E-RIHS
EUROPEAN RESEARCH INFRASTRUCTURE FOR HERITAGE SCIENCE

E-RIHS PP

CALL: H2020-INFRADEV-2016-2

TYPE OF ACTION: CSA

GA n.739503

D.5.3 Data Curation Policy

Lead Authors: Holly Wright & Julian Richards

With contributions from: Olivia Foster, Jessica Hendy, Carrie Wright, Paul Northrup, E. Troy Rasbury & Ashley Coutu

Deliverable nature	Report (R)
Dissemination level	Public
Contractual delivery date	2020-05-30
Actual delivery date	2020-05-13
Version	1.0
Total page number	50
Keywords	ERIC, heritage science, data types, data management, open data

Abstract

This report reviews the issues concerning data curation for heritage science. It provides a policy framework to be implemented by E-RIHS, but the report is designed to be of use to all those with interests in data within the heritage science domain. E-RIHS provides an opportunity to ensure the long term preservation and increased re-use of heritage science data within broader European frameworks. The report follows the structure of the FAIR principles (Findable, Accessible, Interoperable and Re-usable) but interprets them in the context of heritage science. Examples are given with reference to a substantial appendix which covers a broad range of heritage science data types. It is recommended that E-RIHS researchers should be required to complete a Data Management Plan as a condition of support for their usage of an E-RIHS facility, and this, along with the information contained in this report, will assist E-RIHS researchers, facilities and repositories to follow our thirty-three recommendations.

Document information

Project number	739503	**Acronym**	E-RIHS PP
Full title	European Research Infrastructure for Heritage Science – Preparatory Phase		
Project url	www.e-rihs.eu		
Document url			
EU Project Officer	Maria Theofilatou		

Deliverable	**Number**	D.5.3	**Title**	Data Curation Policy
Work Package	**Number**	WP5	**Title**	Access and Interoperability Policies

Date of delivery	**Contractual**	M36	**Actual**	M
Status	Version 0.0		☐ Draft ☒ Final	
Nature	☐ prototype ☒ report ☐ demonstrator ☐ other			
Dissemination level	☒ Public ☐restricted			

Authors (Partner)	Holly Wright & Julian Richards, with contributions from: Olivia Foster, Jessica Hendy, Carrie Wright, Paul Northrup, E. Troy Rasbury & Ashley Coutu			
Responsible Author	**Name**	Holly Wright	**Email**	holly.wright@york.ac.uk
	Partner	UCL (ADS)	**Phone**	+441904323967

Abstract (for dissemination)	This deliverable reviews a range of methodologies in use across Heritage Science and the data types these methodologies produce, along with policy recommendations for these types.
Keywords	ERIC, heritage science, data types, data management, open data

Version Log

Issue Date	Rev. no.	Author	Change
24/10/2019	0.5	Foster, O	Draft of Appendix
30/11/2019	0.6	Wright, H., Hendy, J., Wright, C., Northrup, P., Rasbury, E. C., Coutu, A.	Additional Appendix content
19/12/2019	0.7	Richards, J. D., Wright, H.	Addition of policy analysis
28/01/2020	0.8	Richards, J.D. Wright, H.	Edit and addition of abstract and conclusion
11/05/2020	0.9	Bertrand, L	Review and comments
12/05/2020	1.0	Wright, H	Final edit

Table of contents

Executive Summary

This report reviews issues concerning data curation for heritage science. The intention is to provide a policy framework to be implemented by E-RIHS, but the report is designed to be of use to all those with interests in data within the heritage science domain. This is a broad and heterogenous area, including materials analysis, dating methods, archaeological science, biomolecular archaeology, synchrotron methods, and conservation science, amongst others. E-RIHS provides an opportunity to ensure the long term preservation and increased re-use of heritage science data within broader European frameworks. The report follows the Framework provided by the FAIR principles (Findable, Accessible, Interoperable and Re-usable) but interprets them in the context of heritage science. Examples are given with reference to a substantial appendix which covers a broad range of heritage science data types. E-RIHS is also involved in the development of the DARIAH Heritage Data Reuse Charter, which complements the FAIR principles and will create both principles and mechanisms to which heritage science practitioners should adhere. It is recommended that E-RIHS researchers should be required to complete a Data Management Plan as a condition of support for their usage of an E-RIHS facility, and this, along with the information contained in this report, will assist E-RIHS researchers, facilities and repositories to follow our thirty-three recommendations:

Findability:

1. E-RIHS repositories will need to assign persistent identifiers to datasets, and E-RIHS users should reference these PIDs in their research outputs.
2. E-RIHS should provide information about best practices in data citation to the heritage science research community and repositories, allowing users to easily cite the data, e.g. by using a standardised button which says 'How to cite this dataset'.
3. E-RIHS users should register for an ORCID.
4. E-RIHS should build communities to develop relevant metadata schemas and standards for heritage science.

Accessibility:

5. E-RIHS should ensure Heritage Science research data is easily accessible and retrievable with well-defined access conditions, using standardised communication protocols.
6. E-RIHS should work to create and sustain appropriate E-RIHS repositories.
7. E-RIHS repositories should obtain appropriate certifications.
8. E-RIHS researchers should consider legal requirements, discipline-specific policies and ethics protocols when applicable.
9. E-RIHS researchers should work to make their data Open Access whenever possible.
10. If data cannot be made Open Access, the metadata should be, which at least allows data discovery.
11. E-RIHS repositories should make (meta)data publicly accessible and harvestable by e.g. search engines, vastly improving accessibility.
12. E-RIHS should use standardised protocols to enable greater interoperability.

13. E-RIHS should maintain and publish a registry of protocol endpoints as part of DIGILAB.
14. E-RIHS should support new and developing repositories and provide best practice guidance to ensure they take the form most optimal for re-use within the E-RIHS data ecosystem.
15. E-RIHS should support repositories that make data associated with publications more accessible.

Interoperability

16. E-RIHS should support interoperability standards that are both human and machine-readable.
17. E-RIHS should promote active standards-based user development communities for heritage science
18. E-RIHS should publish the metadata models in use by the heritage science community as part of the resources in DIGILAB.
19. E-RIHS should document the technical specifications of metadata models, including defining the classes and properties, including those which are mandatory and recommended.
20. All data files held in E-RIHS repositories should be in an open, international, standardised file format to ensure long-term interoperability in terms of usability, accessibility and sustainability.
21. Researchers should use preferred formats which are recommended by E-RIHS wherever possible and are independent of specific software, developers or vendors.

Re-usability

22. E-RIHS researchers should ensure heritage science research data is ready for future research and future processing.
23. E-RIHS researchers and laboratories should ensure research data is systematically documented.
24. E-RIHS researchers and laboratories should ensure they maintain adequate version control for research data.
25. E-RIHS researchers should follow a precise and consistent file naming convention wherever possible.
26. E-RIHS repositories should develop guidelines recommending standardised preferred file formats that are widely used in the Heritage Science community.
27. E-RIHS repositories should develop metadata requirements that include information about provenance of samples, name of the laboratory, methodology and equipment.
28. To permit the widest reuse possible of (meta)data, it should be clear who the (meta)data rights holder is and what license applies.
29. E-RIHS should adopt the Creative Commons licencing framework, and map other frameworks used within E-RIHS to it.
30. Metadata should be made available under a CC-0 licence by default.
31. Datasets should be made available under a CC-BY licence by default.

Data Management Planning

32. Completion of a data management plan should be a requirement for E-RIHS support for access to E-RIHS facilities.
33. E-RIHS should adopt the PARTHENOS DMP template as the default.

1. Introduction

Heritage Science is a broad area, which can be subdivided into groups by both method and field of application, but are inherently interdisciplinary. As set out in the E-RIHS Scientific Vision:

> *Collaboration is essential to avoid duplication of investments and fragmentation of research efforts. E-RIHS will foster a culture of interdisciplinarity, exchange and cooperation associating researchers from the required disciplines on equal level (experimental sciences, arts, humanities and social sciences or digital sciences). Newly generated knowledge will be co-created by users and platform scientists. Multidisciplinary interactions within E-RIHS will be an intrinsic element of its identity, as it is of defining importance to heritage science. E-RIHS will provide access to expertise and competences as well as to instruments or databases (Bertrand et al., 2018).*

Whilst there has been a strong tradition of digital data sharing within Archaeology (Richards, 2017) this has been less widely adopted within other fields of heritage science, partly because of a lack of suitable infrastructure, but also possibly because researchers may not appreciate the benefit of sharing. This is now changing as a result of data policies from funding bodies at both European and national level, which require scientists to make available the data which underpin their research results, and E-RIHS can play an important role in making these changes efficient for the community as a whole.

1.1 Data Curation Policy - FAIR principles

Current international best practice for research data management includes implementation of the FAIR Principles (Wilkinson *et al.*, 2016) as policy for data repositories. The FAIR Principles are a guidance framework which says all research data should be Findable, Accessible, Interoperable and Re-usable. A variety of easy to understand guidance on implementing the FAIR principles for cultural heritage has recently been published by the Parthenos project, and can be freely downloaded in several European languages (PARTHENOS *et al.*, 2019). E-RIHS is also involved in the development of the DARIAH Heritage Data Reuse Charter (https://datacharter.hypotheses.org/), which complements the FAIR principles and will create both principles and mechanisms to which heritage science practitioners should adhere (Romary, no date).

Within the general principles are actionable sub-principles, as set out by GoFAIR Initiative (*FAIR Principles - GO FAIR*, no date), *in* collaboration with the European Open Science Cloud (EOSC). The principles and sub-principles include:

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessible

Once the user finds the required data, she/he needs to know how they can be accessed, possibly including authentication and authorisation.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure.

Progress has been made in heritage science across these areas over the last 20 years, but this progress has not been evenly distributed across the four principles. For example, more work has been carried out on making data interoperable, as evidenced by the adoption of ISO standards for cultural heritage such as the CIDOC CRM, than determining whether it is reusable. Of particular interest is the CRMsci, an extension to the CIDOC CRM for mapping metadata around scientific observations (*CRMsci*, no date). One of the main advantages of the FAIR principles is it creates a framework that sets out all four main principles as being of equal importance, and needing to be addressed in concert. It shows that **future policies must work to ensure all FAIR principles are being considered, rather than only focussing on implementing the aspects that are simple or do not require consensus**. The following sections are not meant to be comprehensive, but rather raise points of particular policy interest for E-RIHS within the context of each of the FAIR principles.

2. Findable

For findability, persistent, unique identifiers (e.g. DOIs, URIs, PURLs, URNs and ORCIDs) have become much more standard, but rich metadata optimised for interoperability and re-use continues to be lacking. This can be further hampered by lack of agreement and/or understanding within the heritage domain of what constitutes appropriate metadata. Heritage science research data should be easy to find by both humans and computer systems, and based on mandatory descriptions of the metadata that allows the discovery of datasets.

2.1 Persistent Identifiers

Data cannot be accessed, used, or re-used if it cannot be located. Given the lifespan of a project-based online data resource is typically less than five years, and the lack of sustainable repositories for heritage science data (Wright and Richards, 2018), the use of persistent identifiers is crucial. Persistent identifiers allow resources to be found, irrespective of where they are. If research data have a persistent identifier and are cited in accordance with community standards, the corresponding data objects or datasets are more easily found.

The most recognisable persistent identifier is the Digital Object Identifier (DOI). The concept of DOIs were created in the late 1990s within the publishing community, but its development was a recognition of the importance of non-publication digital resources and the growth of multimedia. The key characteristic of DOIs are the form of persistence they take:

> *A DOI name is permanently assigned to an object to provide a resolvable persistent network link to current information about that object, including where the object, or information about it, can be found on the Internet. While information about an object can change over time, its DOI name will not change. A DOI name can be resolved within the DOI system to values of one or more types of data relating to the object identified by that DOI name, such as a URL, an e-mail address, other identifiers and descriptive metadata. (Digital Object Identifier System Handbook, no date).*

For example, a dataset created as part of a project is uploaded into a "community" within the open research sharing platform *Zenodo*. This allows the minting of a DOI for that dataset, but in the highly unlikely event of Zenodo ceasing to function, the dataset could be moved to a new platform using the same DOI. All previous citations of the dataset will continue to resolve to the correct resource, even though it has moved. This is especially important as heritage science works to increase the academic recognition of the role of citation for datasets, ensuring proper attribution for the data creator.

A persistent author identifier (e.g. VIAF, ISNI or ORCID) also helps to create linkages between datasets, research activities, publications and researchers and allows recognition and discoverability. The most common author identifier is ORCID: *"ORCID's vision is a world where all who participate in research, scholarship, and innovation are uniquely identified and connected to their contributions across disciplines, borders, and time*" (*Our Mission | ORCID*, no date). *It* allows researchers to disambiguate themselves from

other researchers with similar names, and as researchers now tend to work across multiple institutions and countries over the course of their career, it allows them to build a research profile irrespective of their employer or affiliation. This is an important response to the more casualised work environment faced by researchers across academia, as researchers themselves become part of the wider digital infrastructure.

2.2 Rich metadata

Metadata is essential in making data findable, especially the metadata used for citing and describing data. A metadata schema is a list of standardised elements to capture information about a resource, e.g. a title, an identifier, a creator name, or a date. Using existing metadata schemas also ensures international standards for data exchange are met.

E-RIHS policy recommendations for findability:

1. E-RIHS repositories will need to assign persistent identifiers to datasets, and E-RIHS users should reference these PIDs in their research outputs.
2. E-RIHS should provide information about best practices in data citation to the Heritage Science research community and repositories, allowing users to easily cite the data, e.g. by using a standardised button which says 'How to cite this dataset'.
3. E-RIHS users should register for an ORCID.
4. E-RIHS should build communities to develop relevant metadata schemas and standards for heritage science.

3. Accessible

Heritage science data is currently held in a largely fragmentary way. With some notable exceptions, it is either not available in a standards-based format, or not accessible outside of the relevant research group. This is discussed further in the survey of E-RIHS partners carried out as part of E-RIHS D3.3 *Data Management Policy*.

3.1 Repositories

Both the research and funding communities have shown considerable interest in adopting the FAIR principles as policy since their introduction in 2016, which the developers have cited as being aspirational rather than proscriptive. They attribute the rapid and broad uptake of the FAIR principles to their community-based approach with initiatives such GoFAIR, and the Research Data Alliance encouraging broad engagement which has resulted in the funding community taking note (Wilkinson *et al.*, 2019). This engagement includes the European Commission, which has convened an Expert Group on FAIR Data. This group has recently published its Action Plan *Turning FAIR into Reality (Directorate-General for Research and Innovation, 2018)*, *setting* out the coordinated, simultaneous interventions necessary, which will be supported through EOSC. This includes the alignment of standards by member states, but also in concert with initiatives in the US, Australia and across Africa as part of larger global agreements. The report includes 27 recommendations for FAIR implementation, but of particular interest for accessibility is:

> ***Rec. 14: Provide strategic and coordinated funding***
>
> *Funders should adopt a coordinated approach to supporting core infrastructure and services, building on existing investments where appropriate. Funding should be tied to certification schemes, sustainable business models and other community-vetted indicators that demonstrate viability (Directorate-General for Research and Innovation, 2018, p. 55).*

This recommendation shows an understanding of the need to move away from a project-by-project approach to funding, and towards more long-term investment in data infrastructures. Robust implementation of the FAIR principles for heritage science requires data be available in a persistent and sustainable way, and this can only be done through the creation or ongoing development of appropriate E-RIHS repositories.

A certified repository offers a trustworthy home for datasets. Certification is a guarantee that data are stored safely, and will be available, findable and accessible over the long-term. Examples of certification standards are CoreTrustSeal, Nestor seal and ISO 16363 certification. Providing robust guidance on how to use a resource is also an important aspect of accessibility. When

depositing data in a data repository, it should be clear how to use the resource. Repositories may also provide different levels of access, and it is important to provide information about the access options a data depositor can choose.

When choosing an access option, E-RIHS researchers should consider legal requirements, discipline-specific policies and ethics protocols when applicable. Open Access is preferable where possible. When personal data is collected researchers should consider whether it contains any information which might lead to participants' identities being disclosed, to what use participants consented during their participation, and which measures need to be taken to protect personal data. If data cannot be published in an Open Access format, the metadata should be, allowing data discovery. In some cases it may be necessary to specify an embargo period. This allows at least the description of the dataset to be published even when the data is under embargo and not accessible until a later date.

By using standardised exchange protocols, E-RIHS repositories should make (meta)data publicly accessible and harvestable by e.g. search engines, vastly improving accessibility. E-RIHS should therefore use standardised protocols such as SWORD, OAI-PMH, ResourceSync and SPARQL. Metadata schemas should be converted into XML or RDF. E-RIHS should maintain a registry for protocol endpoints, the paths from which research data can be accessed, and publish them. The planned activities within DIGILAB are complementary to this type of registry, so it would be an ideal place to maintain it.

3.2 Repository Types

As part of the Integrated Platform for the European Research Infrastructure ON Cultural Heritage (IPERION CH) a questionnaire was circulated across the European cultural heritage sector to better understand the types of data resources potentially available, and create an overview of best practice. The number of databases proved to be too few, resulting in a change of focus more toward individual datasets, but in either case, the result was even fewer were found to be available online, which concurs with the findings of E-RIHS D3.3. The paucity of online resources meant it was not possible to implement the original purpose of the questionnaire; to create a web directory of online resources, but it did allow information about the resources to be categorised and made searchable within a database (Bertrand *et al.*, 2016, p. 7).

E-RIHS repositories may take various forms, including those that are method specific and those that are discipline or material specific. Method specific repositories focus on data derived from a particular heritage science method, such as microscopy or palaeogenomics. Discipline or material specific repositories focus on a particular research area or type of material under analysis, irrespective of the methods used to create and interpret the data, such as Archaeology or stained

glass. This is an important distinction, as it affects interoperability and re-use greatly. If data is too heterogeneous, it limits the interoperability at the item level, therefore limiting re-use. At the same time, data that is method specific, but the research areas are too heterogenous to answer useful research questions, will also have limited re-use. Neither type is preferable over the other, but it will be important for E-RIHS to support new and developing repositories and provide best practice guidance to ensure they take the form most optimal for re-use within the E-RIHS data ecosystem.

The data model developed for the IPERION CH database shows a similar distinction is already developed with cultural heritage. The entities set out in the data model are divided into two categories: *MaterialName* and *TechniqueName. MaterialName* includes items such as Ivory-Bone or Textiles, whereas *TechniqueName* includes items such as Gas Chromatography - Mass Spectrometry (GC-MS) and Raman Spectrometry (Bertrand *et al.*, 2016, pp. 27–30). Significant progress has been made in creating repositories that are method or technique specific that are of interest to heritage science such as:

- The ProteomeXchange Consortium: established to provide globally coordinated standard data submission and dissemination pipelines involving the main proteomics repositories, and to encourage open data policies in the field (*Proteome Exchange*, no date).
- MassIVE (Mass Spectrometry Interactive Virtual Environment): a community resource to promote the global, free exchange of mass spectrometry data. MassIVE datasets can be assigned ProteomeXchange accessions to satisfy publication requirements (*MassIVE*, no date).
- GenBank: an annotated collection of all publicly available DNA sequences, and part of the International Nucleotide Sequence Database Collaboration (*GenBank*, no date).

Progress has also been made in developing more specialised repositories that combine both a specific method and a cultural heritage topic, such as:

- IsoArcH: an Open Access and collaborative isotope database for bioarcheological samples (*IsoArcH*, no date)
- OAGR (Online Ancient Genome Repository): an Open Access repository for ancient human DNA data (*OAGR*, no date).
- IsoMemo: a Big Data initiative bringing together isotopic data from archaeology, ecology, and environmental & life sciences (*IsoMemo.com*, no date)

These examples show a range of ways to make heritage science data openly available online, in ways that make the datasets more interoperable, but include many approaches. Some are

informal and community-based, whereas others are initiatives supported at the national and international level. Some include the ability to cross-search at the item-level, while others can only be searched at the collection level and the data must be made interoperable by the (re)user. Some use standardised protocols to make the data available such as SWORD, OAI-PMH, ResourceSync and SPARQL, facilitating greater interoperability and the ability to serve machine-readable data, while others include spreadsheets for download.

By promoting and supporting best practice across the heritage science sector, E-RIHS can play an important role in supporting the creation of repositories by E-RIHS partners that are Open Access, machine harvestable and readable, using standards and protocols that facilitate interoperability. E-RIHS should also choose to support repositories that provide supplementary data that underpin publications. Accessibility is more than just providing data, it also makes publications more accessible if the underlying data can be consulted quickly and easily. E-RIHS can also use DIGILAB as a place to maintain and publish a registry for protocol endpoints for heritage science.

E-RIHS policy recommendations for accessibility:

5. E-RIHS should ensure Heritage Science research data is easily accessible and retrievable with well-defined access conditions, using standardised communication protocols.
6. E-RIHS should work to create and sustain appropriate E-RIHS repositories
7. E-RIHS repositories should obtain appropriate certifications
8. E-RIHS researchers should consider legal requirements, discipline-specific policies and ethics protocols when applicable.
9. E-RIHS researchers should work to make their data Open Access whenever possible
10. If data cannot be published in an Open Access format, the metadata should be, which allows data discovery.
11. E-RIHS repositories should make (meta)data publicly accessible and harvestable e.g. by search engines, vastly improving accessibility.
12. E-RIHS should use standardised protocols to enable greater interoperability.
13. E-RIHS should maintain and publish a registry for protocol endpoints as part of DIGILAB.
14. E-RIHS should support new and developing repositories and provide best practice guidance to ensure they take the form most optimal for re-use within the E-RIHS data ecosystem.
15. E-RIHS should support repositories that make data associated with publications more accessible.

4. Interoperable

To speed up discovery and uncover new insights, heritage science research data should be easily combined with other datasets for use by humans as well as computer systems. This includes the creation of metadata models in the form of controlled vocabularies (wordlists), thesauri (wordlists with hierarchical relationships) and ontologies (complex, non-hierarchical data structures primarily associated with graph data applications) across many domains, and the long-term development of machine-readable controlled resources for cultural heritage. Controlled vocabularies, thesauri and ontologies are a key aspect of interoperability. These frameworks provide the structure to which data must be mapped so that it may be cross-searched, and therefore interoperable.

4.1 Controlled vocabularies, thesauri and ontologies

The primary ontology for cultural heritage is the CIDOC-CRM. The CRM was developed by a special interest group under the aegis of CIDOC, the International Council for Documentation; a committee of the International Council of Museums (ICOM) (*Home | CIDOC CRM*, no date). As the CRM was designed to be a high level ontology, it does not include properties specific to heritage science, however, methodologically specific extensions of the CRM continue to be developed by various domains. Most relevant for heritage science is the CRMsci extension, which was designed as *"a global schema for integrating metadata about scientific observation, measurements and processed data in descriptive and empirical sciences such as biodiversity, geology, geography, archaeology, cultural heritage conservation and others in research IT environments and research data libraries"* (*Home | CRMsci*, no date).

CRM extensions still may not be sufficient for mapping specific entities within heritage science to the CRM, so work continues on creating standardised controlled vocabularies and thesauri reflecting specific heritage science domains. An interesting example is *IsoMemo*, which is working towards creating a repository for isotope data, but is starting from the point of building consensus around standards within the community before implementing any interface (*IsoMemo.com*, no date), ensuring interoperability, investment from the isotope research community, and a far better potential outcome for re-use.

While by far the most common, this top-down approach where a research community comes together to create a standard to which everyone agrees to map their data, is not without limitations. Researchers may feel that using a standard to structure their data may result in a loss of nuance or may not sufficiently represent the presence of uncertainty within the data. A contrasting example of a bottom-up approach, where consensus is built through the creation of the standard, is PeriodO (*PeriodO*, no date). PeriodO is intended to address the very difficult

problem of temporal heritage data, which is always contingent on place. For example, the dates associated with the Bronze Age in England are not the same as the Bronze Age in Crete. There may also be disagreement among researchers about what constitutes correct date ranges for a time period. Rather than forming consensus prior to the creation of a standard, as is used in the top-down approach, PeriodO uses a transparent system of attestation, which includes the researcher or organisation who developed the temporal designations. This allows the forming of consensus over time, rather than prior to the creation of the standard. It also means the standard is more dynamic and able to accommodate changing understandings within the research domain, whereas a top-down approach may expand, but typically remains static and changes can be difficult to incorporate.

The description of metadata elements should follow community guidelines that use open, well-defined and well-known controlled vocabularies. Such vocabularies describe the exact meaning of the concepts and qualities that the data represent. These vocabularies also provide a bridge between generalised ontologies like the CIDOC CRM and the specific data to be made interoperable. One of the most important developers of vocabulary resources for heritage science is The Getty Research Institute. The Getty has created a series of authoritative thesauri developed using community guidelines used across data resources in art, architecture, archaeology, history and art history. The most important are The Union List of Artist Names (ULAN), the Art & Architecture Thesaurus (AAT), the Getty Thesaurus of Geographic Names (TGN), the Cultural Objects Name Authority (CONA), and the Iconography Authority (IA) (Getty Research Institute, no date).

In keeping with repository types discussed in Section 5.2, controlled vocabularies, thesauri and ontologies should not only reflect the topics associated with heritage science, but also the methods. An example would be the Chemical Methods Ontology (CMO) which *"describes the methods used to collect data in chemical experiments, such as mass spectrometry and electron microscopy...analysis, such as sample ionisation and chromatography...It also describes the instruments used in these experiments, such as mass spectrometers and chromatography columns"* (Batchelor, no date).

E-RIHS should publish the metadata models in use as part of the resources in DIGILAB. It should document technical specifications and define classes (groups of things that have common properties) and properties (elements that express the attributes of a metadata section as well as the relationships between different parts of the metadata). For metadata mapping purposes, it should list the mandatory and recommended properties. Using a data standard backed by the Heritage Science community will increase the ability to share, reuse and combine data collections.

4.2 File formats

The Appendix to this deliverable provides examples of a broad range of methods employed in Heritage Science, structured according to four groupings:

- Material Characterisation Methods
- Dating Methods
- Biomolecular Methods
- Synchrotron Methods

It is by no means a comprehensive survey of heritage science techniques, but gives illustrative examples from different areas to better understand the types of data produced by heritage science methods, and the types of file formats associated with those methods. While scientific fields may have different constraints, it is possible to understand which formats may or may not be suitable for long-term preservation, even though they may be the most appropriate for data creation, development and dissemination. If recognised at the creation stage of a project, planning can be undertaken for the later conversion of problematic files (*Planning for the Creation of Digital Data*, no date). *The* Appendix is an attempt to begin understanding the scale of the problem. Moving towards standardised file formats independent of proprietary software, developers or vendors helps to ensure long-term interoperability in terms of usability, accessibility and sustainability.

The Appendix structure and headings were an attempt to trace a project workflow from inception through to data collection, processing, and long-term storage. These included:

Overview/Planning

- Purpose - why is the data being created? Are there limitations to the approach or subject?
- Expected reuse/intended audience - are the intended outputs limited or restricted in any way?
- Needs (dissemination/preservation) - are there existing requirements to share or preserve the data?
- Preparation - are there documentation and procedures relating to location, data collection, equipment, and testing?

Collection/Creation

- Raw data - identify the ‘raw data’ alongside any initial (in-device or manual) data cleaning that is undertaken.
- Format options - what formats are available at these early stages. Are these native or open formats, standards-compliant, etc.?

- Equipment settings - is there documentation for the settings or environments used in data collection or processing? Are these files self-generated or require manual documentation? Do documented protocols exist?

Processing (post-acquisition)

- Policy and protocols - as above during the collection phase, are there policy or protocol documents available for the post-acquisition phase that may help users understand and reuse data?
- Intermediary datasets - are there intermediate or transitory sets of data (i.e. between the raw and final dataset) that should be saved?
- Other versions of data - are there multiple outputs for different purposes (dissemination, preview, graphical, etc.)?

Long-term Curation

- Files / Formats - is the final data suitable for long-term preservation (format)?
- Data Selection - has the final dataset had working or draft files removed?
- Structure - is there a meaningful structure to the final dataset? Are relationships between files (raw, derived, etc.) clear? Is the relevant documentation included?

While it was relatively easy to gather information within the Appendix about the purpose of the data, very little information exists about the expected re-use or intended audience for the data beyond the needs of the researcher creating the data. Much of the instrumentation used within heritage science requires the use of proprietary software created by the manufacturer, which results in a huge range of file types in use for raw data. For example, X-ray Crystalline Powder Diffraction (XRD and XRPD) alone is associated with over 70 file formats, a small number of which are listed here:

Bruker/Siemens raw data (*.raw)
Bruker/Siemens DIFFRAC AT peak (dif) data (*.dif)
DRON-3
G670 raw data (*.gdf)
GNR raw data (formerly Ital Structures raw data) (*.esg)
HDF5 (*.h5)
INEL raw data (*.dat)
Jade/MDI/SCINTAG raw data (*.mdi)
JEOL ASCII Export raw data (*.txt)
PANalytical/Philips raw data (*.rd)
PANalytical/Philips raw data (*.udf)
PANalytical/Philips peak data (*.udi)
Rigaku raw data (*.raw)
SCINTAG raw data (*.raw, *.rd)
Seifert

Shimadzu raw data (*.raw)
Siemens raw data (*.uxd)
Sietronics XRD scan data (*.cpi)
Stoe Raw data (*.raw)
Stoe Peak File (*.pks)
TXRD text export (*.txt)
XPowder raw data (*.plv)
XRDML Scan raw data (*.xrdml)

Heritage science methods also exhibit a wide range of differences with regard to raw vs. processed data. It may be necessary for the researcher creating the data to process it from its raw state into an intermediary or transitory state before it can be re-used by another researcher. At the same time, the amount of processing may already begin to affect the results before it can be made re-usable, resulting in either data that researchers don't want to re-use, or the need for transparency in how the data was processed. Each heritage science method needs to explore where best practice lies with regard to the level of processing that is optimal for re-use, along with the other questions listed above.

E-RIHS policy recommendations for interoperability:

16. E-RIHS should support interoperability standards that are both human and machine-readable.
17. E-RIHS should promote active standards-based user development communities for heritage science
18. E-RIHS should publish the metadata models in use by the heritage science community as part of the resources in DIGILAB.
19. E-RIHS should document the technical specifications of metadata models, including defining the classes and properties, including those which are mandatory and recommended.
20. All data files held in E-RIHS repositories should be in an open, international, standardised file format to ensure long-term interoperability in terms of usability, accessibility and sustainability.
21. Researchers should use preferred formats which are recommended by E-RIHS wherever possible and are independent of specific software, developers or vendors.

5. Reusable

A consistent approach is critical for re-use, the final and most challenging FAIR principle. The emphasis on clear and accessible data usage licenses is of particular importance to heritage science, as one of the greatest barriers to re-use is lack of understanding and clarity around data use permissions. Of equal importance is the sub-principle of showing detailed provenance for a data resource. As all research areas work to increase the academic recognition of the role of citation for datasets, not just the publications that use them, but the datasets themselves, ensuring proper attribution for the data creator is paramount.

According to the FAIR principles effective data re-use is underpinned by:

- Good data quality
- Clear usage licences

In order to maximise the potential for data re-use E-RIHS should therefore adopt these data policies.

5.1 Data quality

Data quality is a shared responsibility of the researcher, the laboratory facility, and the digital repository, and each must play a role is ensuring these guidelines are followed:

- Research data should be ready for future research and future processing, making it self-evident that findings can be replicated and new research effectively builds on already acquired, previous results.
- To make clear what can and what cannot be expected in a dataset or repository, data should be systematically documented. Being transparent about what's in the data and what isn't, facilitates trust and, consequently, data reuse.
- Following a precise and consistent naming convention - a generally agreed upon scheme to name data files - makes it significantly easier for future generations of researchers to retrieve, access and understand data objects and datasets. Even if it is not possible to choose the file naming output auto-generated from instrumentation, it should be possible to include suffixes that can identify what the file contains to a human reader.
- By using standardised file formats that are widely used in the Heritage Science community, reusability is increased.

To maintain data integrity it is important that research data when first created should be identical to the research data which are accessed later on. To ensure data authenticity, checks for data integrity should be performed. Researchers and laboratories should Implement a method for version control. The guarantee that every change in a revised version of a dataset is correctly

documented is of integral importance for the authenticity of each dataset. Metadata should also include information about provenance of samples, and the name of the laboratory, methodology and equipment used in processing and analysis (see Appendix 1 for information required for sample methods).

5.2 Usage licences

In order to facilitate data re-use it is also important that the ownership of any data collected at an E-RIHS facility is clear, and that the researcher and laboratory make clear (a) who can have access to the data; (b) what they can do with it; and (c) how the ownership (including the researcher, laboratory and E-RIHS facilitation) should be acknowledged.

E-RIHS will generally require metadata about datasets to be available according to an open licence, and the general assumption will also be that the dataset itself should be available under a similar licence. However, it is accepted that in some circumstances there may be legal, cultural or intellectual reasons why this cannot always be the case (including, for example, where there is a need to protect the location of vulnerable heritage assets), or where the researcher wishes to publish their results first. In such cases it is important that the researcher and laboratory agree an embargo period with the E-RIHS repository. This should be stipulated in the Data Management Plan (see Section 6 below).

To improve interoperability, E-RIHS should adopt licensing frameworks which are already widely known and used. In some cases it is appreciated that researchers, laboratories, or repositories may need to adopt a bespoke licensing framework but E-RIHS should adopt the Creative Commons Framework as its default framework. Where facilities use different licences these should be mapped to Creative Commons. It is recommended that by default, metadata for E-RIHS datasets should be available under CC-0, as this will allow it to be harvested by a variety of aggregators, such as EOSC. Datasets should be made available under a CC-BY licence as the default

E-RIHS policy recommendations for re-use:

22. E-RIHS researchers should ensure heritage science research data is ready for future research and future processing.
23. E-RIHS researchers and laboratories should ensure research data is systematically documented.
24. E-RIHS researchers and laboratories should ensure they maintain adequate version control for research data.
25. E-RIHS researchers should follow a precise and consistent file naming convention wherever possible.

26. E-RIHS repositories should develop guidelines recommending standardised preferred file formats that are widely used in the Heritage Science community.
27. E-RIHS repositories should develop metadata requirements that include information about provenance of samples, name of the laboratory, methodology and equipment.
28. To permit the widest reuse possible of (meta)data, it should be clear who the (meta)data rights holder is and what license applies.
29. E-RIHS should adopt the Creative Commons licencing framework, and map other frameworks used within E-RIHS to it.
30. Metadata should be made available under a CC-0 licence by default.
31. Datasets should be made available under a CC-BY licence by default.

6. Data Management Plans

The most effective way of ensuring that E-RIHS researchers adhere to the FAIR principles is to require them to create a Data Management Plan (DMP) before undertaking research at any E-RIHS facility. Whilst this does not, by itself, enforce adherence to the principles, the process of completing the template forces researchers to consider and address all the considerations outlined in this data policy document. There are various standardised pro forma for DMP and some researchers may be required to abide by the requirements of their home institution or funder. However, for E-RIHS we recommend the adoption of the DMP template developed within PARTHENOS as the default template where no other format is required. This was explicitly created as a light-touch approach appropriate for the heritage community and requires the researcher to answer a series of check-box questions under six headings:

- Data summary
- FAIR data
- Allocation of resources
- Data security
- Ethical aspects
- Other

See the PARTHENOS project website (http://www.parthenos-project.eu/portal/dmp) to download the form.

E-RIHS policy recommendations for Data Management Plans

32. Completion of a data management plan should be a requirement for E-RIHS support for access to E-RIHS facilities.
33. E-RIHS should adopt the PARTHENOS DMP template as the default.

7. Recommendations

Findability:

1. E-RIHS repositories will need to assign persistent identifiers to datasets, and E-RIHS users should reference these PIDs in their research outputs.
2. E-RIHS should provide information about best practices in data citation to the Heritage Science research community and repositories, allowing users to easily cite the data, e.g. by using a standardised button which says 'How to cite this dataset'.
3. E-RIHS users should register for an ORCID.
4. E-RIHS should build communities to develop relevant metadata schemas and standards for heritage science.

Accessibility:

5. E-RIHS should ensure Heritage Science research data is easily accessible and retrievable with well-defined access conditions, using standardised communication protocols.
6. E-RIHS should work to create and sustain appropriate E-RIHS repositories
7. E-RIHS repositories should obtain appropriate certifications
8. E-RIHS researchers should consider legal requirements, discipline-specific policies and ethics protocols when applicable.
9. E-RIHS researchers should work to make their data Open Access whenever possible
10. If data cannot be published in Open Access, the metadata should be, which at least allows data discovery
11. E-RIHS repositories should make (meta)data publicly accessible and harvestable by e.g. search engines, vastly improving accessibility.
12. E-RIHS should use standardised protocols to enable greater interoperability.
13. E-RIHS should maintain and publish a registry for protocol endpoints as part of DIGILAB.
14. E-RIHS should support new and developing repositories and provide best practice guidance to ensure they take the form most optimal for re-use within the E-RIHS data ecosystem.
15. E-RIHS should support repositories that make data associated with publications more accessible.

Interoperability

16. E-RIHS should support interoperability standards that are both human and machine-readable.
17. E-RIHS should promote active standards-based user development communities for heritage science
18. E-RIHS should publish the metadata models in use by the heritage science community as part of the resources in DIGILAB.

19. E-RIHS should document the technical specifications of metadata models, including defining the classes and properties, including those which are mandatory and recommended.
20. All data files held in E-RIHS repositories should be in an open, international, standardised file format to ensure long-term interoperability in terms of usability, accessibility and sustainability.
21. Researchers should use preferred formats which are recommended by E-RIHS wherever possible and are independent of specific software, developers or vendors.

Re-usability

22. E-RIHS researchers should ensure heritage science research data is ready for future research and future processing.
23. E-RIHS researchers and laboratories should ensure research data is systematically documented.
24. E-RIHS researchers and laboratories should ensure they maintain adequate version control for research data.
25. E-RIHS researchers should follow a precise and consistent file naming convention wherever possible.
26. E-RIHS repositories should develop guidelines recommending standardised preferred file formats that are widely used in the Heritage Science community.
27. E-RIHS repositories should develop metadata requirements that include information about provenance of samples, name of the laboratory, methodology and equipment.
28. To permit the widest reuse possible of (meta)data, it should be clear who the (meta)data rights holder is and what license applies.
29. E-RIHS should adopt the Creative Commons licencing framework, and map other frameworks used within E-RIHS to it.
30. Metadata should be made available under a CC-0 licence by default.
31. Datasets should be made available under a CC-BY licence by default.

Data Management Planning

32. Completion of a data management plan should be a requirement for E-RIHS support for access to E-RIHS facilities.
33. E-RIHS should adopt the PARTHENOS DMP template as the default.

8. References

Batchelor, C. (no date) *Chemical Methods Ontology*. Royal Society of Chemistry. Available at: https://github.com/rsc-ontologies/rsc-cmo (Accessed: 20 January 2020).

Bertrand, L. *et al.* (2016) *First edition of the web directory of IPERION CH instruments and databases*. IPANEMA. Available at: https://hal.archives-ouvertes.fr/hal-02138440/ (Accessed: 19 January 2020).

Bertrand, L. *et al.* (2018) *First version of the E-RIHS scientific vision*. Available at: http://www.e-rihs.eu/wp-content/uploads/2018/09/E-RIHS_D9.1.pdf.

CRMsci (no date). Available at: *http://www.cidoc*-crm.org/crmsci/home-1 *(Accessed:* 13 May 2020).

Digital Object Identifier System Handbook (no date). Available at: *https://www.doi.org/hb.html* (Accessed: 12 January 2020).

Directorate-General for Research and Innovation (2018) *Turning FAIR data into reality : final report and action plan from the European Commission expert group on FAIR data*. Publications Office of the European Union. Available at: https://publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF (Accessed: 6 December 2018).

FAIR Principles - GO FAIR (no date) *GO FAIR*. Available at: *https://www.go*-fair.org/fair-principles/ (Accessed: 12 January 2020).

GenBank (no date). Available at: *https://www.ncbi.nlm.nih.gov/genbank/* (Accessed: 19 January 2020).

Getty Research Institute (no date) *Getty Vocabularies as LOD*, *Getty Research Institute*. Available at: https://www.getty.edu/research/tools/vocabularies/lod/ (Accessed: 19 January 2020).

Hendy, J. *et al.* (2018) 'A guide to ancient protein studies', *Nature ecology & evolution*, 2(5), pp. 791– 799. doi: 10.1038/s41559-018-0510-x.

Home | CIDOC CRM (no date). Available at: *http://www.cidoc*-crm.org/ *(Accessed:* 12 January 2020).

Home | CRMsci (no date). Available at: *http://www.cidoc*-crm.org/crmsci/ *(Accessed:* 12 January 2020).

IsoArcH (no date) *IsoArcH - An open-access and collaborative isotope database for bioarcheological samples*. Available at: *https://www.isoarch.eu/* (Accessed: 19 January 2020).

IsoMemo.com (no date). Available at: *https://isomemo.com/* (Accessed: 12 January 2020).

MassIVE (no date). Available at: *https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp* (Accessed: 19 January 2020).

OAGR (no date) *Online Ancient Genome Repository*. Available at: *https://www.oagr.org.au/* (Accessed: 19 January 2020).

Our Mission | ORCID (no date). Available at: *https://orcid.org/about/what*-is-orcid/mission *(Accessed:* 12 January 2020).

PARTHENOS *et al.* (2019) *PARTHENOS Guidelines to FAIRify data management and make data reusable*. doi: 10.5281/zenodo.3368858.

PeriodO (no date) *Periods, Organized*. Available at: *http://perio.do/en/* (Accessed: 19 January 2020).

Planning for the Creation of Digital Data (no date). Available at: https://guides.archaeologydataservice.ac.uk/g2gp/CreateData_1-0 (Accessed: 13 May 2020).

Proteome Exchange (no date). Available at: *http://www.proteomexchange.org/* (Accessed: 19 January 2020).

Richards, J. D. (2017) 'Twenty Years Preserving Data: A View from the United Kingdom', *Advances in Archaeological Practice*. Cambridge University Press, 5(3), pp. 227–237. doi: *10.1017/aap.2017.11.*

Romary, L. (no date) *Cultural Heritage Data Reuse Charter: the Mission Statement*, *Heritage Data Reuse Charter*. Available at: *https://datacharter.hypotheses.org/77* (Accessed: 10 May 2020).

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

Wilkinson, M. D. *et al.* (2019) 'Evaluating FAIR Maturity Through a Scalable, Automated, Community-Governed Framework', *bioRxiv*. doi: 10.1101/649202.

Wright, H. and Richards, J. D. (2018) 'Reflections on Collaborative Archaeology and Large-Scale Online Research Infrastructures', *Journal of Field Archaeology*. Routledge, 43(sup1), pp. S60–S67. doi: 10.1080/00934690.2018.1511960.

9. Appendix: Heritage science methods

Written and compiled by Olivia Foster, Jessica Hendy, Carrie Wright, Paul Northrup, E. Troy Rasbury & Ashley Coutu

This Appendix provides examples of a broad range of methods employed in Heritage Science, structured according to four groupings:

- Material Characterisation Methods
- Dating Methods
- Biomolecular Methods
- Synchrotron Methods

It is by no means a comprehensive survey of heritage science techniques but gives illustrative examples from different areas.

9.1 Material Characterisation Methods

9.1.1 Spectroscopy and Material Analysis

Material characterisation is the study of chemical and mineralogical composition of materials, and their vibrational and thermal analysis. It seeks to support research in the areas of geology, archaeology, heritage, chemistry, and pharmacy, as well as provide services in the fields of construction, civil engineering and nanomaterials. It is used to conduct non-invasive analyses, without the need for the prior physical and chemical preparation of the sample, and without altering the surface on which the analysis is carried out. There is also portable equipment for field use.

This laboratory has various analytical techniques for the characterization of solids, such as: i) x-ray crystalline powder diffraction, ii) x-ray fluorescence spectrometry (XRF), iii) Raman spectroscopy, iv) infrared spectroscopy and v) thermal analysis.

9.1.1.1 X-ray Crystalline Powder Diffraction (XRD and XRPD)

Overview / Planning

X-ray diffraction (XRD) and X-ray powder diffraction (XRPD) is used for phase identification of crystalline materials and to provide information on unit cell dimensions. Applications include identification of materials and components, their manufacture or origins, and any changes due to time or other influences. Samples may be powdered material, chips, cut sections, or intact objects (of limited size). For many samples no preparation is required and measurement is non-destructive. The non-destructive technique allows for the rapid (< 20 min) identification of an unknown crystalline material with minimal sample preparation required. Analysis may operate on the mm to um scale. Available in the laboratory (more limited) or at synchrotron facilities (more advanced capabilities).

XRPD has traditionally been used in fields such as geology, environmental science, material science, and engineering to identify unknown substances and in more recent years, portable X-ray diffractometers have enabled the rapid analysis of materials in the field.

Applications of XRPD in heritage science include the characterisation of crystalline materials, the identification of fine-grained minerals such as clays and mixed layer clays and the measurement of sample purity. In archaeology for example, XRPD can be used to identify pigments or to assess the polymorph of a carbonate, including those of biogenic origin such as mollusc shells prior to radiocarbon dating or stable isotopes measurements.

X-ray diffractometers are widely available and data interpretation is relatively straightforward. Manufacturers of X-ray diffractometers include Olympus, Bruker, Malvern Panalytical, Rigaku, ThermoFisher and Seifert. This technique is also available using synchrotron energy.

Collection / Creation

X-ray crystalline powder diffraction is undertaken using an X-ray diffractometer, with machines consisting of three components: an X-ray tube (generator), a sample holder and an X-ray detector. The analyzed material is finely ground, homogenized, and average bulk composition is determined. It is also possible to take a series of XRD spectra along a line so you have the dimension of space to take into account in a dataset (this also applies in XRF, FTIR, Raman, and most spectroscopies). Raw data usually 2D images or 1D intensity as a function of angle at a given wavelength. Settings typically predetermined for the specific instrument, with variable parameters recorded with data. Industry-standard formats.

The diffractometer processes and converts the x-ray signal to a count rate, with data being saved in the instrument manufacturers 'raw' state as a binary file.

The outputted data type depends largely on the diffractometer manufacturer (e.g. Siemens, Philips, Rigaku, Stoe, Seifert) and diffraction data file formats of different diffractometer manufacturers include:

Bruker/Siemens raw data (*.raw)
Bruker/Siemens DIFFRAC AT peak (dif) data (*.dif)
DRON-3
G670 raw data (*.gdf)
GNR raw data (formerly Ital Structures raw data) (*.esg)
INEL raw data (*.dat)
Jade/MDI/SCINTAG raw data (*.mdi)
JEOL ASCII Export raw data (*.txt)
PANalytical/Philips raw data (*.rd)
PANalytical/Philips raw data (*.udf)
PANalytical/Philips peak data (*.udi)
Rigaku raw data (*.raw)
SCINTAG raw data (*.raw, *.rd)
Seifert

Shimadzu raw data (*.raw)
Siemens raw data (*.uxd)
Sietronics XRD scan data (*.cpi)
Stoe Raw data (*.raw)
Stoe Peak File (*.pks)
TXRD text export (*.txt)
XPowder raw data (*.plv)
XRDML Scan raw data (*.xrdml)

There are around 70 file formats for spectral data, with some being used for multiple methods of spectrometry (e.g. the MAS/ EMSA standard file format for XRF/EDS spectra after ISO 22029 (*.msa)) (Spectragryph).

Data conversion software allows the raw data to be saved as a text (.xye?) or excel file (xml) and there are a number of 'free' XRD formats available, such as: DBWS Raw data (.rfl, .dat) and ASCII profile (.dat, .dif, .pro).

Processing (post-acquisition)

The output from XRD experiments is typically presented in the form of an x-y plot such as a diffractogram (a plot of diffracted X-Ray intensities versus the scanning angle 2θ as abscissa), the quality of which is dependent on both sample preparation and machine settings. Quantitative Phase Analysis of XRD data may be undertaken using methods such as the rietveld method. Additionally, Analysis of XRD data requires access to a standard reference file of inorganic compounds. A database of powder diffraction patterns is maintained by the International Centre for Diffraction Data (ICDD) and the Powder Diffraction File (PDF) may be used to identify substances based on the results of X-ray diffraction. Raw 2D data sets are combined and condensed into 1D intensity as a function of d-spacing, which can be matched to database entries for known phases. 1D data is usually reported and archived.

Long-term Curation

Synchrotron data is 1D, typically archived with any publication.

Resources

https://serc.carleton.edu/research_education/geochemsheets/techniques/XRD.html
https://www.iucr.org/resources/data/meeting-reports/metadata-workshop (workshop on metadata for raw data from X-ray diffraction)
http://pd.chem.ucl.ac.uk/pd/welcome.htm
https://myscope.training/legacy/xrd/introduction/
https://www.crystalimpact.com/match/help/idh_import_diffraction-data_file-formats.htm (list of diffraction data file formats)
https://www.iucr.org/resources/other-directories/software?result_42405_result_page=D (crystallographic software list)
https://www.geo.arizona.edu/xtal/geos306/fall11-11.htm
https://openei.org/wiki/Portable_X-Ray_Diffraction_(XRD)
https://www.researchgate.net/figure/XRD-data-for-an-iridium-powder_fig2_317584101 (example of XRD data for an iridium powder presented in a report)

9.1.1.2 X-ray Fluorescence Spectroscopy (XRF)

Overview / Planning

X-ray Fluorescence (XRF) imaging is used to determine distributions, co-localisation, and relative concentration of major and trace elements. Applications include identification of materials and their constituents, their physical structure, and fingerprinting to indicate provenance. Samples may be cut into sections, chips, or left as small intact objects. For most samples no preparation is required and measurement is non-destructive. Analysis may operate on the mm to nm scale. Available in the laboratory (bulk or very limited) or predominantly at synchrotron facilities. XRF is a well-established method of analysis both in laboratory and fieldwork settings and is one of the most widely used methods of analysing geological materials due to the low cost of sample preparation and ease of use. Portable XRF (pXRF) devices are increasingly being used in fields such as archaeology, with handheld devices making it possible to collect XRF data in the field. On the opposite scale, electron particle accelerators generating synchrotron radiation, which can be converted to X-rays of various energies, also provide a means of analysing archaeological and cultural heritage samples.

Manufacturers of XRF instruments include Bruker, Olympus, ThermoFisher and MalvernPanalytical.

Examples of synchrotron facilities available for research: PUMA beamline, SOLEIL Synchrotron, France, Diamond Light Source, UK and National Synchrotron Light Source II, Brookhaven National Laboratory, USA.

Collection/Creation

Laboratory instrument output of XRF experiments is measured as the number of counts of element-specific fluorescent X-ray energies received in an XRF instrument detector. This indicates which elements are present in a sample and raw data can be numerical and graphical, for example a spectrum graph. Synchrotron XRF output is usually 2D images with fluorescence spectrum (or array of element-specific intensities) at each pixel, or 3D computed microtomography. Settings typically predetermined for the specific instrument/beamline, with variable parameters recorded with data. Synchrotron facilities typically run using extremely tight security data protocols. Raw data is typically in native format, which is not readily transported or usable off-site, with archiving at the facility.

File formats for XRF laboratory instrument data include:

MAS/ EMSA standard file format for XRF/EDS spectra after ISO 22029 (*.msa)
Thermo Noran: WinTrace XRF spectra (*.spc)

File formats for XRF synchrotron beamline data include:

ASCII X, Y and Z

Processing (post-acquisition)

In order to undertake quantitative analysis of XRF data (i.e. to determine the absolute quantity of an element present in a sample), calibrations are required to convert raw qualitative data into quantitative data (e.g. fundamental parameters and empirical calibrations). This can be undertaken using software available both from the instrument used and by using free / open-source XRF data analysis software. Calibrations are created using samples with known concentrations of elements to create a calibration curve that relates the specific known concentrations to peak heights. Synchrotron XRF, raw data is typically processed into 2D images for each chemical element analysed, which can, in some cases, be manually processed into quantitative concentrations. These images are usually reported and archived.

Long-term Curation

Spectrum graphs

Example of published / archived XRF data (Migration of XRF Data to Janus):
http://www-odp.tamu.edu/publications/tnotes/tn37/tn37_20.htm

Resources

https://serc.carleton.edu/research_education/geochemsheets/techniques/XRF.html
https://www.horiba.com/en_en/x-ray-fluorescence-spectroscopy-xrf/
https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-0465-2_1305
https://www.bruker.com/products/x-ray-diffraction-and-elemental-analysis/handheld-xrf/how-xrf-works.html
https://www.bruker.com/products/x-ray-diffraction-and-elemental-analysis/handheld-xrf/archaeometry.html
https://www.bruker.com/products/x-ray-diffraction-and-elemental-analysis/handheld-xrf/xrf-data-primer-quantitative-semi-quantitative-qualitative.html
https://www.rigaku.com/downloads/journal/Vol6.1.1989/latour.pdf (n.b. 1989)
https://www.thermofisher.com/blog/mining/better-together-xrf-and-xrd/
https://www.researchgate.net/publication/316219806_Combining_XRD_and_XRF_analysis_in_one_Rietveld-like_fitting
https://www.sciencedirect.com/science/article/pii/S0305440314004440
http://elementsmagazine.org/archives/e9_1/e9_1_dep_theelementstoolkit.pdf
Archiving spectroscopic data: https://www.beilstein-institut.de/download/…

9.1.1.3 Raman Spectroscopy

Overview/Planning

Raman Spectroscopy is a chemical analysis technique which provides detailed information about chemical structure, phase and polymorphy, crystallinity and molecular interactions. Raman spectroscopy is used for the analysis of solids, powders, liquids, gels, inorganic and organic materials, pigments, ceramics and gemstones.

Manufacturers of Raman spectrometers include Mettler Toledo, Horiba and ThermoFisher.

Collection/Creation

Raman spectroscopy data is outputted as a text file with a single data pair (x, y) written to each line; tab, space or comma delimited. It is also possible to have spatially-resolved datasets.

File formats for raman spectroscopy data include:

.txt, Electron Microprobe Data File (excel)
Enspectr RaPort Raman spectra files (*.esp)
Gemlab: GL Gem Raman (*.fak, *.rruff)
Horiba: Labspec Raman spectra (*.ngs)
UV/VIS, NIR, IR, fluorescence, Raman spectra in JCAMP-DX 4.24/5.00 standard (*.dx, *.jdx)
NT-MDT Spectral Instruments: Raman spectra (*.mdt)
Perkin Elmer: UV/VIS, NIR, FTIR, fluorescence & Raman spectrometers (*.sp)
RRUFF Raman mineral database (*.rruff)

In combination with Raman spectroscopy data, images of samples can also be generated.

Processing (post-acquisition)

Raw raman spectroscopy data may be preprocessed to decrease the amount of noise and to enhance discriminating features.

The general spectrum profile created by Raman spectroscopy can be used to identify a material through comparison with Raman spectral libraries, for example the Raman Mineral Database (RRUFF). Spectrum files can be processed using software (e.g. CrystalSleuth), which compares a Raman pattern against the RRUFF project database. A calibration process can then be undertaken to determine the relationship between peak intensity and concentration, allowing measurements to be made to analyse the concentration of components.

Resources

http://rruff.info/ - RRUFF: Database of Raman spectra & X-ray diffraction data for minerals
https://www.horiba.com/en_en/raman-imaging-and-spectroscopy/
https://www.mt.com/gb/en/home/applications/L1_AutoChem_Applications/Raman-Spectroscopy.html#overviewaf
https://link.springer.com/article/10.1140/epjti/s40485-015-0018-6

9.1.1.4 Infrared Spectroscopy

Overview/Planning

Infrared spectroscopy is the analysis of infrared light interacting with a molecule and is used to characterise and identify organic and inorganic materials. The technique is applied in fields such as microarchaeology, for example to study the preservation of organic remains and human cultural behaviours such as pyrotechnology. It is perhaps the most useful technique for the study of all cultural heritage materials, as it is possible to rapidly identify both organics and inorganics (e.g. pigments + binders). It is also useful in degradation studies when a material is e.g. heated

for a specific amount of time and it is possible to see the transformations which occur in the spectra.

Infrared spectroscopy is considered a 'microdestructive' method of analysis, as miniscule quantities of an analysed substance are required during sample preparation. Samples can be ground and mixed with KBr (which is transparent to IR light) - or the sample surface can be analysed directly using the ATR (attenuate reflectance) mode.

Manufacturers of infrared spectroscopy include PerkinElmer and ThermoFisher.

Collection/Creation

The Thermo/Galactic (.spc) format (https://www.peakspectroscopy.com/fileconversion.html), The SPC file format is a file format for storing spectroscopic data, including infrared spectra, Raman spectra and UV/VIS spectra. Data may also be outputted in csv format (e.g. datasets)

The Fourier transform data-processing technique may be used to turn this raw data into the desired result (i.e. the sample's spectrum).

Processing (post-acquisition)

The pre-processing of infrared spectroscopy data is usually undertaken in order to accentuate the desired spectral features (e.g. scatter-correction methods and spectral derivatives).

Comparison with a spectra database is frequently undertaken as part of the data processing and analysis process.

Resources

https://www.sciencedirect.com/science/article/pii/S2352409X17302705
https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Spectroscopy/Vibrational_Spectroscopy/Infrared_Spectroscopy
https://arrow.dit.ie/cgi/viewcontent.cgi?article=1028&context=biophonart
Free spectral libraries: https://www.effemm2.de/spectragryph/down_databases.html

9.1.1.5 Thermal Analysis

Overview/Planning

Thermal analysis is the application of a precision controlled temperature program that allows quantification of a change in a material's properties with change in temperature. Thermal analysis has applications in the characterisation of pottery (e.g. firing temperature and/or presence of mineral phases).

Collection/Creation

Raw data may be outputted as thermal image data. It is possible to extract raw binary thermal from a thermal image file.

Processing (post-acquisition)

Melting onset temperatures may be matched to the known melting points of standards analyzed by DSC.

Resources

https://www.semanticscholar.org/paper/FT-IR-X-RAY-DIFFRACTION-AND-THERMAL-ANALYSIS-TO-THE-EMAMALINI-ELRAJ/f6cfae26add321f2ee73a6a94b1ef6f4790431b7
https://www.ncbi.nlm.nih.gov/pubmed/29436511

9.2.1 Microscopy

Microscopy and Micro-Computed Tomography Laboratories boast a wide range of equipment enabling the microstructural and elemental characterization of all types of materials (for example, biological, organic and inorganic materials), with practice-oriented analysis, quality control and basic and applied research.

Microscopy area: There are different types of microscopes, making it possible to distinguish four main areas within the field of Microscopy: i) optical, fluorescence and metallographic microscopy; ii) particle analysis; iii) confocal laser microscopy; iv) scanning electron microscopy; and iv) petrography. This wide range of scientific equipment makes it possible to study multiple parameters for different types of materials: from the analysis and automated classification of particles to quality control of electronic components.

List of file formats by the Open Microscopy Environment: https://docs.openmicroscopy.org

9.1.2.1 Optical, Fluorescence and Metallographic Microscopy

Overview/Planning

Optical microscopes use visible light and a system of lenses to magnify images of small objects. A fluorescence microscope is an optical microscope that uses fluorescence and phosphorescence instead of, or in addition to, scattering, reflection, and attenuation or absorption, to study the properties of organic or inorganic substances. Fluorescence microscopes range from relatively straight-forward wide-field microscopes to highly specialised spectral-imaging confocal microscopes. Confocal laser scanning microscopes may also be used for 3D imaging using software reconstruction.

Metallographic analysis can be used as a tool to help identify a metal or alloy, to determine how an object was manufactured, to identify the temperatures the metal may have been subject to during or after manufacture, to locate and characterize imperfections such as voids or impurities.

Collection/Creation

Structures observed in the microscope are often recorded photographically and by observing and noting characteristics such as shape, size, color and distribution of surface features, pores, grains (metallurgy) or inclusions (e.g. pottery).

Images from optical microscopes can be captured digitally as micrographs. Micrographs (or photomicrographs) are widely used in microscopy and information that may be obtained from a micrograph includes: the behavior of material under different conditions, the phases found in the system, failure analysis, grain size estimation and elemental analysis.

Raw microscopy data can be outputted in formats such as:

Raster images (tiff, mrc, cryoEM (industry standard in cryo-electron microscopy and electron tomography (ET). CCP-EM.)

Digital Micrograph file format. an image processing program produced commercially by Gatan (DMF2, DMF3, DMF4 - info).

Metallographic microscopy samples may be examined in 'as-polished' or 'as etched' condition (destructive sampling). It may be beneficial to photograph an object in the unetched/unpolished condition prior to sampling.

Processing (post-acquisition)

For destructive sampling microscopy techniques (metallographic microscopy), it may be beneficial to photograph / record an object prior to sampling (e.g. in an unetched condition).

The pixels in an image can be averaged or `binned' in order to increase the signal-to-noise ratio.

Long-term Curation

Ideally, all photographs / micrographs should be supplied with a scale bar

Examples of archived micrograph data can be found in the Electron Microscopy Public Image Archive (EMPIAR): https://www.ebi.ac.uk/pdbe/emdb/empiar/ (with experimental metadata xml). (although n.b. This is for Electron Microscopy).

The Open Microscopy Environment (OME): a consortium of universities, research labs, industry and developers producing open-source software and format standards for microscopy data. The OME data model for biological imaging: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2005-6-5-r47

A suggested data-standard has been proposed for fluorescence microscopy images, with the aim of increasing data fidelity, ease future analysis and facilitate objective comparison of different datasets, experimental setups, and essays: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10726/2323660/Introducing-a-data-standard-for-fluorescence-microscopy--increasing-data/10.1117/12.2323660.short?SSO=1

Resources
https://www.microscopyu.com/techniques/fluorescence/introduction-to-fluorescence-microscopy
https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10726/2323660/Introducing-a-data-standard-for-fluorescence-microscopy--increasing-data/10.1117/12.2323660.full
https://ocw.mit.edu/courses/materials-science-and-engineering/3-094-materials-in-human-experience-spring-2004/laboratories/manual_suppl.pdf
https://jcs.biologists.org/content/120/10/1703
EMPIAR deposition manual: https://www.ebi.ac.uk/pdbe/emdb/empiar/deposition/manual/

9.1.2.2 Particle Analysis

Overview/Planning
Particle analysis is used to produce quantitative data on the size and distribution of particles in a sample of solid materials, suspensions, emulsions or aerosols. One of the most common methods for analysing particle size is laser diffraction, with other methods including sieving and dynamic light scattering.

Image analysis during particle diffraction provides more data values and options. Measuring each particle allows for calculating and reporting of particle size results.

Applications in environmental archaeology, for example sediment analysis. The particle size distribution of a soil sample for example, may indicate the conditions under which the strata or sediment were deposited.

Collection/Creation
Particle analysis data may be presented numerically as a size class percentages of the total weight of the sample. Data obtained from laser diffraction data may be presented as % of the sample volume. Results may also be graphed as a percentage curve or cumulative percentage curve.

Processing (post-acquisition)
The data output will consist of a statistical distribution of particles of different sizes. It is common practice to represent this distribution in the form of either a frequency distribution curve, or a cumulative (undersize) distribution curve.
Long-term Curation

Resources
https://www.atascientific.com.au/basic-principles-of-particle-size-analysis/
https://historicengland.org.uk/images-books/publications/environmental-archaeology-2nd/environmental_archaeology/
https://www.cif.iastate.edu/sites/default/files/uploads/Other_Inst/Particle%20Size/Particle%20Characterization%20Guide.pdf

9.1.2.3 Confocal Laser Microscopy (CLSM)

Overview/Planning

Confocal laser scanning microscopy (CLSM) is a technique for obtaining high-resolution optical images with depth selectivity and allows protein localisation in specific cellular compartments. The technique offers several advantages over optical microscopy, such as the ability to remove light or glare, reduce background information and to collect serial optical sections from thick specimens.

Collection/Creation

Software packages that accompany commercial confocal instruments can produce composite and multi-dimensional views of optical section data acquired from z-series image stacks. This can be used to create either a single 3D representation of a sample or a video sequence compiled from multiple views of the specimen.

Processing (post-acquisition)

3D data
Information on the processing and analysis of data:
https://www.ncbi.nlm.nih.gov/pubmed/24974025

Resources

https://www.olympus-lifescience.com/en/microscope-resource/primer/techniques/confocal/confocalintro/
https://www.researchgate.net/publication/267406328_Laser_scanning_confocal_microscopy_application_to_stone_tool_function_and_other_archaeological_problems
https://www.sciencedirect.com/science/article/pii/S0305440309002234
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4878693/
https://cmrf.research.uiowa.edu/confocal-microscopy
https://onlinelibrary.wiley.com/doi/full/10.1111/jmi.12424
https://www.ncbi.nlm.nih.gov/pubmed/24974025

9.1.2.4 Scanning Electron Microscopy (SEM)

Overview/Planning

Scanning Electron Microscopy (SEM) uses focused beams of electrons to render high resolution, three-dimensional images. These images provide information on a sample's topography, morphology, and composition when the source is EDX.

Collection/Creation

SEM data is usually collected over a selected area of the surface of the sample, and a 2-dimensional image is generated that displays spatial variations in these properties.

Data may be outputted as raster images and stack formats. File formats for SEM data include:

The Digital Micrograph format (Gatan) (.dm4)
.mrc (mrc-stacks)
.rec

.st
EDX spectra (elemental composition) with SEM-EDX

Processing (post-acquisition)
3D scanning electron microscopy

After data collection there are several preprocessing steps that can be undertaken to extract and analyse features of interest. This often involves the transformation of the file format from a system/software-specific format to a more widely compatible file format or a stack format. (e.g. conversion from .dm4 to .mrc).

Resources
https://www.atascientific.com.au/sem-imaging-applications-practical-uses-scanning-electron-microscopes/
https://serc.carleton.edu/research_education/geochemsheets/techniques/SEM.html
https://bitesizebio.com/32766/data-analysis-for-three-dimensional-volume-sem/
https://onlinelibrary.wiley.com/doi/abs/10.1111/jmi

9.2 Dating Methods

9.2.1 Potassium Argon

Overview/Planning
Potassium (40K) decays through branched decay to 40Ar. Although the half-life is 1.25 billion years, this technique is one of the most widely used chronometers for archeology because with mass spectrometry it is possible to measure very small changes precisely. Typically the potassium (K) is not directly measured. Instead, unknowns, along with monitor standards with known K concentrations and ages, are sent to a nuclear reactor where 39Ar is created by neutron bombardment. The ratio 40Ar/39Ar can then be measured and a factor called the J value is applied based on the 39Ar produced in the monitor standard. Volcanic ash deposits often have minerals such as K feldspar (sanidine) that provide reliable ages of ash deposition. The intended audience is anyone who cares about the ages of sedimentary rocks, fossils, or artifacts. The data will continue to be used unless another measurement shows that it is flawed. Minerals have to be separated from rocks. Then they are cleaned and packaged to be sent to a neutron reactor. When they are returned to the lab, they can be measured in a mass spectrometer in one of two ways. They could be put in a furnace and step heated if they are large enough to produce a series of steps. When there are several steps that give the same ratio (age), this is a plateau, and it is taken as the age of the mineral. For individual grains which are typically too small for multiple heating steps, they can be fused by a CO2 laser and analyzed as one step. In either case, the gases have to be cleaned of other gases through an extraction line where getters of various types capture the gases that are not wanted, such as CO2, O2, and H2O. After the gases are cleaned in the extraction line, they are let into the mass spectrometer and ratios are measured from time zero until the gas is used up. The choice of time zero is important and has been shown to bias

results between labs. It is important to monitor for blanks which limit the precision in the data. It is also important to measure standards to determine signal intensity per volume as well as details of mass fractionation in the mass spectrometer. Typically multiple aliquots of each sample are run and data are presented in probability plots. With the reported age being the mean and the uncertainty the dispersion in the ages.

Collection/Creation

Raw data are signal intensities for each isotope, isotope ratios, and baseline measurements. Monitor standards are irradiated with samples and are run in the same batches as samples. These measurements not only test the reliability of the mass spectrometer, but also provide a means to calculate the J value for the unknowns. There are a number of widely used monitor standards, such as Fish Canyon, Alder Creek. There are several companies that make noble gas mass spectrometers (Thermo Fisher, IsotopX, and Nu Instruments) and there are numerous vintage machines that are still operational. Typically the instruments do not come with software for data collection and reduction. Many nobel gas mass spectrometers are run with in-house software routines. There are several programs that are available for a cost including Mass Spec (by Al Deino, Berkeley Geochronology Center) and Pychron (by Jake Ross, New Mexico Bureau of Geology). Earthtime, an international effort to improve precision in geochronology and to address discrepancies between geochronometers (particularly Ar/Ar and U/Pb), was started in 2003. This effort revealed big differences between labs that were likely a result of different protocols. While each lab could report 0.1% uncertainties (one sigma), there were several percent differences between labs. A tremendous effort on the part of the Ar community has largely closed these gaps, but there is still an open discussion as to best practices. It is important to report details of data collection and data reduction, including which monitor standards were used and what criteria were used to weed out data. Another relatively recent development is that instead of peak hopping in a secondary electron multiplier (SEM), mass spectrometers are being fitted with multiple ion counters and Faraday cups so that the ion beams for each isotope can be collected simultaneously. Thus another parameter that should be noted is types of collection.

9.3 Biomolecular Methods

9.3.1 Palaeoproteomics: Zooarchaeology by Mass Spectroscopy (ZooMS; proteomics)

Zooarchaeology by Mass Spectrometry

ZooMS is a minimally destructive method to determine animal identification using peptide mass fingerprinting of collagenous materials. Collagen is made up of amino acids, and differences in the sequence of these amino acids result in genus and sometimes species-specific peptide profiles. ZooMS works by using an enzyme, most commonly trypsin, to cut the protein at specific places (lysine and arginine for trypsin) creating peptide fragments that are then comparable to reference peptide mass fingerprints.

In archaeological or heritage applications, the data is typically being created to identify the animal origin of artefacts made of collagenous materials, such as bone, antler, ivory, teeth, and skin

(parchment and leather). If the object being analysed was found in an archaeological context, the ZooMS results are often used in combination with the zooarchaeological analysis of other animal bones found on site. The limitations to the approach are primarily based on being able to destructively sample the material, although buffer-solution based methods have been used. In the case of parchment the triboelectric effect is exploited, typically using PVC erasers (the so called dry cleaning method used in conservation studios). These non-destructive methods have only been consistently applied to the identification of skin (leather, parchment). In the case of bone and antler it has also proved possible to recover signals from biomolecules adhered to the plastic bags in which samples are stored or from the tubes used to prepare radiocarbon.

The other limitation is whether there are proteins preserved in the sample, as the technique will not work if the artefact has undergone significant loss and degradation of the organic component. Collagen is preserved for much longer and in many more variable burial environments than ancient DNA, meaning that it is possible to extract collagen from much older and a wider range of samples. The method is being developed further in Denmark in the following directions:-

using materials with greater triboelectric potential which can generate charge without the need to stroke the surface, adding an MS/MS component of target ions using MRM, enhanced ionisation methods.

The intended audience is typically the archaeological and heritage community who are interested in what the identification means for the archaeological site, or for larger patterns of trade and human interaction with animals in the past. There is an increasing demand for provision of commercial analyses from units, which is being provided on an ad hoc basis. Thus, the outputs are primarily in academic publications and at conferences, but there is currently no global database of all of the ZooMS samples which are being sampled and identified.

There is a community requirement to develop a shared database of collagen sequence and target masses. This is currently being developed and will be hosted on the @palaeome.org platform.

Common software tools. The most widely used tool mMass was last developed in 2014, and some of the packages it uses are no longer supported.

There are different ways of disseminating raw ZooMS data, and at present, there is no existing requirement to share or preserve the data, as each laboratory group has their own repository data. There are few examples of raw data being made accessible. Indeed attempts to get raw data submitted to ProteomeXchange (a standard way for submitting mass spectrometry-based proteomics data to public-domain repositories) have been repulsed as ProteomeXchange only supports submission of MS/MS (MS2) data, while ZooMS is MS1 only.

In terms of publication, each journal has specific rules about where raw data is stored, and therefore, it is very dependent on the journal where the data is published. Examples of data outputs include graphs of the mass spectra with the corresponding masses of the peptides, raw text files with lists of the peptide masses, as well as raw text files with lists of peaks which were used to identify each sample.

Proteomics

Palaeoproteomics differs from ZooMS in that instead of generating fingerprints that can be matched to particular species it generates protein sequence information. Using tandem mass spectrometry, it generates data which enables the identification of sequences of amino acids, giving creating specificity to protein identification. Protein sequence information is used to gain evolutionary insight into extinct and extant animals and humans, insights into dietary consumption, health and disease.

Like ZooMS, the intended audience are those working within archaeology, conservation biology, bioanthropology and the heritage sector. Currently, much proteomic analysis is prohibitively expensive to be utilized in the commercial archaeological sector.

It is vital that proteomics data are distributed to these stakeholders, as well as to other researchers. This is vital during the review process to assess the validity of the data interpretation.

While there are numerous protein and proteomics laboratories around the world, only a handful specialise in the analysis of ancient material. This is due to the recent development of the technique, the need for strict contamination controls - meaning that modern and ancient work must be kept separate - and the prohibitive cost of LC-MS/MS instrumentation. Often, laboratories will send ancient protein extracts to core facilities who specialize in the running of LC-MS/MS instrumentation.

Collection/Creation

Zooarchaeology by Mass Spectrometry

Matrix-assisted laser desorption/ionisation mass spectrometry (MALDI-ToF) - this machine transforms biomolecules in the sample into ions, and then the time of flight analyser separates these ions. So, the resulting raw data file is the conversion of these ion counts from the MALDI into masses based on time of flight. Each mass spectrometer manufacturer has software (such as Bruker MALDI-ToF FlexAnalysis), which can batch process a run and analyse the raw data output. The raw data (in an *fid file format) needs to be analysed in a program which does a range of things, including picking/identifying the masses of the peaks in the spectra and is used to visualise the overall shape and intensity of the peaks and also calibrate the spectra. Two programs are primarily used: an open source program called m-Mass which takes text files and then processes those as *msd files, or a software program provided by the manufacturer of the MALDI (e.g. Bruker FlexAnalysis) saves processed files in a *lrd file format.

The raw data can also be converted (using an R code) into a text file so that it is more user-friendly across platforms. This text file is a list of the masses from smallest to largest. Each sample is run in triplicate, so there are 3 separate files associated with each sample. In the program used to pick the masses of the peaks and to analyse the spectra, the user typically analyses each replicate individually before averaging the 3 files together. Therefore, in this case, a new file is created with the averaged spectra file.

There is also a plate map file that has to accompany each run, which is either in PDF or JPEG format, depending on how it has been scanned (as it is typically paper in the lab). The samples are spotted onto a 384-well plate and this plate is what is physically put into the MALDI for analysis, as the laser shoots each well (or spot) on the plate. Therefore, the plate map is a key of which sample has been spotted onto each well on the plate.

Proteomics

Raw proteomics data are comprised of spectra, consisting of detected masses and their intensity. Some raw file formats are proprietary, such as Thermo .raw files. Typically, conversion software picks masses to create peak lists, which are analyzed further using a range of tools.

Often, negative and positive controls are run alongside samples to check for contamination from the laboratory environment (in the case of the former) and the efficacy of the extraction procedure (in the case of the latter).

Processing (post-acquisition)

Zooarchaeology by Mass Spectrometry

The final, processed, data are typically in the file format based on the program used to pick the masses of the peaks in the spectra, so either in mMass or FlexAnalysis. The text files, however, are the easiest files to make accessible/available in the supplementary information of journal articles or for online databases. The data consists of 1008 text files, representing analogue to digital conversion of ion counts detected by Matrix Assisted Laser Desorption Ionisation Time of Flight Mass Spectrometry converted into mass based upon time of flight.

Proteomics

Numerous tools exist for the analysis of proteomic data. Often, these involve a database-searching approach, whereby generated results are compared with reference databases. Selection of data analysis strategies and reference databases is highly linked to the research question and nature of the generated data. Results files may be a list of identified sequences in a fasta format or xml format.

Typically, three forms of proteomics data are shared using existing platforms; raw files, peak-picking files and results files. Several platforms exist for the sharing of these file types, including the ProteomeXchange and Massive.

Recently, standards for data generation and data archiving have been outlined in *A guide to ancient protein studies* (Hendy *et al.*, 2018).

Long-term Curation

Zooarchaeology by Mass Spectrometry

https://archaeologydataservice.ac.uk/archives/view/zooms_eu_2015/index.cfm

Good example of good practice here: text files of the parchment paper linked onto ADS

Proteomics

It is vital that proteomic data from studies focussing on archaeological material are archived. This is useful because it means that data can be reanalysed using alternative data analysis pipelines and expanded databases as they are developed, and it means that other users can replicate data analysis strategies in their own research. Given that this data analysis strategy is specialised and files often large in size, it is vital that metadata, methodologies and data interpretations are also effectively communicated alongside primary data distribution.

Resources

https://www.spectroscopyeurope.com/article/zooms-collagen-barcode-and-fingerprints
https://link.springer.com/referenceworkentry/10.1007/978-1-4419-0465-2_2418
https://www.tandfonline.com/doi/full/10.1080/05704920902717872
Deamidation: https://github.com/franticspider/q2e

9.3.2 Isotope ratio mass spectroscopy (IRMS)

Overview / Planning

Isotope Ratio Mass Spectroscopy (IRMS) is used to measure the relative abundance of a specific element's isotopes in analysed materials. IRMS uses the differences in mass between isotopes of an element as part of investigating the geological, biological and environmental processes that may alter isotope relative abundances. These processes may favor lighter over heavier isotopes, or vice versa, resulting in isotope ratio differences between a starting and end process material. This is known as isotope fractionation, which for stable isotope analysis typically involves unidirectional kinetic fractionation.

Isotope data in the raw form is given as isotope ratios, heavy to light isotopes, which are then inputted into the delta formula and presented with a delta notation in professional publications. The delta formula is: □X ‰ = [(RX – Rstandard)/Rstandard] x 1000, where the □ symbol expresses the ratio of heavy to light isotopes "R" of a sample "X" relative to the ratio of heavy to light isotopes "R" in a standard. The delta notation is in parts per thousand or per mil, which is denoted as ‰. Delta values, per mil, put a numerical value to the enrichment or depletion of a sample relative to a reference standard of known isotopic composition.

It is expected that isotope ratio data will be of interest to a wide range of researchers from multiple fields, including archaeology, environmental studies, geology and exogeology, atmospheric sciences, biology, palaeontology, among others. There is significant interest in a wide range of isotope systems (carbon, nitrogen, oxygen, strontium, lead, calcium, boron, neodymium, sulphur, iron, uranium, etc.) that commonly overlap between disciplines, which makes data useful beyond any one field. A central archive dedicated to isotope data would be extremely helpful and allow researchers to browse in one location instead of having to search based on publication source.

There are different instrument approaches for IRMS. Gas evolution is suitable for samples that release H2, CO2, CO, N2 and SO2 when dissolved in acid. TIMS ionises samples through an electrical current applied to a filament loaded with a sample. MC-ICPMS uses a plasma source to ionise samples, as does SIMS.

Gas source mass spectrometry (light isotope systems that work using gas evolution methods: hydrogen, carbon, nitrogen, oxygen and sulphur). Instrument manufacturers include: Thermo Scientific, Sercon and Nu Instruments. Raw data is typically available as .csv, which is frequently imported into a software package like Excel for further processing.

Thermal ionization mass spectrometry (TIMS) (heavier isotope systems and ones that do not work with gas evolution methods: strontium, calcium, etc.). Instrument manufacturers: Isotopx, Thermo Scientific and Nu Instruments. Raw data is typically available as .xls, which is frequently imported into a software package like Excel for further processing.

Multiple collector inductively coupled plasma mass spectrometry (MC-ICPMS) (heavier isotope systems and ones that do not work with gas evolution methods: boron, strontium, calcium, lead, neodymium, iron, uranium, etc.). Instrument manufacturers: ThermoFinnigan, Nu Instruments. Raw data is typically available as .csv and .txt, which is frequently imported into a software package like Excel for further processing.

9.4 Synchrotron Methods

There is overlap between some synchrotron methods and laboratory material characterisation methods, but synchrotron radiation techniques are notable due to the high brilliance of the beamline compared to X-rays generated by conventional X-ray tubes, high levels of polarization, exceptional signal-to-noise ratio, wide tunability in energy/wavelength through monochromatisation (synchrotron X-rays cover energy ranges of <1 keV [soft X-rays;] and >5 keV [hard X-rays]; a few synchrotrons are also able analyse in the “Tender” energy range [1-5 keV; elements Na through Ca]), pulse light emission durations at or below one nanosecond, etc. The result for researchers is, scanning at µm scale, the detection of trace quantities of chemical components and high resolution spatial mapping of both the sample chemical content and chemical distribution in/on the sample. Archaeometric appropriate spectroscopy techniques include:

X-ray diffraction (XRD; crystallinity and amorphous structures);

- X-ray absorption spectroscopy (XAS; identifies an element in a solid sample as well as the coordination structure of atoms with the analysed modulations producing;
- X-ray absorption near-edge structures (XANES);
- X-ray fluorescence (XRF; chemical element identification and quantification).

An increasing interest in using synchrotron radiation techniques to investigate archaeological and cultural heritage materials/objects will likely result in experimenting with different synchrotron X-ray energies and spectroscopic techniques in the future. Synchrotron radiation also offers unparalleled opportunities for sample imaging (µm scale), including tomography and X-ray imaging. XAS, XRD and XRF can be combined with imaging to map a sample’s chemical composition and chemical distribution. Europe has 15 synchrotron radiation facilities, with the addition of Schengen Agreement partners, both EU and non-EU, the total is 18. The United States

of America, a frequent source for research collaboration, has 15 synchrotron energy based research facilities.

9.4.1 X-ray absorption spectroscopy (XAS)

Overview / Planning

X-ray Absorption Spectroscopy (XAS) is an element-specific technique used to determine oxidation states, chemical speciation, and local molecular structure of major and trace elements. Applications include identification of materials and components (crystalline or non-crystalline), their chemical makeup, and chemical changes over time. Samples may be powders, cut sections, or chips/small pieces. For many samples no preparation is required and measurement is non-destructive. Analysis operates on the mm to um scale. Available almost exclusively at synchrotron facilities.

Collection/Creation

Raw data is usually a simple intensity spectrum as a function of photon energy. Settings typically predetermined for the specific instrument, with variable parameters recorded with data. Raw data usually in plain text format, and archived at the facility.

Processing (post-acquisition)

Raw data typically processed using a variety of industry-standard mathematical packages, including peak fitting, fourier transform, etc., and compared to reference standards and/or previously published data, to determine the targeted chemical information. These comparisons and variously processed data are typically published.

Long-term Curation

Raw data archived according to facility policies, which vary. Ideally, raw and processed data typically archived with any publication.

9.4.2 X-ray Crystalline Powder Diffraction (XRD and XRPD)

Overview/Planning

X-ray diffraction (XRD) and X-ray powder diffraction (XRPD) is used for phase identification of crystalline materials and to provide information on unit cell dimensions. Applications include identification of materials and components, their manufacture or origins, and any changes due to time or other influences. Samples may be powdered material, chips, cut sections, or intact objects (of limited size). For many samples no preparation is required and measurement is non-destructive. The non-destructive technique allows for the rapid (< 20 min) identification of an unknown crystalline material with minimal sample preparation required. Analysis may operate on the mm to um scale. Available in the laboratory (more limited) or at synchrotron facilities (more advanced capabilities).

XRPD has traditionally been used in fields such as geology, environmental science, material science, and engineering to identify unknown substances and in more recent years, portable X-ray diffractometers have enabled the rapid analysis of materials in the field.

Applications of XRPD in heritage science include the characterisation of crystalline materials, the identification of fine-grained minerals such as clays and mixed layer clays and the measurement of sample purity. In archaeology for example, XRPD can be used to identify pigments or to assess the polymorph of a carbonate, including those of biogenic origin such as mollusc shells prior to radiocarbon dating or stable isotopes measurements.

X-ray diffractometers are widely available and data interpretation is relatively straightforward. Manufacturers of X-ray diffractometers include Olympus, Bruker, Malvern Panalytical, Rigaku, ThermoFisher and Seifert. This technique is also available using synchrotron energy.

Collection/Creation

X-ray crystalline powder diffraction is undertaken using an X-ray diffractometer, with machines consisting of three components: an X-ray tube (generator), a sample holder and an X-ray detector. The analyzed material is finely ground, homogenized, and average bulk composition is determined. It is also possible to take a series of XRD spectra along a line so you have the dimension of space to take into account in a dataset (this also applies in XRF, FTIR, Raman, and most spectroscopies). Raw data usually 2D images or 1D intensity as a function of angle at a given wavelength. Settings typically predetermined for the specific instrument, with variable parameters recorded with data. Industry-standard formats.

The diffractometer processes and converts the x-ray signal to a count rate, with data being saved in the instrument manufacturers 'raw' state as a binary file.

The outputted data type depends largely on the diffractometer manufacturer (e.g. Siemens, Philips, Rigaku, Stoe, Seifert) and diffraction data file formats of different diffractometer manufacturers include:

Bruker/Siemens raw data (*.raw)
Bruker/Siemens DIFFRAC AT peak (dif) data (*.dif)
DRON-3
G670 raw data (*.gdf)
GNR raw data (formerly Ital Structures raw data) (*.esg)
INEL raw data (*.dat)
Jade/MDI/SCINTAG raw data (*.mdi)
JEOL ASCII Export raw data (*.txt)
PANalytical/Philips raw data (*.rd)
PANalytical/Philips raw data (*.udf)
PANalytical/Philips peak data (*.udi)
Rigaku raw data (*.raw)
SCINTAG raw data (*.raw, *.rd)
Seifert
Shimadzu raw data (*.raw)

Siemens raw data (*.uxd)
Sietronics XRD scan data (*.cpi)
Stoe Raw data (*.raw)
Stoe Peak File (*.pks)
TXRD text export (*.txt)
XPowder raw data (*.plv)
XRDML Scan raw data (*.xrdml

There are around 70 file formats for spectral data, with some being used for multiple methods of spectrometry (e.g. the MAS/ EMSA standard file format for XRF/EDS spectra after ISO 22029 (*.msa)) (Spectragryph).

Data conversion software allows the raw data to be saved as a text (.xye?) or excel file (xml) and there are a number of 'free' XRD formats available, such as: DBWS Raw data (.rfl, .dat) and ASCII profile (.dat, .dif, .pro).

Processing (post-acquisition)

The output from XRD experiments is typically presented in the form of an x-y plot such as a diffractogram (a plot of diffracted X-Ray intensities versus the scanning angle 2θ as abscissa), the quality of which is dependent on both sample preparation and machine settings. Quantitative Phase Analysis of XRD data may be undertaken using methods such as the rietveld method. Additionally, Analysis of XRD data requires access to a standard reference file of inorganic compounds. A database of powder diffraction patterns is maintained by the International Centre for Diffraction Data (ICDD) and the Powder Diffraction File (PDF) may be used to identify substances based on the results of X-ray diffraction. Raw 2D data sets are combined and condensed into 1D intensity as a function of d-spacing, which can be matched to database entries for known phases. 1D data is usually reported and archived.

Long-term Curation

Synchrotron data is 1D, typically archived with any publication.

Resources

https://serc.carleton.edu/research_education/geochemsheets/techniques/XRD.html
https://www.iucr.org/resources/data/meeting-reports/metadata-workshop (workshop on metadata for raw data from X-ray diffraction)
http://pd.chem.ucl.ac.uk/pd/welcome.htm
https://myscope.training/legacy/xrd/introduction/
https://www.crystalimpact.com/match/help/idh_import_diffraction-data_file-formats.htm (list of diffraction data file formats)
https://www.iucr.org/resources/other-directories/software?result_42405_result_page=D (crystallographic software list)
https://www.geo.arizona.edu/xtal/geos306/fall11-11.htm
https://openei.org/wiki/Portable_X-Ray_Diffraction_(XRD)
https://www.researchgate.net/figure/XRD-data-for-an-iridium-powder_fig2_317584101 (example of XRD data for an iridium powder presented in a report)

9.4.3 X-ray absorption near-edge structures (XANES)

9.4.3.1 X-ray Fluorescence Spectroscopy (XRF)

Overview/Planning

X-ray Fluorescence (XRF) imaging is used to determine distributions, co-localisation, and relative concentration of major and trace elements. Applications include identification of materials and their constituents, their physical structure, and fingerprinting to indicate provenance. Samples may be cut into sections, chips, or left as small intact objects. For most samples no preparation is required and measurement is non-destructive. Analysis may operate on the mm to nm scale. Available in the laboratory (bulk or very limited) or predominantly at synchrotron facilities. XRF is a well-established method of analysis both in laboratory and fieldwork settings and is one of the most widely used methods of analysing geological materials due to the low cost of sample preparation and ease of use. Portable XRF (pXRF) devices are increasingly being used in fields such as archaeology, with handheld devices making it possible to collect XRF data in the field. On the opposite scale, electron particle accelerators generating synchrotron radiation, which can be converted to X-rays of various energies, also provide a means of analysing archaeological and cultural heritage samples.

Manufacturers of XRF instruments include Bruker, Olympus, ThermoFisher and MalvernPanalytical.

Examples of synchrotron facilities available for research: PUMA beamline, SOLEIL Synchrotron, France, Diamond Light Source, UK and National Synchrotron Light Source II, Brookhaven National Laboratory, USA.

Collection/Creation

Laboratory instrument output of XRF experiments is measured as the number of counts of element-specific fluorescent X-ray energies received in an XRF instrument detector. This indicates which elements are present in a sample and raw data can be numerical and graphical, for example a spectrum graph. Synchrotron XRF output is usually 2D images with fluorescence spectrum (or array of element-specific intensities) at each pixel, or 3D computed microtomography. Settings typically predetermined for the specific instrument/beamline, with variable parameters recorded with data. Synchrotron facilities typically run using extremely tight security data protocols. Raw data is typically in native format, which is not readily transported or usable off-site, with archiving at the facility.

File formats for XRF laboratory instrument data include:

MAS/ EMSA standard file format for XRF/EDS spectra after ISO 22029 (*.msa)
Thermo Noran: WinTrace XRF spectra (*.spc)

Processing (post-acquisition)

In order to undertake quantitative analysis of XRF data (i.e. to determine the absolute quantity of an element present in a sample), calibrations are required to convert raw qualitative data into quantitative data (e.g. fundamental parameters and empirical calibrations). This can be undertaken using software available both from the instrument used and by using free / open-

source XRF data analysis software. Calibrations are created using samples with known concentrations of elements to create a calibration curve that relates the specific known concentrations to peak heights. Synchrotron XRF, raw data is typically processed into 2D images for each chemical element analysed, which can, in some cases, be manually processed into quantitative concentrations. These images are usually reported and archived.

Long-term Curation

Spectrum graphs

Example of published / archived XRF data (Migration of XRF Data to Janus):
http://www-odp.tamu.edu/publications/tnotes/tn37/tn37_20.htm

Resources

https://serc.carleton.edu/research_education/geochemsheets/techniques/XRF.html
https://www.horiba.com/en_en/x-ray-fluorescence-spectroscopy-xrf/
https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-0465-2_1305
https://www.bruker.com/products/x-ray-diffraction-and-elemental-analysis/handheld-xrf/how-xrf-works.html
https://www.bruker.com/products/x-ray-diffraction-and-elemental-analysis/handheld-xrf/archaeometry.html
https://www.bruker.com/products/x-ray-diffraction-and-elemental-analysis/handheld-xrf/xrf-data-primer-quantitative-semi-quantitative-qualitative.html
https://www.rigaku.com/downloads/journal/Vol6.1.1989/latour.pdf (n.b. 1989)
https://www.thermofisher.com/blog/mining/better-together-xrf-and-xrd/
https://www.researchgate.net/publication/316219806_Combining_XRD_and_XRF_analysis_in_one_Rietveld-like_fitting
https://www.sciencedirect.com/science/article/pii/S0305440314004440
http://elementsmagazine.org/archives/e9_1/e9_1_dep_theelementstoolkit.pdf
Archiving spectroscopic data: https://www.beilstein-institut.de/download/...