalised. A search for “post holes” reveals a large number of spelling inconsistencies and references in multiple columns (including Columns H, J, L:“Interpretation temp,” and O:“Interpretation”). A related challenge to the varying levels of finality is that traces of processes and transformations are present in different columns of the spreadsheet under similar and quasi-similar headings (e.g., “Interpretation,” “Interpretation_temp”). The exact relation between “Interpretation” and “Interpretation_temp” appears impossible to establish by solely consulting the dataset. Sometimes only “Interpretation” or “Interpretation_temp” is given; sometimes both are given. Based on the information in the dataset, the “Interpretation_temp” seems to be used as a temporary (e.g., preliminary, unspecific) alternative to “Interpretation,” and in the absence of “Interpretation,” an interpretation that needs no elaboration. Similar diversity characterises references to processes and actions in the dataset. Interestingly enough, informal discussions with the developers of the Intrasis system suggest that there can be different views among developers and users of how users appropriate and populate different fields of the database and how straightforward it is to query the database. Moreover, paradata can be identified in both explicit (e.g., “Find collecting method: Trowel,” data unit 39) and less explicit statements as subordinate clauses to other expressions. For instance, “Interpretation: Probably a depression left by a foundation (sill) stone in the north wall of the north house, southern half was investigated” (data unit 18) reports on the extent of the excavation.

The in-progress nature of the data indicates that the interpretations made and the terms used do not have the evidentiary value of final classifications. The technical dimension of having to deal with information with different levels of finality is that it is difficult to combine data and to manage explicit utterances of uncertainty such as question marks and approximations (e.g., “Uncertain interpretation due to slightly unclear boundaries,” data unit 53). The different degrees of uncertainty and finality are also difficult to compare when different notations are used. The differences in certainty and finality suggest that a user of the data needs to know how to deal with the uncertainties and interim interpretations made. The reading of this dataset in isolation underlines the need to know the contextual background on the methods used (Koesten et al. 2020; Kim, Yakel, and Faniel 2019), but emphasises also the need to gain insight into the local ways of documenting the knowledge-making process in this particular fieldwork project (for example, what the “temp” extension might indicate in “Interpretation_temp”), preferably by getting in contact with the dataset creators.

Yet another technical challenge is that paradata may be contained in alphanumeric structures that are not readily readable in the export format at hand. For example, the flattening of the original relational Intrasis database to a two-dimensional table structure in this export version means that the Columns M and N that contain relational attributes and their textual descriptions (e.g., “over,” “under,” “belongs to”) are visually legible. The final data package will contain both a flattened version of the data (similar to the analysed spreadsheet file) for simple manual/visual data exploration, as well as a relational version in a geopackage that allows a user to re-create the original relational structure of the database in full. The intention is to help users with different backgrounds to make the best use of the data, including to explore the grouping of objects, processes, and interpretations (as in Column O).

Table 3: Different configurations of detailedness made up by the number of informative elements and their respective systematicity

Column H, object_ name excerpt	KMP cell content	Value for data reliability assessments
"F4141 Jetton? 1700c" (data unit 23,040)	Several (4) informative elements, low degree of systematicity	Detailed information, format that requires modulation of evidence value before processing: • Object description uncertain ("?") • Object dating general (to century)
"F4142 Tinplate" (data unit 23,039)	Few (2) informative elements, high degree of systematicity	Detailed information, ready-to-process format

Epistemological Challenges and Opportunities

A major epistemological challenge of interpreting the dataset relates to difficulties in making sense of the vocabulary. Partly, the nomenclature (i.e., terms) used to refer to particular methods, activities, and interpretations varies. In addition, a wide array of alphanumeric instances (e.g., codes, references) and terms are used for descriptive purposes in each data unit. For example, "object_id" (Column H) sometimes can contain a NULL (i.e., undefined), be populated with a single data value, or be populated with a combination of several elements. Adding to the challenge is that these elements originate from different systems (e.g., the enumeration of finds, the standard for naming types of ceramics, conventions for dating) and are used with different levels of systematicity. Thus, as exemplified in Table 3, comparable cells in different data units can either be detailed in the sense that they contain several informative elements but with a low degree of systematicity or be detailed in the sense that they contain few informative elements but with a high degree of systematicity, and all varieties in between.

A comparable problem relates to expressions of certainty and their hierarchical relation (i.e., whether "possibly" is more or less certain than "might be"). Reading the investigation report helps to a certain degree in this respect because it gives an overview of the site and the actors involved in the work.

A parallel challenge to the diversity of nomenclature relates to the different levels of completeness of the data, which means that the individual data points are not a priori comparable to each other and that they do not necessarily represent the latest verdict in a given matter. Their evidentiary value is relative to the stage of the investigation process when they were recorded and eventually revised. However, timestamps are not available for data inserts and changes in the analysed file.

The complexity of the stage of the data is further complicated by the presence of multiple "interpretations" and "preliminary interpretations" in different columns. Interpretations are not final, and the varying presence of "interpretations" or "preliminary interpretations" in Column L makes it difficult to establish the evidential status of each of the data points. In some cases, interpretations are formulated as questions to colleagues, such as "NN: Does this ditch have a relation to house 44?" (data unit 230), which could be an indication of a very preliminary assumption but also of a more final interpretation if the documenting archaeologist considers the colleague NN as an authority in the matter. In some cases, Column O contains a post-analysis interpretation that explicitly reinterprets, corrects, and clarifies earlier descriptions and interpretations of the documented feature.

As a whole, the epistemological challenges and opportunities relating to the interpretation of the analysed dataset can be traced back to its technical and epistemological heterogeneity. It contains both direct evidence and traces of processes and actions. Sometimes a method or action is documented explicitly, whereas in many cases the data functions as a trace that indicates that a particular action probably took place. The descriptions of find retrieval methods in Column L can largely be taken as evidence of how a particular feature was investigated, whereas the varying use of "Tolkning" and "Tolkning_temp" attributes provide at most traces of how the interpretation process proceeded during the investigation. As Geiger and Ribes (2011) note, the epistemological utility of traces lies less in their evidential value than in how they are a part of the dataset and the broader assemblage of sources relating to the particular archaeological site and investigation project. The analysed dataset might not be as "highly-standardized" as some other "sociotechnical infrastructure[s] of documentary practices" (Geiger and Ribes 2011, 5). Its heterogeneity limits the possibilities of weaving together hard evidence of the traces; however, at the same time, by being less "purified" (Nadim 2021), it means that the traces themselves are rich and diverse, and have more nuances than would be possible within a highly standardised system.

Using Paradata for Reliability Assessments

Even if the analysis so far shows that there are multiple technical and epistemic challenges to overcome when identifying paradata and analysing identified paradata as evidence, it also points to how the existing paradata, with its imperfections, can be used for understanding the making—or, rather, becoming—of the dataset and for assessing the level of reliability of the data.

Knowledge Organisation Paradata

In the KOP, we distinguish three discrete categories where the identified paradata categories can be useful: assessments of structural, temporal, and terminological variation.

A review of *structural variation* in terms of the internal (i.e., coherence of data between single units in particular columns) and external (i.e., the coherence of data with documentation guidelines and with other datasets) consistency of the KOP can help to determine the reliability of individual data points—like when a feature is described in one unit as a sand layer and in another as a clay layer—within the dataset as well as with other datasets. A similar analysis of KMP provides comparable evidence of the degree and variation of uncertainties during the process of making the database. The presence of multiple levels of preliminary and non-preliminary interpretations can similarly signal uncertainties that arose during the investigation process and how they were resolved. The interpretation of these cues does, however, require explicit consideration. The absence of preliminary interpretations or inconsistent reinterpretations can indicate both data reliability problems but also features that are, respectively, easy or difficult to identify. Moreover, the patterns of what attributes have been included and what information they contain seem to follow distinct patterns, which are very likely traceable back to different individuals with individual interests and documentation ideals (Faniel et al. 2021; Börjesson 2016). Identifying and applying such structural patterns to analysing and comparing the documentation approaches and their outcomes could make it possible to assess the reliability of interpretations between individual archaeologists. A parallel possibility would be to investigate the consistency between Columns J (subclass) and O (description).

Even if the work-in-progress nature of the data causes both technical and epistemic challenges, the *temporal variation* of tentative, preliminary, and final interpretations is indicative of the data creation process. The characteristics of the variations as coincidental or rule-bound could be further explored by cross-analysing differences in KOP with information in the original Intrasis database events table that tracks all imports and changes. Such cross-analysis would be crucial to test the evidentiary value of seemingly patterned variations (e.g., does "Interpretation_temp" precede "Interpretation" chronologically?). However, as a response to the current difficulty of fathoming data creation and editing process information by ocular spreadsheet reading, wiki-based approaches foregrounding processuality have been proposed for archaeological documentation (e.g., Huvila 2012).

Even if it would be impossible to provide a comprehensive reconstruction of the entire data creation process based on the dataset, it is possible to compare it with (likely more final) information in the published report. Such a comparison could provide indications of how and when preliminary interpretations and approximations have been confirmed or otherwise considered authoritative. Patterns of similar expressions and utterances could be indications of higher or lower authoritativeness of specific individuals or types of observations. In some cases, a preliminary interpretation ("Interpretation_temp") of a feature as a "post hole" that remains "post hole" (in "Interpretation") could indicate an interpretation that is reliable from the start, whereas a preliminary interpretation ("Interpretation_temp") of a feature as a "pit" that becomes a "carcass pit" (e.g., data unit 2718) could indicate an increasing reliability and specificity of the interpretation of the particular pit.

Finally, *terminological variation* in the paradata can be indicative of both the interpretative processes of individual archaeologists and the documentation activity itself. Archaeologists' choices of terminology in describing both things and activities can reveal the epistemic underpinnings of their interpretations and how they progress. A preliminary interpretation that is replaced by a more specific interpretation (hole > specific type of hole) suggests a different type of process than an interpretation that remains open (e.g., "Uncertain interpretation due to vague boundaries," data unit 53) or is changed from a broader category of features to another, more specific category. The variation in terminology used by different archaeologists can similarly be indicative of the competencies, analytical thinking, and work processes of individual archaeologists and subsequently be a measure of how reliable the data is. If interpretations coming from one individual differ considerably from those of others, it might be an indication of their either higher or lower reliability. Also, changes in terminology between preliminary and non-preliminary interpretations can be indicative of uncertainty or increasing certainty.

Knowledge-Making Paradata

In contrast to the KOP, which provides cues about the reliability of data through largely unwanted and unplanned variation and disconnects, KMP provides data creators' version of explicitly identified and disclosed assessments, doubts, and justifications of the data's trustworthiness. As a whole, the dataset contains relatively little KMP, and it seems likely that the explicit descriptions follow a logic of selective black-boxing of obvious and non-essential information and transparency concerning information that is considered important to communicate either for personal or social reasons (Huvila, Sköld, and Börjesson 2021). Even if the dataset contains a fair amount of KOP that is readable from the data itself, a thorough understanding of the KMP appears to require a cross-reading of the specific KMP in the dataset and a generic description of work procedures for the project, if available in, for instance, the fieldwork report. Here the limitations of the analysed tabular excerpt of the full Intrasis database also become explicit. The omission of user log and event registries of the full Intrasis database in producing the analysed spreadsheet means also that a lot of potentially valuable KMP is excluded from the data for the sake of making the dataset simpler and to protect personal data. The inclusion of log and event registries would open up other KMP identification opportunities, but the data processing would then require compliance with the General Data Protection Regulation (GDPR).

Even if the lack of standardisation in the dataset was identified as a problem in both technical and epistemological senses, the rich nuances in the heterogenous lacework of traces can be advantageous from the KMP perspective. We suggest that a higher number of traces and expressions of uncertainty points to a more thorough investigation process and pondering of a particular interpretation than the presence of a single settled interpretation. This is obviously not a direct indication a priori of a better evidential value of such interpretations but, read in context, a potentially useful reliability indicator. From a technical perspective, the obvious drawback is the difficulty of automating the analysis of interpretation traces and expressions of uncertainty.

Discussion: Paradata Supporting Data Literacy

In contrast to the widespread tendency to consider data as objective, unproblematic, and apolitical evidence, earlier research has repeatedly demonstrated that it is both messy and—what Gitelman (2013) aptly described by declaring that raw data is an oxymoron—contextual. This was apparent also in the analysed dataset, which is perhaps best described as being "partially digested." However, even if its messiness appeared as a hurdle, especially concerning the significance of traces and the overlap and variety of descriptions as (a source of) paradata, the analysis highlights that data can also be too clean. A messy dataset contains much more paradata than a normalised one and provides many more cues to assess its reliability. A challenge is how to make use of those cues and overcome the problem that datasets are too amoeba-like and messy to be easily approachable in comparison to other information sources—for instance, reports, as discussed by Huvila (2016). Moreover, these challenges multiply with ambitions to combine data that is often heterogeneous and partially incomplete in archaeology and other legacy fieldwork data–dependent disciplines. In worst cases, the data can be scattered around the world and parts of it might not be available or accessible at all (e.g., Sobotkova 2018; Stilborg 2021).

With a close reading of the sources of paradata in an excavation dataset and an estimation of the evidentiary value of the identified paradata sources, this article refines the understanding of necessary steps for examining a dataset and assessing its suitability for aggregation. Drilling down into the traces of knowledge organisation and knowledge-making in the dataset reveals the methodological efforts needed to make use of legacy data, as pointed to in previous calls for resources for legacy data reuse (Sobotkova 2018; Kansa and Kansa 2021; cf. Yan et al. 2020). An improved understanding of the efforts and the required knowledge and skills to extract paradata from datasets is also vital for a better understanding of how data should be described at creation time and to what extent.

A closely related practical question is how much time it makes sense to invest in cleaning datasets. Tidy datasets, including strictly standardised linked open data, have irrefutable advantages in enabling interoperability and data (re)use in research endeavours that build on shared ontological and epistemological premises. However, even if the current paradigmatic opinion supports extensive cleaning of data and there is a broad consensus that messy data is one of the key obstacles of data reuse (Richards et al. 2021), the current analysis highlights the multiple adverse effects of data cleaning. Excessive cleaning does not need to be intentional but can be a result of forcing systematicity onto observations and interpretations that are not conclusive enough to be systematised. Many technical and epistemological problems can be addressed at (re)use time, and many of the inconsistencies and challenging aspects of the data provide opportunities for a more thorough understanding of the data and how it came into being. Therefore, we are inclined to suggest that resources could be more well spent in describing data, terminology, and data creation procedures at the time of creation rather than cleaning the data beyond the needs of its primary use.

Drawing on the close reading of the dataset analysed in this article, we also want to direct attention to the colloquialisms (i.e., non-formalised expressions such as questions to colleagues and stream-of-consciousness-like reasoning) in text strings. Based on the high frequency of these types of utterances, we assume that they are playing a significant role for the data creators. If the goal is a dataset free from uncertainties, these colloquialisms could be framed as problematic. However, in an epistemological sense and from a knowledge-making perspective, these data points can be seen as heuristic zones in the dataset, where systematic observations and interpretative reasoning meet and are negotiated. With further research attention to data colloquialisms in field science—for example, analyses of the traces of interpretative processes in fieldwork data—there would be a potential to better understand the role of unstructured data for knowledge-making processes, the dynamics between unstructured and structured data, and how process paradata could be harvested from unstructured data.

Based on the analysis, a crucial characteristic of effective paradata is its capability to uncover transformations and, in a broader sense, change. KOP does this through the crookedness of the data itself: mistakes, errors, incompleteness, and omissions. KMP does the same through explicitly created descriptions and utterances of different kinds. The major difference between the two is that the KOP is system-bound and essentially an involuntary by-product of documenting in a particular system and according to a specific scheme, whereas KMP is consciously created to provide an explanation of a transformation. Both are political but, on the one hand, the politics of the metalevel descriptions are primarily functioning on the infrastructural level through the socio-technic-informational meshwork of the archaeological field practice and its infrastructures. On the other hand, the politics of KMP are more explicit and explicitly social, and determined by the data creators themselves. The explicitness of the politics of KMP and the political nature of the data also mean that it entails ethical issues that are different from those of KOP, including the processing of personal data.

A key prerequisite for effectively exploiting KOP is to develop the means to compile profiles of individual data creators and stages in the documentation process by identifying patterns in terminological and structural variation in the data. This would allow users to compare the documentation process over time (how interpretations and documentation practice have changed, how and if uncertainties have been solved), to compare how data created by specific individuals differs from each other, and, subsequently, to understand the evolution of data and its dependability. Another key task is to identify elements and alphanumerical structures in the dataset that have potential informational value for specific potential users. However, when interpreting the terminology used, it is necessary to be mindful of the differences between individuals who have contributed data to the dataset because even personal ideals may affect externally standardized documentation (Faniel et al. 2021; Börjesson 2016). Another issue relates to the tracing of named or anonymous individuals. Traces of individual archaeologists' interactions with the database exist in a log file in the original Intrasis database but have been removed from the export file to avoid personal data processing. Even if definite indications of the individual agency have arguable advantages, the analysis of the variation of the use of descriptive terms and patterns of filling in attribute information could still provide enough information on personal preferences and ideals and biases linked to professional specialisation profiles or stage of professional training.

An advisable next step in developing the paradata extraction approach tested in this study is to operationalize the iterative close reading approach by computer-aided methods, for example by natural language processing (NLP). The pre-analysis data structuring and cleaning in such processes would depend on the goals of the analysis. Within the analysed dataset, it is possible that the individual data units can be too heterogeneous and as such unreliable for a local or regional analysis conducted by a secondary user, even if the dataset would be cross-read with all available evidence. In contrast, depending on demands regarding the type and level of reliability, the data can be sufficient for analyses on a broader cross-regional or global scale without refinement. Such methods development based on analyses of cell content—for example, expressions indicating survey methods—would target KMP. However, the analysis presented in this paper proves the value of parallel manual analysis of KOP to identify the multiple places (e.g., multiple columns) within a data structure where similar or overlapping data can be found, like the traces of interpretative processes found in Columns L and O, and consequently what columns a NLP-supported analysis should target.

Our study has obvious limitations. Observations made on the basis of a single dataset retrieved from one database does not allow us to infer how Intrasis users in general use the system, how field scientists document their field observations, or what additional paradata categories could be present in other databases. In the analysis, we have looked for paradata only in the dataset itself. Compared to the analysed dataset and archaeological data in general, investigation reports contain more information on, for instance, methods and, to varying degrees, work processes (Huvila, Sköld, and Börjesson 2021). There are also other potential sources of paradata, including administrative documents and retrieved finds. These were consciously left out of the present study since the focus was on investigating what can and cannot be seen specifically in the data itself.

Although the above-highlighted technical and epistemological challenges in identifying paradata probably represent only a part of all conceivable hurdles, we are inclined to believe that the present analysis can still be helpful in clarifying what kinds of challenges should be expected when dealing with a relatively well-formed and systematically developed fieldwork dataset. The analysis also contributes to a better understanding of only partly digested datasets—actually, most of them—that have not been created with reuse in mind. This is, again, something that often characterises legacy fieldwork data that has not been collected as part of a concerted effort to accumulate uniform data.

Conclusions

Our findings show that a fieldwork dataset contains information on how it came into being (i.e., paradata) in the forms of both traces and direct evidence. The analysed dataset included KOP relating to how a fieldsite was transformed into structured information objects and archaeological information, and KMP as explicit and implicit descriptions of interpretative processes, how the data was created, and the fieldwork conducted. Different levels of finality and completeness of the data in the dataset; internal inconsistencies in nomenclature; vocabulary; and references between data units posed challenges from both technical and epistemological perspectives. This "roughness," however, also worked as an opportunity by revealing traces of the data-in-making. In addition, the structural, temporal, and terminological variations in the data provide cues to assess the reliability of the data: who did what, how specific the data is, how certain the interpretations appear to be and how they may have changed. As a whole, even if the lack of systematicity caused problems, its apparent advantages raised concerns about whether data can be made too clean. Instead of essentially purifying data (cf. Atici et al. 2013), it can be more useful to describe what the data is about, list the vocabulary used in the data, and explain how data units and concepts relate to each other.

Moreover, the findings point to the need to better understand different levels of completeness in archaeological data. First, there is a colossal difference between *systematising* and *making systematic*. We see in the dataset a "sufficient systematisation" that enabled archaeologists to draw necessary conclusions during the fieldwork project and report writing. The resulting level of systematicity that was apparently enough for a human reader did not, however, equal the machine-readable "systematic documentation" required in the Urdar project. Similarly, considering Hodder's (1989) remarks on archaeological data and writing, it is conceivable that such systematic documentation is not necessarily best suited for a human reader. We also suggest that there is a difference between systematising knowledge organization (creating the database structure) and systematising knowledge-making (populating data structure with content). In the analysed dataset, there are examples of everything from users employing fairly structured terminology to users populating the data structure in the least systematic way that the structure allows.

A key implication of the varied degrees of structure is that there is a dire need to broaden the understanding of systematicity in archaeology and comparable field sciences. Systematicity is a gradient and, as the analysed dataset shows, not always the ultimate goal of those who create datasets. Systematicity has a function, but only in a certain phase of the knowledge-making process. Messiness or systematicity does not make data (un)reusable a priori but rather reusable in two very particular senses. The scalarity of systematicity calls into question not only whether it is always reasonable to have the goal of creating FAIR data when documenting fieldwork (cf. Huvila 2012) but also whether interoperability and reusability should be perpetual goals. To quote Hodder's famous expression about where archaeological interpretations are made, making data truly interoperable and reusable would require that it is made such at "trowel's edge" (Hodder 1997, 693)—i.e., when first entered in a database or scribbled on a piece of paper. Considering the apparent prevalence and social usefulness of certain messiness, it is apparent that making data FAIR from the field has both advantages and disadvantages that have to be weighed against each other.

Another key implication of the scalarity of systematicity is that data reuse and aggregation needs to be seen as a research activity that requires methodological framing (see Kansa and Kansa 2021; Sobotkova 2018). Reading the type and grade of systematicity and extracting data provenance and processing information are undertakings that have a place within that methodological framing; they require explicit effort, skills, and a meticulously developed approach to succeed and are a part of the data literacy (see Kansa and Kansa 2021a) of a competent data reuser.

The findings of this study also have implications for digital preservation and repository practice. Our analysis shows the importance of supporting data descriptions that spell out what the data is about (aboutness) and lay out the context of the used vocabulary and terminology, including eventual data colloquialisms. It also points to the importance of the relations of data units and concepts when seeking to understand the underpinning principles of knowledge organisation and knowledge-making. When aggregating data, it is important to compare the knowledge organisation (as in KOP) and knowledge-making (as in KMP) principles used in the different datasets to assess whether the datasets can be aggregated for specific purposes or not. Finally, if strict structural and terminological standards are used to enable machine-readability of an aggregated dataset, it is crucial to keep the expressions of uncertainties, stages of interpretation, and questions inscribed in the original data to allow reusers to understand, exploit, and be aware of how unFAIR (or MEAN—see Huvila 2017) they potentially are.

Acknowledgements

We would like to thank the archaeologists who originally compiled the excavation database and the Urdar project for sharing the data export. We are also grateful to Dr. Angus Graham for assisting in translating the Swedish archaeological terminology to English. The two anonymous reviewers provided valuable feedback.

This article is part of the project CAPTURE that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 818210) and project Urdar funded by Riksbankens Jubileumsfond (In19-0135:1).

Competing Interests

The authors declare that they have no competing interests. The editors would like to note that Isto Huvila is an editorial board member of *KULA* but that this article went through the same submission process, including anonymous peer review, as all other research articles.

References

Allison, Penelope. 2008. "Dealing with Legacy Data - an Introduction." *Internet Archaeology* 24. https://doi.org/10.11141/ia.24.8.

Atici, Levent, Sarah Kansa, Justin Lev-Tov, and Eric C. Kansa. 2013. "Other People's Data: A Demonstration of the Imperative of Publishing Primary Data." *Journal of Archaeological Method and Theory* 20 (4): 663–81. https://doi.org/10.1007/s10816-012-9132-9.

Bhargava, Rahul, Erica Deahl, Emmanuel Letouzé, Amanda Noonan, David Sangokoya, and Natalie Shoup. 2015. "Beyond Data Literacy: Reinventing Community Engagement and Empowerment in the Age of Data." Data-Pop Alliance White Paper Series. New York: Internews Center for Innovation and Learning and the MIT Media Lab Center for Civic Media. https://datapopalliance.org/item/beyond-data-literacy-reinventing-community-engagement-and-empowerment-in-the-age-of-data/. Archived at: https://perma.cc/7AGR-245E.

Boozer, Anna Lucille. 2014. "The Tyranny of Typologies: Evidential Reasoning in Romano-Egyptian Domestic Archaeology." In *Material Evidence: Learning from Archaeological Practice*, edited by Robert Chapman and Alison Wylie, 92–109. Abingdon: Routledge.

Börjesson, Lisa. 2016. "Beyond Information Policy: Conflicting Documentation Ideals in Extra-Academic Knowledge Making Practices." *Journal of Documentation* 72 (4): 674–95. https://doi.org/10.1108/JDOC-10-2015-0134.

Börjesson, Lisa, Olle Sköld, and Isto Huvila. 2021. "Paradata in Documentation Standards and Recommendations for Digital Archaeological Visualisations." *Digital Culture & Society* 6 (2): 1. https://doi.org/10.14361/dcs-2020-0210.

Brown, Hannah, Helen Goodchild, and Søren M. Sindbæk. 2014. "Making Place for a Viking Fortress. An Archaeological and Geophysical Reassessment of Aggersborg, Denmark." *Internet Archaeology* 36. https://doi.org/10.11141/ia.36.2.

D'Andrea, Andrea, and Kate Fernie. 2013. "CARARE 2.0: A Metadata Schema for 3D Cultural Objects." In *2013 Digital Heritage International Congress (DigitalHeritage)*, 137–43. https://doi.org/10.1109/DigitalHeritage.2013.6744745.

Ellis, Steven J. R. 2008. "The Use and Misuse of 'Legacy Data' in Identifying a Typology of Retail Outlets at Pompeii." *Internet Archaeology* 24. https://doi.org/10.11141/ia.24.4.

European Commission. n.d. "Open Science." Accessed May 21, 2021. https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science_en. Archived at: https://perma.cc/N22U-X2JF.

Faniel, Ixchel, Anne Austin, Sarah Whitcher Kansa, Eric Kansa, Jennifer Jacobs, and Phoebe France. 2021. "Identifying Opportunities for Collective Curation During Archaeological Excavations." *International Journal of Digital Curation*. https://doi.org/10.2218/ijdc.v16i1.742.

Faniel, Ixchel, Eric Kansa, Sarah Whitcher Kansa, Julianna Barrera-Gomez, and Elizabeth Yakel. 2013. "The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse." In *JCDL '13: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 295–304. https://doi.org/10.1145/2467696.2467712.

Faniel, Ixchel M., Rebecca D. Frank, and Elizabeth Yakel. 2019. "Context from the Data Reuser's Point of View." *Journal of Documentation* 75 (6): 1274–97. https://doi.org/10.1108/JD-08-2018-0133.

Federal Geographic Data Committee. n.d. "Content Standard for Digital Geospatial Metadata (CSDGM)." Accessed May 28, 2021. https://www.fgdc.gov/metadata/csdgm-standard. Archived at: https://perma.cc/3JMX-Z3AJ.

Furner, Jonathan. 2021. *Information Studies and Other Provocations*. Sacramento: Litwin Books.

Geiger, R. Stuart., and David Ribes. 2011. "Trace Ethnography: Following Coordination through Documentary Practices." In *2011 44th Hawaii International Conference on System Sciences (HICSS)*, 1–10. https://doi.org/10.1109/HICSS.2011.455.

Gitelman, Lisa, ed. 2013. *"Raw Data" Is an Oxymoron.* Cambridge, MA: MIT Press.

Gustafsson, Anders, and Björn Magnusson Staaf. 2001. "Rapport Om Rapporter — En Diskussion Kring Kvalitetsbedömningar Av Arkeologiska Rapporter." Report 2001:3. Stockholm: RAÄ.

Heitman, Carrie, Worthy Martin, and Stephen Plog. 2017. "Innovation through Large-Scale Integration of Legacy Records: Assessing the 'Value Added' in Cultural Heritage Resources." *Journal on Computing and Cultural Heritage* 10 (3): 1–10. https://doi.org/10.1145/3012288.

Hodder, Ian. 1989. "Writing Archaeology: Site Reports in Context." *Antiquity* 63 (239): 268–74.

Hodder, Ian. 1997. "Always Momentary, Fluid and Flexible': Towards a Reflexive Excavation Methodology. *Antiquity* 71 (273): 691–700.

Huvila, Isto. 2006. *The Ecology of Information Work: A Case Study of Bridging Archaeological Work and Virtual Reality Based Knowledge Organisation.* Åbo: Åbo akademis förlag.

Huvila, Isto. 2012. "Being Formal and Flexible: Semantic Wiki as an Archaeological e-Science Infrastructure." In *Revive the Past: Proceedings of the 39th Conference of Computer Applications and Quantitative Methods in Archaeology*, edited by Mingquan Zhou, Iza Romanowska, Zhongke Wu, Pengfei Xu, and Philip Verhagen, 186–97. Amsterdam: Amsterdam University Press. https://doi.org/10.1017/9789048516865.

Huvila, Isto. 2016. "Awkwardness of Becoming a Boundary Object: Mangle and Materialities of Reports, Documentation Data, and the Archaeological Work." *The Information Society* 32 (4): 280–97. https://doi.org/10.1080/01972243.2016.1177763.

Huvila, Isto. 2017. "Being FAIR When Archaeological Information Is MEAN: Miscellaneous, Exceptional, Arbitrary, Nonconformist." Presentation at the Centre for Digital Heritage Conference 2017, Leiden, June 14–16, 2017. http://www.istohuvila.se/node/526.

Huvila, Isto. 2020a. "Information-Making-Related Information Needs and the Credibility of Information." *Information Research* 25 (4): paper isic2002. https://doi.org/10.47989/irisic2002.

Huvila, Isto. 2020b. "Use-Oriented Information and Knowledge Management: Information Production and Use Practices as an Element of the Value and Impact of Information." *Journal of Information & Knowledge Management* 18 (4): 1950046. https://doi.org/10.1142/s0219649219500461.

Huvila, Isto, Olle Sköld, and Lisa Börjesson. 2021. "Documenting Information Making in Archaeological Field Reports." *Journal of Documentation* 77 (5): 1107–27. https://doi.org/10.1108/JD-11-2020-0188.

ISO. 2014. "ISO 19115-1:2014 Geographic information – Metadata – Part 1: Fundamentals." https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/37/53798.html.

Kansa, Eric, and Sarah Whitcher Kansa. 2021. "Digital Data and Data Literacy in Archaeology Now and in the New Decade." *Advances in Archaeological Practice* 9 (1): 81–85. https://doi.org/10.1017/aap.2020.55.

Kersel, Morag M. 2015. "STORAGE WARS: Solving the Archaeological Curation Crisis?" *Journal of Eastern Mediterranean Archaeology and Heritage Studies* 3 (1): 42–54. https://doi.org/10.5325/jeasmedarcherstu.3.1.0042.

Kim, Jihyun, Elizabeth Yakel, and Ixchel M. Faniel. 2019. "Exposing Standardization and Consistency Issues in Repository Metadata Requirements for Data Deposition. *College & Research Libraries.* https://doi.org/10.5860/crl.80.6.843.

Koesten, Laura, Pavlos Vougiouklis, Elena Simperl, and Paul Groth. 2020. "Dataset Reuse: Toward Translating Principles to Practice." *Patterns* 1 (8). https://doi.org/10.1016/j.patter.2020.100136.

Larsson, Åsa M., and Daniel Löwenborg. 2020. "The Digital Future of the Past - Research Potential with Increasingly FAIR Archaeological Data." In *Re-Imagining Periphery: Archaeology and Text in Northern Europe from Iron Age to Viking and Early Modern Periods*, edited by Charlotta Hillerdal and Kristin Ilves, 61–70. Oxford: Oxbow.

Löwenborg, Daniel, Maria Jonsson, Åsa Larsson, and Johan Nordinge. 2021. "A Turn Towards the Digital. An Overview of Swedish Heritage Information Management Today." *Internet Archaeology* 58. https://doi.org/10.11141/ia.58.19.

Montoya, Robert D., and Katherine Morrison. 2019. "Document and Data Continuity at the Glenn A. Black Laboratory of Archaeology." *Journal of Documentation* 75 (5): 1035–55. http://doi.org/10.1108/JD-12-2018-0216.

Nadim, Tahani. 2021. "The Datafication of Nature: Data Formations and New Scales in Natural History." *Journal of the Royal Anthropological Institute* 27 (S1): 62–75. https://doi.org/10.1111/1467-9655.13480.

Olson, Carina, and Yvonne Walther. 2007. "Neolithic Cod and Herring Fisheries in the Baltic Sea, in the Light of Fine-Mesh Sieving : A Comparative Study of Subfossil Fishbone Form the Late Stone Age Sites at Ajvide, Gotland, Sweden and Åland, Finland." *Environmental Archaeology* 12 (2): 175–85.

Richards, Julian D., Ulf Jakobsson, David Novák, Benjamin Štular, and Holly Wright. 2021. "Digital Archiving in Archaeology: The State of the Art. Introduction." *Internet Archaeology* 58. https://doi.org/10.11141/ia.58.23.

Roskams, Steve. 2001. *Excavation*. Cambridge: Cambridge University Press.

Roy, Sohon, Felienne Hermans, Efthimia Aivaloglou, Jos Winter, and Arie van Deursen. 2016. "Evaluating Automatic Spreadsheet Metadata Extraction on a Large Set of Responses from MOOC Participants." In *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, 135–45. https://doi.org/10.1109/SANER.2016.98.

Sabo, Katalin Schmidt, Magnus Andersson, Mats Anglert, Caroline Arcini, Adam Bolander, Torbjörn Brorsson, Annica Cardell, Bo Knarrström, Per Lagerås, Linda Rosendahl, Fredrik Strandmark, Marie Svedin, and Håkan Svensson. 2013. *Arkeologisk Undersökning 2010 Örja 1:9 Skåne, Landskrona Kommun, Örja Socken, Örja 1:9, Fornlämningarna Örja 9, 35, 40, 41 Och 42*. Vol. 2013: 68. UV Rapport. Lund: RAÄ.

Secci, Massimiliano, Carlo Beltrame, Stefania Manfio, and Francesco Guerra. 2019. "Virtual Reality in Maritime Archaeology Legacy Data for a Virtual Diving on the Shipwreck of the *Mercurio* (1812)." In "Multidisciplinary Study of the Sarno Baths in Pompeii," special issue, edited by Lara Maritan, Caterina Previato, and Filippo Lorenzoni, *Journal of Cultural Heritage* 40 (November–December), 169–76. https://doi.org/10.1016/j.culher.2019.05.002.

Silverman, David. 2013. *A Very Short, Fairly Interesting and Reasonably Cheap Book about Qualitative Research*. London: SAGE Publications. https://doi.org/10.4135/9781526402264.

Sobotkova, Adela. 2018. "Sociotechnical Obstacles to Archaeological Data Reuse." *Advances in Archaeological Practice* 6 (2): 117–24. https://doi.org/10.1017/aap.2017.37.

Stilborg, Ole. 2021. "A Study of the Representativity of the Swedish Ceramics Analyses Published in The Strategic Environmental Archaeology Database (SEAD)." *Fornvännen* 116 (2): 89–100.

The Swedish Research Council. 2017. *God forskningssed*. Stockholm: Swedish Research Council. https://www.vr.se/analys/rapporter/vara-rapporter/2017-08-29-god-forskningssed.html.

Tkaczyk, Dominika, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Łukasz Bolikowski. 2015. "CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature." *International Journal on Document Analysis and Recognition (IJDAR)* 18 (4): 317–35. https://doi.org/10.1007/s10032-015-0249-8.

Ullah, Isaac I. T. 2015. "Integrating Older Survey Data into Modern Research Paradigms: Identifying and Correcting Spatial Error in 'Legacy' Datasets." *Advances in Archaeological Practice* 3 (4): 331–50. https://doi.org/10.7183/2326-3768.3.4.331.

Voss, Barbara L. 2012. "Curation as Research. A Case Study in Orphaned and Underreported Archaeological Collections." *Archaeological Dialogues* 19 (2): 145–69. https://doi.org/10.1017/S1380203812000219.

Warner, Julian. 2010. *Human Information Retrieval*. Cambridge, MA: MIT Press.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18.

Wylie, Alison. 2017. "How Archaeological Evidence Bites Back: Strategies for Putting Old Data to Work in New Ways." *Science, Technology, & Human Values* 42 (2): 203–25. https://doi.org/10.1177/0162243916671200.

Yan, An, Caihong Huang, Jian-Sin Lee, and Carole L. Palmer. 2020. "Cross-Disciplinary Data Practices in Earth System Science: Aligning Services with Reuse and Reproducibility Priorities." *Proceedings of the Association for Information Science and Technology* 57 (1). https://doi.org/10.1002/pra2.218.

How to cite this article: Börjesson, Lisa, Olle Sköld, Zanna Friberg, Daniel Löwenborg, Gísli Pálsson, and Isto Huvila. 2022. Re-purposing Excavation Database Content as Paradata: An Explorative Analysis of Paradata Identification Challenges and Opportunities. *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 6(3). https://doi.org/10.18357/kula.221

Submitted: 23 June 2021 **Accepted:** 25 October 2021 **Published:** 27 July 2022

KULA: Knowledge Creation, Dissemination, and Preservation Studies is a peer-reviewed open access journal published by University of Victoria Libraries.

OPEN ACCESS