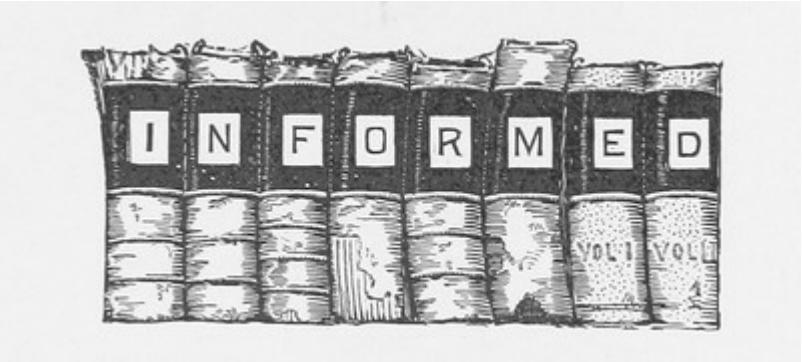
# Grey Data

written by Jeremy Huggett | 20/06/2022




In recent years, digital access to unpublished archaeological reports (so-called 'grey literature') has become increasingly transformational in archaeological practice. Besides being important as a reference source for new archaeological investigations including pre-development assessments (the origin of many of the grey literature reports themselves), they also provide a resource for regional and national synthetic studies, and for automated data mining to extract information about periods of sites, locations of sites, types of evidence, and so on. Despite this, archaeological grey literature itself has not yet been closely evaluated as a resource for the creation of new archaeological knowledge. Can the data embedded within the reports ('grey data') be re-used in full knowledge of their origination, their strategies of recovery, the procedures applied, and the constraints experienced? Can grey data be securely repurposed, and if not, what measures need to be taken to ensure that it can be reliably reused?

Although the archaeological definition of grey literature is a fairly narrow one and represents only one of the 16 main types and 55 subtypes of grey literature defined by Pejšová et al. (2011), for instance, it is nevertheless associated with the same concerns about quality and accessibility associated with grey literature more generally (e.g. Evans 2015). One of the emphases of the past twenty years has been improving the accessibility of grey literature reports in recognition of the fact that they have become – or have the potential to be – the key resource for creating knowledge about the past. Many European countries have accumulated tens of thousands of grey literature fieldwork and associated reports: for example, in the UK the Archaeology Data Service's Digital Library contains over 64,000 grey literature reports with many more being added each month, while around 60,000 grey literature reports were produced in Dutch archaeology up to 2017 with approximately 4000 added each year (Brandsen and Lippok 2021, 1).

The characterisation of grey data I'm using here is similar to that used by Edwards and Wilson who defined grey data in archaeology as "that collected and published as part of commercial archaeology, usually associated with the planning process" (2015, 4). By this definition, grey data includes data embedded in grey literature reports, as well as the data that underpin the creation of those reports in the first place, although as we'll see, connecting those two is not as straightforward as it might be. Grey data inherits many of the concerns encountered with grey literature more generally – in some respects, it can be seen as having parallels with domestic greywater in the sense that can be seen as contaminated, unclean, but potentially useful nonetheless.

Simply improving accessibility to grey literature (and grey data) through online repositories is not

the whole story, however. For example, the Rural Settlements and Landscapes of Roman Britain project aimed to realise the research potential of development-led Roman archaeology in England and Wales using grey literature, most of which was in digital format. Nevertheless, Fulford and Holbrook calculated that it took around ten person years to interrogate around 3500 grey literature and published reports relating to around 2500 sites, mostly work done since 1990, and they describe it as "likely to be a once-in-a-generation event" (Fulford and Holbrook 2018, 215). One of the main reasons behind this was the variable quality and lack of consistency in the data that they encountered in the reports (2018, 216). Many did not contain key information – for example, one or more of the total area excavated, the sampling and retention strategies employed, the quantification of assemblages of finds and environmental remains, the classification of artefact categories, etc. could be missing (2018, 224). Others have experienced similar problems: for example, in her work to create a dataset of burial mounds in SE Bulgaria, Sobotkova found that even after a year of work, the data extracted from the PDFs was not clean or flawless, with errors in transcription and formatting, imprecise locations, ambiguous phrasing, etc. In the end, missing dimensions or spatial definitions led to 1/3rd of her original dataset being rejected (Sobotkova 2018, 120).

In a recent paper (Huggett 2022), I sought to look at the data journeys behind grey literature reports. I was interested in trying to understand how data that was originally collected in the field was transformed into the final report, as a means of establishing confidence in the ability of the data in those reports to be reused. As Sobotkova observed in her account of her attempts to extract data from grey literature, "I suspect much fuzziness had already been removed in the process of writing the reports and that problems were much more pervasive in reality." (2018, 120). The table below summarises the results of my small survey of 15 cases from 2021 across a number of different commercial units and a mix of evaluations, excavations, and watching briefs.

| | Maps/Plans | | Context tables | | Finds tables | | Photos | |
|---|---|---|---|---|---|---|---|---|
| | Report | Archive | Report | Archive | Report | Archive | Report | Archive |
| Eval | ✓ | ✓ | ✓ | ✓ | × | × | ✓ | ✓ |
| Eval | ✓ | × | ✓ | × | ✓ | × | ✓ | ✓ |
| Eval | ✓ | × | ✓ | × | ✓ | × | ✓ | ✓ |
| Eval | ✓ | × | ✓ | × | ✓ | × | ✓ | ✓ |
| Eval | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Eval | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Eval | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Exc | ✓ | × | × | × | ✓ | × | ✓ | × |
| Exc | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Exc | ✓ | × | ✓ | × | ✓ | × | ✓ | ✓ |
| Exc | ✓ | × | ✓ | × | ✓ | × | ✓ | × |
| Exc | ✓ | × | × | ✓ | × | ✓ | ✓ | ✓ |
| Exc | ✓ | × | ✓ | × | ✓ | × | ✓ | ✓ |
| WB | ✓ | × | ✓ | × | ✓ | × | ✓ | ✓ |
| WB | ✓ | × | × | ✓ | ✓ | ✓ | ✓ | ✓ |

While most grey literature reports in the study included digitally created maps and plans and digital photos, some are missing summary tables of contexts and small finds. However, there's a degree of

consistency across the reports which might almost be described as representing some kind of template, which, given the reporting requirements laid down by professional bodies (e.g., CIfA 2020a,b) is not too surprising. But when it comes to the associated archives, the picture is much more mixed, with maps, plans, context and finds tables commonly absent. Surprisingly, the archives of several excavations contained only images although one archive was completely missing any of these datasets, containing only some text files derived from the report. Insofar as you can generalise from this limited data, for some reason evaluations appear to be represented more completely than any of the other categories of intervention. These results broadly matched those reported by Richards (2017, 229-230), who found that few project archives met minimum standards, frequently only including the text of the report and some photographs. It would be easy to argue that this points to a worrying degree of unreliability and poor quality which raises questions about the reuse of these kinds of data. However, it's more likely to be a picture of a collection in flux, and it may well be that the missing elements have yet to be added into the system (bearing in mind that the year selected was 2021), or that they have been added but are not yet linked to the parent reports, and a combination of both might be the case here. I've not looked at whether selecting an earlier year would dramatically change the results, but one might hope so.

Even with the limitations of this study, though, it is interesting to consider the results alongside the data-related problems identified by Fulford and Holbrook. In combination, they make it difficult to assess the journey between site data and report and understand something of the translations that take place during the post-excavation process. That has to leave open questions about the reliability and quality of the grey data extracted from grey literature reports in the absence of the ability to comprehensively revisit the original site data archives themselves. Understanding the character of the data is key to having confidence in how it can be extracted and reused in a robust and reliable way. A hands-on approach to grey literature and its associated grey data, while time-consuming, does at least represent a close engagement with the material and enables a clearer appreciation of its origins, inconsistencies, and variabilities. It's a form of 'slow' archaeological analysis: a positive friction that helps reveal the nature of the data and enable its thoughtful and aware reworking. However, the move to semi-automated and ultimately automated approaches to extracting data from grey literature introduces a more arms-length, remote relationship, mining data as a resource, setting aside local constraints and imperfections in the pursuit of large-scale big data-style analyses, and the results here suggest that the quality and reliability of the data in advance of its extraction may be over-estimated.

*[This post is part of a presentation given to GRASCA, the Graduate School in Contract Archaeology at Linnaeus University on 7th June 2022. Thanks to Cornelius Holtorf and colleagues for their invitation and generous hospitality]*

### References

Brandsen, A. and Lippok, F. 2021. 'A burning question – Using an intelligent grey literature search engine to change our views on early medieval burial practices in the Netherlands', *Journal of Archaeological Science*, 133, 105456. https://doi.org/10.1016/j.jas.2021.105456

CIfA 2020a. 'Standard and Guidance for an Archaeological Watching Brief'. Chartered Institute for Archaeologists. https://www.archaeologists.net/codes/cifa

CIfA 2020b. 'Standard and Guidance for Archaeological Field Evaluation'. Chartered Institute for Archaeologists. https://www.archaeologists.net/codes/cifa

Edwards, B. and Wilson, A.T. 2015. ‘Open Archaeology: Definitions, Challenges and Context’,in Wilson, A.T. and Edwards, B. (eds) *Open source archaeology: ethics and practice*. pp, 1-5. Berlin: De Gruyter Open. https://doi.org/10.1515/9783110440171-002

Evans, T. 2015. ‘A Reassessment of Archaeological Grey Literature: semantics and paradoxes’, *Internet Archaeology*, 40. https://doi.org/10.11141/ia.40.6

Fulford, M. and Holbrook, N. 2018. ‘Relevant Beyond the Roman Period: Approaches to the Investigation, Analysis and Dissemination of Archaeological Investigations of the Rural Settlements and Landscapes of Roman Britain’, *Archaeological Journal* 175(2), 214–230. https://doi.org/10.1080/00665983.2017.1412093

Huggett, J. 2022. ‘Data Legacies, Epistemic Anxieties, and Digital Imaginaries in Archaeology’, *Digital* 2(2), 267–295. https://doi.org/10.3390/digital2020016

Pejšová, P., Mynarz, J. and Simandlová, T. 2011. ‘A linked-data vocabulary of grey literature document types: Version 1.0’, in Thirteenth International Conference on Grey Literature (GL13), Washingon, DC. Available at: http://invenio.nusl.cz/record/81435 (Accessed: 25 May 2022).

Richards, J. 2017. ‘Twenty Years Preserving Data: A View from the United Kingdom’. *Advances in Archaeological Practice 5(3)*, 227–237, https://doi.org/10.1017/aap.2017.11

Sobotkova, A. 2018. ‘Sociotechnical Obstacles to Archaeological Data Reuse’, *Advances in Archaeological Practice* 6(2), 117–124. https://doi.org/10.1017/aap.2017.37