Heterotoki: Non-Structured and Heterogeneous Terminology Alignment for Digital Humanities Data Producers

Marion Lamé, Perrine Pittet, Federico Ponchio, Béatrice Markhoff, Emilio Sanfilippo

Heterotoki: Non-Structured and Heterogeneous Terminology Alignment for Digital Humanities Data Producers

Marion Lamé
Laboratoire Archéologie et Territoires
CITERES, UMR7324, Université de Tours
Tours, France
marion.lame@univ-tours.fr

Perrine Pittet
Intelligence des Patrimoines
CESR UMR 7323, Université de Tours
Tours, France
perrine.thuringer@univ-tours.fr

Federico Ponchio
ISTI - CNR
CNRS
Pisa, Italia
federico.ponchio@isti.cnr.it

Béatrice Markhoff
LIFAT EA 6300
Université de Tours
Blois, France
beatrice.markhoff@univ-tours.fr

Emilio M.Sanfilippo
Intelligence des Patrimoines - Le Studium
CESR UMR 7323, Université de Tours
Tours, France
emilio.sanfilippo@univ-tours.fr

Abstract

In this paper, we present an online communication-driven decision support system to align terms from a dataset with terms of another dataset (standardized controlled vocabulary or not). *Heterotoki* differs from existing proposals in that it takes place at the interface with humans, inviting the experts to commit on their definitions, so as to either agree to validate the mapping or to propose some enrichment to the terminologies. More precisely, differently to most of existing proposals that support terminology alignment, *Heterotoki* sustains the negotiation of meaning thanks to semantic coordination support within its interface design. This negotiation involves domain experts having produced multiple datasets.

1 Introduction

Data aggregation projects deal with heterogeneous, multiplatform and interdisciplinary datasets. Such projects aim at federating database systems and at managing distributed, heterogeneous, and autonomous databases.

In: Carlo Meghini, Antonella Poggi (eds.): Proceedings of ODOCH 2019 – Open Data and Ontologies for Cultural Heritage – Rome, Italy, 3 June 2019, published at http://ceur-ws.org

In this framework, data come from projects belonging to different fields, provided by different information systems, architectures and technologies, with heterogeneous modelling approaches and different formats. Therefore, these heterogeneous data cannot fully interoperate without being syntactically, structurally and semantically integrated. In such configurations, linking data requires correct and complete ontological models, terminological alignment work and quality control evaluation of the whole infrastructure. Since the '90s scientific literature has accumulated a great number of contributions to foster datasets integration. In the bioinformatics field, a major ontological model is the Gene ontology [42], a network of biological classes describing molecular functions, cellular locations, and processes gene products may carry out.

Regarding alignment, among the most famous ones is the international medical terminologies with the creation of the SNOMED Clinical Terms in 1999 and its current 311,000 concepts [41]. Focusing on quality evaluation, the Ontology Alignment Evaluation Initiative [30] works on semi-automated evaluation of such ontological alignments on scientific datasets.

Benefitting from the wide experience of scientific initiatives, Digital Humanities literature has addressed, more recently, the subject of datasets integration, from a cultural heritage standpoint, through heritage data aggregation platforms (e.g. ARIADNEplus[4], MASA Platform[29], GAMS Platform[19], Isidore[27], HeritageS platform[23]). Such platforms generally host cultural (and sometimes natural) heritage datasets from interdisciplinary research projects. Their goal is to make these data interoperable and thus accessible online by both humans and agents, through web portals, web applications and/or APIs, for research, public engagement or education purposes. However, datasets integration in Digital Humanities, especially in the field of Cultural Heritage, may differ, for some aspects, from scientific datasets integration.

Cultural Heritage, which involves scholarly fields such as History and Archaeology, studies past societies. Research projects in this field produce data on objects of the past at different historical periods, whereas scientific research projects generally focus on some constants of the actual reality of the world, sometimes compared with data from the past. Therefore, in Cultural Heritage, alignment of datasets to standard models asks the question of the temporary as well as cultural distance between concepts behind such data and concepts of standard models. Particular attention has to be paid to take into account concept drift during the alignment process. That is why domain experts for heritage play a crucial role in datasets alignment and this role still cannot rely only on automatization, whereas automatic matching appears partially applicable on scientific datasets alignment. Such alignment implies a mandatory manual step of semantic coordination [12] among domain experts. Also, domain experts can rely on existing methodological research and alignment tools with standard terminologies for their datasets. At the same time, little research has been conducted on alignment of heritage datasets providing, within their interface, support on semantic coordination for domain experts with a focus on heritage data.

This paper addresses the problem of digitally supporting semantic coordination in a terminology alignment process for heritage domain experts in a simple and intuitive way. The rest of this paper is organized as follows. Section 2 presents the state of the art on heritage data, terminologies alignment with a focus on semantic coordination, Section 3 presents our approach and proposes a methodology to address this issue, Section 3.1 presents *Heterotoki*, the online alignment decision support system which implements this methodology. Section 3.2 compares our proposal with related works. Section 4 illustrates it with two case studies. Section 5 summarises the main contributions and introduces future works.

2 Heritage data, Terminologies Alignment and Semantic Coordination

Heritage Data

Digital Humanities projects have been producing many heterogeneous datasets for years. According to [28], Cultural Heritage is a field encompassing a wide range of content that varies drastically by type and properties, but is still semantically richly interlinked. Currently, this content mainly resides in closed databases, distributed nationally and internationally in different locations, and commonly organized by content type – separate databases are typically used for different contents, such as books, artifacts, videos, music etc. Data providers produce metadata to foster digital editing of their data for project applications[37], such as websites with search engines and catalogues. These metadata are used to annotate heritage data and to ease their exploitation by the target applications. In the rest of this paper, these metadata will be referred to as source and target terminologies.

Terminologies Alignment

Terminology is understood in this paper in a broad sense as defined by Guarino in [21]. It can be a flat thesaurus as well as a formal ontology. The alignment process is assumed as a semantic mapping action between terms from a terminology belonging to a source dataset and terms belonging to a target terminology that is usually a Semantic Web based vocabulary.

Alignment of such source terminology addresses two main issues regarding semantic integration with Semantic Web technologies. First, and especially for the disciplines that study cultures from the past, historical data are a representation of an interpretation of the knowledge of a past reality which cannot be, for some aspects, concretely verified nowadays. Therefore, the semantics of the observed past reality and the semantics of the observing reality do not match. Second, the study of a primary source is a practical work, through which domain experts and data providers produce practical metadata.

The domain covered by their data is narrower than the scholarly domain itself and data are produced for specific uses related to their research problems and processes. Consequently, the practical specificity of the produced metadata and the theoretical genericity of the existing target controlled vocabularies used for semantic integration do not match either. For instance, the French concept of "blanc-manger" (in English "blancmange") in cooking recipes of the Middle Ages, which is known by medievalists as a dish prepared with white meat and / or fish has disappeared in modern recipes. In addition, experts are not agreeing on the type of animal protein used as ingredient, and this is part of both a scholarly and a semantic challenge. Instead, the term "blanc-manger" today refers to a sweet dish prepared with milk. Therefore, to annotate medieval recipes we cannot use the term "blanc-manger"[11] found in ontologies such as FoodOn, as it refers to the actual sweet milk dish concept.

These two issues often lead scholarly experts and data providers to avoid reusing standard controlled vocabularies, because the meaning of the terms do not reflect the concepts of the observed reality or the practical specificity of their data. To bridge these gaps, data providers build their own terminology. However, even if data producers are experts in their domain, they are not computer scientists. Therefore, metadata are generally poorly structured and weakly formalized. Depending on their data modelling skills, metadata can be flat list of terms, or more rarely taxonomies and thesauri. However, to make these data interoperable in an interdisciplinary framework, alignment of these homemade terminologies to standard vocabularies is mandatory.

Semantic Coordination

Even though various research efforts towards the definition of formal theories have been proposed over the years, see for instance [12], the Cultural Heritage community still lacks mature technologies to tackle the problem of terminology alignment. Borrowing similar methodological background, semantic coordination is understood here as a task during which domain experts analyse meanings and definitions of both source and target terms, facing the issue of finding an agreement on the meaning of heterogeneous semantic models. In particular, experts needs appropriate interfaces to validate the quality of alignments. Therefore, a tool for communication-driven decision support designed for semantic coordination is needed.

3 The *Heterotoki* Online Tool

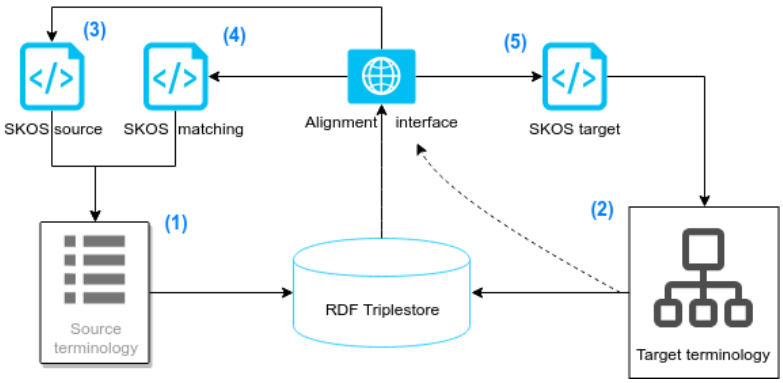

Fig. 1. *Heterotoki* architecture

Heterotoki is a communication-driven decision system in the sense that it enables cooperation, supporting one or more persons working on a shared task [38]. This task consists in creating detailed scholarly mapping relations for some dynamic and operational Linked Data, without ever achieving utopian domain-wide agreement on common vocabularies. In addition, *Heterotoki* feeds cross-searchable repository resource and keeps them updated when experts edit the mapping relations they created in the course of their project. The scholarly alignment does not have to occur at the end of the scholarly work anymore and does not have to be definitive to be operational. Such alignments, and their operational Linked Data, can also occur in the course of this scholarly work and be part of the research process. Doing so, teams also enhance hot spots of scholarly interests or needs, such as other situations not well-known or even more complex than the "blanc-manger" given as an example above. It also provides basic features such as user authentication, to support collaborative work and to keep track of progress, import as well as export facilities. Heterotoki is available at the url http://heterotoki.isti.cnr.it.

3.1 Workflow

Overview

Usually alignment tools take into account an implicit shared meaning between the source and the target, on the contrary, the core of *Heterotoki* usage lies in some semantic coordination validated by experts' consent.

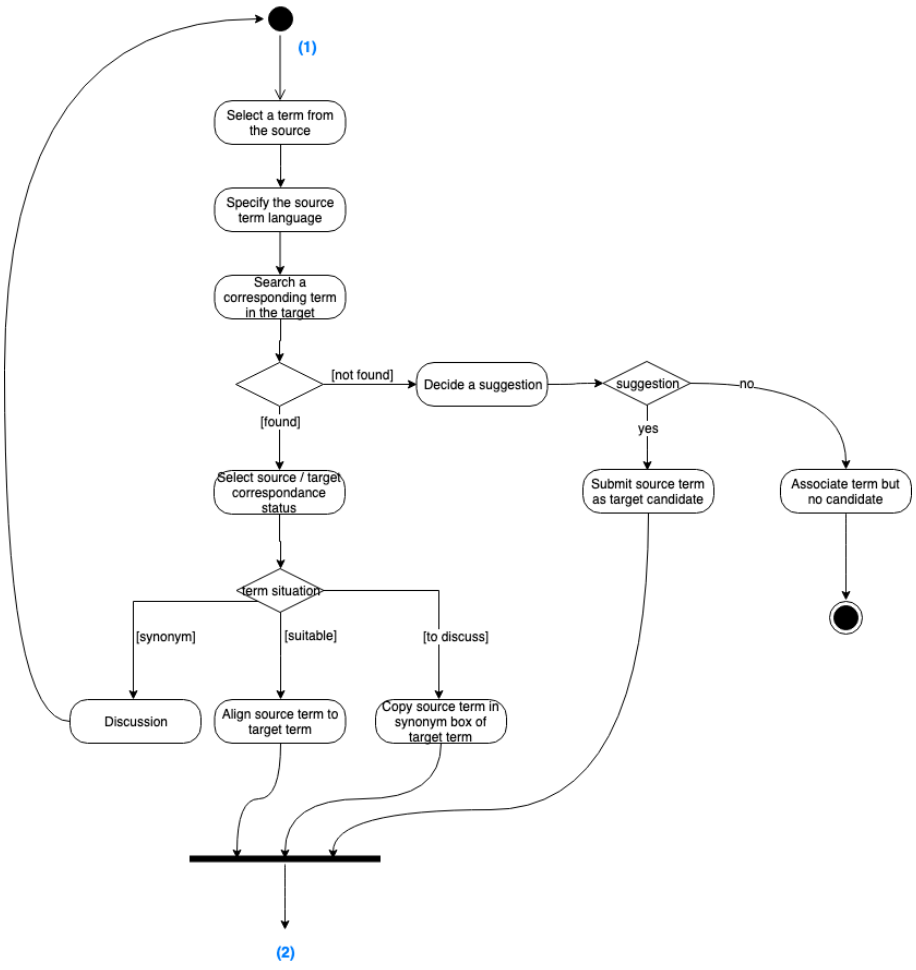

Fig. 2. *Heterotoki* terminology alignment workflow (part 1): term matching

The input for the source terminology can be a flat terminology: a list of terms and definitions exported in CSV, or a more structured terminology in SKOS format or some XML, see Fig. 1 (1). The target terminology is imported in the triplestore in RDF/XML or preferably triples (Turtle) or alternatively directly interrogated using a public API, see Fig. 1 (2). For each term in the source terminology, the required actions are:

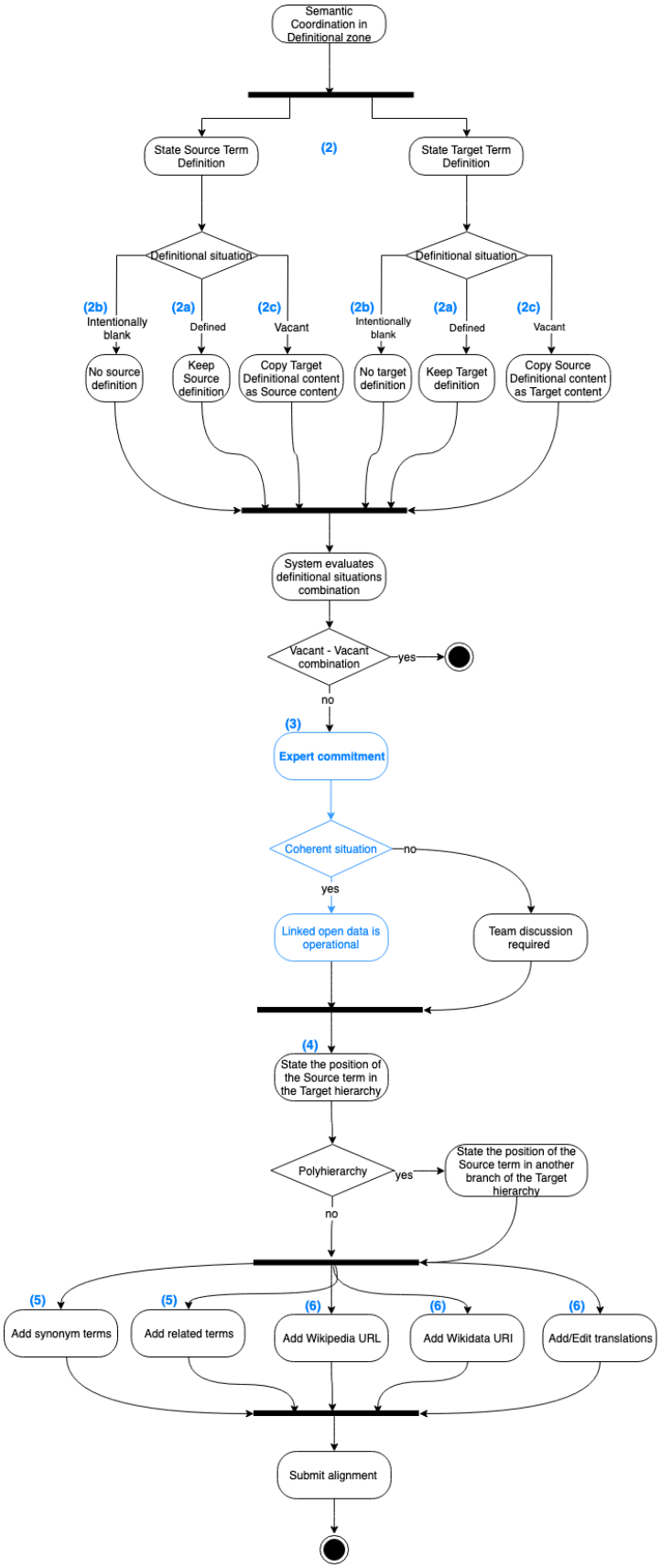

Fig. 3. *Heterotoki* terminology alignment workflow (part 2): semantic coordination.

1. Find a match in the target terminology or suggest one if missing, see Fig. 2;
2. In the definitional zone of the layout, work on definitional blocks (the source and the target one), possibly suggest changes (semantic coordination, see Fig. 3 (2); each definitional block, i.e. the one for the source term and the one for the target, can be left in in one of the following three states:
   (a) Definition in natural language, see Fig. 3 (2a);
   (b) Left intentionally blank - whatever the scholarly reasons are (e.g., a very rich or complex term, see Fig. 3 (2b);
   (c) Incorporation of the content of state (a) or state (b) by mirroring the state of the other terminology, see Fig. 3 (2c) and for an example of layout, with a (b)-(a) combination, see Fig. 6.
3. The expert validates the semantic agreement between the two definitional blocks, declaring the overall situation as *coherent*. This commitment of the expert(s) leads to an operational linked open data (LOD) that can be updated or is reversible at any time for any good scholarly reason, see Fig. 3 (3);
4. Validate hierarchical situation, considering polyhierarchy, if necessary, see Fig. 3 (4);
5. Check for synonyms and related terms, see Fig. 3 (5);
6. Validate translations, sometimes provided by Wikidata, if requested by the project, see Fig. 3 (6).

The output for the source terminology remains a simple list of terms and the corresponding mapping (CSV, JSON) or alternatively some more structured data (in SKOS), see Fig. 1 (3) and (4). The output of the target terminology contains the edited target (in SKOS), see Fig. 1 (5). The terminology maintainers can thus modify and re-import these terminologies.

The output for the users keeps track of the work done and to be done, and categorizes the alignment situation into **seven situations**:

1. Discuss about the two terms (most often morphological differences);
2. Clarify the source definition because an understanding of the desired alignment is necessary;
3. Deal with an inconsistency between the term and the definition in the target;
4. Harmonize the definitional situation by discussing the definitions and by establishing a semantic coordination;
5. Discuss both terms and global definitional situation and resolve inconsistencies (between the terms of the two terminologies and the semantics provided by the definition or definitions or the absence of a definition);
6. Analyze and solve some difficulties in the positioning of the source term in the target structure (with or without polyhierarchy);
7. No specific problem has been identified. Either the term is a candidate to enrich the target or the alignment is validated.

Combinations of these seven situations produce **four distinct team actions** around alignment work to facilitate decision-making, task redistribution and automatic updating of source databases and enrichment of target terminology. In addition, the alignment status information is found both in the lookup table and in the alignment record itself. These four actions are as follows:

1. Re-elaboration: request to handle a problematic situation;
2. Consultation with the target terminology teams: this includes a set of semantic issues concerning the target terminology;
3. Consultation with the source terminology team: this includes a set of issues requiring upstream scholarly reflection;
4. Consensus: this includes validated situations, namely alignment (with or without synonymy) or proposal for enrichment with a new candidate.

3.2 Comparison With Related Tools

Terminology alignment has been widely addressed in Computer Science. More than thirty terminology alignment tools, all of which are not compared here, are still available and in use today (against more than fifty tools in 2014). Many of them are referenced and briefly described on the web page [34]. Among them, we have selected nine existing terminology editors and/or mapping tools usually used in Digital Humanities projects: 3M [1] Ariadne Vocabulary Matching Tool [3] and [10], BBTalk [8], Cultuurlink [15], Ginco [20], OnaGUI [31], OntoME [32], SKOS Shuttle [40] and Vista [6]. We compare these nine tools according to the following eight characteristics:

1. Imports non-formalized data such as data in CSV format.

2. Supports collaborative work, i.e., to allow multiple users to work on the same alignment in progress, keeps track of the status of the proposals and logs the work done.
3. Uses CIDOC-CRM with extensions enabling thesaurus editing, even though lacking specific support for SKOS.
4. Edits terminology: new properties can be added to the source (and/or target) terminology.
5. Focuses on automatic mapping: the tool is designed for large scale automatic mapping and manual editing is a last resort.
6. Supports mapping between SKOS thesauri.
7. Supports full SKOS: not all the tools with editing capabilities support multilingual labeling, *skos:altLabel* or hierarchy, while not all the tools with mapping capabilities support properties such as *skos:closeMatch* or *skos:narrowMatch*.
8. Enables for semantic coordination.

	3M	**OntoMe**	**SKOS Shuttle**	**BBTalk**	**Ginco**	**OnaGui**	**Cultuurlink**	**ARIADNE**	**Vista**	**Heterotoki**
Imports lists										y
Collaborative	y	y	y	y	y					y
CIDOC-CRM	y	y								
Editing	y[1]	y[1]	y	y[2]	y	y				y
Automatic						y	y			
Mapping	y	y		y[3]	y	y	y	y	y	y
Full SKOS			y		y	y				y
Semantic Coord.										y

Table 1. Comparison of alignment tools. Notes: 1) 3M and OntoME can edit ontologies not specialized for SKOS, hence not user friendly for non-ontology experts. 2) Edited proposal to the target BBT ontology only. 3) Mapping only towards BBT

Table 1 presents a synthetic comparison of the tools according to the aforementioned characteristics. *Heterotoki* is compared with tools offering both terminology editing and mapping (like 3M, OntoME, BBTalk, Ginco, Onagui), management of non-formalized data (like Ginco, Cultuurlink), support of collaborative work (like Ginco, SKOS Shuttle, BBTalk) and full SKOS matching relations (like SKOS Shuttle). On the other hand, *Heterotoki* does not support automatic mapping (supported by OnaGUI and Cultuurlink), since it is designed for manual intervention. The tool supports also the alignment with SKOS terminologies but does not support alignment to ontologies like CIDOC-CRM (supported instead by OntoME and 3M). *Heterotoki* was designed for domain experts to formalize and to enrich the expressivity and structure of their own terminology by giving a standard SKOS formalism and semantics within the alignment. This step is for a first interoperability achievement in the heritage data semantic integration process. It intervenes before a secondary ontology alignment step, which is meant to enforce the formalization and expressivity of the alignment for interdisciplinary and logical based inference purposes. As such, *Heterotoki* can work in tandem with 3M to produce a first interoperable and well-defined XML schema input to 3M. Therefore, it semantically prepares heritage data before aligning them to CIDOC-CRM. The last characteristic, namely, semantic coordination, is only handled in *Heterotoki* for now. Even in heritage dedicated tools like 3M, Vocabulary Matching Tool of Heritage Data (used, e.g., in ARIADNEplus), BBTalk, and OntoME, no digital support is intended for semantic coordination validated by experts.

As said in Guarino *et al.* [22], logical languages are eligible for the formal, explicit specification of knowledge, and, thus, for ontologies. Many ontology alignment techniques and tools can be automated because of their high level of formalization [17, 16]. Many mapping tools that align from one data structure to another one exist. In particular they align from CSV files, XML documents or relational databases to ontologies. Examples include Datalift [39] for CSV files, 3M (Memory Manager Mapping) [1] for XML documents, and Ontop [13] for relational databases. BBTalk [8] aligns a source terminology with a specific target terminology, i.e., the *Backbone* meta-thesaurus. In this case, the choice of words is important, but differently from the task *Heterotoki* is dealing with, the coordination takes place at a more abstract and formal level. In contrast, *Heterotoki* is at the beginning of the formalization process which enables data producers to generate a knowledge graph, to contribute to it and benefit from it in return. Vista [6] is more similar to *Heterotoki*, but it is designed for negotiation between terminologists rather than data-producers.

4 Case Studies

Several scholarly editings use *Heterotoki* or its fork called *OpenTermAlign* at different stages of their project, some of them within the ARIADNEplus framework[4]: OUTAGR (Inventaire de l'Outillage Agricole Gallo-Romain) [35], I-CERAMM (Information sur la CÉRAmique Médiévale et Moderne) [24] both supported by the Laboratoire Archaeologie et Territoires, *Tesserarum Sisciae Sylloge* by the Archaeological Museum in Zagreb, for its textual terminology in Latin about Roman textile combined with archaeological terminology. All these projects aim at opening and linking dynamically their scholarly edited and controlled terminology with larger knowledge organization systems of any kind. We present here two case studies. One on textual terminologies extracted from medieval manuscripts, CoReMA (Cooking Recipes of the Middle Ages) [14], the other one, AERBA (Atlas des Établissements Ruraux de la Beauce Antique) [2], uses *OpenTermAlign*, which is designed for archaeological datasets working with multidisciplinary terminologies. *OpenTermAlign* is available at the url http://opentermalign.humanum.fr.

4.1 CoReMa and Alignment Between Terminologies Based on Textual Primary Sources

CoReMA is a Franco-Austrian research program aiming at analyzing and understanding transmission of culinary recipes from the Middle Ages, putting an interdisciplinary focus on the cross-cultural research of medieval cooking recipes and their interrelation. The project studies the transmission of cooking recipes in French and German speaking countries, which includes more than 80 manuscripts and ca. 8000 recipes up to now. It implies analysis of their origin, their relation, and their migration through Europe. The partners provide the expertise to collect, to edit, and to analyze these multilingual texts following up-to-date methodology. For machine aided analysis, a recipe corpus and its metadata are modelled according to international humanities and technical standards. The recipes are enriched with content annotation with terminologies for ingredients, cooking processes, etc.

Within this annotation process of the transcriptions in TEI (Text Encoding Initiative), one of the main tasks of the French team is to extract Ancient French and Latin terms from recipes transcriptions and to align them to actual French language as historians' working language. In other words, given these two terminologies (T), T1 (Ancient French within the scope of a certain historical context) and T2 (contemporary French used by historians as their working language), about the same domain D (in this case culinary tradition), CoReMA domain experts need to establish some correspondence between terms in T1 and terms in T2. Amongst their modeling requirements, it is relevant to attribute to concepts a temporal dimension in order to possibly consider their change over time. In the case of the Ancient French of CoReMA, we apply the code FRO – French, Old (842-ca.1400) – possibly combined with FRM – French, Middle (ca.1400-1600) – according to ISO639-2 standard.

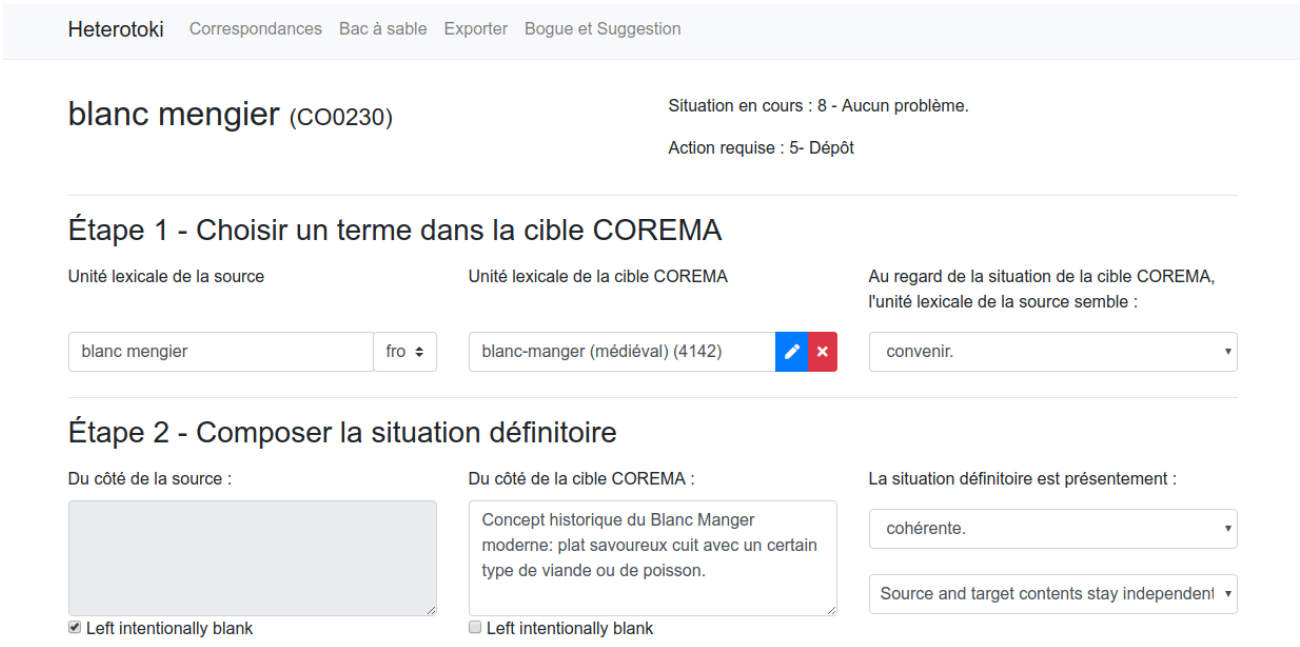

Fig. 4. CoReMa example of semantic coordination with the term "blanc-manger" within *Heterotoki*.

In addition, when dealing with concepts extracted from ancient vocabularies, their definitions have to be provided (as far as possible) according to the larger conceptual structures in which they were originally conceived [9]. Accordingly, when defining a certain concept, its intended meaning – with respect to an ancient vocabulary – has to be preserved and, as much as possible, based on primary sources of information.

Let us consider the example of "blanc-manger" mentioned in Section 2. The concept used in Middle Ages culinary texts and the concept nowadays used are defined in different and incompatible terms, hence they are disjoint and cannot be mapped via taxonomical links (e.g., subsumption). However, an historian may want to relate them to express that "blanc-manger" nowadays used *historically originates from* the concept "blanc-manger" found in ancient culinary text, and labelled with "FRO" language code.

At the current state of the *Heterotoki*, the two concepts can be mapped via the primitive *skos:closeMatch* in order to express a weak link between them – in step 4 "Multi-relational Enrichment (polyhierarchy and others)", text area "Associated terms"[1], see Fig. 5. Further work is however necessary to specify and strengthen the intended semantic of this SKOS modeling element when tuned to historic studies.

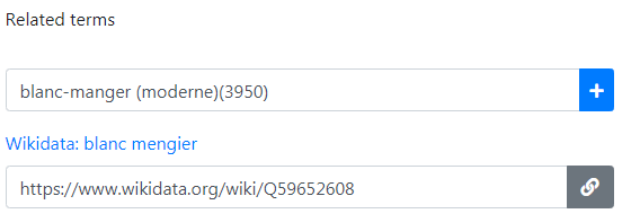

Fig. 5. Related match to the modern concept of "blanc-manger" within CoReMa and exact match to the medieval concept in Wikidata.

4.2 AERBA and Alignment Between Terminologies Based on Archaeological Primary Sources

OpenTermAlign (OTA) is a customized version of *Heterotoki*. OTA has been implemented as part of the collaborations and developments of the Consortium Mémoires des Archéologues et des Sites Archéologiques (MASA), in order to make archaeological data interoperable. Created in 2014, the MASA Consortium is certified by the TGIR Huma-Num, a French "Very Large Research Infrastructure". The consortium combines several French teams and institutions working in the field of archaeology. Its upcoming platform is called *OpenArchaeo*.

The OTA tool has been customized using the 24 official languages of the European Union. It has been tested using a source terminology consisting of a cluster of 600 index terms from four different archaeological databases. (ArSol [5], AERBA [2], OUTAGR [35] and I-CERAMM [24]). It has also been tested with a specialized target terminology that serves as a reference for French-speaking archaeology, the PACTOLS thesaurus [36] set up by the Fédération et Ressources sur l'Antiquité network (FRANTIQ [18]. The development of the thesaurus began in 1987 and now includes more than 50,000 concepts (each with an ARK - Archival Resource Key - URI), organized into seven domains. The names of the domains give the thesaurus its name PACTOLS (Peuples, Anthroponymes, Chronologie, Toponymes, Œuvres, Lieux, Sujets). The thesaurus complies with the ISO 25964 standard for multilingual thesauri and their organization and interoperability. Its alignment with the DARIAH Backbone meta thesaurus [7], with GeoNames, and with Wikidata is a work in progress.

In the case of AERBA, and other LAT projects, an expert works with the same number of choices limited to the four steps of *Heterotoki*. For example, in step 1 (see Fig. 6), the term may be a new candidate for the target terminology, compatible with an existing term in the target terminology, a synonym of an existing term in the target terminology, incompatible with the target terminology, or simply associated with a position without candidating and without improving the target terminology. For the term "grange" ("barn" in English), in step 2 "Definitional zone", the expert coordinates the semantics making two choices:

– Deliberately not defining AERBA term "grange" within the scope of AERBA and deliberately not using the PACTOLS's definition within AERBA ("Source et target demeurent indépendants l'un de l'autre").

[1] It is worth noting that *skos:related* does not have a precise meaning and is indeed used with different intended meanings in alignment practices.

Fig. 6. AERBA example of semantic coordination with the term "grange" within the *OpenTermAlign* fork of *Heterotoki*.

– Committing in linking both datasets by describing the overall definitional situation as *coherent*, exactly as it is.

Doing so, AERBA and PACTOLS are connected, updated and, at the same time, the semantic relationship between both is under scholarly control. The result of the mapping is shown online by OTA in SKOS format. This is then read by the AERBA website which links their resources dynamically to PACTOLSl's ARK and reflects changes in the mapping. The list of candidates for new terms in PACTOLS is available, also in SKOS format, to *OpenTheso* [33], the software used by the FRANTIQ community to validate and enrich the PACTOLS thesaurus.

The result of the mapping is exposed online by OTA in SKOS format which is then read by the AERBA website to dynamically link their resources to ARK of PACTOLS and reflects the changes in the mapping. At the same time the list of candidates for new terms in PACTOLS is available, in SKOS format, for OpenTheso [33], the software used to validate and enrich the PACTOLS thesaurus by the library science community federated by Frantiq.

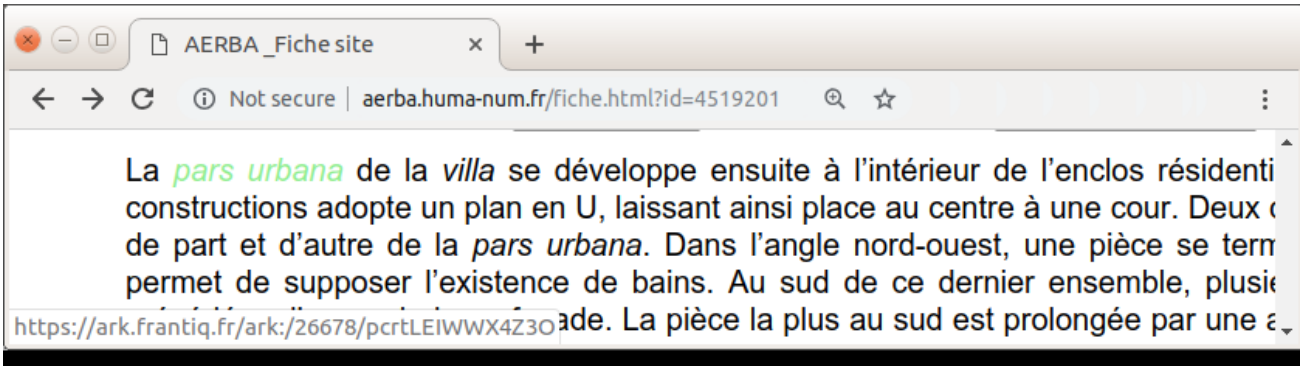

Fig. 7. AERBA result of dynamically updated linked open data with PACTOLS by *OpenTermAlign*. Example with the term "pars urbana".

Out of the 600 aligned archaeological terms, 300 candidates are proposed for the enrichment of PACTOLS. Of the remaining 300 terms, about 70% do not pose any particular problem and the remaining 30% require team consultations to adjust, clarify and achieve a satisfactory and functional alignment both semantically and technically. Again, such validation (or non-validation) enables an operational and dynamic LOD that works as an open (or a closed gate) and facilitates library science work while the scholarly project is following its own rhythm and respect the natural uncertainty inherent to any research work.

OpenTermAlign respects some archaeologists' requirements by: 1) Allowing the preservation and characterization of source terminology in foreign languages, especially those of the society under study (in this case, Latin); 2) Harvesting the translations of terms in the corresponding Wikidata concept pages (for multilingual enrichment of the source and target terminologies); 3) Interacting with the endpoint of the target multidisciplinary thesaurus (live query); 4) At the local level, installing the same thesaurus (to address specific network problems); 5) Harvesting its ARK identifiers.

The OAT is built on *Heterotoki* and has been tested with other target terminologies of the thesaurus type offering the same technical guarantees. Functionalities enabling alignment with several target vocabularies are under study (including *Iconclass* [25] and *Inventaire Général* [26]).

5 Conclusion

In this paper we presented *Heterotoki*, an online communication-driven decision support system. The tool is explicitly designed for heritage data providers who need to align their terminologies to formalize SKOS vocabularies for semantic integration in the context of data aggregation projects. *Heterotoki* does not intend to replace existing ontology alignment frameworks such as 3M. *Heterotoki* should be used as a preliminary step in the structural and semantic heterogeneity resolution process. Such step enable heritage data providers to produce well-defined and well-formalized terminologies. This preliminary step is mandatory to ensure the original semantics of their data are preserved during the process, even more if their terminologies are weakly structured. The semantic coordination and commitment within the interface helps domain experts in managing heritage data specificity, such as temporal or cultural concept drift.

The development of *Heterotoki* is in progress and the tool is currently used by both archaeological and historical data providers in the context of two data aggregation projects: (1) the MASA platform, and (2) HeritageS, managed by the ARD Intelligence des Patrimoines research project and supported by the CESR laboratory (Centre d'Études Supérieures de la Renaissance, CNRS). Other data providers from different cultural and natural heritage fields are expected to use *Heterotoki* in the context of the HeritageS project. Having a larger pool of diversified users will provide us with a pertinent amount of feedbacks, especially on the semantic coordination question. Benefitting from these feedbacks, future work will focus on leveraging the collaborative and descriptive aspects of the semantic coordination, the alignment evaluation, and on improving the description of the changes management for the target terminology. For this last aspect, the XML format is not suitable to express changes to a SKOS thesaurus. A more formal way to propose changes could be expressed as two lists of triples to be removed and added from the triple store. The proposed changes will be expressed in SPARQL language since most standard controlled vocabulary tools have a SPARQL endpoint. Each SPARQL query is accompanied by a human-readable description of the proposed change so that the target terminology administrators can validate it.

Acknowledgements

This research has been funded mainly by a postdoctoral grant from UMR 7324 CITERES – Laboratoire Archéologie et Territoires, Université de Tours, CNRS and partially by the ARIADNEplus project and ARD Intelligence des Patrimoines. It also received help from the FRANTIQ network and support from the Parthenos project. This work benefited from the scholarly content of the digital *Atlas des Établissements Ruraux de la Beauce Antique* (AERBA), directed by Alain Ferdière and Alain Lelong. Thanks are due as well to *Cooking Recipes of the Middle Ages*, program CoReMa ANR FWF (CESR / ZIM-ACDH), led by Bruno Laurioux and Helmut W. Klug. The authors would also like to acknowledge Olivier Marlet and Rémi Ossant for their scholarly digital editing of AERBA within the framework of the *OpenArchaeo* platform, Denise Ardesi and Corentin Poirier Montaigu, for sharing their user experience and feedback on cooking terminologies of the Middle Ages. Ceri Binding, Blandine Nouvel, Miled Rousset, and Magali Mangin provided expertise in knowledge organization, technical collaboration and professional terminology.

References

1. 3m, memory manager mapping. https://github.com/isl/3MEditor
2. Aerba, atlas des établissements ruraux de la beauce antique. https://masa.hypotheses.org/aerba
3. Ariadne vocabulary mapping. https://heritagedata.org/vocabularyMatchingTool/
4. Ariadneplus. https://ariadne-infrastructure.eu
5. Arsol, archives du sol. http://arsol.univ-tours.fr
6. Axaridou, A., Konsolaki, K., Theodoridou, M., Kozlov, A., Haase, P., Doerr, M.: Vista: Visual terminology alignment tool for factual knowledge aggregation. In: Proceedings of the Third International Workshop on Semantic Web for Cultural Heritage co-located with the 15th Extended Semantic Web Conference, SW4CH@ESWC 2018, Heraklion, Crete, Greece, June 3, 2018. (2018), http://ceur-ws.org/Vol-2094/paper2.pdf
7. Backbone thesaurus. http://www.backbonethesaurus.eu
8. Bbtalk, backbone thesaurus management tool. http://www.backbonethesaurus.eu/BBTalk
9. Betti, A., van den Berg, H.: Modelling the history of ideas. British Journal for the History of Philosophy **22**(4), 812–835 (2014)
10. Binding, C., Tudhope, D.: Improving interoperability using vocabulary linked data. International Journal on Digital Libraries **17**(1), 5–21 (2016)
11. Blancmange class in foodon. http://purl.obolibrary.org/obo/FOODON%5F03317017
12. Bouquet, P., Serafini, L., Zanobini, S.: Semantic coordination: A new approach and an application. In: Proceedings of the Second International Conference on Semantic Web Conference. pp. 130–145. LNCS-ISWC'03, Springer-Verlag, Berlin, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39718-2_9, http://dx.doi.org/10.1007/978-3-540-39718-2_9
13. Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., Rodriguez-Muro, M., Xiao, G.: Ontop: Answering sparql queries over relational databases. Semantic Web **8**(3), 471–487 (2017)
14. Corema, cooking recipes of the middle ages. https://corema.hypotheses.org
15. Cultuurlink, connecting cultural heritage. http://cultuurlink.beeldengeluid.nl/
16. Da Silva, J., Revoredo, K., Baiao, F.A., Euzenat, J.: Interactive ontology matching: using expert feedback to select attribute mappings. In: 13th ISWC workshop on ontology matching (OM). pp. 25–36. No commercial editor. (2018)
17. Euzenat, J., Shvaiko, P., et al.: Ontology matching, vol. 18. Springer (2007)
18. Frantiq, fdration et ressources sur l'antiquit, https://www.frantiq.fr
19. Gams, geisteswissenschaftliches asset management system. https://gams.uni-graz.at/
20. Ginco, gestion informatise de nomenclatures collaboratives. https://github.com/culturecommunication/ginco
21. Guarino, N.: Ontology and terminology. In: Institute for Cognitive Science and Technology. National Research Cuncil Trento (2006)
22. Guarino, N., Oberle, D., Staab, S.: What is an ontology? In: Handbook on ontologies, pp. 1–17. Springer (2009)
23. Heritages, chantier transversal. https://intelligencedespatrimoines.fr/chantier-transversal/
24. I-ceramm, information sur la céramique médiévale et moderne. http://iceramm.univ-tours.fr
25. Iconclass, http://www.iconclass.org
26. Inventaire général, http://data.culture.fr/thesaurus/page/ark:/67717/T96
27. Isidore, search assistant in humanities and social sciences. https://isidore.science/about
28. Mäkelä, E., Hyvönen, E., Ruotsalo, T.: How to deal with massively heterogeneous cultural heritage data–lessons learned in culturesampo. Semantic Web **3**(1), 85–109 (2012)
29. Mémoires des archéologues et des sites archéologiques (masa). https://masa.hypotheses.org/
30. Ontology alignment evaluation initiative. http://oaei.ontologymatching.org/2018/results/index.html
31. Onagui. https://github.com/lmazuel/onagui/wiki
32. Ontome, ontology management environment. http://ontologies.dataforhistory.org/
33. Opentheso, gestionnaire de thésaurus multilingue. https://masa.hypotheses.org/99
34. Other editor and alignment tools. http://www.mkbergman.com/2129/30-active-ontology-alignment-tools/
35. Outagr, inventaire de l'outillage agricole gallo-romain. http://outagr.huma-num.fr/Outagr/
36. Pactols, peuples, anthroponymes, chronologie, toponymes, œuvres, lieux, sujets, https://pactols.frantiq.fr
37. Pierazzo, E.: Digital scholarly editing: theories, models and methods. Routledge (2016)
38. Power, D.J.: Decision support systems: concepts and resources for managers. Greenwood Publishing Group (2002)
39. Scharffe, F., Bihanic, L., Képéklian, G., Atemezing, G., Troncy, R., Cotton, F., Gandon, F., Villata, S., Euzenat, J., Fan, Z., et al.: Enabling linked data publication with the datalift platform. In: Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence (2012)
40. Skos shuttle, thesaurus management as a (collaborative) service. https://skosshuttle.ch/
41. Snowmed ct worldwide. https://www.snomed.org/snomed-ct/sct-worldwide
42. The gene ontology, http://geneontology.org