Semantic Web 12 (2021) 279–294
DOI 10.3233/SW-200416
IOS Press

A challenge for historical research: Making data FAIR using a collaborative ontology management environment (OntoME)

Francesco Beretta
Laboratoire de recherche historique Rhône-Alpes, CNRS – Université de Lyon, 14 avenue Berthelot, 69363 Lyon cedex 07, France
E-mail: francesco.beretta@cnrs.fr

Editors: Antonis Bikakis, University College London, UK; Beatrice Markhoff, University of Tours, FR; Alessandro Mosca, Faculty of Computer Science, Free University of Bozen-Bolzano, IT; Stephane Jean, University of Poitiers – ENSMA, FR; Eero Hyvönen, University of Helsinki, Aalto University, Finland
Solicited review: Three anonymous reviewers

Abstract. This paper addresses the issue of interoperability of data generated by historical research and heritage institutions in order to make them re-usable for new research agendas according to the FAIR principles. After introducing the *symogih.org* project's ontology, it proposes a description of the essential aspects of the process of historical knowledge production. It then develops an epistemological and semantic analysis of conceptual data modelling applied to factual historical information, based on the foundational ontologies *Constructive Descriptions and Situations* and DOLCE, and discusses the reasons for adopting the CIDOC CRM as a core ontology for the field of historical research, but extending it with some relevant, missing high-level classes. Finally, it shows how collaborative data modelling carried out in the ontology management environment OntoME makes it possible to elaborate a communal fine-grained and adaptive ontology of the domain, provided an active research community engages in this process. With this in mind, the *Data for history* consortium was founded in 2017 and promotes the adoption of a shared conceptualization in the field of historical research.

Keywords: FAIR principles, historical research data interoperability, Cultural Heritage, factoid data model, Constructive Descriptions and Situations, DOLCE, CIDOC CRM, OntoME, *dataforhistory.org*

1. Introduction

The FAIR principles, "make data Findable, Accessible, Interoperable, and Re-usable",[1] stem from the vision inherent to the *open science* movement of being able to re-use data generated by research in the context of new research agendas: "There is an urgent need to improve the infrastructure supporting the re-use of scholarly data" [36]. Researchers are therefore invited not only to publish articles and books, but also to provide the data that has enabled them to establish their research results.[2] While the 'F', 'A' and 'R' articles of the FAIR principles are relatively easy to implement – as they refer to "technical" recommendations about the persistence of identifiers, the provision of rich metadata, data access rules and their user/re-user licences, etc., the 'I' (*Interoperable*) in FAIR poses a significant challenge. This is particularly true for historical

[1] Cf. *Guidelines on FAIR Data Management in Horizon 2020*, Version 3.0, 26 July 2016, as well as https://www.force11.org/group/fairgroup/fairprinciples.

[2] See for instance the journals *Scientific data* published by the Nature group and *Research Data Journal for the Humanities and Social Sciences* published by Brill.

research, and more broadly for data produced in the field of Cultural Heritage and heritage institutions.

The first paragraph of the 'I' article advises researchers, during the production of data, "[to] use a formal, accessible, shared, and broadly applicable language for knowledge representation[3]". This principle may be further clarified by noting the established definition of ontology in the computer science sense: "An ontology is a formal explicit specification of a shared conceptualization of a domain of interest" [21]. It is therefore a question of adopting, for a given academic discipline, a broadly shared data model expressed using a formalization that is compatible with technologies used on the semantic web. This principle also applies to the second paragraph which refers to controlled vocabularies (concept taxonomies, gazetteers, authority files). These are an indispensable complement to an ontology understood as a conceptual model of the world. Finally, the third paragraph of the article recommends the use of explicit and standardized terms when referring to other resources.[4]

We may wonder to what extent this vision may be applied to data produced by historical research and, more broadly, data issuing from the field of cultural heritage (galleries, libraries, archives, museums). Indeed, given the vast wealth of data produced in these two fields, the relevance of making data interoperable is clear: so that one community may benefit from the data produced by another and vice versa, thereby improving the quality and the volume of data available, both in terms of research and the documentation of items being conserved.

However, since data production will by default be linked to a specific line of inquiry, doesn't this render it non-useable for other research agendas? Aren't the vocabulary and concepts inevitably linked to a particular historical era or discipline, and therefore unable to be transposed into other fields? Given these observations, how is it possible to adopt a quasi-universal *conceptualization*, an ontology, in order to ensure the interoperability of the data produced by historians? The vision of promoting the re-use of data stemming from various research agendas and programmes, in interaction with those from heritage institutions, raises questions that are both scientific and semantic, and calls for an in-depth consideration of the process of *knowledge production* in the field of historical research and of methods of semantic data modelling tailored to achieve these objectives.

These issues, already discussed in the field of digital humanities [15] and notably in digital history [3], have become more pressing in recent years due to the proliferation of semantic web technologies [28] and projects producing huge amounts of data, such as the *Time machine* large-scale research initiative[5] [29]. They have been pertinent since the beginnings of the *symogih.org* project (*Système modulaire de gestion de l'information historique*), first developed in 2008 by the Digital history research team at the *Laboratoire de recherche historique Rhône-Alpes* (LARHRA – CNRS/Universités de Lyon et Grenoble). The *symogih.org* project sprang from the desire to pool the data produced by the researchers of the same laboratory, yet active in different research areas, within a collaborative virtual research environment (VRE) in order to share the data to conduct further research and be later re-used by doctoral students and researchers in new projects.

From the very start, collaborative conceptual modelling has therefore been at the core of this endeavour. In addition, the development of the semantic web called for an extension of this vision and a commitment to making the data publicly available, seeking at the same time to ensure interoperability. This evolution has required a confrontation with semantic methods and technologies and has led to the initiative of a *Data for history consortium*, launched in 2017. This is a proposal addressed to the entire community of data producers and/or consumers, in the field of historical research and cultural heritage, for discussion and common work on semantic methodologies, and the development a *communal ontology* that would facilitate data interoperability.

The paper is organized as follows. The second part will present the *symogih.org* project's ten years' experience in data pooling and re-use, but also the limitations of the adopted collaborative pattern-based modelling approach. In the third part, the phases in the production of historical knowledge will be highlighted, starting with the discussion of the *factoid* model developed by the prosopography projects at King's College London. Part four will propose an epistemological and semantic analysis of the process of conceptual data modelling applied to factual historical information,

[3] https://www.go-fair.org/fair-principles/i1-metadata-use-formal-accessible-shared-broadly-applicable-language-knowledge-representation/

[4] https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/

[5] https://www.timemachine.eu/

based on very relevant inputs from the foundational ontologies *Constructive Descriptions and Situations* (*c.DnS*) and DOLCE.

In part five, the reasons for creating an extension of the CIDOC CRM for the field of historical research will be presented, based on the experience of the *symogih.org* project and integrating modelling elements of c.DnS and DOLCE. Part six will present OntoME, an "Ontology management environment" for collaborative and dynamic management of the aforementioned extension, and the *Data for History* consortium, created in 2017 in order to promote the development of a community sharing and maintaining a common, extensible conceptualization of historical research data as a prerequisite of their interoperability and re-use. In conclusion, the questions that remain open will be summarized and the conditions for the realization of this vision outlined.

2. The *symogih.org* project and its collaborative data model

The *symogih.org* project first came into being in 2008 when several historians from the LARHRA sought to pool the structured data acquired during their research in order to enable it to be re-used in subsequent projects. This approach follows the rationale of data curation, understood as the enrichment and gradual improvement of research data in order to guarantee its quality, accessibility and preservation.[6] For example, the data produced over the course of the *Patrons de France* project,[7] which was financed for three years by the French *Agence nationale de la recherche* and focused on French businessmen (XIXth–XXth centuries), continues to be enriched and used by researchers and students, notably as part of the SIPROJURIS project[8] which focuses on law professors in France from 1804 to 1950. These two projects each had their own dedicated website, but the collection of data was based on a single collaborative information system, the *symogih.org* virtual research environment (VRE), that encouraged the exchange and re-use of the data.

The VRE was designed to be modular in order to integrate new modules based on existing and standardized technologies (e.g. a system of spatial data management, GEO-LARHRA,[9] and an environment for semantic annotation and text editing in XML/TEI formats[10]), or services made available by other organisations. It is thus possible to easily realize heuristic analyses of data using the RStudio instance deployed by the French humanities research infrastructure HumaNum.[11]

Each project can have a dedicated website to present its own data in a customized layout while a general *symogih.org* website publishes all the data present in the repository under the *Creative Commons Attribution-ShareAlike 4.0 International* licence. In addition, a SPARQL endpoint makes it possible to directly query the portion of data that the researchers have decided to publish in RDF format using the ontology that will be described below.[12] According to the linked open data (LOD) rationale, the instances present in the VRE are connected to authority files and public reference bases. This happens manually in the graphical interface or, as it was tested in a pilot alignment experiment carried out using the IdRef[13] authority file, with a semi-automatic workflow.[14] This alignment enabled the enrichment of the SIPROJURIS data with the list of each professor's publications by retrieving them from the records of the ABES SUDOC library catalogue.[15]

A growing number of projects, both French and European (over 60 users and around 15 projects), used or are using this VRE to produce and pool their data. Two PhD theses were successfully defended which had produced their data in the *symogih.org* VRE.[16] Despite the fact that this infrastructure evidently ensures the long-term availability of the data beyond the projects' funding period, its connection to international authority files and its re-usability for new projects, various institutional and research policy issues have jeopardized the maintenance of the VRE at LARHRA. A know-how transfer agreement between the CNRS and the

[6] https://en.wikipedia.org/wiki/Data_curation
[7] http://www.patronsdefrance.fr/
[8] http://siprojuris.symogih.org/
[9] http://geo-larhra.ish-lyon.cnrs.fr/ – cf. [9].
[10] http://xml-portal.symogih.org/ – cf. [10].
[11] A few examples on https://frama.link/phn-shiny.
[12] http://symogih.org/?q=rdf-publication
[13] https://www.idref.fr/
[14] This project in collaboration with François Mistral, head of authority control at the *Agence bibliographique de l'enseignement supérieur* (ABES), used data from the SIPROJURIS project, cf. https://punktokomo.abes.fr/2019/09/10/labes-soutient-la-recherche-en-humanites-numeriques-2-retours-sur-une-cooperation-fructueuse-avec-le-larhra/.
[15] Cf. e.g. http://siprojuris.symogih.org/siprojuris/enseignant/44315 ("Bibliographie externe" tab).
[16] http://symogih.org/?q=news

KleioLab company (Basel) has made it possible to develop a new VRE, Geovistory.[17] *symogih.org* project's data could be transferred to the new VRE in the future, and be re-used for new research agendas. Meantime, a RDF export of the whole dataset will be deposited on a long term preservation platform. The use of explicitly defined identifiers, well defined and interlinked resources and an explicit, documented data model will allow re-use of this data which comply with the FAIR principles but will be frozen.

Much more relevant for this paper is the *symogih.org* project's vision about a generic and open modelling approach on which we will focus now. From the outset of the programme, particular attention was paid to conceptualizing an open data model capable of being adapted to any type of historical information, regardless of the research topic or period being studied. At the same time the model was designed in connection with existing authoritative models such as GEDCOM, a *de facto* standard for exchanging genealogical data. This approach aimed to guarantee interoperability between the data created in the VRE and those from other producers.

On the methodological side, the standards in the field of database modelling where applied [2,5,33]. On the epistemological side, a generic approach was chosen as the foundation of the information system in order to implement two fundamental principles in digital history research. Firstly, a clear separation was established between the production of data and the research agenda that spurred its collection. Even if any data collection originates from a line of inquiry, it is nonetheless necessary to model the information stored in the research environment in the most objective manner possible in order to enable its re-use for new research. This method makes it possible to avoid the bias often introduced in data production when coding replaces neutral fact-based modelling. The latter is the condition of data re-usability in the future. Coding and analysis are carried out in a second, distinct stage, when the produced data are queried using SQL or SPARQL in order to answer the questions relevant for the research agenda. This might involve reconstituting the proceedings of a trial, creating a spatial representation of a series of events, or comparing the careers of people belonging to specific groups or classes [6]. These classifications should not be defined *a priori*, and projected into the data during its production, but emerge as the result of the heuristic process that goes along with the analysis and interpretation of the data.

Secondly, it is essential to proceed with information fragmentation; i.e. undertaking the process of breaking down complex situations into elements that correspond to simple, independent propositions which ideally cannot be further broken down themselves [11,16]. This fragmentation process must be explicitly documented, identifying the meaning of each proposition as well as the role of each object involved. With this approach, a distinction should be drawn between an event or phase in its entirety, such as a congress or a battle, and the multiple events and situations represented by the activities, or presence / absence of various persons at various moments of the same event, since the duration and continuity of each person's presence may be varied and significant for any historical reconstitution.

In order to achieve these goals, the *symogih.org* project used a generic database model for the implementation of the VRE, separating it from its semantic component: this was implemented in the form of modelling patterns that were themselves stored as data instances. The information system thereby enables researchers to create new aspects of the model to match the needs of new research agenda, and contribute to enriching the semantics of the whole VRE, without modifying the database model [32]. The modelling patterns were discussed by the users' community, then validated in order to become usable by everyone and published on the main *symogih.org* website.[18] More than 150 such modelling patterns were produced in the course of the project, covering different aspects of social, economic, intellectual and religious life. The more generic ones, that have proven to be of interest to several projects, happened to be used intensively, such as the one that models the carrying out of a social function by a person.[19] Other ones have been used only in specific research contexts.

This collaborative perspective open to semantic enrichment of the data model explains the early interest of the *symogih.org* project in the semantic web [8]. A rewriting of the VRE generic model using the RDFS vocabulary was begun in 2013 in order to reproduce it as an ontology, to enable data publication on a SPARQL endpoint and to allow its alignment with resources available in the infrastructure of libraries, archives and museums.

[17] https://www.geovistory.com/

[18] http://symogih.org/?q=types-of-informations-list

[19] http://symogih.org/?q=type-of-information-record/7

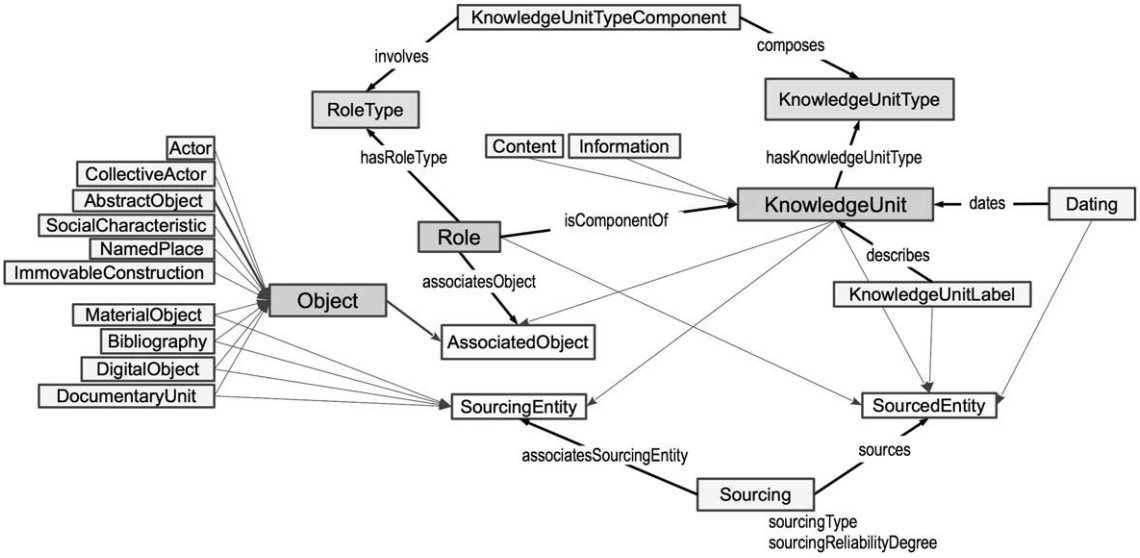

Fig. 1. Ontology of the symogih.org project – version 0.2.1.

Figure 1 represents the core elements of the ontology. In the centre are the three main classes: *Object, Role* and *KnowledgeUnit*. The first covers all entities which have a distinct identity that is stable in time despite any transformation of their characteristics or appearances. This refers to physical objects (such as a person, house or manuscript) or abstract objects (such as a concept, a bibliographical record or an occupation). The *KnowledgeUnit* class models a piece of information understood as a representation of the relations between objects, situated in time and space. As indicated above, the information should be fragmented and designed in the most objective manner possible in order to enable its re-use in new research contexts.

By way of example, let us consider the proposition "In 1592, Galileo Galilei was hired by the University of Padua, where he taught mathematics until 1610." We may extract from it a piece of information or knowledge unit that represents the interrelationship, during a given period of time, of a person (Galileo Galilei), an organisation (the University of Padua) and a discipline (mathematics). A pattern that models the 'teaching' is created as an instance of the *KnowledgeUnitType* class and published on the *symogih.org* website.[20] This pattern specifies the meaning of the data produced and enables VRE users to understand the data semantics. The participation of each object in the information unit is modelled with the class *Role* which reifies its association with the information and thus makes it possible to specify the nature of the participation with a *RoleType* and to eventually add qualifications, metadata, etc. to the *Role* instances.

It should be noted that some other piece of information could have been extracted or deduced from the same proposition, such as the fact that Galileo now resided in the city of Padua, or that he was hired by the University, or that he held the title of professor regardless of whether or not he was effectively teaching. Although the data should be produced in the most factual manner possible in order to be re-usable, the choice of the kind of information that has to be extracted from the source depends on the line of enquiry. Data are always constructed according to the research agenda and the methodology of the discipline. It is therefore crucial for the sake of data interoperability to explicitly document this process. This is realized thanks to the definition of modelling patterns as instances of the *KnowledgeUnitType* class.

The strength of this pattern-based approach lies in the documentation of the conceptual modelling process that allows researchers to understand the abstraction underlying the available data by reading the patterns' definitions, additional help being provided by graphical entity-relationship diagrams that were added to each pattern in the VRE. This is as useful at the time of data production as it is at the time of querying pooled data and re-using it for new research. The *symogih.org* approach, which has been validated and improved through its use in numerous individual or

[20] http://symogih.org/resource/TyIn97

collective projects (see above), thus differs substantially from the usual practice of research projects in history, using, in the worst case, spreadsheets shared in the cloud, or local databases established *ad hoc*, with poorly documented data models. It allows effective re-use of previously produced data for new research agendas.

However the application of the RDFS semantics to the *symogih.org* data model showed the limits of the adopted pattern-modelling method from two points of view: the lack of coherence in the semantics of the model and the absence of formalization. The model was produced in a flat, unorganized way, re-using the same *role types* (i.e. *properties*) in different patterns with usually slight, but sometimes significant differences in meaning. This was mainly due to the failure to use the concept of inheritance in pattern-design and the lack of a clear methodology to assess the ontological consistency of the patterns, and even of the whole generic model. The need to learn more about semantic methodologies and to start a more in-depth comparison with existing models and ontologies became urgent in order to enhance the collaborative modelling effort and promote its adoption by a larger community.

3. The factoid model and the production of historical knowledge

Before addressing the alignment with existing ontologies, some considerations will be presented about the production of historical knowledge. Since the beginning of the *symogih.org* project an essential distinction was introduced between those statements which model factual information (for example the fact that Galileo taught at Padua) and those which reproduce the content of a document literally, so to speak, with each source providing different points of view on the same "fact", e.g. on the date and circumstances of it or on the interpretation thereof.

This distinction is also underlined by John Bradley and Michele Pasin in an article that publishes the *factoid* data model, developed in the context of prosopography projects for the Middle Ages undertaken by the Department of Digital Humanities at King's College London. On one side, they explain, there are "*states of affairs*", on the other side is what the sources assert regarding these same facts: "The factoid approach prioritizes the sources, rather than our historians' reading of them" [14].

In other terms, the factoids tend to model the content of the sources, while the "information" as defined by the *symogih.org* project attempts to describe factual information about the past, the "facts". In order to account for this distinction, essential in historical research, "contents" have been introduced into the *symogih.org* VRE since 2010 as sub-class of the *KnowledgeUnit* class (Fig. 1). These are built so as to be analogous to the "information" units, i.e. by using the same patterns, but they have a substantially different epistemological status: the "contents" (much like factoids) model the *assertions of the source* about facts, including the full range of uncertainties, contradictions and ambiguities it may hold, while "information" models the *assertions made by the historian* after having applied the critical method to the sources' content, in order to establish factual information. As an expression of the content of the source, factoids may also be directly annotated within the transcription of a document, for example by using the XML format according to the standards of the *Text encoding initiative*,[21] and by then proceeding to semantic annotation in connection with a shared reference base [7,22].

With regard to its generic modelling approach, the structure of the factoid data model is comparable to the one of the *symogih.org* project: a factoid is qualified by a type; roles indicate which objects are related to the factoid; the semantics of a role are specified using a type [14]. It is not the structure, therefore, but the epistemological status of the data that marks the difference between the two models. In order to go from one to the other level of information we must apply the methods of historical criticism, such as conjecture, inference, contextualization, etc., with the aim of verifying the reliability and degree of veracity in each assertion made by the source, then aggregating the content of the various sources into a single information unit which is intended to reproduce, to some degree of certainty, *factual information* (Fig. 2). The latter can be defined as a representation in form of data of the "facts" as they probably happened and are reconstituted with the apparatus of historical criticism.

This process of aggregation and changing the epistemological status of the information is generally required in order to meet the needs of historical research agendas. Indeed, as a general rule, when data is submitted for processing and analysis it is essential to

[21] http://www.tei-c.org/

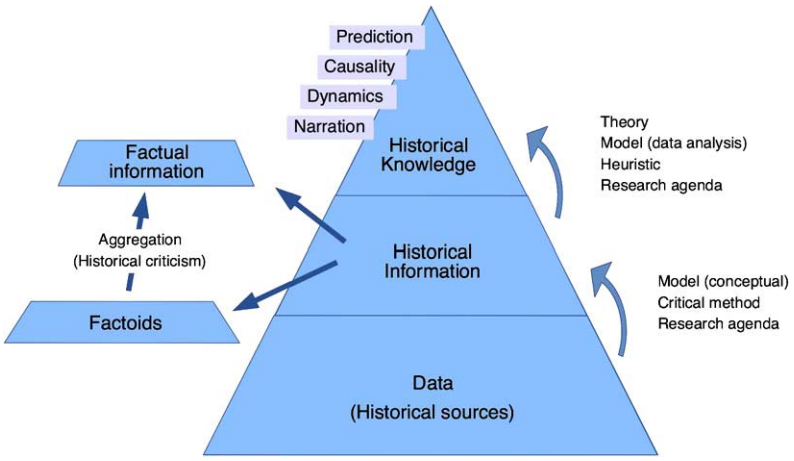

Fig. 2. The process of historical knowledge production.

have at our disposal information that is consistent, non-redundant and not contradictory in terms of the representation of the same state of affairs, and this in view of avoiding distortions in the results of analysis.

For example, we cannot compare the careers of a population of university teachers, like Galilei, if the data available does not contain unique information for each career segment, but rather several mentions of each segment issuing from different sources. In this case, data aggregation is essential prior to analysis: it is necessary to transform mentions of events into factual information and to indicate, as far as possible, its degree of probability in relation to the sources available.

Instead of modelling the complex process of aggregation of factoids, the *symogih.org* project proposed a simplified, pragmatic approach (thus following the usual practice in historical research) adding some specific properties to the *Sourcing* class in order to document the origin of information (cf. Fig. 1). Several sources, or even instances of the factoids already created in the VRE, could be provided as a sourcing for one factual information unit. This method enables other historians to have a rough, but effective measure of the value and reliability of the data in regard to representing "facts".

It could be tempting, in the context of the semantic web and open world assumption, and given the need of integrating data produced by different research projects, to adopt the factoid model on a larger scale and, by analogy, to apply it to all the data coming from various projects, even if it was produced as a representation of historical factuality. As an expression of the point of view of different research projects, the assertions of historians about a fact can themselves be considered as factoids [35].

Despite its apparent practical usefulness, this approach fails to take into account the essential epistemological difference between the status of information as provided by the sources and the one as it is reconstructed through the application of the critical method. From an epistemological point of view, this is tantamount to emptying the historical discipline of its substance. In order to carry out factual information integration, a much better approach consist in providing metadata about the origins and reliability of the shared information, e.g. using the PROV-O ontology,[22] in order to facilitate the merging of the data. This relevant but complex issue will be not discussed in this paper as it is not its focus.

After analysing the difference between factoids and factual information, it is now a question of assessing their place at the heart of the process of historical knowledge production. To this end, we will undertake a reinterpretation of the Data-Information-Knowledge pyramid (DIK) [31], developed by information management, from the point of view of the epistemology of historical research (Fig. 2). According to the generally practiced workflow of knowledge production in history, after intensively reading existing literature, a research agenda is outlined aiming at revising established interpretations of historical phenomena or inspecting new ones. One or more lines of inquiry are then developed in order to clarify the scope of the investigation. They guide the choice of the sources that

[22] https://www.w3.org/TR/prov-o/

will be analysed. In this perspective, data in the pyramid is not to be intended as digital data but, in a general sense, as including all kinds of historical sources (handwritten or printed documents, artefacts, archaeological sites, oral transmissions, paintings, recorded sounds, digital images, etc.).

Sources are the vestiges of the lives, beliefs and activities of past people and societies, and they make it possible to reconstruct life courses, social and economic phenomena, or intellectual evolutions. Out of them factoids will be produced if the point of view of the source is relevant, or factual information if the research agenda aims at reconstituting biographies, structural phenomena, etc. Information extraction will be performed according to the lines of inquiry previously defined by applying to the sources the critical method (intended here in a broader sense). As soon as a sufficient amount of information is available different heuristic methods will be deployed according to the lines of inquiry in order to find relevant facts or structural patterns that provide an answer to the research agenda, in line with existing theories and interpretations, or challenging them.

Historical knowledge can be conceived as the result of this process and can take various forms ranging from narration, to the reconstruction of dynamics of past societies, to the search for the causes of events or, more rarely, to the prediction of possible social phenomena. It is important to point out that historical *knowledge*, even if it can be expressed as information, e.g. in a book or in an ontology, has an epistemological status fundamentally different from historical *information*. This is not just because of the foundational role of *historical distance* in research, in order to distinguish between the reality of the past and the representation by the historians in a different cultural context. But also because of the methodological complex and relevant step that is taken from information to knowledge as shown in reinterpreting the DIK pyramid from the perspective of historical research.

From this point of view, and especially if we aim at conceiving and implementing virtual research environments, and achieving data interoperability, the core issue is about the best way of sharing factual historical information. In earlier times, information was put down on paper, possibly using index cards. Today, if digital data is produced, the tools of conceptual modelling are adopted. In both cases, index cards and databases, it is necessary to apply a conceptualization, be this an unconscious or a well reflected one.

If we carefully consider the process just outlined, it becomes evident that historical information is the result of a *construction* carried out according to the epistemological criteria of the discipline, to the needs of the research agenda and to the chosen modelling methodology. Interoperability and data re-use will only be possible if there is at least a minimal agreement on a conceptualization that attempts to express the best possible approximation of the factuality of the past and allows, at the same time, to apply different research agendas in order to produce new knowledge.

4. Factual historical information and foundational ontologies: c.DnS and DOLCE

In the new context of the semantic web, foundational ontologies can play a major role in helping to achieve this task. In particular, the *Descriptions and situations* ontology (DnS), developed by the WonderWeb project in the early 2000s [12,13,18,20,24–29,31–33,35,36], and especially its reformulation in line with the constructivist paradigm (c.DnS) [23], provides an extremely useful conceptual framework allowing clarification of the epistemological status of historical information, as defined above. Combined with DOLCE (*Descriptive Ontology for Linguistic and Cognitive Engineering*), it provides, in the point of view of the *symogih.org* experience, the basis for a robust conceptualization of historical information, including the modelling of social life.

The *Descriptions and Situations* ontology (DnS) is based on the distinction peculiar to philosophy between *flux*, the flow of events in history, and *logos*, the human discourse about them, the "intentionality". In a similar perspective, the classical distinction between "material object" and "formal object" shows that the latter, although always in reference to the former, is constructed by the observer from the own point of view. The reference to reality is given by the effective hold of the formal object on the material object, which must be collectively verified in its practical application [1]. In DnS, *descriptions* are conceptualizations shared by agents that are used to isolate and define *situations* in the *flux*, the states of affairs in the world, present and past ones. Situations are portions of states of affairs that are "carved up by virtue of a description": facts, legal cases, technical actions, etc. which cognitive agents recognise in the *flux* as corresponding to a description they share. According to this vision, a situation has to *satisfy* a description [24].

The DnS ontology develops two principles that are essential for our purpose. First, it situates information models, i.e. conceptualizations, in relation to the scientific discipline or general perspective they stem from. This approach is thoroughly developed in *Constructive DnS* (c.DnS) which provides a formal analysis of "knowledge collectives that share an epistemic workflow in order to exchange or modify knowledge". According to the principle of "epistemological layering", the same portion of states of affairs can be "re-described" by different situations that satisfy different descriptions. Second, descriptions and situations are reified and belong, with the concepts they define, in the same domain of discourse of the entities they re-describe. The reification allows the modelling of descriptions, on the one hand, and the situations and social entities they describe, i.e. agents, collectives, relationships, etc., on the other hand, in the same "domain of quantification or logical level" [23].

If we apply the first c.DnS principle to the workflow of historical knowledge production, we can interpret the process of modelling factual information as the one of creating descriptions that are related to the specific research agenda, and consistent with the principles of the discipline, and allow to "carve out" situations in states of affairs that, if collected in adequate quantity, are suitable to answer research questions. Different research agendas, or different scientific disciplines, will produce different conceptualizations (i.e. descriptions) of the same portion of states of affairs that have to be consequently interpreted and situated in their epistemological context.

If we apply this view to the *symogih.org* project's modelling patterns, it results that the instances of the *KnowledgeUnitType* class (Fig. 1) are *descriptions* that allow users of the VRE to "carve out" *situations*, i.e. instances of the *Information* class. As a general rule, the capacity of these descriptions (the conceptualizations in the model) to express the historical factuality can be evaluated by considering their ability to be re-used and actionable by different research agendas and disciplines. But at the same time the "epistemological layering" principle will impose limits on re-use of data and in some cases it will require reinterpretation and rewriting of *descriptions* in different scientific contexts. The presupposition for succeeding in this is obviously a clear documentation of the definition and intended use of descriptions. i.e. database documentation.

If c.DnS thus allows a precise epistemological and semantic analysis of the *symogih.org* modelling patterns, its second principle, introducing social entities as an expression of shared intentionality, is even more relevant and shows its full potential if associated with DOLCE, a foundational ontology designed as a means of studying the "ontological categories underlying natural language and human common-sense" [12,13,26–29,31–33,35,36],[23] in order to create an "ontology of social reality" [13]. This was first undertaken in *Dolce+* and DOLCE *light plus* (DLP) [27],[24] and then expressed with a new, more user-friendly vocabulary, integrating c.DnS, in *DOLCE Ultra Light* (DUL)[25] [25].

Using the concepts provided by DLP we can analyse the *symogih.org* ontology and gain useful insights that can be easily generalized to any representation of factual historical information. The *symogih.org Object* class (*sym:Object*) (cf. Fig. 1) is clearly equivalent to the *endurant* DLP class (*dlp:Endurant*), grouping entities, like persons, physical objects, concepts, etc., that "are wholly present at any time they are present" even if their properties (e.g. color, dimension, etc.) can change over time. The pieces of factual information represented by the *sym:Information* class need more careful inspection. In this ontology, time is essentially related to this class and never to the *sym:Object* class. *sym:Information* is conceived as expressing relationships among *sym:Object* instances that are situated in time and space. Thus, at first glance, this class appears to be equivalent to *dlp:Perdurant*, modelling "events, processes, phenomena, activities and states": they happen in time, and *dlp:Endurant* instances *participate* in them. This is certainly true for *situations* described by the *sym:KnowledgeUnitType* class as events or phases directly related to time *and* space, i.e. spatio-temporal phenomena, like a birth or a fight. But if you consider membership of a group, ownership of a painting, or having social roles like being a king or the pope, despite the fact that the duration of these social situations is limited in time, and therefore relative to time, these are not strictly speaking spatio-temporal phenomena, and therefore do not belong to the *dlp:Perdurant* class.

This is where all the importance of the DnS ontology comes into play, particularly in its integration with DOLCE. On the one hand, DLP introduces social objects as subclasses of *dlp:Endurant*, allowing modelling of social concepts, organizations, roles, plans, etc., which belong to the sphere of "intentionality",

[23] Cf. http://www.loa.istc.cnr.it/dolce/overview.html.
[24] DLP (version 3.9.7), http://ontologydesignpatterns.org/ont/dlp.
[25] http://ontologydesignpatterns.org/ont/dul/

and use them in the ontology. On the other hand, the notion of time-indexed qualities, in DOLCE, and of time-indexed relations or roles, in DnS/DUL [25], makes it possible to integrate the temporal dimension in the description of qualities of the objects as well as of their relations in intentionality and the *social* space. This is essential for modelling the domain of discourse of historical research wherein factual information must always be related to time, but not necessarily to *geographical* space.

At the end of this analysis, it appears that the *sym:Information* class grouped and mixed up three distinct types of time-related situations that have to be clearly distinguished: *perdurants*, on the one hand, and *time-indexed qualities* ("quale" relations in DOLCE) and *time-indexed social relations*, on the other hand. The most relevant difference between the two groups seems to be the *direct* reference with (geographical) space in perdurants (at least through the physical entities that participate in them), which is absent for the time-indexed situations insofar as they refer in their description only to time and intentionality. If we consider the example of Galileo's teaching in Padua (see above), from the same proposition we can extract two pieces of information, depending on the epistemological perspective adopted by the researcher: the teaching as a *process* situated in time and space (*dlp:Perdurant*), involving students, experiments, classes in different buildings, etc., if the interest is in teaching practice; the position of "professor of mathematics" as time-indexed role (*dlp:Situation*) defined by the University's statutes, if the interest is in positions and careers.

5. Extending the CIDOC CRM for historical research

At this point, one could imagine rewriting the *symogih.org* modelling patterns using the DLP/DUL categories, as well as the modelling methodology developed by the *ontologydesignpatterns.org* project, in order to cope with the lack of coherence in the semantics and absence of formalization we have noted above (part 2). Given, however, that one of the essential principles of interoperability is to use as far as possible existing wide-spread domain ontologies, it seemed more sensible to approach the CIDOC CRM and check the feasibility of integrating the *symogih.org* model to this quasi-standard ontology in the cultural heritage domain. The CIDOC CRM was created in the 1990s as an object-oriented data model aiming at achieving semantic interoperability of museum data. It has been defined as a "formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information" [18] and offers several advantages in terms of interoperability: it has been an ISO norm since 2006;[26] it has an active community maintaining and developing it;[27] it is used by renowned institutions such as the British Museum[28] and the Getty Conservation Institute;[29] it has a useful set of extensions modelling related domains like bibliographical and archaeological data;[30] it was formalized in first-order logic [20].

An active discussion with the *Special interest group* (SIG) that maintains the CIDOC CRM was started in August 2016, a first alignment workshop was held in Heraklion in November 2016 in order to analyse the characteristics of the *symogih.org* ontology from the point of view of the CIDOC CRM, and the LARHRA is now actively participating in the development of the CIDOC CRM conceptual model (SIG), notably with regard to a dedicated extension to create a *Model for social phenomena* (CRMsoc).[31]

The two ontologies, the *symogih.org* and the CIDOC CRM, have many similarities which makes it easy to align the *sym:Object* sub-classes and some of the *sym:Information* modelling patterns to the classes and properties in the CIDOC CRM. Both ontologies aim at modelling factual information and have an approach based on atomization [17]. The mechanism of specialization trees with multiple property inheritance brings order in the somewhat anarchic *symogih.org* modelling patterns by clarifying the meaning of the properties (*sym:RoleType*, Fig. 1). Useful hints are provided by the *crm:E77 Persistent Item* class, corresponding to *dlp:Endurant* and *sym:Object*, as well as by its subclasses that help improve the modelling of material, immaterial and active entities. The integration with the FRBRoo extension helps to conceptualize historical sources and their intellectual content. The treatment of time and its uncertainty adopts, in both ontologies, a similar approach defining internal and external fuzzy borders of time-spans. The CIDOC CRM furthermore

[26] ISO21127:2006, renewed in 2014: ISO21127:2014.
[27] http://www.cidoc-crm.org/
[28] https://lod-cloud.net/dataset/british-museum-collection
[29] http://www.getty.edu/conservation/our_projects/field_projects/arches/arches_overview.html; http://www.getty.edu/conservation/our_projects/science/integrating_data/technology.html.
[30] http://www.cidoc-crm.org/collaborations
[31] http://www.cidoc-crm.org/crmsoc/fm_releases

adds so-called Allen operators [4], modelled as properties, in order to express relative position of events in the flow of time, or their overlapping or inclusion, without knowing their precise position in the time reference space [26]. This is very useful for historical research.

During this alignment process, still ongoing, some relevant issues also appeared, especially regarding the modelling of stative temporal events and social phenomena, that deserve careful consideration. If information in the Cultural Heritage domain is partly about characteristics of preserved objects, in the material sense, it is also about the history of their origins, change in ownership, historical and cultural contextualization. In the early stages of its development, the CIDOC CRM excluded the explicit treatment of temporal situations or states, and restricted its domain of discourse to events. This choice was dictated on the one hand by practical considerations, as it facilitates the integration of data from different producers. But, on the other hand, it stems from the epistemological approach within which the conceptualization of the CIDOC CRM, its *description* in the sense of c.DnS, has been carried out: "This approach was inspired by considerations of modern physics". It is assumed that observation of events is less difficult than that of states, and less dependent on contextual properties [19].

In factual historical information management however, the reference to phases is not only relevant because testimonials about the states are often provided in the sources, and not about the events producing them, but also because active, long lasting activities or relationships, can hardly be modelled as events. DLP provides in this respect the useful classes of *state*, a phase in which no changes happen in the considered aspects, or *process*, a dynamic phase in which according to the principle of homeomericity all relevant parts of the perdurant can be defined with the same description. All kinds of commercial relationships, teaching activities, but even wars and similar kinds of complex processes, can be more suitably described, from the point of view of historical research, as phases than as events.

As DLP points out, these kinds of states are considered to differ from *situations* "because they are not assumed to have a description on which they depend". This is true, on the one hand, in the point of view of domain modelling and the example of the union of a man and a woman, can illustrate this fundamental distinction. If we consider the union as a sentimental relationship that begins with being in love, continues in the marriage and leads to having children, we are in the perspective of a spatio-temporal phenomenon that can be conceptualized as a stative subclass of perdurant, assuming that the common modelling property is the dynamic (love) relationship between two persons. If you take the union in the sense of marriage as defined by law, the fact of being married is a *situation*, i.e. a time-indexed relationship based in the intentionality of agents situated in a specific society.

But, on the other hand, if we take the epistemological point of view of c.DnS, and generalize it, we could even conceive perdurants as situations defined by descriptions "that assert the constraints by which a state of a certain type is such, and in this case, becomes a situation".[32] In other words, the boundary between *perdurants* and *situations* would become blurred and be dependent on the research agenda. We can however adopt the first option, restricting the scope to domain modelling, and rely on the definition of social life as an expression of intentionality shared by rational agents, as provided by DnS, in order to find guidance in distinguishing between *phases* and *situations*.

In this perspective the CIDOC CRM could be enriched, in a suitable extension, by adding a *Phase* class as subclass of *crm:E4 Period* to express states in the sense of long-lasting, homeomeric spatio-temporal phenomena, as counterpart of the *crm:E5 Event* class. Furthermore, a *Time-indexed social situation* class could be added, in the same extension, as a subclass of *crm:E1 Temporal entity* and as a sibling of the *crm:E4 Period* and *crm:E3 Condition State* classes. This kind of times-indexed *situations* correspond to aspects of social life that exist only in intentionality (and not directly in physical space) and is defined by laws, rules of organizations, informal social norms, i.e. *descriptions* that the *situations* have to *satisfy*.

It is important to notice that the relationship with a (geographical) place only occurs, in the CIDOC CRM, at the level of the *crm:E4 Period* class, and that therefore *sensu stricto* this is the root class for modelling the spatio-temporal phenomena, and the equivalent to *dlp:Perdurant*. The *crm:E1 Temporal entity* class, given its properties, only models an absolute or relative position in time, and is therefore suitable as a parent of the *crm:E3 Condition State* class, somehow equivalent to the time-indexed *quale* relation in DOLCE, holding between the quality of a physical object and its value, and likewise as a parent of the envisaged *Time-indexed*

[32] DLP 3.9.7, OWL serialization.

social situation. The latter would add to the CIDOC CRM, in a suitable extension, a more developed conceptualization of social time-indexed phenomena like ownership or membership that are currently expressed with timeless properties.[33] The projection of a social perspective is present in the CIDOC CRM, specifically in the *crm:E72 Legal Object* class, but without explicit reference to time; a real limitation from the point of view of historical research.

6. OntoME and the data for history consortium

In order to produce and share a consistent, but extensible conceptualization of historical information, to connect it to existing ontologies and standards, and to align existing project data models to it, a collaborative web environment is needed, and also a research community using it. This view led us to undertake in 2016 a detailed consideration of the tools available, offering functionalities for both the alignment of ontologies and collaborative discussion of data modelling. Having evaluated the existing tools, particularly WebProtégé[34] [34], it seemed opportune (given the limitations they then had regarding our requirements), to establish a new online application in the form of an ontology management environment named OntoME.[35]

OntoME is a classical web application built on top of a PostgreSQL database, using the PHP framework Symfony for building the front end and the APIs.[36] This technology choice corresponds to the know-how gained in the LARHRA Digital Research Team during ten years of the *symogih.org* project. OntoME is still in *beta* version but already used in production by different projects (see below). Once a minimum viable product is operational (most likely by the end of 2020), the code will be switched to *open source*. However, the aim is not to distribute a new application but to invite interested projects and researchers to use the existing instance collaboratively. Indeed, the development of the application is in line with the intention to extend into a wider community the modelling experience that was presented above.

OntoME is designed for allowing collaborative conceptual modelling and integrates basic modelling elements compatible with the object-oriented modelling (UML class diagram) as well as with RDFS and OWL2 DL. Namely it includes classes and properties with cardinalities. Instances are not implemented in the sense of data production but will be available for modelling purposes in a second phase of development. At the moment, a prototype workflow for the management of controlled vocabularies in relation with OntoME classes is in test phase using *Opentheso*, developed by Miled Rousset.[37] This is part of a French ANR funded project of data FAIRification [30].

Classes can be arranged in a multiple subclass hierarchy with inheritance of properties. These can take other classes or primitive values (in the form of corresponding classes) as ranges. Modelling properties like rdfs:subclassOf, owl:equivalentClass or OWL relations and constraints of properties are directly implemented in the application and can be managed in the GUI. SHACL constraints could be added if useful for users and projects.[38] Hiding the "technical" modelling aspects of the ontology by "embedding" them in the graphical interface is conceived as a way of helping non-expert users to focus on conceptual modelling as a central point in the construction of a shared conceptualization. The target audience are therefore specialists from different scientific disciplines interested in modelling and semantic interoperability. Prototypes of OWL/RDF APIs are available and allow experts to directly inspect classes and properties of one or more namespaces in Protégé, and verify their consistency with reasoning tools. Modelling constraint and ontology consistency checks could be directly introduced into the OntoME application and added to the checks already implemented concerning reference namespaces control, inheritance, etc. Although OntoME is freely available, advanced use requires the creation of a user account in order to benefit from customised access and additional functionalities. A dashboard allows someone to filter the ontologies, and more specifically the versions of them, that are visible from the user's point of view. In this way the user can navigate and inspect only classes and properties of ontologies of interest, and compare different versions of them. The application revolves around three pillars: projects, namespaces and profiles. Users have differ-

[33] Respectively *crm:P51 has former or current owner* and *crm:P107 has current or former member* (CIDOC CRM 6.2).

[34] https://protege.stanford.edu/products.php#web-protege

[35] https://ontome.dataforhistory.org/

[36] The public documentation of the project is to be found on the Data for history Forum, http://forum.dataforhistory.org/forum/39.

[37] https://www.mom.fr/ressources-numeriques/opentheso

[38] Shapes Constraint Language (SHACL), https://www.w3.org/TR/shacl/.

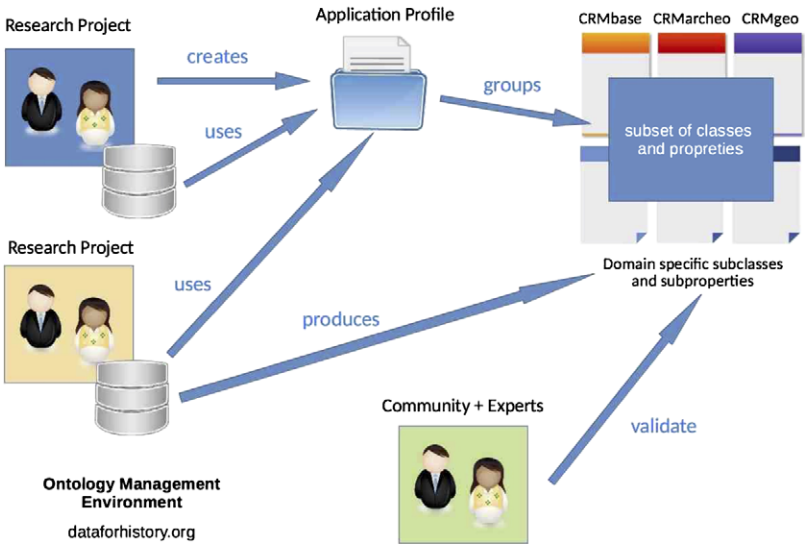

Fig. 3. OntoME (ontology management environment) use cases.

ent editing rights in different projects; namespaces allow the management of different versions of the same ontology; profiles are used to select classes and properties from different ontologies, and configure a data model to be used for data wrangling or production.

As Fig. 3 shows, OntoME enables the application of several use cases, adapted to the needs of different users and projects. It is, first of all, a learning space, enabling non-experts to get to grips with data models and ontologies more rapidly than they would by reading written documentation, thanks to the display of inherited properties for each class and graphic representations designed to facilitate navigation. This is particularly true for the CIDOC CRM and family of extensions (FRBRoo, CRMgeo, CRMarcheo, etc.) of which an integrated overview is possible in OntoME. Other standard ontologies will be imported on user request. A graphical tool for importing models and ontologies will be developed as soon as funding becomes available.

A project can manage one or more profiles, and use them for data wrangling or production. Profiles are sub-sets of classes and properties that can be exported via an API to be used for data production in distributed information systems. Customized labels and definitions in different languages can be added in profiles to existing classes and properties in order to improve their usability by non-experts. The VRE Geovistory[39] (and the same is possible for any other interested data production infrastructure) gets the conceptual model used in production directly from OntoME through the JSON API. Four ongoing research projects funded by public research agencies currently use Geovistory as VRE, based on OntoME profiles. As mentioned above, a migration of *symogih.org* VRE data to Geovistory is underway, based on the alignment of the original information patterns with the new classes and properties in the namespace(s) for historical research (see below).

A project can also create a new namespace and produce in it more specialized classes and properties, dedicated to its field of study, while also connecting them to available standard ontologies through the tree of class specialization. In order to do so, the project's members have access to a dedicated namespace of which they are the sole master, yet they have different editing rights. New classes and properties will enable researchers to produce data that will be interoperable at a higher level of abstraction. As experience in coaching several research projects shows, if data integration needs a higher level of abstraction, data production requires by contrast more specialized classes that suit the research agenda. These can be edited in a specific project's namespace and then added to profiles to be used in production. Two EU funded projects, SILKNOW[40] and READ-IT,[41] currently use OntoME for data modelling.

[39] https://geovistory.com/

[40] https://silknow.eu/ – https://data.silknow.org/.

[41] https://readit-project.eu – https://ontome.dataforhistory.org/namespace/38.

If a project already has its own database, it could import or replicate the existing model in OntoME and align it with standard ontologies, then export the alignment generated and use it to wrangle the data, before publishing it in RDF according to a new, interoperable model. Scheduled data exporting, or rewriting in real time, would enable users to continue generating data in the original project information system, all while transposing it on the fly into an interoperable format. A prototype process of integrating data from different sources concerning Dutch maritime history in the 17th and 18th century, remodelled using OntoME, was carried out in 2019 in the Huygens ING/CLARIAH Geovistory pilot project and presented at the Time Machine Conference 2019.[42] A workshop on FAIRification of data produced by different research projects is currently ongoing at the initiative of the LARHRA Digital research team.[43]

In this same vision, the *symogih.org* VRE modelling patterns mentioned above, and other relevant elements of the ontology, have been imported into a dedicated namespace in OntoME,[44] in order to provide their alignment with the CIDOC CRM. Instead of doing this directly, given the considerations developed above (part 5), a different alignment strategy was chosen. A new project devoted to *Historical data management and interoperability* (*HistDMI*)[45] has been created aiming at integrating to the CIDOC CRM the modelling experience of the *symogih.org* project, revisited thanks to the foundational ontologies DOLCE and c.DnS, and their serialization in DLP/DUL.

In the main project namespace, missing high-level classes like *Phase* or *Time-indexed social situation* have been added. Also other high-level classes, less explicit in the CIDOC CRM, are created in order to facilitate the ontology's use. For example, we observe an inconsistent treatment in different projects of the class adopted to model geographical places due to the misleading name of the class *crm:E53 Place*, defined as a portion of a purely abstract reference space, while the physical geographical location intended as portion of the Earth surface should be modelled as a *crm:E26 Physical Feature*. This led to the introduction of class *hist:C8 Geographical Place* in the new namespace, enabling clarification of the concept for non-experts. Other namespaces in sub-projects will treat different aspects of (historical) social and intellectual life, and provide classes and properties included in the general hierarchy but with an appropriate level of specialization suitable to different research agendas. The modelling will build upon the *symogih.org* project's experience, and benefit from the contribution of interested co-editors, in reference to other existing models. The alignment of the *symogih.org* VRE patterns will be done with the respective classes and properties of the HistDMI project. Data will be then wrangled accordingly and imported into the Geovistory VRE, thus opening a new data life cycle.

In the vision that inspired the development of OntoME, all these use cases, and the workflows they involve, have ideally to be carried out in the dynamic context of a research community: its members, or semantic technology experts, can evaluate and discuss the new classes and properties, or the application profiles, in a process that will enable them to progressively improve their quality and be extended to more specific research fields, while providing interoperability and integrating the standards adopted by Cultural Heritage institutions. Indeed, it can be observed today that several projects have adopted the CIDOC CRM,[46] or other reference ontologies, and enrich them with local extensions or interpretations which do not necessarily converge with those of other research groups, or which are perhaps never published. This situation represents a major obstacle to research data interoperability.

It seemed therefore appropriate to launch an initiative to federate these efforts. Following two French workshops (Lyon, June 2016; Brest March 2017) a *Data for History* consortium was officially constituted in an international workshop held in Lyon, in November 2017; in attendance were around thirty historians, art historians, archaeologists and information science specialists from six European countries. After some other meetings (Lyon, May 2018; Galway, December 2018;[47] Leipzig, April 2019) the first *Data for History* conference and members' meeting (first planned in May 2020) will be held in Berlin in May 2021. The consortium operates a public forum and a mailing list, both of which are open to anyone upon request.[48]

[42] http://forum.dataforhistory.org/node/150 – https://halshs.archives-ouvertes.fr/halshs-02314003 – https://ontome.dataforhistory.org/profile/8.

[43] https://frama.link/xqbLMetB

[44] https://ontome.dataforhistory.org/namespace/2

[45] https://ontome.dataforhistory.org/project/8

[46] E.g. https://doc.biblissima.fr/ontologie-biblissima – https://masa.hypotheses.org/500.

[47] https://eadh2018.exordo.com/programme/session/49

[48] http://dataforhistory.org/

7. Conclusion

If we come back to the question raised at the beginning – can we apply the vision of the FAIR principles to data arising from historical research and, more broadly, those in the field of Cultural Heritage, and promote their interoperability with a view to their being re-used for new research? – the preceding considerations show that, in the point of view of the *symogih.org* project's experience, the main issue to be addressed involves the sharing of a communal, formalized and extensible conceptualization for factual historical information. Such a collaborative ontology cannot be written in the abstract but it requires the roll-out of a process that can only work under certain conditions. I shall summarize these in three points while highlighting the work still to be done.

Firstly, the foundations of this endeavour were laid by the development of CIDOC CRM which, from the point of view of historical research, can be considered (with some of its extensions) as the core ontology for the domain. However, if we carefully analyse the process of historical knowledge production, and apply to it an epistemological and semantic analysis with the aid of the c.DnS and DOLCE foundational ontologies, it appears that CIDOC CRM has to be integrated and completed from two points of view. On the one hand, given the different epistemological perspectives in Cultural Heritage and historical research, some high-level classes will have to be added in an extension dedicated to historical research data; namely a *Phase* class, representing DOLCE stative perdurants, and a *Time-indexed social situation* class, expressing situations concerning social life defined by human, collective intentionality. On the other hand, the integration of the more than 150 modelling patterns from the *symogih.org* project, and those produced by all other interested research projects, stemming from sub-domain data production and re-use, requires the creation of several extensions devoted to different sub-domains and historical research agendas, and, at the same time, alignment with higher abstraction level classes and properties, for the sake of interoperability.

Secondly, this undertaking evidently requires the implementation of an infrastructure to match the vision of elaborating a common fine-grained and adaptive ontology, and one which is easy for projects to use. OntoME, coupled with a controlled vocabularies management tool like Opentheso, constitutes the prototype of the envisaged infrastructure but a lot of work remains in terms of formalization of this complex modelling process. This is particularly true in the sense of applying description logic and other formalisms to verify the consistency of the ontology, and integrating this process into the OntoME application in order to allow real-time checks. To this end it will be important to collaborate with experts and seek together the necessary funding.

Thirdly and finally, this process will only be successful if a community of users is formed and built up, driven by a genuine desire to share data and modelling expertise, in full awareness of how useful the issues of ontology and controlled vocabularies can be for research. The *Data for History* consortium is one such initiative that makes a demand and need visible, laying down the foundation of a network. We must hope that the holders of projects and platforms, notably those which are in the domain, will have the foresight to get on board with this dynamic so that it can truly flourish and contribute to realize the FAIR principles in historical research.

References

[1] E. Agazzi, *Scientific Objectivity and Its Contexts*, Springer, Cham, 2014. doi:10.1007/978-3-319-04660-0.

[2] J. Akoka and I. Comyn-Wattiau, *Conception des bases de données relationnelles*, Vuibert, Paris, 2001.

[3] J. Alerini and S. Lamassé, Données et statistiques. L'avenir du travail en ligne pour l'historien, in: *Les historiens et l'informatique. Un métier à réinventer*, École française de Rome, Rome, 2011, pp. 171–187.

[4] J. Allen, Maintaining knowledge about temporal intervals, *Communications of the ACM* **26** (1983), 832–843. doi:10.1145/182.358434.

[5] L. Audibert, *Bases de données: de la modellisation au SQL*, Ellipses, Paris, 2009.

[6] F. Beretta, Exploration du site web *scholasticon.fr*: une application de la méthode SyMoGIH (Système modulaire de gestion de l'information historique), in: *La prosopographie au service des sciences sociales*, CEROR, Lyon, 2014, pp. 289–310.

[7] F. Beretta, Pour une annotation sémantique des textes: le projet *symogih.org* et la *Text encoding initiative*, *Bruniana & Campanelliana* **22**(2) (2016), 453–465. doi:10.19272/201604102005.

[8] F. Beretta, L'interopérabilité des données historiques et la question du modèle: l'ontologie du projet SyMoGIH, in: *Enjeux numériques pour les médiations scientifiques et culturelles du passé*, Presses Universitaires de Paris Nanterre, Paris, 2017, pp. 87–127.

[9] F. Beretta and C. Butez, Un SIG collaboratif pour la recherche historique, *Géomatique Expert* **91** (2013), 30–35 / **92** 48–54.

[10] F. Beretta and R. Letricot, Le portail XML du projet symogih.org: un projet d'édition numérique collaborative de sources et d'informations historiques, in: *Ecrilecture augmen-*

tée dans les communautés scientifiques, ISTE Editions, London, 2017, pp. 125–143.

[11] F. Beretta and P. Vernus, Le projet SyMoGIH et la modellisation de l'information: une opération scientifique au service de l'histoire, *Les Carnets du LARHRA* **1** (2012), 81–107.

[12] S. Borgo and C. Masolo, Foundational choices in DOLCE, in: *Handbook on Ontologies*, Springer-Verlag, Berlin/Heidelberg, 2009, pp. 361–381. doi:10.1007/978-3-540-92673-3_16.

[13] E. Bottazzi, C. Catenacci, A. Gangemi and J. Lehmann, From collective intentionality to intentional collectives: An ontological perspective, *Cognitive Systems Research* **7**(2–3) (2006), 192–208. doi:10.1016/j.cogsys.2005.11.009.

[14] J. Bradley and M. Pasin, Factoid-based prosopography and computer ontologies: Towards an integrated approach, *Literary and Linguistic Computing* **30**(1) (2015), 86–97. doi:10.1093/llc/fqt037.

[15] A. Courtin and J.-L. Minel, Propositions méthodologiques pour la conception et la réalisation d'entrepôts ancrés dans le Web des données, in: *Enjeux numériques pour les médiations scientifiques et culturelles du passé*, Presses Universitaires de Paris Nanterre, Paris, 2017, pp. 53–86.

[16] J.-P. Dedieu, Les grandes bases de données: una nouvelle approche de l'histoire sociale: le système Fichoz, *HISTORIA. Revista de Faculdage de Letras da Universidate do Porto s.3* **5** (2004), 101–114.

[17] M. Doerr, The CIDOC CRM. An ontological approach to semantic interoperability of metadata, *AI Magazine* **24**(3) (2003), 75–92. doi:10.5555/958671.958678.

[18] M. Doerr, Ontologies for cultural heritage, in: *Handbook on Ontologies*, 2nd edn, Springer, Berlin, 2009, pp. 463–486. doi:10.1007/978-3-540-92673-3_21.

[19] M. Doerr, J. Hunter and C. Lagoze, Towards a core ontology for information integration, *Journal of Digital Information* **4**(1) (2003).

[20] M. Doerr and C. Meghini, A first-order logic expression of the CIDOC conceptual reference model, *International Journal of Metadata, Semantics and Ontologies* **13**(2) (2018), 131–149. doi:10.1504/IJMSO.2018.098393.

[21] J. Domingue, D. Fensel and J.A. Hendler (eds), *Handbook of Semantic Web Technologies. Vol. 1. Foundation and Technologies*, Springer, Berlin/Heidelberg, 2011. doi:10.1007/978-3-540-92913-0.

[22] Ø. Eide, Ontologies, data modelling, and TEI, *Journal of the Text Encoding Initiative* **8** (2014–2015). doi:10.1093/llc/fqp010.

[23] A. Gangemi, J. Lehmann and C. Catenacci, Norms and plans as unification criteria for social collectives, *Autonomous Agents and Multi-Agent Systems* **17** (2008), 70–112. doi:10.1007/s10458-008-9038-9.

[24] A. Gangemi and P. Mika, Understanding the semantic web through descriptions and situations, in: *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, Springer, Berlin/Heidelberg, 2003, pp. 689–706. doi:10.1007/978-3-540-39964-3_44.

[25] A. Gangemi and V. Presutti, Dolce+D&S Ultralite and its main ontology design patterns, in: *Ontology Engineering with Ontology Design Patterns*, IOS Press, Amsterdam, 2016, pp. 81–103. doi:10.3233/978-1-61499-676-7-81.

[26] J. Holmen and Ch.-E. Ore, Deducing event chronology in a cultural heritage documentation system, in: *Making History Interactive. Computer Applications and Quantitative Methods in Archaeology*, Arcaeopress, Oxford, 2010, pp. 122–129.

[27] C. Masolo, S. Borgo, A. Gangemi, N. Guarino and A. Oltramari, WonderWeb deliverable D18 ontology library (final), Laboratory for Applied Ontology, Trento, 2003.

[28] A. Meroño-Peñuela, A. Ashkpour, M. van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach and F. van Harmelen, Semantic technologies for historical research: A survey, *Semantic Web* **6** (2015), 539–564. doi:10.3233/SW-140158.

[29] R. Ramón (ed.), *Big Data: Analyse zum digitalen Wandel von Wissen, Macht und Ökonomie*, Transcript, Bielefeld, 2014. doi:10.14361/transcript.9783839425923.

[30] M.-O. Rousset, F. Beretta et al., HisArc-RDF: Prototyping an operating chain, related to the linked open data, on structurally and semantically heterogeneous archaeological data sets (poster), in: *Linked Pasts 5: Back to the (Re)Sources*, Bordeaux, 2019, https://halshs.archives-ouvertes.fr/halshs-02413859.

[31] J.E. Rowley, The wisdom hierarchy: Representations of the DIKW hierarchy, *Journal of Information Science* **33**(2) (2007), 163–180. doi:10.1177/0165551506070706.

[32] T. Segaran, C. Evans and J. Taylor, *Programming the Semantic Web*, O'Reilly, Beijing, 2009. doi:10.5555/1696488.

[33] C. Soutou, *UML 2 pour les bases de données*, Eyrolles, Paris, 2007.

[34] T. Tudorache, C. Nyulas, M.A. Musen and N.F. Noy, WebProtégé: A Collaborative Ontology Editor and Knowledge Acquisition Tool for the Web, *Semantic Web* **4**(1) (2013), 89–99. doi:10.3233/SW-2012-0057.

[35] G. Vasold, M. Schlögl and G. Vogeler, Data exchange in practice: Towards a prosopographical API, Preprint, 2019. doi:10.17613/yw4h-5f09.

[36] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg et al., The FAIR guiding principles for scientific data management and stewardship, *Scientific Data* **3** (2016), 160018. doi:10.1038/sdata.2016.18.