Information needs on research data creation

Lisa Börjesson, Isto Huvila and Olle Sköld

Abstract

Introduction. *Researchers' data related information needs are growing. This paper reports the findings of a study with archaeologists and cultural heritage professionals focussing on data reuse related meta-information needs.*
Methods. *Interviews with (N=)10 archaeologists and cultural heritage professionals.*
Analysis. *Qualitative coding and content analysis.*
Results. *Four types of paradata needs (data on processes, e.g. data creation) are identified, including 1) scope, 2) provenance, 3) methods and 4) knowledge organisation and representation paradata. Knowledge organisation and representation paradata has been least explored both in research and practises so far. The findings point to a need to develop the understanding of the needs and means of documentation of knowledge organisation and representation.*
Conclusions. *The findings contribute to the data literacy of researchers producing and using data descriptions, and to the study of how paradata can be created and used. Further, the findings indicate that distance-to-data is a significant parameter in determining whether information needs are continuous or discrete. Further, the most likely type of reuse should guide the level and type of paradata. Finally, the findings underline that in spite of the comprehensiveness of available meta-information, it will be incomplete. Complementary means — including collaboration with data creators and meta-information extraction approaches — are needed to increase information reusability.*

Keywords*: research data, data management, data reuse, information needs, paradata*

Introduction

Researchers' information needs have been one of the long-standing interests in the information behaviour and practises research with seminal studies dating back to the early second half of the 20th century (Gannon-Leary, et al., 2007). However, in contrast to the needs relating to scholarly and scientific literature, data needs and research data-related information needs have started – with some exceptions (for a review of early work, see Friedrich, 2020) – to attract significant research attention only fairly recently (e.g. Friedrich, 2020; Gregory, et al., 2019). In this area, researchers' needs relating to secondary data they are reusing in their own research is a central subset of their data-related needs. So far, much of the emerging research on data-related information needs has focused on data discovery and descriptions (Friedrich, 2020; Gregory, et al., 2019), whereas especially specifics of data creation and manipulation processes have remained less researched. Still, similarly to how catering for the general information needs is crucial for research, addressing data and process-related information needs by providing access to appropriate forms of metadata (data on data) and paradata (data on processes, see Huvila, et al., 2021) is of equal importance.

This paper explores researchers' information needs relating to the creation and manipulation of research data. The objective is to provide new knowledge about researchers' information-related information needs or *meta-information-needs* in situations of data reuse. We report findings from an interview study with (N=10) archaeologists and cultural heritage professionals creating and reusing data. The study addresses the gap in the earlier information needs research relating to data creation and processing. Simultaneously it pushes forward the state-of-the-art in information behaviour research by shedding light on an earlier understudied aspect of information needs relating to the creation and manipulation of information at hand. The results are significant for researchers and data curators engaged in documenting data creation and processing and for system designers providing infrastructures for data descriptions.

Literature review

Researchers' information needs

While much of the early research on researchers' information needs focused on scientists and engineers (Case and Given, 2016), the more recent studies, especially from the 1980s onwards, have steadily increased the understanding of the diversity of information needs across disciplines. Since the 1990s, studies of the specific needs of interdisciplinary scholarship (Gannon-Leary, et al., 2007) and in new cross-disciplinary fields such as digital humanities have nuanced the picture even more (Toms and O'Brien, 2008; Warwick, 2012). Researchers need a lot of different types of information for various aspects of their work, from primary and secondary research material to literature and information about methods and tools (Toms and O'Brien, 2008).

As information needs in general (Borlund and Pharo, 2019), also researchers' information needs are contextual to the situation and task at hand (Ingwersen and Järvelin, 2005, p. 273 Fig. 6.7) and differ from one field to another (Bowker, 2000). Studies have shown that a key factor that influences information practices—and subsequently—needs is how researchers know what they know in their scholarly field. Knorr-Cetina calls such "amalgams of arrangements and methods" that make up how things are known in a particular *epistemic culture* (Knorr-Cetina, 2003, p. 1). Information studies research has shown that such field-specific differences influence how and what kind of information researchers seek (Bates, 1996a) and how they use information sources (Fry and Talja, 2005; Roos, 2016). For example, humanities researchers tend to search for information more often using geographical and chronological terms and proper names than scientists (Bates, 1996b). Cross-disciplinary fields and epistemic cultures are characterised by information practices and needs related to crossing disciplinary boundaries, exploration and translation (Fry, 2006). Another characteristic of information needs and how they are satisfied in scholarly contexts is the widespread reliance on informal information exchange. This led to the now famous styling of everyday wildcat scholarly networks as *invisible colleges* (Crane, 1972).

Data-related information needs

The recent surge of data-intensive research and datafication of scholarly episteme across a large number of disciplines has made apparent that effective reuse requires that the reuser knows enough about the data to be able to determine, for example, its fitness for the planned use and how to use it (Bishop et al., 2019), how it has been used before, and where to find it (Chapman et al., 2019; also Koesten et al., 2019). Such information is not always available and as Chapman et al. (2019) criticise, the development of data search systems is often driven by available metadata rather than users' information needs. According to David (1991), users of survey data need information about the completeness and validity (evaluation) of the data, ambiguities, errors, portability of data, the design and execution of data collection, and whether earlier analysis results can be verified using the data. Evidence-based studies of data users and their information needs in different disciplines have generally referred to similar issues: reusers need information about where to find data (Friedrich, 2020), its integrity and quality (Faniel, et al., 2016), and relevance for planned use (Bishop et al., 2019), original research questions, and instruments used in data-making (Gregory et al., 2019). A parallel line of research has inquired into how and if particular types of metadata and descriptions (incl. keywords, abstracts, methods information and data descriptions) help to make data reusable. In Murillo's (2016) study methods information, attribute table, and data description were the most useful pieces of information. Further, it has been found that data needs to be searchable across different metadata schemes (Tenopir, et al., 2011) but also hospitable to field-specific needs of describing it (Thomer, et al., 2017). Regarding information on data-making, studies of data reusers (for reviews Chapman et al., 2019; Gregory et al., 2019) have pointed to the need for contextual information in general, and in particular on scope and framing of research (Faniel and Jacobsen, 2010), provenance (Faniel, et al., 2019), data research and data-making procedures (Miksa, et al., 2014), and, for example, methods used in data creation (Chao, 2015; Koesten, et al., 2019).

Friedrich (2020) found that experienced survey researchers tend to work with more complex data than novices and suggests that with more complex data and data uses, more detailed and reliable information about data is needed. In contrast, the combined findings of Faniel, et al. (2012) and Yoon (2016) suggest the opposite: that, in general, novices might need more and more specific information while experts can cope with less. Similarly, the complexity (Faniel and Jacobsen, 2010) and type of the task (data used as information vs as an ingredient) and decision where information is needed influence the amount and level of the necessary information on data (Chapman, et al., 2019). Data-related information needs also differ between research fields. Humanities researchers have stressed the importance of sufficient semantic information on, for instance, cultural meanings and historical change in the data (Geser and Selhofer, 2014) and comprehensive documentation of data to complement full-text searching that often fails with typically messy humanities data (Golub, et al., 2020). Similarly to how knowing about geographical places, chronology and proper names is a central general information need among humanities researchers (Bates, 1996b), it is equally prominent also in their data-related needs (Kumpulainen, et al., 2019).

A major problem with addressing data-related information needs is that they cannot always be met using data descriptions (Borgman and Bourne, 2021). Sometimes metadata is not detailed enough or does not '*inspire the confidence needed for reuse*' (Gregory et al., 2019, p. 426); it is missing or too difficult to interpret (Koesten, et al., 2017). Even when contextual metadata is provided, like in the publication of data papers, different journals have developed different standards for the contextual information requested from paper authors (Kim, 2020). Simultaneously, however, earlier research underlines that paradata and metadata for satisfying users' '*genuine*' (Papenmeier, et al., 2021) information needs can be derived not only from purposefully created descriptions but also from the structure and contents of datasets (Börjesson, et al., in press; Thomer and Wickett, 2020).

Theory

Our investigation builds on four fundamental empirically-grounded postulations that point to an enmeshed nature of information needs. First, there are different levels of specificity and explicitness of information needs. Taylor's (1968) popular model enumerates four, from visceral (unarticulated) to

compromised (articulated as a question). Second, information needs are contextual, and change over time (Bothma et al., 2013). Third, not all needs are satisfied. Sometimes if a need appears difficult to satisfy, the needy individual might change it according to what is available (Friedrich, 2020). Fourth, information (and data) use is not always precluded by an explicit need. It can begin and be complemented by appropriating what is available (Huvila, 2019). Serendipitous discoveries are a fundamental aspect of scholarship in general (Ford and Foster, 2003). In many disciplines, perhaps especially in historical sciences (Martin and Quan-Haase, 2016), serendipitous acquisition of information and data is the rule rather than an exception.

Taken together, we posit that the interrelatedness of information needs to human needs and other information needs leads to a necessity to consider and understand meta-level information needs, or *meta-information-needs*, as a specific analytical category of needs of meta-information, or *'information about the information'* (Higgins, 1999, p. 132). Such needs can be related to source credibility (as in Higgins, 1999 study), relevance, usefulness, or pertinence assessment but also, for instance, to making information or data useful for a specific purpose and to transposing observations, knowledge or a particular type of data or information to another type of information, data or knowledge. Correspondingly, such needs can be satisfied using different informational means. When the meta-information-needs concern processes, a core category of such means are different categories of paradata (i.e. data about processes), which are measured in this paper against the corresponding meta-information-needs to explicate what types of meta-information (cf. Higgins, 1999) researchers need to make research data actionable.

Material and methods

The analysis of researchers' research data creation related information needs is based on ten semi-structured interviews (A-J) with archaeologists and cultural heritage professionals. We used purposive theoretical sampling to ensure as much breadth as possible with regard to engagement with research data, subject specialisations and career stages (Robinson, 2014). Seven interviewees are from Nordic countries, and the remaining three are from the UK. Seven of them have research positions, two professional positions, and one is an independent researcher. While two of the interviewees with academic positions are full professors, four are mid-career, and one is a doctoral student. Subject specialisations include ceramic analysis, classical archaeology, environmental archaeology, geoarchaeology, historical archaeology, landscape archaeology, mediaeval archaeology and monuments protection. Characteristic to all interviewees is that they create and reuse data that is standardised to a certain extent in combination with less structured data like interpretative remarks in free text notes (for an example of an illustration of semi-structured research data see Börjesson, et al., in press). Thus as a whole, the data created and reused by the informants are semi-structured rather than structured or *tidy* (Tierney and Cook, 2020)**.**

The interviews followed an interview guide with questions on data creation and data reuse developed on the basis of earlier research on researchers' needs for contextual information (Faniel, et al., 2019), with particular focus on the creation and use of paradata about research data (Kvale, et al., 2014). By paradata the interviews referred to data on the means (procedures, tools, activities) by which a body of information came into being. Interviews covered both already existing forms of paradata and such currently non-existing forms that would be anything from crucial to useful when reusing data. The interview material, including consent forms and interview guides, will be kept in the CAPTURE project archive at Uppsala university until 2034 (Börjesson and Sköld, 2021).

The interview format allowed the interviewees to express needs at different levels of specificity, both explicitly (e.g. "*I need to know how outlier values were treated in the statistical analysis*"), and implicitly by reasoning (e.g. "*Since I didn't know how outliers had been treated in the shared dataset, I had issues knowing how to combine the dataset with my newly generated data*") (cf. Taylor, 1968). Moreover, interviewees expressed both their own information needs as data users and their views on what needs should be met by the information they supply as data creators. In line with a constructivist grounded theory approach with the researcher as an active co-creator of understanding of the social (Charmaz, 2014), in this case information-related information needs, both explicitly and implicitly

expressed needs were coded as needs, as were the interviewees own needs and the needs they assumed others to have. Thus, we engaged in an interpretative coding aiming at outlining a theory of researchers' information-related information needs. The types of needs in the broader information needs category are interpreted as four types of needs pertaining to the scope, provenance, methods, and knowledge organisation and representation paradata (Morse, 2008).

Analysis

According to the analysis, the types of paradata needed to satisfy the information needs on data creation for data reuse fall into the four major categories: *scope paradata* (what does the data cover), *provenance paradata* (where does the data come from), *methods paradata* (how was the data generated) and *knowledge organisation and representation paradata* (how is the data structured and communicated). The needs are interrelated. It often takes several types of paradata to read a dataset, e.g. provenance paradata on the research project where a dataset was generated give scope paradata explaining what the dataset includes and does not. However, the analytical separation of needs helps to highlight the last of the four needs: knowledge organisation and representation paradata, i.e. the need to know how knowledge generated using research methods has been *transposed* into data. As the results show, the transposition of knowledge into data includes choices with significant bearing on reusers' understanding of data and, ultimately, its reusability.

Scope paradata

Table 1: Information needs on research data creation met by scope paradata

Information needs on research data creation	***Interview references***
Where to find data	Repositories, collections (B, C, E, H, J), Individual researchers' holdings (G)
Aspects affecting data coverage	How region/topic has been surveyed (B, C, E), Policies directing reporting (C), Practical circumstances affecting coverage (H, J), Update interval (G, I)
Aspects affecting data quality	Survey methods used (B), Range of detail within dataset (G)

Scope paradata addresses information needs relating to data reuse that concern what data covers and not. Needing to know the scope of data can be relevant both generally, e.g. to know what data exists about a certain region, and within the bounded context of a dataset, i.e. to know exactly what that dataset covers.

The first need, needing to know what data exists for a certain location or time period, includes knowing which repositories are available. As interviewees demonstrate, institutional data repositories are one resource while additional sources include, for example, physical document or object collections, personally held unreported data and data from specific projects never submitted to institutional repositories. An illustrative example is archaeological sites investigated by multiple research projects with different approaches to documentation and archiving (H). All sources are vital to estimate the representativity of the dataset at hand and to make balanced interpretations in relation to the totality of knowledge about, e.g. a place or time period.

The second need concerns the need to know what a dataset covers, what affects its coverage and the data quality. It includes knowing how data has been gathered. As research methods tend to change during and between projects, it is also crucial to know when data was gathered. For example, knowing that one part of an archaeological site has been surveyed on the ground and that a neighbouring area

was surveyed remotely should affect interpretations of the spread of finds over the two areas (B). Further, knowing how heritage policy has directed finds reporting over time is also important, e.g. when the policy changed from voluntary to mandatory finds reporting and how the change has been implemented in practice. Yet another aspect is the database update intervals that affect database coverage vis-a-vis data gathered.

Provenance paradata

Table 2: Information needs on research data creation met by provenance paradata

Information needs on research data creation	***Interview references***
Disciplinary and timebound origin	Disciplinary-bound ways of dating and classifying objects (C, D, I), Methods and technologies in use (G, H), Type of review (D)
Epistemological culture	Observational vs interpretative data (C), Signposting (C), Quality range indicators available (C)
Rationale of data generation	Individual researcher's purposes and preferences (C, G, H, I), Institutional context of data generation (D), Resource related limitations (I), Intention to make data a primary source or a reference source (G)

Data reuse related information needs met by provenance paradata concern the origins of data. While scope paradata tells about the coverage of a dataset, provenance paradata explains the reasons underpinning the specific scope.

Provenance paradata needs expressed in the interviews can be categorised into information needs on disciplinary and timebound origin, epistemological culture, and the rationale of data generation. The reason for pairing disciplinary origin with timeboundedness is that several interviewees noted how methods and technologies used in, e.g. field archaeology or evolutionary biology differ between specific periods of time. Disciplinary origin also encompasses disciplinary-bound ways of dating and classifying, for example, in biology, the ways of dividing families of species into species at a certain point in time (D).

The analysis shows that epistemological culture while being a relevant factor in all categories of paradata, emerges as especially impactful in relation to provenance paradata. Here, epistemological culture is intertwined with disciplinary origin but less with conscious engagement with specific methods and practices and more with the epistemological underpinnings of data generation, i.e. what can be known and communicated as knowledge. Epistemological underpinnings imbue the tacit habits of data generation. For instance, the interviewees bring up the differences between generating observational or interpretative data, factual and performative data, and data that describes objects or the process an object has been subjected to.

In addition to disciplinary and epistemological influences, the rationale of the particular undertaking of generating data is important. Both the individual researchers' aims and goals and the institutional context of data generation can influence what and how data, e.g. on an object, has been generated. One example is how data granularity differs when a ceramics shard is catalogued in a collections database at a museum as opposed to a specialised research database.

Yet another difference in the rationales of database creation is that sometimes it remains unclear if a database was intended as a primary data source or as a reference source to primary data. If a database creator expects users to access the primary data elsewhere, some of the primary data might be missing from the database, while the user might assume that all available information is presented.

Methods paradata

Table 3: Information needs on research data creation met by methods paradata

Information needs on research data creation	***Interview references***
Data generation methods	How location data has been generated (C, I), Field methodology (H), Sampling context (D, F), Analysis method used (F)
Technical information, e.g. on equipment	Instrument (F), Instrument conditions (F), Calibration information (F), Filtering and manipulation (J), Precision level (F), Correction program used (F)
Decision-making information	Data structuring (C), Date classifications (D), Type classifications (F), Treatment of outliers (F), Sources of contamination (F, I)

While provenance paradata explains the context of the data generation, methods paradata delve deeper into the procedures of generating data. Data reuse related information needs addressable by methods paradata concern research methods, information on techniques and decision-making.

Data generation methods information is often, to a certain extent, covered in traditional methods descriptions, e.g. how location data has been generated or what field methodology, sampling context or analysis method was used. A joint concern among interviewees is the granularity of methods descriptions. The interviewees' own insights into the profusion of possible methods, e.g. for generating location data, lead to questions about the exact procedures used in data generation. For location data, there is firstly a difference between using a GPS device and a map to generate coordinates, and secondly, what those coordinates denote, e.g. a spot or an area.

Technical information concerns such details as equipment and instruments used, instrument condition and calibration, eventual data filtering, corrections and manipulations prior to analysis, and their level of precision. Calibration information is vital, for instance, with the equipment used to analyse the geochemistry of glass shards. There is a standardised routine for calibrating the equipment by running tests on shards with known composition, but the frequency for running tests can differ, and calibration information is not always published along with the data. The lack of this information complicates detailed comparative analyses.

Decision-making information concerns such details that are not covered by method and technical information alone, for example, how the data originator decided to treat outliers, possible sources of contamination in the field or in the lab, structure the data, and classify datings and types. A possible information need that a data reuser might have is, for example, how a sampling context has been deemed clean enough – free from sources of contamination – to be used as the basis for dating a specimen found in the same layer of soil. This need can arise if the dating of a species becomes questioned but cannot always be met if a dataset lacks an account of the composition of materials in the layer and an evaluation of possible sources of contamination.

Knowledge organisation and representation paradata

Table 4: Information needs on research data creation met by knowledge organisation and representation paradata

Information needs on research data creation	***Interview references***
Rationale for representation of information vs non-information	Difference between negative result and no result (B, F, H)
Subsets within data	Legacy formats (A), Different quality (D), Granularity (H), Typologies used (H)
Standards structuring data	Intention (F), Precision (C), Visualisation (E)
Semantics for representation	Internal codes (A), Granularity (C, H), Ambiguities (C), Legacy terms (D), Standardisation (D, H), Relative frames of references (F)
Rationale of location and dating data	Intention (D, F), Precision (D, E, F), Margin of error (D, I), Relation to structured ways of representing place and time (F, H)
Relations between data entities	Relation between find and dating (D)

In comparison to provenance and methods paradata, knowledge organisation and representation paradata concerns how data is structured and communicated. The interviewees express several different needs related to the presentation of data, including the need to know how missing information is represented, how to differentiate between different subsets of data, standards used for structuring, the semantics for representation, the rationale behind location and dating data, and the relations between data entities, like a find and a dating.

Needing to know how missing information is represented ties back to scope paradata in the sense that knowing if an empty cell means "no data gathered" or a negative result is crucial for evaluating what the dataset covers and not and how much work it takes to prepare the dataset for analysis. Taking the empty cell as an example, the absence of data can mean a range of different things, like "no data was gathered", "data was gathered but does not indicate value X", "data was gathered but not yet ingested into dataset", "data was gathered but does not meet database creator's quality criteria" etc. If all empty cells in a dataset denote the same, the problem can be solved by using one command (e.g. "fill all empty cells with a code for 'no data gathered'"). If empty cells in a dataset denote different types of information, the data reuser may need to go back to the primary source to fill all cells with values representing each specific reason for data absence.

For similar reasons, interviewees express a need to identify subsets within the data to understand why coverage or quality differs in different parts of the dataset. The explanation might be that a certain category of data was not collected before a certain date in the project or that less precise location data has been recorded in the absence of more precise location data.

Standards, the type of standards and the possible lack of standards used for structuring data are of interest to understand the rationale behind the database structure as, for example, completely unique to the project or related to other projects or data collection endeavours. Similar needs are expressed with regards to the guiding principles behind the alphanumeric forms used to populate the structure, like the need to know what classification standards – both legacy standards and current ones – are used and

with what degree of strictness, the frames of references used to generate descriptive data, and the meaning of non-standardised codes in a dataset.

Finally, interviewees express a need to understand the rationales behind the representations of location and dating data and relations between data entities in the dataset. This need ties back to the need for methods paradata on the procedures used to generate for example a dating, but adds another layer of information need in that, with the example of a dating, the dating can be transposed into data as a year, a timespan, a time period classification, and given with or without caveats. If we also add the potential variations in types of relation between data entities, like a date and an object, it is evident that knowledge organisation and representation paradata would be informative. The dating data can, for example, be the dating of a sample taken from the object, a dating of a sample taken from the finds context of the object (e.g. the soil surrounding the object), or classification based on a typology of similar dated objects, or an estimation based on the data originator's experience from handling the same type of objects. The absence of declaration of relation between data entities like a dating and an entity can reduce the dataset's value to data reusers.

Discussion

The aim of this paper is to provide new knowledge on researchers' information needs relating to the creation and manipulation of research data they reuse in their research as an example of an information-related information need or a *meta-information-need*. Four categories of paradata needed to meet the information needs on research data creation for research data reuse emerge in the analysis: scope, provenance, methods, and knowledge organisation and representation paradata.

Certain limitations need to be taken into account when interpreting the findings. Even if the study population represents a group of researchers and professionals with subject specialisations engaging with data in various ways, the information needs are likely to vary from one discipline to another, meaning that the results of this study are not directly applicable to all fields of research. Another inherent bias in the material relating to interviews is that the needs recalled by the interviewees might not always correspond completely with their 'actual' needs when they occur. Also, due to the limited number of interviews included N(=10) the results should be considered exploratory rather than final.

While the results largely overlap with and confirm previous studies of information needs in data reuse (e.g. Gregory, et al., 2019), the findings expand the knowledge on meta-information-needs relating to data creation and manipulation, especially in three respects: I. knowledge organisation and representation as an emerging meta-information need, II. distance-to-data as factoring into the contextuality of meta-information needed, and III. meta-information-needs as needs that are only partially satisfiable.

I.

Information needs to be satisfiable with scope paradata, i.e. what a dataset covers, which are well documented in previous research (David, 1991; Faniel and Jacobsen, 2010; Friedrich, 2020). This includes both where to find data and what the found data covers both in relation to the empirical phenomena and to how the phenomena have been surveyed up to date. The need for provenance information (or paradata), including the framing of research and original research questions, has similarly been documented earlier (Faniel, et al., 2019; Faniel and Jacobsen, 2010; Gregory et al., 2019). Both scope and provenance information describe the dataset itself but are at the same time information about the history of the research that produced the data and was conducted in, e.g. a specific region, on a specific topic, by a certain group of researchers etc. Traditionally, the reviews of previous research focus on research results. What would be required to fulfil the need of meta-information on scope and provenance would be research reviews focusing on where and how research has been conducted. A pertinent parallel question is how and how much of this information can be expressed as structured metadata or paradata.

The need for methods information, also identified by for example Gregory, et al. (2019) and Miksa, et al. (2014), is the category of information needs with the greatest similarity to traditional methods descriptions in research literature and the primary need for extended methods descriptions in the data paper genre caters to (Kim, 2020). While data collection and generation methods traditionally have been accounted for in greater detail in research publications, the analysis indicates that the methods for knowledge organisation, like dataset structuring and data representation, require as much explanation. The interviewed researchers—who all had experience of both using and creating data—reflected on and posed questions about the structuring of data, eventual subsets within a dataset, the rationales for representations of knowledge as alphanumeric content – in sum, how knowledge has been transposed into the dataset format. While the need to know about data collection and generation is a standard scholarly trope known and acknowledged by the most researchers (as documented in the literature e.g. Gregory, et al., 2019; Miksa, et al., 2014), it seems that the need to know about knowledge organisation and representation is less recognised among researchers without explicit knowledge of extensive data-intensive research. Therefore there is a risk that while this meta-information need remains unrecognised and unsatisfied, as Friedrich (2020) suggests of unmet data needs, the explicitly or implicitly needy individuals might change their understanding of what is needed—and as we are inclined to suggest on the basis of the present study, ignore the need with possibly significant consequences to their research.

Apart from underlining the need to consider how to make researchers aware of potentially crucially meta-information and its function in their work and explaining principles for knowledge organisation and representation, for instance, as an aspect of data literacy (cf. Schneider, 2013), the observation points also to what Kumpulainen, et al. (2019) suggest of information needs: that they form a bridge between people's cognitive space and the space of research documentation. In a similar but broader sense of scholarly collectives, information and especially meta-information-needs form a bridge between the immaterial space of epistemic cultures and research data they deal with.

As a whole, to advance both practical and theoretical state-of-the-art of semi- and non-structured knowledge organisation and representation, there is a dire need for a more critical analysis of knowledge organisation and representation schemes and their use (see for example Börjesson, et al., in press; Kjellman, 2013). This is a theoretical exercise to explore the assumptions underpinning semantic representations of knowledge as data, but also a journey into the purposive and non-purposive knowledge organisation and representation design of all formats, e.g. structured documents, software, information systems, infrastructural resources, through which observations transpose into data.

II.

Similarly to what is known of information needs in general, the analysis shows that also meta-information needs are contextual. They emerge in a situation where a person tries to interact with data and analogously to general information needs (e.g. Bates, 1996b; Fry and Talja, 2005), they vary between research fields and epistemic cultures. Partly, as Huvila (2020) suggests, distance to the community where the data originated—and the analysis here shows the data itself—is an important parameter in deciding when contextual needs shift. A general interest in data generates *continuous needs* to know more to understand the data while a specific interest in a specific dataset generates *discrete needs* to know specific details about data for particular analytical purposes or to be able to draw conclusions at a particular level of certainty.

The contextuality of meta-information-needs challenges ambitions to codify meta-information into structured paradata similar to structured metadata. The contextuality calls for alternative approaches to generate, identify, extract and use meta-information.

III.

The analysis also points to—again similarly to information needs in general—that data related information needs and meta-information-needs might not always be possible to meet. Sometimes

information has not been documented or is lost. Sometimes the level of required detail of information is not good enough for unanticipated uses, for instance, when an interviewee criticised a data creator for omitting *how* a particular archaeological context was deemed clean enough for sampling material. Further, some aspects of data creation, for example, epistemological influences, are difficult to document if they lack clearly articulated and articulable descriptions and definitions.

Further, as earlier studies have suggested, information is not necessarily detailed enough or "inspire the confidence needed for reuse" (Gregory et al., 2019), and sometimes the information is not easy to interpret if it is out of context (Koesten et al., 2017). The direct consequence of this is that researching information and perhaps especially meta-information-needs leads inexorably to discovering countless needs that cannot be met and a realisation that all datasets are useless for innumerable purposes. Attempting to meet these needs leads to an endless need to produce and keep more information, and further, as Gant and Reilly (2017) underline, new layers of meta-information: meta-metadata to describe metadata and para-paradata to describe paradata. The findings of the present study on insufficient meta-information calls for further research on how to assess and decide what data *cannot* be reused for. Such knowledge would support the development of ethics in data reuse in the wake of Open science.

As disheartening this might sound and to undermine the relevance of describing and preserving data at all, it is worth remembering that even if sometimes some needs cannot be met using data descriptions (cf. Borgman and Bourne, 2021), they can be satisfied by collaborating with data creators (Pasquetto, et al., 2019). Moreover, while data descriptions may not fulfil reusers meta-information-needs, descriptive information may serve as clues to how data was generated and structured, and form a basis for serendipitous insights (cf. Ford and Foster, 2003) and situational appropriation (cf. Huvila, 2019) of that information. As such, it can guide meta-information generation and extraction at the point of data reuse for the *discrete* data reuse purposes (Börjesson, et al., in press). Further, as Faniel and Jacobsen (2010) remind us, data needs to be comprehensive but not flawless to be useful. Similarly, even if it can be analytically useful to consider meta-information-needs, it does not mean that they are necessarily solved with meta-information or meta-meta-information.

Conclusions

This paper identifies four categories of paradata corresponding with types of meta-information-needs researchers pertaining to data creation processes: 1) scope, 2) provenance, 3) methods, 4) knowledge organisation and representation paradata. As the last category is the least explored both in the information studies research and research data management practice so far, the findings point to the need to develop the understanding of the needs and means of effective documentation of knowledge organisation and representation both in theory and practice. This applies not least to the data literacy of researchers producing and using such descriptions and how such information is and can be created and used in practice. The findings suggest further that distance-to-data is a significant parameter in determining whether information needs are *continuous* or *discrete*. Similarly, it is suggested that the most likely type of reuse (e.g. reference, comparison, data aggregation) should guide determining the level and type of paradata. Finally, the findings underline that in spite of the comprehensiveness of available meta-information, it is always incomplete. As a consequence, many meta-information-needs remain impossible to satisfy, and complementary means—including collaboration with data creators—are needed to make information reusable.

Acknowledgements

We would like to thank the interview study participants for taking their time to give us insights into their work. We are also grateful to the conference organisers and the two anonymous reviewers who 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).

About the authors

Dr Lisa Börjesson works as a researcher at the Department of ALM at Uppsala University in Sweden. Her research focuses on research information, including research information management systems, data descriptions, data publishing and use. She can be contacted at lisa.borjesson@abm.uu.se

Isto Huvila is professor in Information Studies at the Department of ALM, Uppsala University in Sweden. Huvila chaired the recently closed COST Action ARKWORK and is directing the ERC funded research project CAPTURE. His primary areas of research include information and knowledge management, information work, knowledge organisation, documentation, and social and participatory information practices. He can be contacted at isto.huvila@abm.uu.se

Dr Olle Sköld is a senior lecturer at the Department of ALM and the director of Uppsala University's Master's Programme in Digital Humanities. His research is characterised by a broad interest in the ALM field, research data creation and use, and digital humanities. He can be contacted at olle.skold@abm.uu.se

References

Börjesson, L., Friberg, Z., Sköld, O., Löwenborg, D., Pálsson, G., & Huvila, I. (in press). Re-purposing excavation database content as paradata—An explorative analysis of paradata identification challenges and opportunities. *KULA: Knowledge Creation, Dissemination, and Preservation Studies*.

Börjesson, L., & Sköld, O. (2021). The making and use of paradata: An interview study. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-455730 (Archived by the Internet Archive at https://web.archive.org/web/20220531125810/http://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A1601832&dswid=1554)

Bates, M. J. (1996a). Learning about the information seeking of interdisciplinary scholars and students. *Library Trends*, *45*(2), 155-164.

Bates, M. J. (1996b). The Getty end-user online searching project in the humanities: Report no. 6: Overview and conclusions. *College & Research Libraries*, *57*(6), 514-523. https://doi.org/10.5860/crl_57_06_514

Bishop, B. W., Hank, C., Webster, J., & Howard, R. (2019). Scientists' data discovery and reuse behavior: (Meta)data fitness for use and the FAIR data principles. *Proceedings of the Association for Information Science and Technology*, *56*(1), 21-31. https://doi.org/10.1002/pra2.4

Borgman, C. L., & Bourne, P. E. (2021). Why it takes a village to manage and share data. *Harvard Data Science Review (under Review)*. https://doi.org/10.48550/arXiv.2109.01694

Borlund, P., & Pharo, N. (2019). A need for information on information needs. *Information Research*, *24*(4), paper colis1908. (Archived by the Internet Archive at https://web.archive.org/web/20191216154208/http://informationr.net/ir/24-4/colis/colis1908.html)

Bothma, T. J. D., Bergenholtz, H., & Bergenholtz, H. (2013). 'Information needs changing over time': A critical discussion. *South African Journal of Libraries and Information Science*, *79*(1), 22-34. https://doi.org/10.7553/79-1-112

Bowker, G. C. (2000). Biodiversity datadiversity. *Social Studies of Science*, *30*(5), 643-683. https://doi.org/10.1177/030631200030005001

Case, D. O., & Given, L. M. (2016). *Looking for information: A survey of research on information seeking, needs, and behavior*. Emerald.

Chao, T. C. (2015). Enhancing metadata for research methods in data curation. *Proceedings of ASIS&T-AM*, *51*(1), 1-4. https://doi.org/10.1002/meet.2014.14505101103

Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.-D., Kacprzak, E., & Groth, P. (2019). Dataset search: A survey. *The VLDB Journal*, *29*(1), 251-272. https://doi.org/10.1007/s00778-019-00564-x

Charmaz, K. (2014). Constructing Grounded Theory. London: SAGE.

Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities*. The University of Chicago Press.

David, M. (1991). The science of data sharing. Documentation. In J. E. Sieber. (Eds.), *Sharing social science data. Advantages and challenges* (pp. 91-115). SAGE

Faniel, I. M., Frank, R. D., & Yakel, E. (2019). Context from the data reuser's point of view. *Journal of Documentation*, *75*(6), 1274-1297. https://doi.org/10.1108/JD-08-2018-0133

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW)*, *19*(3), 355-375. https://doi.org/10.1007/s10606-010-9117-8

Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *JASIST*, *67*(6), 1404-1416. https://doi.org/10.1002/asi.23480

Faniel, I. M., Kriesberg, A., & Yakel, E. (2012). Data reuse and sensemaking among novice social scientists. In A. Grove (Eds.), *Proceedings of the American society for information science and technology* (Vol. 49, pp. 1-10). ASIS&T. https://doi.org/10.1002/meet.14504901068

Ford, N., & Foster, A. (2003). Serendipity and information seeking: An empirical study. *Journal of Documentation*, *59*(3), 321-340. https://doi.org/10.1108/00220410310472518

Friedrich, T. (2020). *Looking for data: Information seeking behaviour of survey data users*.(Humboldt-Universität zu Berlin PhD thesis). https://doi.org/10.18452/22173

Fry, J. (2006). Scholarly research and information practices: A domain analytic approach. *Information Processing & Management*, *42*(1), 299-316. https://doi.org/10.1016/j.ipm.2004.09.004

Fry, J., & Talja, S. (2005). *The cultural shaping of scholarly communication: Explaining e-journal use within and across academic fields*. Proc. Am. Soc. Info. Sci. Tech. *41*(1), 20-30. https://doi.org/10.1002/meet.1450410103

Gannon-Leary, P., Bent, M., & Webb, J. (2007). Researchers and their information needs: A literature review. *New Review of Academic Librarianship*, *13*(1-2), 51-69. https://doi.org/10.1080/13614530701868686

Gant, S., & Reilly, P. (2017). Different expressions of the same mode: A recent dialogue between archaeological and contemporary drawing practices. *Journal of Visual Art Practice*, *17*(1), 100-120. https://doi.org/10.1080/14702029.2017.1384974

Geser, G., & Selhofer, H. (2014). *D2.1 First Report on Users´Needs*. ARIADNE. http://legacy.ariadne-infrastructure.eu/wp-content/uploads/2019/07/ARIADNE_D2-1_First_report_on_users_needs.pdf (Archived by the Internet Archive at https://web.archive.org/web/20200831105711/http://legacy.ariadne-infrastructure.eu/wp-content/uploads/2019/07/ARIADNE_D2-1_First_report_on_users_needs.pdf)

Golub, K., Tyrkkö, J., Hansson, J., & Ahlström, I. (2020). Subject indexing in humanities: A comparison between a local university repository and an international bibliographic service. *Journal of Documentation*, *76*(6), 1193-1214. https://doi.org/10.1108/JD-12-2019-0231

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019). Searching data: A review of observational data retrieval practices in selected disciplines. *JASIST*, *70*(5), 419-432. https://doi.org/10.1002/asi.24165

Higgins, M. (1999). Meta-information, and time: Factors in human decision making. *JASIS*, *50*(2), 132-139. https://doi.org/10.1002/(sici)1097-4571(1999)50:2<132::aid-asi4>3.0.co;2-n

Huvila, I. (2019). Genres and situational appropriation of information. *Journal of Documentation*, *75*(6), 1503-1515. https://doi.org/10.1108/jd-03-2019-0044

Huvila, I. (2020). Information-making-related information needs and the credibility of information. *Information Research*, *25*(4), paper isic2002. https://doi.org/10.47989/irisic2002

Huvila, I., Greenberg, J., Sköld, O., Thomer, A., Trace, C., & Zhao, X. (2021). Documenting information processes and practices: Paradata, provenance metadata, life-cycles and pipelines. *Proceedings of the Association for Information Science and Technology*, *58*(1), 604-609. https://doi.org/10.1002/pra2.509

Ingwersen, P., & Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer. https://doi.org/10.1007/1-4020-3851-8

Kim, J. (2020). An analysis of data paper templates and guidelines: Types of contextual information described by data journals. *Science Editing*, *7*(1), 16-23. https://doi.org/10.6087/kcse.185

Kjellman, U. (2013). A whiter shade of pale. *Scandinavian Journal of History,38*(2), 180-201. https://doi.org/10.1080/03468755.2013.769458

Knorr-Cetina, K. (2003). *Epistemic cultures: How the sciences make knowledge*. Harvard University Press.

Koesten, L., Kacprzak, E., Tennison, J. F. A., & Simperl, E. (2017, May). The trials and tribulations of working with structured data. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3025453.3025838

Koesten, L., Kacprzak, E., Tennison, J., & Simperl, E. (2019). Collaborative practices with structured data: Do tools support what users need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1-14. https://doi.org/10.1145/3290605.3300330

Kumpulainen, S., Keskustalo, H., Zhang, B., & Stefanidis, K. (2019). Historical reasoning in authentic research tasks: Mapping cognitive and document spaces. *JASIST*, *71*(2), 230-241. https://doi.org/10.1002/asi.24216

Kvale, S., Brinkmann, S., & Torhell, S.-E. (2014). *Den kvalitativa forskningsintervjun* (Tr [reviderade] upplagan). Studentlitteratur.

Martin, K., & Quan-Haase, A. (2016). The role of agency in historians´ experiences of serendipity in physical and digital information environments. *Journal of Documentation*, *72*(6), 1008-1026. https://doi.org/10.1108/JD-11-2015-0144

Miksa, T., Strodl, S., & Rauber, A. (2014). Process management plans. *International Journal of Digital Curation*, *9*(1), 83-97. https://doi.org/10.2218/ijdc.v9i1.303

Morse, J. M. (2008). Confusing Categories and Themes. *Qualitative Health Research*, *18*(6), 727-728. https://doi.org/10.1177/1049732308314930

Murillo, A. P. (2016). *Data sharing and data reuse: An investigation of descriptive information facilitators and inhibitors*. UNC Chapel Hill.[PhD dissertation]

Papenmeier, A., Krämer, T., Friedrich, T., Hienert, D., & Kern, D. (2021). Genuine Information Needs of Social Scientists Looking for Data. *Proceedings of the Association for Information Science and Technology*, *58*(1), 292-302. https://doi.org/10.1002/pra2.457

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and Reuses of Scientific Data: The Data Creators' Advantage. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.fc14bf2d

Robinson, O. C. (2014). Sampling in Interview-Based Qualitative Research: A Theoretical and Practical Guide. *Qualitative Research in Psychology*, *11*(1), 25-41. https://doi.org/10.1080/14780887.2013.801543

Roos, A. (2016). *Understanding information practices in biomedicine: A domain analytical approach*. Hanken School of Economics.

Schneider, R. (2013). Research data literacy. In S. Kurbanoğlu, E. Grassian, D. Mizrachi, R. Catts, & S. Špiranec (Eds.), *Worldwide commonalities and challenges in information literacy research and practice* (pp. 134-140). Springer. https://doi.org/10.1007/978-3-319-03919-0_16

Taylor, R. S. (1968). Question-negotiation and information seeking in libraries. *College & Research Libraries*, *29*, 178-194.

Tenopir, C., Palmer, C. L., Metzer, L., van der Hoeven, J., & Malone, J. (2011). Sharing data: Practices, barriers, and incentives. *Proceedings of ASIS&T-AM*, *48*(1), 1-4. https://doi.org/10.1002/meet.2011.14504801026

Thomer, A., Cheng, Y.-Y., Schneider, J., Twidale, M., & Ludäscher, B. (2017). Logic-based schema alignment for natural history museum databases. *Knowledge Organization*, *44*(7), 545-558. https://doi.org/10.5771/0943-7444-2017-7

Thomer, A., & Wickett, K. M. (2020). Relational data paradigms: What do we learn by taking the materiality of databases seriously? *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951720934838

Tierney, N. J., & Cook, D. H. (2020). Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. *ArXiv:1809.02264 [Stat]*. http://arxiv.org/abs/1809.02264 (Archived by the Internet Archive at https://web.archive.org/web/20220601000000/https://arxiv.org/abs/1809.02264)

Toms, E. G., & O'Brien, H. L. (2008). Understanding the information and communication technology needs of the e-humanist. *Journal of Documentation*, *64*(1), 102-130. https://doi.org/10.1108/00220410810844178

Warwick, C. (2012). Studying users in digital humanities. In C. Warwick, M. Terras, & J. Nyhan (Eds.), *Digital humanities in practice* (pp. 1-21). Facet.

Yoon, A. (2016). Data reusers' trust development. *JASIST*, *68*(4), 946-956. https://doi.org/10.1002/asi.23730