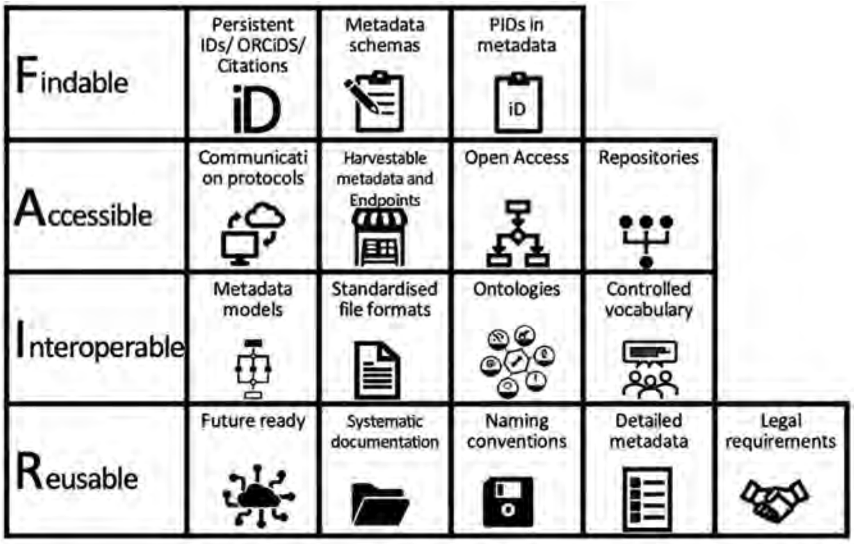
Making archaeological grey literature work harder:

FAIR data and AI as allies in information management

Grey literature (the reports, assessments and technical documents forming archaeology's backbone) represents both the sector's greatest resource and its most persistent information management challenge. Thousands of excavation reports sit in archives, rich with data that could transform our understanding of past societies. Yet accessing and reusing this information remains frustratingly difficult. The question isn't whether we have enough data; it's whether we can find and use what we already have.



Findable	Persistent IDs/ ORCiDS/ Citations	Metadata schemas	PIDs in metadata		
Accessible	Communication protocols	Harvestable metadata and Endpoints	Open Access	Repositories	
Interoperable	Metadata models	Standardised file formats	Ontologies	Controlled vocabulary	
Reusable	Future ready	Systematic documentation	Naming conventions	Detailed metadata	Legal requirements

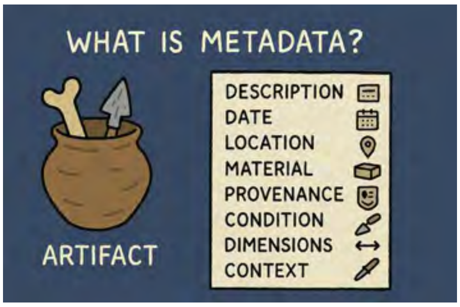
The FAIR data principles as interpreted for this article. Credit: Go-FAIR n.d.

Consider a typical research scenario: someone investigating medieval diet in northern England needs bioarchaeological data from excavation reports. Where do they start? Which archives hold relevant reports? What terminology did different authors use for the same conditions? These aren't edge cases, they're everyday challenges that consume countless research hours and often lead to data simply being regenerated rather than re-used.

Recent research examining heritage data governance within England's High Street Heritage Action Zones programme revealed similar patterns. Despite extensive creation of valuable datasets, the long-term accessibility and reusability of project data was not a focus, meaning the legacy is uncertain. Archive deposits were made in some cases, but whether future researchers could effectively access and understand that information was far from guaranteed. This extends beyond individual projects to affect entire regeneration programmes, where extensive community engagement created highly valuable content, but current practices risk this information becoming effectively invisible within a few years.

The answer lies not in creating more documentation requirements, but in making smarter use of what we produce. Two approaches offer practical pathways forward that archaeological organisations can adopt now, regardless of size or technical capacity: FAIR data principles and artificial intelligence.

The FAIR principles (making data Findable, Accessible, Interoperable and Reusable) provide a framework that translates abstract ideals into concrete actions. This isn't about achieving perfection; it's about incremental improvements that compound over time. Much of it is easily operable. What matters is having a plan: where will materials be deposited, and how will they remain available? *Findable* means creating metadata that helps people discover work. Using effective IDs and ensuring these are in the metadata keeps resources discoverable. *Accessible* doesn't require expensive infrastructure, rather it requires collaborations with archives such as the Archaeology Data Service, encouraging data to become as open as possible. *Interoperable* means using standards that allow data to connect across projects. Simple choices like using Getty vocabularies for object types or open file formats create compatibility without additional work. *Reusable* depends on clear documentation. Future users need to understand methodology, terminology and restrictions. Rather than generic file names like 'Report_Final_v3.pdf', descriptive titles and proper catalogue entries make the difference between data being lost and being discovered five years later.



Metadata and how they apply to archaeology. Credit: Alphaeus Lien-Talks, using GPT

Alphaeus Lien-Talks, Historic Royal Palaces

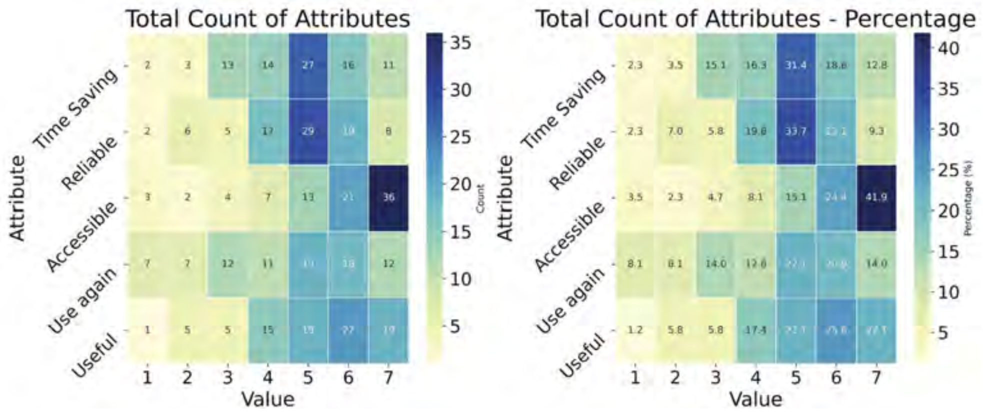
Natural Language Processing offers tangible solutions to grey literature challenges. This isn't distant future technology, it's here now. Recent work with Crossrail excavation reports led to the development of OPES (Osteoarchaeological and Palaeopathological Entity Search), a system that automatically identifies and extracts bioarchaeological information from report text. Using Named Entity Recognition, the system can process hundreds of pages in minutes, identifying skeletal elements, pathological conditions and demographic data that would take days to extract manually.

The results were encouraging. The system achieved strong performance across different report styles and terminologies, successfully handling variations like 'cribra orbitalia' versus 'orbital cribra'. Evaluation with 83 participants across expert, student and public groups demonstrated that AI-assisted search significantly improved people's ability to find relevant information, reducing time spent and creating tools that were useful. Performance metrics showed the system could identify bioarchaeological entities with accuracy rates comparable to human experts, whilst dramatically reducing extraction time.

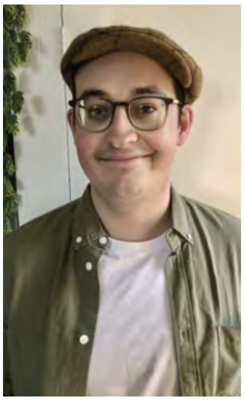
This matters because it demonstrates feasibility, not just possibility. Archaeological organisations don't need massive technical teams to benefit from these approaches. Open-source tools exist, and collaborative development can spread costs and benefits across the sector. The OPES system was built using freely available Python libraries and training methods that could be replicated for other archaeological specialisms.

Practical steps organisations can take now include adopting consistent metadata schemas, depositing materials with trusted digital archives like ADS, and exploring partnerships for developing shared tools. The Computer Applications and Quantitative Methods in Archaeology (CAA) offers one venue for connecting with others working on similar challenges.

By embracing FAIR principles as a roadmap and viewing AI as a practical ally, the sector can make grey literature work considerably harder. The data exists; what's needed is ensuring it remains findable, understandable and useful. The alternative means repeatedly rebuilding knowledge that's already been created. The sector can do better, and increasingly, it has the tools to prove it.



Combined results: (a) *count of respondents for each score 1–7 (Value) with 7 being most agreed with;* (b) *proportions of that score. All criteria showed statistically significant differences between groups (p<0.05). Credit: Lien-Talks, 2026*



Alphaeus Lien-Talks

Alphaeus is a PhD Candidate in Digital Archaeology at the University of York, working with Historic England and the Archaeology Data Service, as well as a Heritage Scientist at Historic Royal Palaces, researching heritage data governance and approaches to making archaeological data more accessible through AI and digital preservation. He chairs the CAA UK national chapter and CAA Special Interest Group on Machine Learning and AI and develops practical tools and frameworks that make archaeological information more findable and reusable. Find out more at www.AlfieTalks.com

References

Lien-Talks, A., 2026. Evaluating Natural Language Processing and Named Entity Recognition for Bioarchaeological Data Reuse [online]. Preprints.org preprint. Posted 10 September 2025. Available at: https://doi.org/10.20944/preprints202509.0822.v1 [Accessed 14 January 2026].