How to handle vagueness and uncertainty in graph-based LOD knowledge modelling. Dealing with archaeological numismatic and ceramological real world data.

Florian Thiery M.Sc.[1], Dr. Allard W. Mees[1], Dr. Karsten Tolle[2], and Dr. David G. Wigg-Wolf[3]

[1]Römisch-Germanisches Zentralmuseum
[2]University of Frankfurt
[3]Römisch-Germanische Komission des DAI

1 Introduction

Strategies for graph-based knowledge modelling as Linked Open Data (LOD) (Berners-Lee, 2006) of real world archaeological data comprising vague and/or uncertain information are widely lacking. Semantically modelled information related to fuzziness and wobbliness[1] in numismatics and ceramic studies can allow reasoning-based analysis of archaeological research data, eventually revealing imperfect inferences. Uniform modelling of uncertainties and vagueness in research data is extremely challenging. In the context of archaeological data these aspects can appear in any phase, starting from the discovery of an object through each further processing step. It might also depend on context- and meta-information, such as who performed the steps, and in the case of archaeological objects also to a great extent on the preservation of the object itself and its archaeological context.

When focusing on modelling uncertainties and vagueness in RDF, we can find different approaches, for example special properties, refined statements, attribute assignments or blank scope nodes, which have been tested in best practices for numismatics (c.f. Metzger (2014); Thiery and Mees (2018); Tolle and Wigg-Wolf (2015)) and ceramics (c.f. Thiery (2013); Thiery and Mees (2021b)), recently by using the Academic Meta Tool (AMT) (c.f. Thiery and Unold (2018); Unold and Thiery (2018); Unold et al. (2019)). The jury is still out on which of these approaches would be the best, since each of them has different advantages and drawbacks. In the domain of archaeology there is certainly no consensus, e.g. for CIDOC CRM the matter is still an open issue (CIDOC-CRM, 2019), and the same is true for Nomisma[2], for example. This results in non-homogeneous modelling even within the same domains, while

[1] As container terms for all related categories of vagueness and uncertainty
[2] http://nomisma.org

transformation rules between modelling concepts for vagueness and uncertainty are missing. Beyond this, uncertain and vague information may simply be removed from the publicly available data or, even worse, not even be captured and stored at all.

This paper discusses and evaluates modelling approaches, challenges, and limits to fuzziness and wobbliness (e.g. uncertainty, vagueness, accuracy, and precision) in research data in the exemplary contexts of numismatics and ceramic research. Both areas are similar, but both follow their own intrinsic research aspects and both are subject to different use wear conditions (corrosion, fragility). By comparing the two domains, we hope to have a broader approach to the modelling possibilities and their effects. The results should serve as a basis for answering the question as to which of the modelling approaches should be preferred, and how mapping between them could be performed. Through the CAA Special Interest Group (SIG) on Semantics and LOUD in Archaeology (SIG-DataDragon[3]) and the planned NFDI initiative NFDI4Objects[4] - e.g. TRAIL 2.2 (Mees et al., 2021) - it then can be evaluated and scaled to other subject domains.

2 Challenges

The above points present challenges that must be tackled: Performance Issues, Shortcuts and Knowledge Entailment.

2.1 Performance Issues

- What is the impact of different modelling approaches on the query performance of current systems?

⟶ We plan to present first benchmark results based on different modelling approaches for some uncertainty situations and the corresponding SPARQL queries. The benchmark will initially be executed on Apache Jena Fuseki[5].

2.2 Shortcuts

- What are the risks of different modelling approaches for existing solutions and users?

⟶ Due to the lack of uncertainty modelling standards, uncertain data are often neglected. Existing solutions and users might not be prepared to deal with uncertainty, especially if the modelling allows so called short cuts. We will present examples of the risk of hidden uncertainties and how they depend on the modelling.

2.3 Knowledge Entailment

- To what extent does fuzziness and wobbliness become visible in the knowledge model, as well as for the user? How useful and trustworthy are automatic procedures (e.g. semantic reasoning) for entailing new knowledge?

⟶ We will present visualisations of knowledge entailment realised by managing vagueness and uncertainty issues in research specific applications.

---

[3] http://datadragon.link
[4] https://www.nfdi4objects.net
[5] https://jena.apache.org/documentation/fuseki2/

3 Use Case: Numismatics

The importance of uncertainty naturally also depends on the quality of the coins. In our case we are dealing with coin finds and not with coins from collections. Finds are often badly preserved and thus generate uncertain situations. Within AFE-RGK we currently record 16,394 coins and have some 24 fields that can be marked as uncertain. This means that if a value is entered the user can indicate that they are uncertain about its accuracy. Some of these fields even facilitate the entry of alternative values. About 23% of the coins entered have at least one field marked in this way, and in total we have 5756 cases.

We are currently collecting and comparing solutions for modelling uncertainty in the numismatic space in order to better understand the differences and their implications. Here we demonstrate how two solutions represent the following information:

> A coin (Coin_5) was certainly struck at the mint called Comama, while for another coin (Coin_4) it is uncertain if it was struck at the same mint.

The first solution (S1) was proposed by the Research Space project in conjunction with CIDOC-CRM (Alexiev, 2012), while the second (S2) uses a blank node within the path to the resource, as we proposed in (Tolle and Wigg-Wolf, 2015). Solution 1 (S1) would use one triple for Coin_5 and six triples for Coin_4. This would mean seven triples in total, as shown in figure 1. Solution (S2) would require one triple for Coin_5 and three triples for Coin_4, a total of four triples (figure 2).

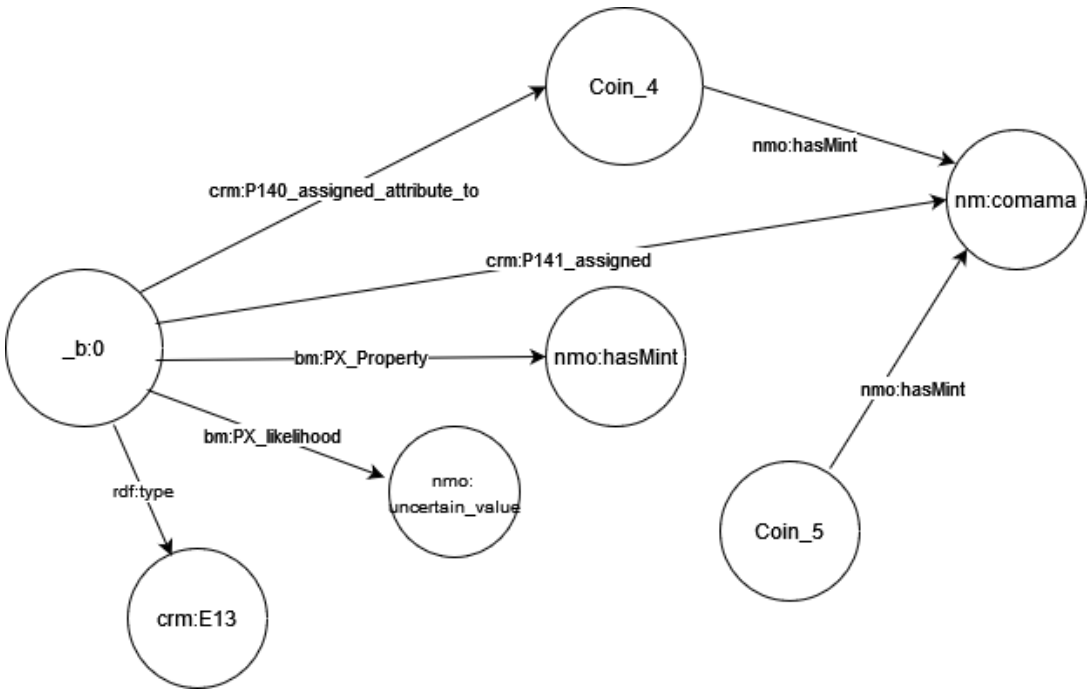

Figure 1: Solution 1 - Graph representation comprising of seven tuples.

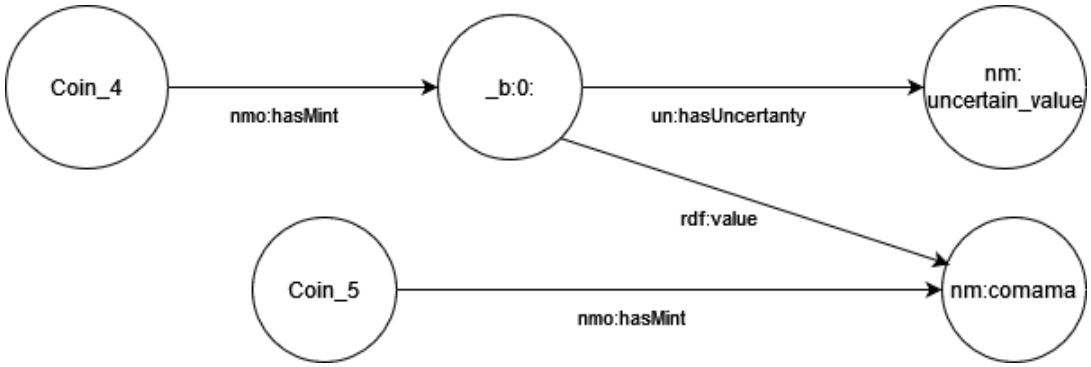

Figure 2: Solution 2 - Graph representation comprising of four tuples.

In either case the representation of the uncertain information will of course require the representation of more triples. For S1 the factor of increase would be twice as high as for S2. We are currently setting up performance tests to analyse the impact of different query types for the solutions. We plan to run these performance tests on Apache Jena Fuseki, although other systems could follow. Since the modelling of uncertain information is not defined, we (and probably others as well) do not export to find this kind of uncertain information. Thus for both solutions the current query to retrieve the information which coins where minted in Comama would be (see listing 1):

```
PREFIX nm: <http://nomisma.org/id/>
PREFIX nmo: <http://nomisma.org/ontology#>
SELECT ?coin WHERE { ?coin nmo:hasMint nm:comama . }
```

Listing 1: SPARQL query Q1

```
Result on S1: Coin_4, Coin_5
Result on S2: Coin_5
```

Listing 2: SPARQL query Q1 results

The problem we see is that solutions like S1 also return the uncertain values, and the user cannot see any differences in the result. Another query to request the mints for the coins in the graph would be:

```
PREFIX nm: <http://nomisma.org/id/>
PREFIX nmo: <http://nomisma.org/ontology#>
SELECT ?coin ?mint WHERE { ?coin nmo:hasMint ?mint . }
```

Listing 3: SPARQL query Q2

```
Result on S1: Coin_4, nm:comama and Coin_5, nm:comama
Result on S2: Coin_4, _b:0 and Coin_5, nm:comama
```

Listing 4: SPARQL query Q2 results

Again, the user cannot distinguish certain and uncertain results for S1. For S2 a blank node (_b:0) is provided where the value is uncertain. But although this could also be due to reasons other than uncertainly, at least the user is informed that there is a value and that they need to write the query in a more specific way to include particular kinds of values (e.g. with uncertainty). Of course SPARQL queries can be generated for each solution such that the result represents the set requested by the user. This will be part of the performance analysis that compares such queries on a semantically equal level.

4 Use Case: Samian Ware

In this Samian Ware Use Case we use the online database Samian Research[6]. The database comprises a quarter of a million potters' stamps on Terra Sigillata. Using LOD-based methods and workflows (Thiery et al., 2020b, #transformation-workflow) we created and published 7.869.468 triples[7] and 306.615 instances under a DPPL License, as well as the underlying ontology as Linked Open Samian Ware (Thiery et al., 2020a).

About 21% of the Samian Ware dataset relating to potters and kiln sites, as well as potsherds attributed to these kiln sites, include strings with expressions containing the keywords AND or OR, indicating vagueness; AND statements combined with a question mark indicate uncertainty:

- Since the same potter may have worked subsequently in different kiln sites (Hartley, 1977), there is a vagueness in the possible attribution of potters to individual sites (Approach A).
- It follows from Approach A that information carriers (potsherds with potters' stamps) also have vague relations to kiln sites (Approach B).
- Vagueness and uncertainties also occur in the context of the determination of vessel fragments (e.g. when only a base fragment of a vessel is preserved) which may be attributable to different possible types of pot forms (Approach C).

To model approaches A, B and C semantically, vagueness is modelled using the Academic Meta Tool (AMT) (Unold et al., 2019). The main idea of AMT is to create a semantically modelled knowledge network using nodes (so called concepts) and weighted edges (named roles) representing a normalised degree of connection between 0 and 1. With this information modelled in the AMT meta-ontology, a domain-specific ontology comprising Role-Chain-Axioms can be created which allows for reasoning using multivalued logics (Unold et al., 2019, section 4.1) (figure 3). AMT uses a quadruple with a blank node, three rdf properties and one special AMT weight property as shown in figure 4.

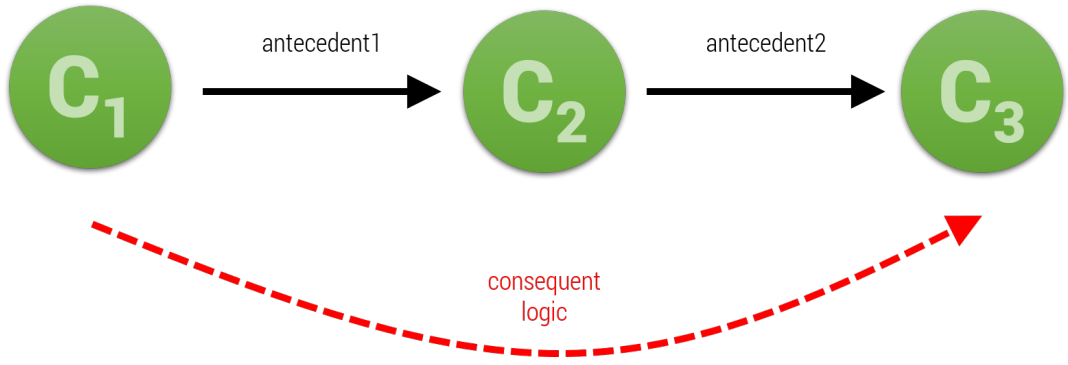

Figure 3: Schematic representation of the role-chain axiom (Rollen-Kettenregel) [Florian Thiery, Martin Unold, CC BY 4.0 via Wikimedia Commons]

6 https://www.rgzm.de/samian
7 status: 4/12/2020

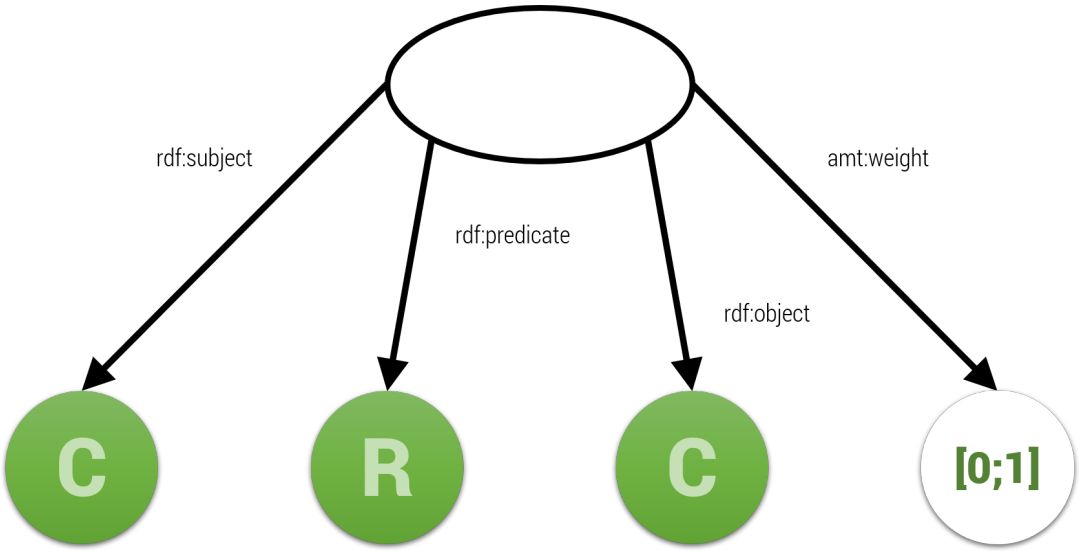

Figure 4: AMT modelling as quadruple [Florian Thiery, Martin Unold, CC BY 4.0 via Wikimedia Commons]

The vagueness and uncertainty expressed in string statements - depending on the keywords AND and OR, as well as a question mark - are transformed into the vagueness-based AMT logic as degree of connection. The resulting formula to calculate the degree of connection - amt weight - is based on the existence of a question mark (maybe in combination with the keyword AND), as well as the total possibilities:

$$amt_{weight} = 100/(n_{questionmark} + n_{possibilities})$$

- no AND or OR: total count of possibilities for all entities IS 1
- AND: total count of possibilities for all entities IS 1
- OR: total count of possibilities IS the count of entities between the OR statements

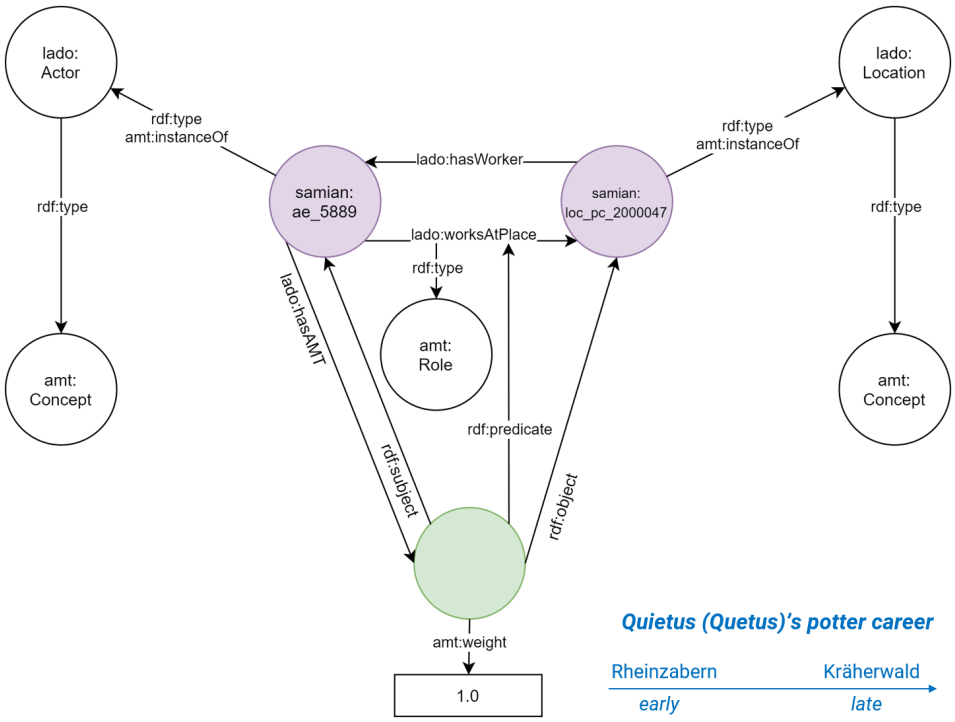

Figure 5: Exemplary schematic representation Approach A [Florian Thiery, Dennis Gottwald CC BY 4.0 via Wikimedia Commons]

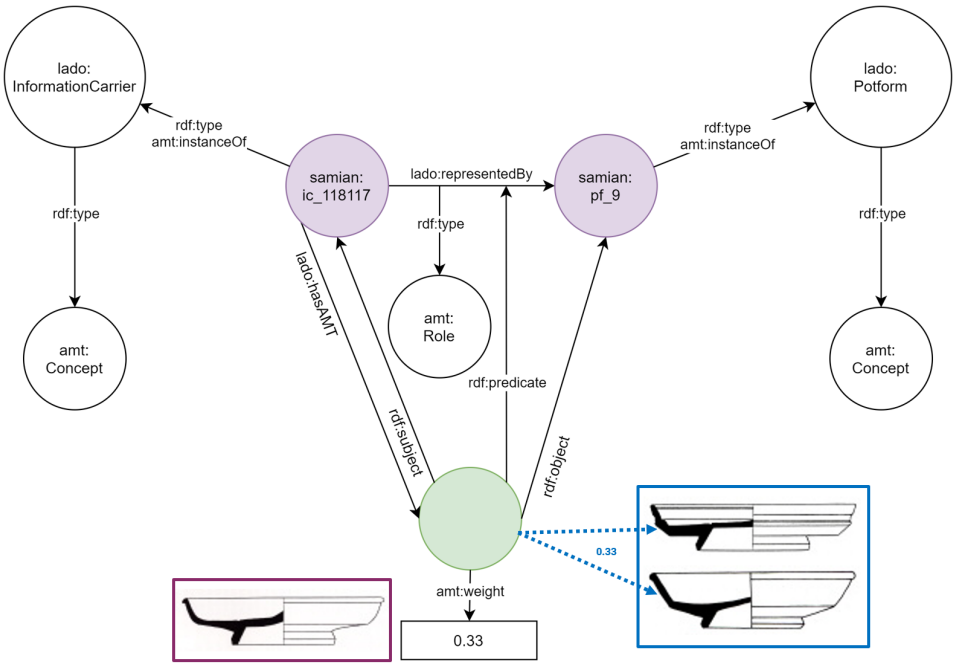

Figure 6: Exemplary schematic representation Approach C [Florian Thiery, Dennis Gottwald CC BY 4.0 via Wikimedia Commons]

Figure 5 shows that the potter Quietus (Quetus) worked in Rheinzabern AND Kräherwald, resulting in a degree of 1 for Rheinzabern. This is based on the fact that Quietus' (Quetus') career started early in Rheinzabern before he moved to Kräherwald. Figure 6 shows that Informationcarrier 118117 can be represented by the Dragendorff pot forms (Dragendorff, 1895) 18 (purple coloured box) OR 15/17 OR 18/31 (blue coloured box), resulting in a degree of 0.33 for 18. This three pot forms, their relations to each other, and to other ceramic typologies are modelled using the Ceramic Typologies Ontology (Thiery and Mees, 2021a). The semantically modelled degrees of connections can thus result in a domain-specific ontology with role-chain axioms:

```
(AE)-[lado:worksAtPlace]->(PC)-[lado:isKilnsiteOf]->(IC)
(AE)=[lado:createdPot/lado:potCreatedBy(ProductLogic)]=>(IC)
```

Listing 5: Role-Chain-Axiom 1

```
(PF)-[lado:worksAtPlace]->(PC)-[lado:potCreatedBy]->(AE)
(PF)=[lado:createdPotWithType(ProductLogic)]=>(AE)
```

Listing 6: Role-Chain-Axiom 2

References

Alexiev, V. (2012). Types and annotations for cidoc crm properties. URL: https://www.researchgate.net/publication/257764696. [Working Paper].

Berners-Lee, T. (2006). Linked Data - Design Issues. URL: https://www.w3.org/DesignIssues/LinkedData.html. [Website].

CIDOC-CRM (2019). Issue 349: Belief Values. URL: http://www.cidoc-crm.org/Issue/ID-349-belief-values. [Website].

Dragendorff, H. (1895). Terra Sigillata. Bonner Jahrbücher 96/97, 18–155.

Hartley, B. (1977). Some Wandering Potters. In J. Dore and K. Green (Eds.), Roman Pottery Studies in Britain and Beyond, Number 30 in British Archaeological Reports, pp. 251–261. Oxford: BAR Publishing.

Mees, A., F. Thiery, K. Tolle, and D. Wigg-Wolf (2021). TRAIL2.2: Evaluation of fuzziness and wobbliness in numismatics and ceramology. DOI: 10.5281/zenodo.5654897. [Working Paper].

Metzger, M. S. (2014). Modeling Uncertainty and Beliefs using Ontologies. URL: http://www.bigdata.uni-frankfurt.de/wp-content/uploads/2021/11/Modeling-Uncertainty-and-Beliefs-using-Ontologies_Melvin_S_Metzger.pdf. [Bachelor Thesis, Goethe-University Frankfurt am Main].

Thiery, F. (2013). Semantic Web und Linked Data: Generierung von Interoperabilität in archäologischen Fachdaten am Beispiel römischer Töpferstempel. DOI: 10.5281/zenodo.292979. [Master Thesis, Fachhochschule Mainz].

Thiery, F. and A. Mees (2018). Taming Ambiguity - Dealing With Doubts In Archaeological Datasets Using Lod. DOI: 10.5281/zenodo.1200111. [Conference Talk, CAA 2018].

Thiery, F. and A. Mees (2021a). Ceramic Typologies Ontology (CeraTyOnt). URL: https://github.com/RGZM/ceramictypologies-lod. [Ontology].

Thiery, F. and A. Mees (2021b). Linked Open Samian Ware – AMT Perspective. DOI: 10.5281/zenodo.5415571. [Working Paper].

Thiery, F., A. Mees, and D. Gottwald (2020a). Linked Open Samian Ware. DOI: 10.5281/zenodo.4305708. [Research Data].

Thiery, F., A. Mees, and D. Gottwald (2020b). Linked Open Samian Ware - Documentation. URL: https://rgzm.github.io/samian-lod/doc/. [Website].

Thiery, F. and M. Unold (2018, January). Academic Meta Tool Ontology - Leonard Edition. URL: http://academic-meta-tool.xyz/ontology. [Website].

Tolle, K. and D. Wigg-Wolf (2015). Uncertainty Handling for Ancient Coinage. In F. Giligny, F. Djindjian, L. Costa, P. Moscati, and S. Robert (Eds.), CAA2014. 21st Century Archaeology. Concepts, methods and tools. Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology., Oxford, pp. 171–178. Archaeopress. OCLC: 907629230.

Unold, M. and F. Thiery (2018). Academic Meta Tool JavaScript Library. DOI: 10.5281/zenodo.3992520. [Software].

Unold, M., F. Thiery, and A. Mees (2019). Academic Meta Tool. Ein Web-Tool zur Modellierung von Vagheit. Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Ausgewählte Beiträge der Tagung 19.-20.01.2018 an der Akademie der Wissenschaften und der Literatur Hg. von Andreas Kuczera / Thorsten Wübbena / Thomas Kollatz. Wolfenbüttel 2019. (ZfdG / Sonderbände), 4. Publisher: Herzog August Bibliothek Version Number: 1.0, DOI: 10.17175/SB004_004.