Data-Driven Community Standards for Interdisciplinary Heterogeneous Information Networks

Mattis thor Straten [1,*], Steffen Strohm [1], Florian Thiery [2], Matthias Renz [1]

[1] Department of Computer Science, Kiel University;
[2] Leibniz-Zentrum für Archäologie & Research Squirrel Engineers;
[*] Corresponding author: mts@cs.uni-kiel

Creating a unified, interoperable representation by integrating diverse datasets is a key challenge in interdisciplinary research. Heterogeneous information networks (HINs) offer a graph-based approach to linking datasets while preserving their semantic structure. This study examines data-driven community standards, with a particular focus on ontologies, to ensure semantic interoperability in cultural heritage HINs.

Using the CIDOC Conceptual Reference Model (CIDOC-CRM) alongside other commonly used or domain-specific ontologies, this study develops the hybrid ArNO ontology for archaeo-natural research data based on the experience gained in NFDI4Objects Task Area 3. The data integration follows Linked Open Data principles to ensure schema-level consistency through ontology-based modelling, while enforcing data-level consistency through terminologies. This work contributes to the development of FAIR-compliant research infrastructures by establishing standards at schema and data level, thereby enhancing the reuse of data across the humanities and natural sciences.

This approach is demonstrated by integrating ancient DNA datasets from the Poseidon framework to enable cross-dataset analysis by linking natural, archaeological, and contextual data. The study also addresses challenges in CIDOC-CRM-based modelling, such as its event-centric nature.

Keywords: Data Integration, Ontologies, Linked Open Data, Heterogeneous Information Networks, Knowledge Graphs

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23914 (CC BY-SA 4.0).

1 Introduction

Interdisciplinary research requires disparate datasets to be integrated into a unified representation, enabling holistic, cross-dataset analyses to uncover hidden patterns and dependencies. This provides a foundation of combined data on which network analysis methods can be applied to derive new insights and solve problems – the second generation of semantic cultural heritage (CH) systems (Hyvönen 2019). Heterogeneous information networks (HINs) provide a flexible, graph-based framework for integrating diverse datasets (e.g. Patel, Paraskevopoulos, and Renz 2018a). They are particularly effective at preserving both the complex relationships of the data and the contextual information in an intuitively understandable graph structure consisting of typed nodes and edges.

Interdisciplinary initiatives related to research data management, such as the German National Research Data Infrastructure (NFDI, see Hartl, Wössner, and Sure-Vetter 2021), are creating knowledge graphs using ontologies, reference models and standards such as the Resource Description Framework to promote FAIR (Wilkinson et al. 2016) research data management and the NFDI vision of applying open science principles. Several domain-focused consortia within the NFDI are working on the interdisciplinary task of FAIRifying research data related to the material remains of human history. All resulting domain-specific knowledge graphs must be interoperable using the Base4NFDI Basic Services[1], such as TS4NFDI and KGI4NFDI. The NFDI4Objects consortium (N4O, see Rummel, Keller, and Fricke 2025; Thiery et al. 2023) addresses these challenges by fostering interdisciplinary collaboration across the humanities, cultural and natural sciences. The initiative covers data spanning approximately 2.6 million years and encompasses various types of objects, including human-made artifacts, biological specimens, and geological specimens. These objects evolve through various transformations, resulting in a so-called object biography, with Task Area 3 "Analytics and Experiments" (TA3) taking the lead in natural science research data. A key challenge in this process is the interdisciplinarity within the CH and archaeological domains (e.g., Strohm, Buenning, and Renz 2023), and the challenges of joint semantic modelling with the humanities, natural and life sciences, and geosciences.

This work presents the Archaeo-Natural Ontology (ArNO), a hybrid ontology for archaeo-natural research data that is developed within TA3. Archaeo-natural data refers to research data produced at the intersection of archaeology and the natural sciences, including anthropology, archaeogenetics, archaeozoology and archaeobotany, among others. The term highlights the necessity of integrative modelling approaches that can represent both cultural-historical context and scientific measurements within a unified semantic framework. ArNO is an application ontology that incorporates classes and properties from multiple foundational (Steller et al. 2025, p. 11) and domain ontologies (Guarino 1998, p. 8). It provides a self-contained schema that is suitable for representing interdisciplinary archaeo-natural research data in an HIN. It is also designed to integrate with existing in-

[1] https://base4nfdi.de/projects; *Visited on May 20, 2025.*

frastructures modelled using the contained ontologies, such as the CIDOC-CRM-based NFDI4Objects Knowledge Graph (e.g., Voß et al. 2024; Voß and Heers 2024).

The aim of this work is to integrate anthropological ancient DNA (aDNA) data from the Poseidon framework (Schmid et al. 2024) into a generalisable format that allows it to be combined with other archaeo-natural and CH data. Therefore, the data representation must transcend the boundaries of the dataset, relating the information it contains to a wider context by matching different instances of a common object (Doerr 2005), such as spatial and temporal information or involved agents.

Using HINs for data integration enables cross-dataset analysis (e.g., Patel, Paraskevopoulos, and Renz 2018b) through common concepts, such as spatial and temporal information, which act as connectors between objects of different origins. Incorrect data integration results in invalid HINs, which may consequently lead to incorrect analysis results. This creates the need for standards at schema and data level to ensure semantically correct data integration. The proposed interoperable ontology provides these standards at the schema level. At the data level, terminologies – from keywords to enriched thesauri – are employed (Thiery and Engel 2016, pp. 259-261).

2 Preliminaries

2.1 Data Integration

In the context of this work, data integration refers to the fusion of datasets at the record level. This involves synchronising different representational instances of the same object in different datasets to create a consistent, interconnected data foundation under a mediated schema. The main challenges are finding corresponding schema elements (Legler and Naumann 2007) and equivalent object descriptions (duplicate detection) in different datasets, in order to create a single, consistent representation of the integrated data from both datasets (Bleiholder and Naumann 2008, p. 2).

2.2 Information Networks

An information network is a directed graph $G = (V, E, \phi, \psi)$, where the set of nodes V represents information objects – also known as entities – and the set of edges E represents relationships between pairs of these information objects – also known as relationships or properties. The functions ϕ and ψ map nodes and edges to their corresponding types, i.e., $\phi : V \longrightarrow \mathcal{A}$ and $\psi : E \longrightarrow \mathcal{R}$, where the set $\mathcal{A}$ contains all object types and the set $\mathcal{R}$ contains all relationship types. When more than one node or edge type is considered, i.e., $|\mathcal{A}| > 1$ or $|\mathcal{R}| > 1$, this information network is called a heterogeneous information network (HIN, see Sun et al. 2011). HINs are represented using task-specific formats, e.g., by using adjacency matrices for efficient querying.

The network schema of an HIN G, denoted as $T_G = (\mathcal{A}, \mathcal{R})$, is a directed graph containing the object types $\mathcal{A}$ as labelled nodes (Sun et al. 2011). Two nodes are connected by a labelled edge for each relationship type in $\mathcal{R}$ that connects them. An example of a simple network schema containing nodes of the types *Sample*, *Site*, *Country* and *Time Period*, as well as the relationship types *found in*, *in country* and *dated as*, can be seen in Figure 1.

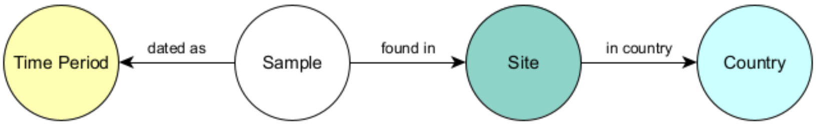

Figure 1: Network schema illustrating relationships among archaeological entities.

HINs are closely related to knowledge graphs (KGs), which are networks designed to store and convey knowledge about the real world. Their nodes represent entities of interest, while their edges capture the relationships between these entities. The underlying network conforms to a graph-based model, such as a directed edge-labelled graph, e.g., RDF (Cyganiak, Hyland-Wood, and Lanthaler 2014), or property graphs (Rodriguez and Neubauer 2010), e.g., Neo4J. Knowledge, which is defined as what is known, can be derived from external sources or inferred from the graph itself. It can consist of simple factual statements or more complex quantified assertions. To represent and reason about such structured knowledge, more expressive formalisms, such as ontologies or rules, can be employed (Hogan et al. 2021). Sun et al. (2022) state that KGs are special cases of HINs with a much richer schema than most HINs considered, and that they can add a hierarchical structure to the flat structure of an HIN at the schema or object instance level. Shi et al. (2017) state that KGs can be considered as HINs. They also note that KGs are too complex to be modelled as an HIN with a simple network schema because they are schema-rich networks that cannot be represented by a simple network schema. Similarly, Zheng, Qu, and Yang (2024) describe KGs as large-scale HINs.

2.3 Data Modelling

The Resource Description Framework (RDF, see Cyganiak, Hyland-Wood, and Lanthaler 2014) represents graph data in the form of subject-predicate-object triples, where the predicate represents a relationship linking the subject to the object. Internationalised Resource Identifiers (IRIs)[2] are the standard way for uniquely identifying entities in RDF.

[2] https://datatracker.ietf.org/doc/html/rfc3987; *Visited on May 20, 2025.*

An ontology is a knowledge organisation system (Zeng 2008) that is often defined from the perspective of a specific domain and specifies terms by describing their relationships with other terms (Hitzler et al. 2012). It is a structured model of the scientific research process that defines general concepts, their internal structure and the relationships between them. Unlike a taxonomy, which organises entities hierarchically, an ontology explicitly represents relationships, enabling semantic networks for exploratory analysis and inference (Hughes, Constantopoulos, and Dallas 2015, p. 15). In computer and information science, an ontology is a technical artefact that describes a particular reality by defining a case-specific vocabulary consisting of explicit assumptions about the intended meaning of words (Guarino 1998, p. 4).

The Web Ontology Language (OWL) builds on RDF by providing a more expressive formalism based on Description Logic, which enables reasoning capabilities. It is used to formally represent ontologies, allowing the definition of object types (classes), relationships between object types (object properties) and associations between objects and literals (data properties) in a machine-interpretable way (Hitzler et al. 2012; Motik, Patel-Schneider, and Parsia 2012). In the context of Linked Open Data (LOD), which aims to provide openly available and interlinked online datasets that are ready for cross-querying (Schmidt, Thiery, and Trognitz 2022), Berners-Lee's (2006) linked data principles model are a well-known guideline.

2.4 Introducing the Used Ontologies

The schema-level ontology for archaeo-natural data, presented in Figure 2, contains seven top-level concept domains: (A) Cultural Heritage as the central domain within N4O, (B) the Application Domain (here, Biology is the first use case), the trinity of (C) Time, (D) Space and (E) Agent, as well as the metadata-focused entities (F) Bibliography and (G) Provenance. Note that the concept domains are intertwined. In bibliographic information, for example, agents are involved as authors, while time is represented by the year of publication.

(A): The CIDOC Conceptual Reference Model (CIDOC-CRM, Version 7.1.3) is "a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous CH information and similar information from other domains" (Bekiari et al. 2021, p. 9). CIDOC-CRM is used to represent, for example, aDNA samples as biological object individuals.

(B): To define biological features as part of the application domain for the Poseidon data, such as tissue, contamination or genetic sex, the Uber-anatomy ontology (UBERON, Version 2025-01-15)[3], the Phenotype And Trait Ontology (PATO, Version v2025-02-

[3] https://www.ebi.ac.uk/ols4/ontologies/uberon; *Visited on May 20, 2025.*

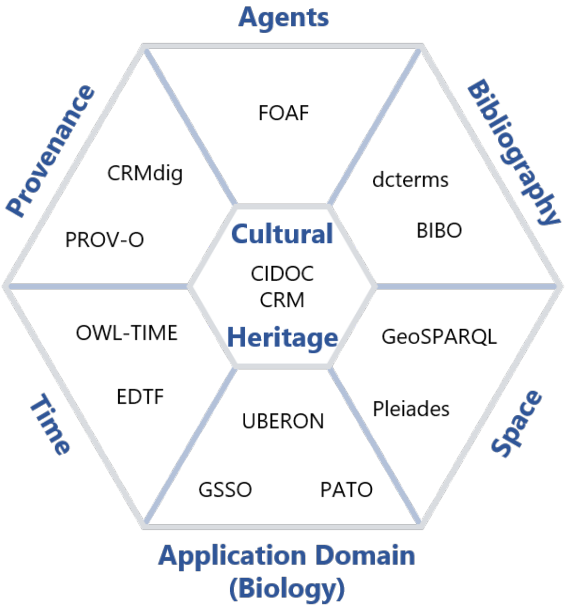

Figure 2: The ontologies grouped by the concept domain they are used for.

01)[4] and the Gender, Sex, and Sexual Orientation ontology (GSSO, Version 2.0.5)[5] are used respectively.

(C): Temporal concepts, such as time positions, durations, and temporal relations (including the ordering of intervals and instants), are represented using the OWL-Time ontology (Cox and Little 2022). The standardised definition of temporal features, such as time periods or years, is achieved using the Extended Date/Time Format (EDTF, Version from February 4, 2019)[6].

(D): Standardised representation of spatial data types (Simple Features) together with spatial query functions are provided by the GeoSPARQL ontology (Car et al. 2024; Car and Homburg 2022, Version 1.1) and used to represent concepts such as (discovery) sites. This enables effective use in geographic information systems and linked data contexts (e.g. Thiery et al. 2021). The humanities' perspective on geolocations, in addition to GeoSPARQL, is incorporated by including the concepts of Pleiades *Place* and *Location* (Elliott 2017, Version 4.0.1).

(E): People and other types of agents, such as organisations, are modelled by the Friend of a Friend (FOAF) Vocabulary Specification 0.99 (Brickley and Miller 2014).

(F): Bibliographic information, such as documents and journals, is represented using the Bibliographic Ontology (BIBO), which facilitates interoperability between data sources (Giasson and D'Arcus 2009, Version 1.3). A standardised set of metadata elements is defined by DCMI Metadata Terms (dcterms) and used to specify the concept of an author in accordance with BIBO (Board 2020, Version from January 20, 2020).

[4] https://www.ebi.ac.uk/ols4/ontologies/pato; *Visited on May 20, 2025.*
[5] https://www.ebi.ac.uk/ols4/ontologies/gsso; *Visited on May 20, 2025.*
[6] https://www.loc.gov/standards/datetime/; *Visited on May 20, 2025.*

(G): The PROV Ontology (PROV-O) is a framework for representing relationships between entities, activities and agents (Lebo, Sahoo, and McGuinness 2013, W3C Recommendation 30 April 2013). It will mainly be used to describe the process of transforming data into RDF triples. CRM$_{dig}$ (Doerr and Theodoridou 2011, Version 4.0) is an extension of CIDOC-CRM to represent the provenance of digital data generated from scientific observations. It is used to represent software executions for analysis.

The ontologies selected for the schema were chosen based on a balance of domain coverage, semantic interoperability, maturity, and uptake within the relevant research communities. While the chosen ontologies are well-established, there are issues with their use, and there are alternatives that warrant consideration. For example, the FOAF vocabulary, which is used to model agents, has not seen significant content updates since 2014 and is therefore considered outdated for modern Semantic Web applications. However, it is still included because many existing datasets are modelled using FOAF, and the aim is to achieve interoperability with these datasets. In the context of biological data, the selected ontologies (UBERON, PATO and GSSO) are all part of the OBO Foundry, a collaborative initiative that aims to create interoperable ontologies related to the life sciences. While integrating these ontologies provides a strong semantic basis for representing anatomical and phenotypic features, further aligning them with OBO or other upper-level categories could enhance cross-domain interoperability. However, such extensions were deemed out of scope for the current use case, which focuses on representing aDNA-related attributes. Regarding provenance modelling, PROV-O and CRM$_{dig}$ were chosen due to their formal W3C endorsement (in the case of PROV-O) and compatibility with CIDOC-CRM-based N4O modelling approach.

Overall, the selected ontologies provide a coherent foundation for the integration of heterogeneous archaeo-natural data. However, ontology evolution must be continuously monitored, particularly with regard to maintenance and adoption. Future iterations of the schema may involve replacing or augmenting inactive vocabularies, such as FOAF, with robust alternatives that are actively maintained.

3 Related Work

The presented approach aims to extend CIDOC-CRM to be applicable to archaeo-natural data, allowing integration with data not based on CIDOC-CRM. There is prior work about combining ontologies, which also include CIDOC-CRM. Relevant approaches in the scope of this work are presented below.

Binding, May, and Tudhope (2008) demonstrate how CIDOC-CRM and its English Heritage extension (CRM-EH) can facilitate the integration of heterogeneous archaeological datasets by transforming relational databases into RDF. They highlight the challenges and benefits of ontology-based data integration, particularly when modelling archaeological processes, ensuring semantic interoperability and improving dataset accessibility. They also emphasise the role of semi-automatic mapping and extraction tools in streamlining the transformation process.

Deicke (2016) explores the application of the CIDOC-CRM for modelling archaeological catalogues, focusing specifically on Late Bronze Age elite burials. The study shows that structured semantic modelling improves data reuse, provides a better understanding of archaeological contexts and enables integration into LOD initiatives The paper emphasises the advantages of using ontologies to structure research for greater transparency and interoperability, particularly in the fields of digital humanities and archaeology.

Nys, Ruymbeke, and Billen (2018) propose a hybrid ontology that integrates GeoSPARQL and OWL-Time with CIDOC-CRM, enhancing spatio-temporal reasoning. Although CIDOC-CRM includes spatial and temporal elements, a lack of standardisation limits its computational reasoning capabilities. To address this issue, the authors link CRM$_{geo}$, an extension of CIDOC-CRM designed for geospatial data, with GeoSPARQL for spatial representation and OWL-Time for temporal relationships. This enables complex spatio-temporal queries and demonstrates how semantic, spatial and temporal reasoning can facilitate knowledge discovery in archaeological datasets. This integration enhances the interoperability and analytical potential of CIDOC-CRM-based systems and is also considered in this work.

Hyvönen (2019) examines the evolution of semantic portals in CH, tracing their development from first-generation systems focused on data aggregation and search to second-generation systems that incorporate tools for data analysis and knowledge discovery. Third-generation systems, which integrate automated knowledge discovery and explainable AI, are then introduced. The paper introduces the Sampo model: a framework for building linked data infrastructures that support digital humanities research by providing interactive tools for filtering, visualising and analysing large datasets.

Koho et al. (2020) extend CIDOC-CRM with military history ontologies to create the WarSampo knowledge graph, which semantically represents spatio-temporal events related to Finland during the Second World War. They introduce CIDOC-CRM subclasses and SKOS[7] vocabularies to ensure interoperability while leveraging the event-based structure of CIDOC-CRM for data harmonisation. The WarSampo KG integrates heterogeneous historical datasets, enabling linked data queries and providing a foundation for semantic applications in digital humanities and military history research.

Tzitzikas et al. (2022) explore the integration of machine learning (ML) into CIDOC-CRM processes, addressing challenges posed by its event-centric structure and complex hierarchy, such as the manual effort required for schema alignment. They analyse key tasks in which ML could enhance CIDOC-CRM-based workflows, such as information

[7] https://www.w3.org/2004/02/skos/; *Visited on May 20, 2025.*

extraction, data transformation and instance matching. The paper reviews successful applications of ML, outlines future research directions and highlights its potential to improve semantic interoperability, particularly when processing unstructured data. It also provides a curated list of publicly available CIDOC-CRM datasets that can be used to train and evaluate ML models.

4 From Data to Standards

4.1 Use Case: Poseidon Framework

The Poseidon Framework (Schmid et al. 2024) aims to enable FAIR and open management of ancient DNA (aDNA) data. It enriches genotype data with metadata and archaeological context data in a structured yet flexible format. The Poseidon Framework consists of three different archives, one of which is the Poseidon Community Archive[8]. It contains publication-wise genotype data to ensure reproducibility and is maintained by the Max Planck Institute for Evolutionary Anthropology and researchers worldwide.

In addition to other information, the Janno file contains standardised contextual information and metadata for genotyped individuals. This information – consisting of over 18,000 rows, each representing a contextualised genetic sample – originates from 194 datasets and their associated bibliographic metadata files. It is incorporated into the constructed HIN.

4.2 Motivating Community Standards

Data linkage is difficult to achieve due the variety of schemas and formats resulting from different requirements and modelling choices (Mountantonakis and Tzitzikas 2019). A naively created Poseidon network schema that only considers the attributes and their descriptions cannot be integrated with records containing similar information represented using different modelling decisions. For example, suppose the object type *Place* is called *Region* or *Location* in an HIN created from other data. There is no way to automatically infer that these node types represent the same object type and should therefore be unified. This highlights the need for standards at the schema level to ensure the interoperability of networks representing the same node types. Adherence to such standards, achieved by creating a common ontology, in well-structured data with the potential for analysis beyond the context of a specific dataset (Deicke 2016). Formal ontologies and terminologies are essential for achieving the precision of human-mediated knowledge (Doerr 2005).

[8] https://github.com/poseidon-framework/community-archive/tree/master; *Visited on May 20, 2025.*

Standards are also needed at the data level to match and map differently represented instances of the same object (Thiery and Mees 2023).

Schema-Level Standards

One of the main barriers to working with linked data is the lack of a semantically consistent representation of object types (Middle 2024). Matching datasets to a global, dataset-overarching network schema provides unified representations of concepts and enables the semantic interoperability of data from different sources. Using ontologies as modelling standards addresses this by providing a unified conceptual framework that enables the semantic integration of network data at schema level. Well-known foundational ontologies or reference models can be used as the basis for such a global network schema, for example, CIDOC-CRM (e.g., Bekiari et al. 2021) for the CH domain or BFO (Otte, Beverley, and Ruttenberg 2022) for NFDI4Culture (Steller et al. 2025, p. 11). CIDOC-CRM offers a formal structure of object and relationship types that supports cross-dataset integration and exploration within the cultural heritage domain (Doerr 2005; Doerr, Light, and Hiebel 2020). It is also employed as the reference model for the application ontology (Guarino 1998, p. 8) of the N4O KG. The Archaeo-Natural Ontology (ArNO), which is presented in this work, is thus based on CIDOC-CRM to ensure the HIN's schema-level consistency with other CH datasets.

However, CIDOC-CRM alone is insufficient for representing the anthropological, archaeobotanical and archaeozoological data considered in TA3, no for ensuring interoperability with datasets from domains other than CH. Using CIDOC-CRM alone for the semantic modelling of spatial information (e.g., Padfield et al. 2019; Stein and Balandi 2019; Doerr 2003) raises conceptual issues in the context of cultural heritage (see Voß 2025). The CIDOC-CRM class *E53 Place* covers extents in the natural space in a purely physical sense, independent of temporal phenomena and matter. In contrast, *E27 Site* covers pieces of land or sea floor represented by photographs, paintings, and maps. This does not allign the Pleiades logic of *Place* – a descriptive and conceptual context for geographic or toponymic information – and *Location* – a record information about measurable points on the Earth's surface. To be interoperable with the CH domain and to be able to model the presented view of spatial concepts, this work uses CIDOC-CRM as a reference model, integrating domain-specific extensions, such as archaeology.link's LADO (e.g. Thiery and Mees 2025). This intertwines CIDOC-CRM classes with the Pleiades Place and Location concepts and GeoSPARQL to represent spatial objects.

Therefore, modelling all archaeo-natural data directly with CIDOC-CRM alone does not address the domain-specific and interdisciplinary requirements of TA3. To bridge this gap, the hybrid ArNO ontology (Nys, Ruymbeke, and Billen 2018) is being developed, combining classes and properties from the ontologies introduced above (see Section 2.4). This allows for a fine-grained representation of the considered archaeo-natural data and enables integration with datasets from CH and other domains that use some of the same ontologies. ArNO is being developed on GitHub[9] (Straten and Thiery 2025). Subclasses and custom properties are used to accurately represent the Poseidon data (see Figure 3) while ensuring interoperability with other systems based on the used ontologies (Koho et al. 2020). All object types shown in the network schema Figure 3 are subclasses of classes from the ontologies incorporated in ArNO, particularly the CIDOC-CRM classes.

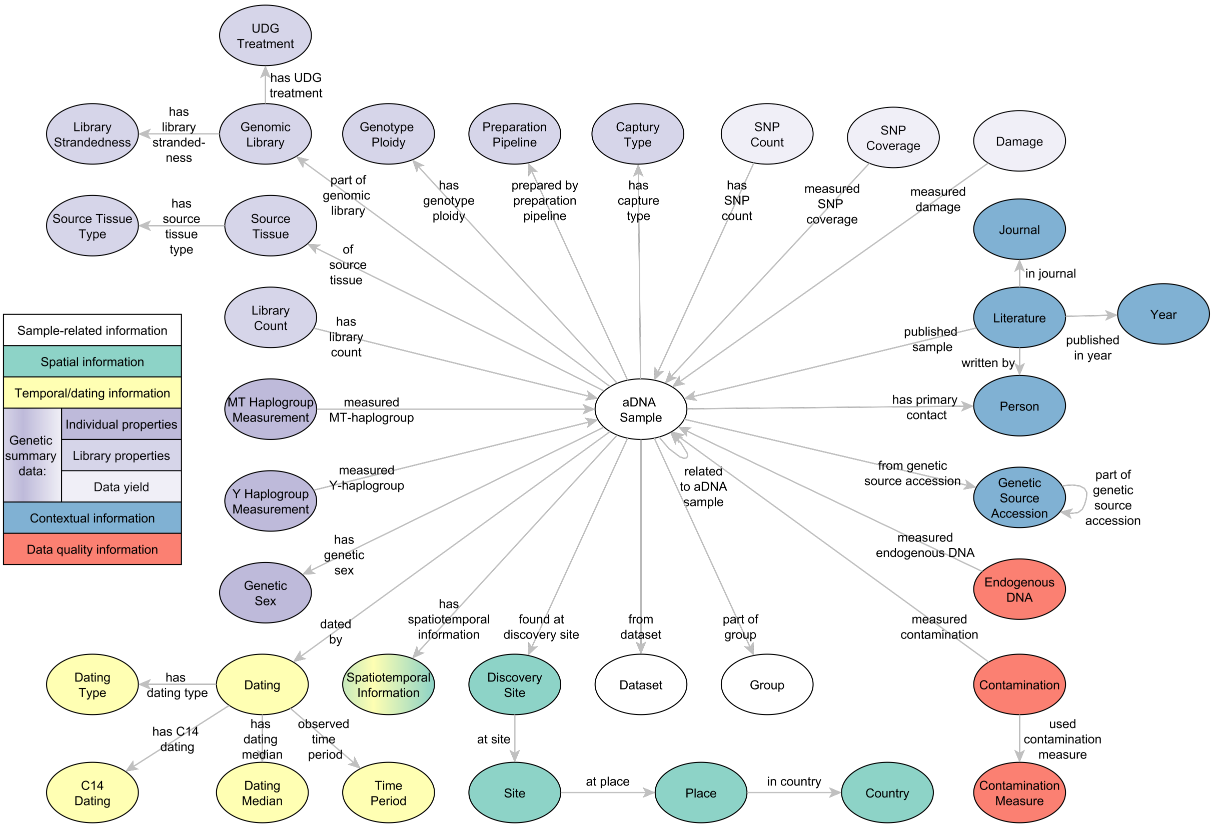

Figure 3: The network schema of the Poseidon network without data properties for better readability.

However, not all relationships are subproperties of properties from the used ontologies, particularly not from CIDOC-CRM. This is because CIDOC-CRM properties are not suitable for subtyping to create the relationship types of the Poseidon HIN, as they are designed to be generally applicable in the CH domain and do not specify the approach used for analysis in the natural and life sciences. As an event-centric approach, CIDOC-CRM typically models a series of events in which an object has been involved, rather than placing the individual object at the centre of the schema (Middle 2024; Bruseker, Carboni, and Guillem 2017; Bekiari et al. 2021). CRM-EH creates “virtual” events that are not part of the considered dataset in order to correctly connect such entities in their network. Currently, ArNO adopts an object-oriented approach by considering CIDOC-CRM as a reference model rather than applying it holistically to model the data. This results in the creation of new types of relationships linking *aDNA Samples* to their respective measure-

[9] https://archaeonatural-cloud.github.io/archaeonatural-ontology/; *Visited on May 20, 2025.*

ments. This avoids definiting an event that contains no additional information, which would artificially increase the complexity of the network and hinder network analysis, resulting in slow and overly complex queries (Koho et al. 2020). This is particularly important for meta path-based HIN analysis (e.g., Sun et al. 2011), where the complexity of the network schema complicates the query definition. The hybrid ontology introduced in this work thus introduces new relationships that are not inherited from the CIDOC-CRM properties. These relationships directly connect the *aDNA Samples* to their respective measurements, without the need to create an event as a connector.

HIN data is stored in a task-specific format, enabling attribute values for nodes and relationships to be stored directly when modelling them similar to property graphs. However, when OWL and RDF are used to represent the network, it becomes necessary to use data properties to represent attributes. It is not possible to represent data properties of object properties (i.e. attributes of relationships). Therefore, an additional class, *aDNA Sample Relation*, had to be introduced in ArNO to enable attributes to be added to the relationship between *aDNA Samples.* An alternative way of representing this information is to use quintuples, which add both attributes to the relation using blank nodes, in a manner similar to the use of quadruples in AMT modelling (Thiery et al. 2022).

Data-Level Standards

In order to successfully integrate diverse datasets into a single HIN, it is crucial to map multiple instances of the same object to the same object representation. The same object may be referred to in different languages, by various identifiers or by conflicting modelling choices, which complicates entity identification and affects entity resolution. Gazetteers, controlled vocabularies and authority files – such as GeoNames[10] or the Integrated Authority File (GND)[11] – provide standardised identifiers and labels that support the harmonisation of different representations of the same object. In contrast, community-driven vocabularies, which are often created during domain-specific database projects (e.g., Thiery and Mees 2024, pp. 68-69), are less rigidly standardised and offer greater flexibility and adaptability to emerging needs. However, they may have limited potential for generalisation.

Different instances that refer to the same object can be resolved by *canonicalisation*, i.e., by transforming both objects into the same representation, or by *binding*, i.e., by establishing an additional relationship between these objects (Mountantonakis and Tzitzikas 2019). As one of the goals of the integrated network presented in this work is to perform holistic network analysis, canonicalisation is used to ensure that all of an object's semantics are captured in a single representative instance for use in network analysis. Binding would result in multiple linked instances of the same object, which would pre-

[10] https://www.geonames.org/; *Visited on May 20, 2025.*
[11] https://explore.gnd.network/en/; *Visited on May 20, 2025.*

vent the semantic integration and incorporation of all modelled information about the object into a single representative instance. If binding is chosen instead of canonicalisation, not all information – e.g., attributes and relationships to other objects – of all instances of the query object can be considered when using a single representative as the query object in network analysis tasks. However, canonicalisation might introduce limitations. By reducing multiple instances to a single representation, important nuances may be lost, e.g., various provenance trails, uncertainty in the original data sources or obfuscating contextual ambiguity. For example, if different datasets refer to an object with varying spatial precision or conflicting temporal attributions, canonicalisation may obscure these discrepancies in favour of a unified representation. Therefore, the trade-off between analytical clarity and the preservation of interpretative richness must therefore be carefully considered. The presented work represents spatial and temporal information as separate classes, enabling such conflicting information to be represented using multiple relationships.

Consistency at the data level is achieved through the use of IRIs where available. Rather than retrieving all relevant bibliographic information directly from the data files themselves, Digital Object Identifiers (DOIs) are used to consistently retrieve it via the Crossref API[12]. This prevents different representations of bibliographic objects from resulting in multiple instances of the same object being created. ISO 3166 country codes are used to ensure the consistent representation of *Countries* in the source datasets. Data-level consistency can be further ensured by providing IRIs for objects obtained using appropriate gazetteers for spatial data. However, the name of a *Site* alone, as it is found in most existing datasets, is not sufficient to correctly map it to an existing IRI using a gazetteer, since several *Sites* may have the same name. Therefore, geographic coordinate information must also be (created and) taken into account to make the mapping more likely to be correct.

The presented framework aims to create a semantically consistent integrated HIN that is interoperable with future datasets from the archaeo-natural domain and related fields. This is achieved by combining well-known ontologies and terminologies. The resulting LOD provides an opportunity to address new research questions and discover previously unknown relationships between objects (Middle 2024) by applying network analysis methods to a shared digital heritage rather than isolated datasets.

5 Discussion

Although CIDOC-CRM enables the representation of detailed and well-structured data, mapping data to it is a time-consuming and manual process that requires expertise, as the quality of the mapping depends on familiarity with both the dataset and the complex

[12] https://api.crossref.org/; *Visited on May 20, 2025.*

ontology (Deicke 2016; Tzitzikas et al. 2022; Middle 2024). Several projects have avoided using CIDOC-CRM due to the high resources required for its correct implementation (Middle 2024). Nevertheless, this study uses CIDOC-CRM as a reference model to ensure interoperability in the CH domain and to facilitate the data integration into the N4O KG. Despite the fact that the data considered includes well-documented metadata and comes in well-structured formats, extensive and repeated consultations with domain experts were required to accurately map it to CIDOC-CRM. Without close collaboration between modellers and data owners, models that are semantically correct but misrepresent the underlying data may be created. Therefore, continuous communication is required for interdisciplinary collaboration to ensure that the structures accurately reflect the data content, enable future integration and meet the needs of researchers.

Representing the network data using RDF enables the use of blank nodes for missing data, which act as connectors. For example, this could be used for missing *Discovery Site* where *Site* information is available (see Figure 3). In traditional HINs, it would be difficult to associate the *Site* information with the *aDNA Sample* if there were no details about the connecting *Discovery Site* (see Figure 3). Therefore, HINs may be based on a star network schema (Sun, Yu, and Han 2009) to prevent the loss of link information. However, this limits the analytical depth obtained through meta-path-based analysis, in which objects are connected by a semantic sequence of edge types. This is because these paths always pass through the central node type of the star network schema when connecting a pair of object types.

In conventional RDF-based KGs, purely spatial or temporal information, e.g., *Discovery Site* or *Time Period*, is usually defined as a literal (attribute) using data properties. However, in order to discover patterns of spatial or temporal similarity in HINs, it is crucial to have objects representing such information. Therefore, the object types of *Discovery Site* and *Time Period* are included in ArNO.

The ArNO ontology is a work in progress, with the Poseidon datasets currently forming its foundation. Other archaeo-natural datasets from TA3, such as the archaeozoological data from the Ossobook framework (Kriegel et al. 2009) or the archaeobotanical data from ArboDat (Kreuz, Elger, and Frenzel 2022), will be integrated into the unified HIN based on ArNO and will be followed by other archaeo-natural data. The WarSampo KG (Koho et al. 2020) has extended its class structure iteratively when considering new datasets. The class structure of ArNO may also change when integrating such datasets in a community-driven manner. However, ArNO's modular structure, the use of CIDOC-CRM as a reference model and its extension using standard ontologies – e.g., for spatial and temporal information – allow for an integrated foundation that only needs to be extended when previously unseen classes or properties are added, for example, from a new application domain. Ontologies such as ArNO are introduced to the community through discussions within the NFDI consortium. These discussions take place in N4O groups such as Temporary Working Groups and Community Clusters, as well as groups that transcend the N4O consortium, such as Memorandum of Understanding groups and NFDI sections.

The mapping of ArNO's classes and properties to top-level ontologies (Guarino 1998, p. 7) and other existing ontologies will be discussed later, with the aim of producing a consistent mapping of the data to other systems, e.g., by including their concepts to ArNO. For example, one could incorporate ETS[13], SOSA[14] or OBOE[15] in ArNO, allowing interoperability with natural data modelled in NFDI4Biodiversity. One could also incorporate BFO to enable interoperability with NFDIcore-based systems[16], aiming to ensure data interoperability across NFDI consortia. After the successful extension and integration into the N4O KG, extended documentation and exemplary data modelling will be provided.

6 Conclusion & Future Work

This work promotes community standards for research data that enhance the interoperability of heterogeneous datasets in an interdisciplinary context. Adherence to schema- and data-level standards ensures alignment with the FAIR principles, guaranteeing that integrated HINs are semantically consistent and interoperable with other datasets. This creates a rich source of reusable data for future research. Using standardised ontologies to enforce schema-level standards results in a unified representation of broad, multidisciplinary resources. This connects disparate datasets and makes connections to the wider Linked Data ecosystem (Middle 2024). By sharing this approach, this work aims to contribute to wider efforts to promote the interoperability and semantic consistency of heterogeneous datasets within CH and other related research fields.

KGs typically focus on structured knowledge representation and reasoning, often leveraging ontologies. In contrast, HINs are typically used for applying network analysis methods to discover new knowledge (e.g., Hogan et al. 2021; Shi et al. 2017). Sun et al. (2022) distinguish KGs and HINs by the fact that KGs may contain a hierarchical structure, whereas HINs are not defined to consider a hierarchical network schema. However, the use of ontologies presented here to standardise HINs at the schema level incorporates a taxonomic hierarchy into the network schema, where the node types are subclasses of ontology classes. Conventional HIN analysis methods also only consider flat, non-hierarchical network schemas and therefore fail to adequately represent hierarchically structured information modelled by ontologies. When ontologies are used to define the HIN network schema and IRIs to identify object instances, the distinction between HINs and KGs becomes less clear, particularly when data models that can accommodate both

[13] https://biodivportal.gfbio.org/ontologies/ETS; *Visited on May 20, 2025.*

[14] https://biodivportal.gfbio.org/ontologies/SOSA; *Visited on May 20, 2025.*

[15] https://biodivportal.gfbio.org/ontologies/OBOE; *Visited on May 20, 2025.*

[16] https://ise-fizkarlsruhe.github.io/nfdicore/; *Visited on May 20, 2025.*

heterogeneous networks and semantic representations are employed. Therefore, the presented standards will lead to greater interoperability between HINs and KGs, especially if standardised formats such as RDF are used for HIN representation. The achieved interoperability could enable the integration of separate HINs into a large, unified KG, and could also facilitate the application of HIN analysis methods to large-scale KGs. However, meta path-based methods require a manageable network schema in order to define relevant paths. Future work could therefore focus on discovering how HIN analysis methods can be applied to KGs and RDF data. Current HIN analysis methods typically use a flat network schema for knowledge discovery. Another promising area for future research will therefore be to extend HIN analysis methods (e.g., Sun et al. 2011; Rossi et al. 2019) to consider hierarchical ontology structures (e.g., Huang 2021).

Future work will also involve integrating the Arbodat and Ossobook datasets into ArNO, primarily through mapping them to existing object and relationship types. If necessary, new classes will be added to ArNO, particularly within the application domain (see Figure 2). The average of the *Discovery Site* coordinates will be computed to obtain representative spatial information for mapping the *Site* and *Location* label to IRIs. Different spatial granularities and inexact spatial matches can also be considered when mapping *Sites* to gazetteers. Future work will also involve incorporating top-level ontologies and ontologies from related domains into ArNO. This will facilitate ontology alignment, enable semantic mapping across domains, and improve compatibility with existing knowledge graph infrastructures.

Acknowledgements

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) PNr. 501836407 (NFDI4Objects), by the state of Schleswig-Holstein within the DatenCampus project (22021016), and MARISPACE-X – Smart MARItime Sensor Data SPACE X (Grant FKZ 68GX21002E) funded by the German Federal Ministry for Economic Affairs and Climate Action.

Authorship Contributions

- Conceptualization: M.t.S., F.T. and S.S.
- Methodology: M.t.S., F.T. and S.S.
- Formal analysis: M.t.S. and F.T
- Investigation: M.t.S. and F.T
- Writing (original draft preparation): M.t.S. and F.T
- Writing (review and editing): M.t.S., F.T. and S.S
- Supervision: M.R.

- Project administration: M.R.
- Funding acquisition: M.R.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

The authors declare no conflicts of interest.

Bibliography

Bekiari, Chryssoula, George Bruseker, Martin Doerr, Christian-Emil Ore, Stephen Stead, and Athanasios Velios. 2021. *Definition of the CIDOC Conceptual Reference Model v7.1.1.* Technical report. The CIDOC Conceptual Reference Model Special Interest Group. https://doi.org/10.26225/FDZH-X261.

Berners-Lee, Tim. 2006. *Linked Data – Design Issues.* Visited on March 24, 2025. https://www.w3.org/DesignIssues/LinkedData.html.

Binding, Ceri, Keith May, and Douglas Tudhope. 2008. "Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction Via the CIDOC CRM", 280–290. ISBN: 978-3-540-87598-7. https://doi.org/10.1007/978-3-540-87599-4_30.

Bleiholder, Jens, and Felix Naumann. 2008. "Data Fusion". *ACM Computing Surveys* 41 (1): 1–41. https://doi.org/10.1145/1456650.1456651.

Board, DCMI Usage. 2020. *DCMI Metadata Terms.* Visited on March 24, 2025. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

Brickley, Dan, and Libby Miller. 2014. *FOAF Vocabulary Specification.* Visited on March 24, 2025. http://xmlns.com/foaf/spec/.

Bruseker, George, Nicola Carboni, and Anaïs Guillem. 2017. "Cultural Heritage Data Management: The Role of Formal Ontology and CIDOC CRM". In *Heritage and Archaeology in the Digital Age: Acquisition, Curation, and Dissemination of Spatial Cultural Heritage Data,* edited by Matthew L. Vincent, Víctor Manuel López-Menchero Bendicho, Marinos Ioannides, and Thomas E. Levy, 93–131. Cham: Springer International Publishing. ISBN: 978-3-319-65370-9. https://doi.org/10.1007/978-3-319-65370-9_6.

Car, Nicholas J., and Timo Homburg. 2022. "GeoSPARQL 1.1: Motivations, Details and Applications of the Decadal Update to the Most Important Geospatial LOD Standard". *ISPRS International Journal of Geo-Information* 11 (2): 117. https://doi.org/10.3390/ijgi11020117.

Car, Nicholas J., Timo Homburg, Matthew Perry, Frans Knibbe, Simon J.D. Cox, Joseph Abhayaratna, Mathias Bonduel, Paul J. Cripps, and Krzysztof Janowicz. 2024. *OGC GeoSPARQL – A Geographic Query Language for RDF Data.* OGC Standard 22-047r1. Open Geospatial Consortium. Visited on March 24, 2025. https://docs.ogc.org/is/22-047r1/22-047r1.html.

Cox, Simon, and Chris Little. 2022. *Time Ontology in OWL.* W3C Candidate Recommendation Draft CRD-owl-time-20221115. World Wide Web Consortium (W3C). Visited on March 24, 2025. https://www.w3.org/TR/2022/CRD-owl-time-20221115/.

Cyganiak, Richard, David Hyland-Wood, and Markus Lanthaler. 2014. "RDF 1.1 Concepts and Abstract Syntax". *W3C Proposed Recommendation,* visited on March 24, 2025. https://www.w3.org/TR/rdf11-concepts/.

Deicke, Aline Julia Elisabeth. 2016. "CIDOC CRM-based modeling of archaeological catalogue data." In *DHC@ MTSR.* Visited on March 24, 2025. https://ceur-ws.org/Vol-1764/4.pdf.

Doerr, Martin. 2003. "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata". *AI Magazine* 24 (3): 75–75. https://doi.org/10.1609/aimag.v24i3.1720.

———. 2005. "The CIDOC CRM, an Ontological Approach to Schema Heterogeneity". In *Semantic Interoperability and Integration,* edited by Y. Kalfoglou, M. Schorlemmer, A. Sheth, S. Staab, and M. Uschold, 4391:1–5. Dagstuhl Seminar Proceedings (DagSemProc). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/DagSemProc.04391.22.

Doerr, Martin, Richard Light, and Gerald Hiebel. 2020. *Implementing the CIDOC Conceptual Reference Model in RDF.* Technical report. CIDOC CRM Special Interest Group. Visited on March 24, 2025. https://cidoc-crm.org/Resources/implementing-the-cidoc-conceptual-reference-model-in-rdf.

Doerr, Martin, and Maria Theodoridou. 2011. "CRMdig: A Generic Digital Provenance Model for Scientific Observation". In *3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP 11).* Visited on March 24, 2025. https://www.usenix.org/legacy/event/tapp11/tech/final_files/Doerr.pdf.

Elliott, Tom. 2017. *Pleiades Data Model.* Visited on March 24, 2025. https://pleiades.stoa.org/help/pleiades-data-model.

Giasson, Frédérick, and Bruce D'Arcus. 2009. *Bibliographic Ontology Specification.* Version 1.3, published on November 4, 2009. Visited on March 24, 2025. http://bibliontology.com.

Guarino, Nicola. 1998. “Formal Ontologies and Information Systems”. Visited on March 24, 2025. https://www.loa.istc.cnr.it/old/Papers/FOIS98.pdf.

Hartl, Nathalie, Elena Wössner, and York Sure-Vetter. 2021. “Nationale Forschungsdateninfrastruktur (NFDI)”. *Informatik Spektrum* 44 (5): 370–373. https://doi.org/10.1007/s00287-021-01392-6.

Hitzler, Pascal, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. 2012. *OWL 2 Web Ontology Language Primer.* W3C Recommendation. World Wide Web Consortium (W3C). Visited on March 24, 2025. http://www.w3.org/TR/owl2-primer/.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia D’amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, et al. 2021. “Knowledge Graphs”. *ACM Computing Surveys* 54 (4): 1–37. https://doi.org/10.1145/3447772.

Huang, Yue. 2021. “Incorporating domain ontology information into clustering in heterogeneous networks”. *WIREs Data Mining and Knowledge Discovery* 11 (4): e1413. https://doi.org/10.1002/widm.1413.

Hughes, Lorna, Panos Constantopoulos, and Costis Dallas. 2015. “Digital Methods in the Humanities”. In *A New Companion to Digital Humanities,* edited by Susan Schreibmann, Ray Siemens, and John Unsworth, 150–170. John Wiley / Sons, Ltd. ISBN: 9781118680599. https://doi.org/10.1002/9781118680605.ch11.

Hyvönen, Eero. 2019. “Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery”. *Semantic Web* 11:1–7. https://doi.org/10.3233/SW-190386.

Koho, Mikko, Esko Ikkala, Petri Leskinen, Minna Tamper, Jouni Tuominen, and Eero Hyvönen. 2020. “WarSampo knowledge graph: Finland in the Second World War as Linked Open Data”. *Semantic Web* 12:1–14. https://doi.org/10.3233/SW-200392.

Kreuz, Angela, Kirsten Elger, and Simone Frenzel. 2022. *Digitising and sharing archaeobotanical data–the ArboDat 2016 Datacentre project.* Visited on March 24, 2025. https://gfzpublic.gfz-potsdam.de/rest/items/item_5017643_2/component/file_5018422/content.

Kriegel, Hans-Peter, Peer Kröger, Henriette Obermaier, Joris Peters, Matthias Renz, and Christiaan Hendrikus van der Meijden. 2009. “OSSOBOOK: database and knowledge-management techniques for archaeozoology”. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management,* 2091–2092. CIKM ’09. Hong Kong, China: Association for Computing Machinery. ISBN: 9781605585123. https://doi.org/10.1145/1645953.1646318.

Lebo, Timothy, Satya Sahoo, and Deborah McGuinness. 2013. *PROV-O: The PROV Ontology.* World Wide Web Consortium (W3C) Recommendation. Visited on March 24, 2025. https://www.w3.org/TR/prov-o/.

Legler, Frank, and Felix Naumann. 2007. “A Classification of Schema Mappings and Analysis of Mapping Tools”. In *Datenbanksysteme in Business, Technologie und Web (BTW 2007), 12. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), Proceedings, 7.-9. März 2007, Aachen, Germany,* volume P-103, 449–463. LNI. GI. ISBN: 978-3-88579-197-3, visited on March 24, 2025. https://dl.gi.de/handle/20.500.12116/31815.

Middle, Sarah. 2024. “Linked Ancient World Data: Implementation, Advantages, and Barriers”. *Digital Classics Online* 10 (1): 16–49. https://doi.org/10.11588/dco.2024.10.104105.

Motik, Boris, Peter F. Patel-Schneider, and Bijan Parsia. 2012. *OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition).* Visited on March 24, 2025. https://www.w3.org/TR/owl2-syntax/.

Mountantonakis, Michalis, and Yannis Tzitzikas. 2019. “Large-scale Semantic Integration of Linked Data: A Survey”. *ACM Computing Surveys* 52:1–40. https://doi.org/10.1145/3345551.

Nys, Gilles-Antoine, Muriel Ruymbeke, and Roland Billen. 2018. “Spatio-Temporal Reasoning in CIDOC CRM: An Hybrid Ontology with GeoSPARQL and OWL-Time”. Visited on March 24, 2025. https://ceur-ws.org/Vol-2230/paper_04.pdf.

Otte, J Neil, John Beverley, and Alan Ruttenberg. 2022. “BFO: Basic formal ontology”. *Applied ontology* 17 (1): 17–43. https://doi.org/10.3233/AO-220262.

Padfield, Joseph, Kalliopi Kontiza, Antonis Bikakis, and Andreas Vlachidis. 2019. “Semantic Representation and Location Provenance of Cultural Heritage Information: the National Gallery Collection in London”. *Heritage* 2 (1): 648–665. https://doi.org/10.3390/heritage2010042.

Patel, Hardik, Pavlos Paraskevopoulos, and Matthias Renz. 2018a. “Data Fusion of Diverse Data Sources: Enrich Spatial Data Knowledge Using HINs”. In *Proceedings of the Fifth International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data,* 13–18. GeoRich’18. Houston, TX, USA: Association for Computing Machinery. ISBN: 9781450358323. https://doi.org/10.1145/3210272.3210275.

———. 2018b. “GeoTeGra: A System for the Creation of Knowledge Graph Based on Social Network Data with Geographical and Temporal Information”. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* 617–620. https://doi.org/10.1109/ASONAM.2018.8508674.

Rodriguez, Marko A., and Peter Neubauer. 2010. “Constructions from dots and lines”. *Bulletin of the American Society for Information Science and Technology* 36 (6): 35–41. https://doi.org/10.1002/bult.2010.1720360610.

Rossi, Ryan A., Nesreen K. Ahmed, Aldo Carranza, David Arbour, Anup Rao, Sungchul Kim, and Eunyee Koh. 2019. *Heterogeneous Network Motifs.* https://doi.org/10.48550/arXiv.1901.10026.

Rummel, Philipp Von, Christin Keller, and Fabian Fricke. 2025. "NFDI4Objects: Warum brauchen wir eine gemeinsame Forschungsdateninfrastruktur für die materiellen Hinterlassenschaften der Menschheitsgeschichte?" *Archäologische Informationen* 47 (NW-DVA 2024 Bochum: Digitale Archäologie): Bd. 47 (2024): Archäologische Informationen. https://doi.org/10.11588/AI.2024.1.110390.

Schmid, Clemens, Ayshin Ghalichi, Thiseas C. Lamnidis, Dhananjaya B. A. Mudiyanselage, Wolfgang Haak, and Stephan Schiffels. 2024. "Poseidon – A framework for archaeogenetic human genotype data management". *eLife* 13. https://doi.org/10.7554/elife.98317.1.

Schmidt, Sophie C., Florian Thiery, and Martina Trognitz. 2022. "Practices of Linked Open Data in Archaeology and Their Realisation in Wikidata". *Digital* 2 (3): 333–364. https://doi.org/10.3390/digital2030019.

Shi, Chuan, Yitong Li, Jiawei Zhang, Yizhou Sun, and Philip S. Yu. 2017. "A survey of heterogeneous information network analysis". *IEEE Transactions on Knowledge and Data Engineering* 29 (1): 17–37. https://doi.org/10.1109/TKDE.2016.2598561.

Stein, Regine, and Oguzhan Balandi. 2019. "Using LIDO for Evolving Object Documentation into CIDOC CRM". *Heritage* 2 (1): 1023–1031. https://doi.org/10.3390/heritage2010066.

Steller, Jonatan Jalle, Tabea Tietz, Linnaea Charlotte Söhn, Harald Sack, Heike Fliegl, Torsten Schrade, Alexandra Büttner, Etienne Posthumus, and Oleksandra Bruns. 2025. *Knowledge Graph-based Research Data Integration for NFDI4Culture and Beyond.* Poster (E-Science-Tage 2025). https://doi.org/10.5281/zenodo.14989010.

Straten, Mattis thor, and Florian Thiery. 2025. "Archaeo-Natural Ontology (ArNO) – v0.1". *Squirrel Papers* 7 (2): L4. https://doi.org/10.5281/zenodo.15095413.

Strohm, Steffen, Hartwig Buenning, and Matthias Renz. 2023. "Implementing a FAIR Information System for Archaeology-Related Interdisciplinary Research". In *2023 IEEE 19th International Conference on e-Science (e-Science),* 1–2. https://doi.org/10.1109/e-Science58273.2023.10254840.

Sun, Yizhou, Jiawei Han, Xifeng Yan, Philip Yu, and Tianyi Wu. 2022. "Heterogeneous information networks: the past, the present, and the future". *Proceedings of the VLDB Endowment* 15:3807–3811. https://doi.org/10.14778/3554821.3554901.

Sun, Yizhou, Jiawei Han, Xifeng Yan, Philip S. Yu, and Tianyi Wu. 2011. "PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks". *Proceedings of the VLDB Endowment* 4 (11): 992–1003. https://doi.org/10.14778/3402707.3402736.

Sun, Yizhou, Yintao Yu, and Jiawei Han. 2009. “Ranking-based clustering of heterogeneous information networks with star network schema”. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 797–806. KDD ’09. Paris, France: Association for Computing Machinery. ISBN: 9781605584959. https://doi.org/10.1145/1557019.1557107.

Thiery, Florian, and Thomas Engel. 2016. “The Labeling System: A Bottom-up Approach for Enriched Vocabularies in the Humanities”. In *CAA2015. Keep the Revolution Going. Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology.* Edited by Stefano Campana, Roberto Scopigno, Gabriella Carpentiero, and Marianna Cirillo, 259–268. Oxford: Archaeopress. ISBN: 978-1-78491-337-3. https://doi.org/10.5281/zenodo.3741958.

Thiery, Florian, Timo Homburg, Sophie C. Schmidt, Jakob Voß, and Martina Trognitz. 2021. “SPARQLing Geodesy for Cultural Heritage – New Opportunities for Publishing and Analysing Volunteered Linked (geo-)data”. *FIG Peer Review Journal* FIG e-Working Week 2021 – Virtually in the Netherlands 21-25 June 2021. https://doi.org/10.5281/zenodo.5639381.

Thiery, Florian, and Allard W. Mees. 2023. “Taming Ambiguity – Dealing with doubts in archaeological datasets using LOD”. *Proceedings of the Computer Applications and Quantitative Methods in Archeology* CAA 2018: Human History and Digital Future (2018). https://doi.org/10.15496/PUBLIKATION-87762.

———. 2024. “Sharing Linked Open Data with domain-specific data-driven community hubs – archaeology.link in NFDI4Objects”. *Archeologia e Calcolatori* 35 (2): 63–74. https://doi.org/10.19282/ac.35.2.2024.08.

———. 2025. “Linked Archaeological Data Ontology (LADO) – v1.1”. *Squirrel Papers* 7 (2): L3. https://doi.org/10.5281/zenodo.15477358.

Thiery, Florian, Allard W. Mees, Bernhard Weisser, Felix F. Schäfer, Stefanie Baars, Sonja Nolte, Henriette Senst, and Philipp Von Rummel. 2023. “Object-Related Research Data Workflows Within NFDI4Objects and Beyond”. In *Proceedings of the Conference on Research Data Infrastructure,* edited by York Sure-Vetter and Carole Goble, volume 1, CoRDI2023–46. Hannover: TIB Open Publishing. https://doi.org/10.52825/cordi.v1i.326.

Thiery, Florian, Karsten Tolle, Allard Mees, and David Wigg-Wolf. 2022. “How to handle vagueness and uncertainty in graph-based LOD knowledge modelling? Dealing with archaeological numismatic and ceramological real world data”. In *Graphs and Networks in the Humanities 2022 (Graphum2022).* Squirrel Papers. https://doi.org/10.5281/zenodo.7184524.

Tzitzikas, Yannis, Michalis Mountantonakis, Pavlos Fafalios, and Yannis Marketakis. 2022. “CIDOC-CRM and Machine Learning: A Survey and Future Research”. *Heritage* 5 (3): 1612–1636. https://doi.org/10.3390/heritage5030084.

Voß, Jakob. 2025. *CIDOC-CRM in RDF Application Profile.* Visited on March 24, 2025. https://nfdi4objects.github.io/crm-rdf-ap/.

Voß, Jakob, and Josef Heers. 2024. *Integration von Forschungsdaten im NFDI4Objects Knowledge Graph.* Zenodo. https://doi.org/10.5281/zenodo.13744338.

Voß, Jakob, Josef Heers, Gerald Steilen, and Anja Gerber. 2024. *N4O Graph: The Knowledge Graph of NFDI4Objects.* Zenodo. https://doi.org/10.5281/zenodo.13946053.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. https://doi.org/10.1038/sdata.2016.18.

Zeng, Marcia. 2008. "Knowledge Organization Systems (KOS)". *Knowledge Organization* 35:160–182. https://doi.org/10.5771/0943-7444-2008-2-3-160.

Zheng, Yuyan, Jianhua Qu, and Jiajia Yang. 2024. "StructSim: Meta-Structure-Based Similarity Measure in Heterogeneous Information Networks". *Applied Sciences* 14 (2). https://doi.org/10.3390/app14020935.