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 – francesco.beretta@cnrs.fr

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. It outlines a methodological approach allowing to integrate data stemming from different lines of inquiry and belonging to different epistemological levels. After introducing the *symogih.org* project's ontology, designed to cope with this issue, it compares it with the factoid model conceived at King's College London and the CIDOC CRM conceptual model, highlighting the specificities of each and their contribution to data interoperability. Finally, it shows how collaborative data modelling carried out in the ontology management environment OntoME makes it possible to elaborate a common fine-grained and adaptive understanding of information, applying domain knowledge to data production. The condition of a positive outcome of this process is that the research community actively engages in the elaboration of a communal ontology, which the *dataforhistory.org* consortium is currently seeking to promote.

Keywords: FAIR principles; historical research data interoperability; Cultural Heritage; CIDOC CRM; Factoid ontology; 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" [20]. 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 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" [15]. It is therefore a question of adopting, for a given academic discipline, a broadly shared data model expressed using a protocol that is compatible with technologies used on the semantic web. This principle also applies to the second paragraph

[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 journal *Scientific data* published by the Nature group.

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

— "(meta)data use vocabularies that follow FAIR principles" — 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 — "(meta)data include qualified references to other (meta)data" — recommends the use of explicit and clearly defined 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, those issuing from the field of Cultural Heritage (galleries, libraries, archives, museums). Indeed, given the vast wealth of data produced in these two fields, the importance 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 this data 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? Moreover, aren't the data extracted from historical sources often uncertain, ambiguous and even contradictory? Given these observations, how is it possible to adopt a quasi-universal language, an ontology, and to ensure the interoperability of the data produced by historians?

These kinds of questions, which are already being asked in the field of digital humanities [12] and notably in digital history [2], have become more pressing in recent years due to the proliferation of semantic web technologies [18] and projects producing huge amounts of data, such as the *Time machine* large-scale research initiative[5]. 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), which sprang from the desire to pool and reuse the data produced by the researchers for new projects within a collaborative virtual research environment (VRE). Extended to the entire community of research projects and heritage institutions seeking to make their data accessible according to FAIR principles, these issues lead us to question the very possibility that data produced by geo-historical information systems designed independently from one another, stemming from various research agendas and programs, can ever be made interoperable.

The second part of this contribution will present the *symogih.org* project's experience of data pooling. The third part will focus particularly on the issue of generic and open data modelling, which was adopted by this project as a condition for interoperability. In part four, the *factoid* model developed by the prosopography projects at King's College London will be presented, while also connecting it to the modelling options undertaken by the *symogih.org* project, as well as to the CIDOC CRM, in order to highlight the specificities of each. Part five will present the reasons having led to the development of a new "Ontology management environment", OntoME, and the creation of the *Data for History* consortium[6] in order to provide the foundations of an infrastructure and community on which the interoperability of historical data could be built. In conclusion, the findings of the methodological analyses will be summarized and the conditions for the success of this endeavour outlined.

2. The *symogih.org* project

The *symogih.org* project first came into being in 2008 when several historians from the Laboratoire de recherche historique Rhône-Alpes sought to pool the structured data acquired during their research, in order to enable these data to be re-used by other researchers. This approach, based on applicable standards in the field of database modelling [1-4-19], follows the rationale of data curation, understood as referring to the enrichment and gradual improvement of research data in order to guarantee the quality, accessibility and preservation of said data[7]. For example, the data produced over the course of the SIP-PAF project[8], 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[9] which focuses on law professors in France

[4] https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/
[5] https://www.timemachine.eu/
[6] http://dataforhistory.org/
[7] https://en.wikipedia.org/wiki/Data_curation
[8] http://www.patronsdefrance.fr/
[9] http://siprojuris.symogih.org/

from 1804 - 1950. These two projects each had their own dedicated website, but the collection of data was based on a single collaborative information system that encouraged the exchange and re-use of the data. The fact that the data from both projects was managed using a single IT system, designed to be modular and generic, ensured the long-term availability of the data following the end of the financing period, and the re-usability of the data for other projects. This architecture also enables researchers to integrate new modules based on existing and standardized technologies, or services made available by other organisations. For example, analyses of data produced in the the *symogih.org* virtual research environment (VRE) can easily be realized in the RStudio instance deployed by Huma-Num[10].

A growing number of projects inside and outside LARHRA, both French and European (currently over 60 users and around 15 projects), are using this VRE to produce and pool their data. These data are made publicly available under the *Creative Commons Attribution-ShareAlike 4.0 International* licence on the *symogih.org* generic website, which represents the central access point, as well as on the various project websites. For projects adopting the *open data* approach, a SPARQL access point enables users to directly query the portion of data that the researchers have decided to publish in RDF format[11]. The structure of the data model applied so far, and publicly documented on the *symogih.org*, website, will be described below. According the linked open data (LOD) rationale, it is also essential to link the instances present in the VRE with authority files and public reference bases. In this field, a pilot alignment experiment was carried out using the IdRef[12] authority base with data from the SIPROJURIS project under the supervision of François Mistral, head of authority control at the *Agence bibliographique de l'enseignement supérieur* (ABES)[13]. The alignment carried out between the IdRefs and the authority files native to the *symogih.org* project enabled the list of each professor's publications to be displayed in the SIPROJURIS files, by retrieving them in real time from the records of the ABES library catalogue[14].

The *symogih.org* VRE also includes a system of spatial data management, GEO-LARHRA[15], and an environment for semantic annotation and text editing in XML/TEI formats[16], which may be accessed from the general website.

3. Generic and open modelling at the heart of the information system

From the outset of the *symogih.org* project, 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, and at the same time reflecting existing standards. As such, we aimed to guarantee interoperability between the data being produced and those from other projects using the same approach, as well as with data from heritage institutions such as museums, *Europeana* and the *Bibliothèque nationale de France*. This approach was essential in order to implement a generic information system that would enable various scientific projects to work in a collaborative VRE.

Two fundamental principles guided the modelling operation for the *symogih.org* project. Firstly, a clear separation was established between the production of data and the research agenda that spurred its collection. Of course, all data production originates in a research agenda. However, the information stored in the research environment must be modelled in the most objective manner possible in order to enable its re-use for new research. The historical dimension of the research is applied whenever the data collected is queried; the data is aggregated based on the research agenda and subject to requests which stem from historians' lines of inquiry. For example, 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 a specific group or class [5].

Secondly, it is essential to proceed with data fragmentation; i.e. undertaking the process of breaking down the information into elements that correspond to simple, independent propositions which ideally cannot be further broken down themselves [13-10].

[10] A few examples on https://frama.link/phn-shiny, knowing that these analyses and data visualizations have a heuristic function and have been set up to be mainly used by researchers in the concerned projects. Cf. https://www.huma-num.fr/services-et-outils – Huma-Num is the French representative of DARIAH and CLARIN.

[11] http://symogih.org/?q=rdf-publication

[12] https://www.idref.fr/

[13] https://punktokomo.abes.fr/2019/09/10/labes-soutient-la-recherche-en-humanites-numeriques-2-retours-sur-une-cooperation-fructueuse-avec-le-larhra/

[14] Cf. e.g. http://siprojuris.symogih.org/siprojuris/enseignant/44315 ("Bibliographie externe" tab).

[15] http://geo-larhra.ish-lyon.cnrs.fr/ – cf. [8].

[16] http://xml-portal.symogih.org/ – cf. [9].

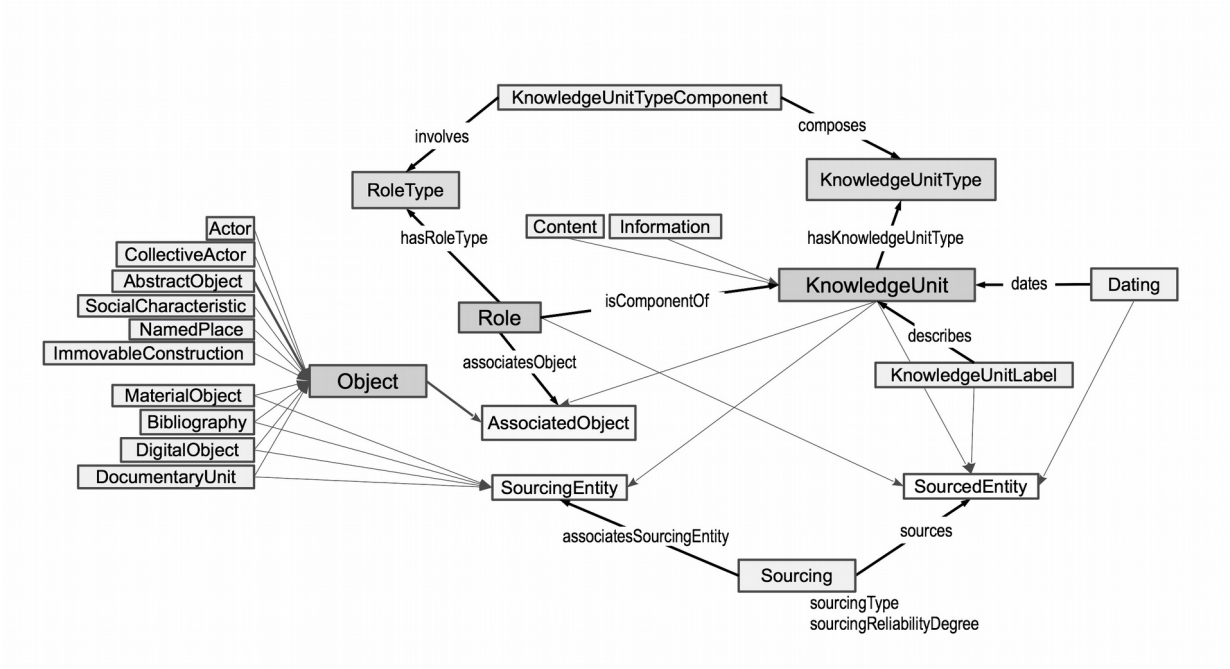

Fig. 1. Ontology of the symogih.org project -version 0.2.1

This fragmentation process must be explicitly documented, identifying the meaning of each proposition as well as the role of each object involved. As such, we would distinguish between an event, such as a congress in its entirety, and the multiple events and situations represented by the presence or absence of various persons at various moments of this event, since the duration and continuity of each person's presence may be varied and significant for any historical reconstitution. Also, in the case of a murder it would be necessary to specify the role of each person involved (victim, assassin, accomplice, etc.) and eventually to break the event down into its constituent phases if these present characteristics that the historian wishes to study. In order to achieve this goal, the project used a generic database model that transforms the model's own instances into data: the information system thereby enables researchers to create new aspects of the model to match the needs of their research agenda. These modelling instances are discussed by the users' community, then validated in order to become useable by everyone. The definitions of the model's instances are published on the main *symogih.org* website in order to provide the meaning of the data published and enable them to be re-used, notably through requests submitted via the SPARQL endpoint[17]. The available data thereby becomes intelligible and re-useable.

Reflecting new developments in semantic technology, a process of rewriting the generic model for the *symogih.org* project was begun in 2013, in order to reconstitute it ontological form and to rethink its alignment with the references in the world of Cultural heritage, such as the CIDOC CRM and its extensions [7]. Central to the simplified representation of the ontology in figure 1 are the two main classes: the *Object* class and the *KnowledgeUnit* class. The first covers all "objects" which have a distinct identity that is stable in the long term, 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 a profession).

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." Within this proposition we may identify a human being, "Galileo Galilei", the discipline of "mathematics", the organisation "University of Padua", and also, implicitly, the place "city of Padua". Each object will be described by a stable identifier, published

[17] Cf. http://symogih.org/?q=type-of-knowledge-unit-classes-tree

in the form of a *Uniform Resource Identifier* (URI), which may be dereferenced on the *symogih.org* site, and via a notice that succinctly expresses its essential characteristics in order to enable other researchers to easily understand what object it refers to. In Figure 1, the *Object* class is broken down into ten sub-classes (e.g. *Actor*, *Collective Actor*, *Abstract Object*, etc.) which were established in the most objective possible manner in order to be suitable for all research contexts.

The second group contains *KnowledgeUnit*s, defined in the ontology of the *symogih.org* project as the assertions of the historian describing a piece of information; i.e. a relation between objects situated in time and space, having virtually existed (and which is therefore not fictional). As indicated above, these assertions are fragmented and designed in the most objective manner possible in order to enable their future re-use in other research contexts. As such, from the proposition given by way of example we may extract an information fragment or Knowledge unit that interlinks, during a given period and for a precise reason a person (Galileo Galilei), an organisation (the University of Padua) and a discipline (mathematics). An instance of the *KnowledgeUnitType* class – 'teaching' in the present example – is defined for each type of information that we wish to store and publish on the *symogih.org* website[18]. This specifies the meaning of the data produced and enables users to understand the relationship between objects that are interlinked within the knowledge unit, with the participation of each one being defined under a precise role (classes *Role* and *RoleType*).

It should be noted that other information may be 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. It is therefore the research agenda applied to a source that enables one to extract various information from it or, in other terms, to build data according to a model: for the sake of data interoperability it is crucial to specify and document this process, which is done thanks to the documentation of instances in the *KnowledgeUnitType* class. The model is therefore not set in stone from the outset, but rather can adapt to various research inquiries, while aiming to provide - thanks to the fragmentation and separation between the inquiry and data production - the maximum level of objectivity. As regards controlled vocabularies, taxonomies may be produced which document, firstly, concepts which are very close to the sources, linked to specific eras or fields; these concepts are then interlinked with more abstract concepts aligned with external reference bases, which ensures interoperability between projects.

[18] See the defintion of the type of information "Enseignement" (Teaching): http://symogih.org/resource/TyIn97

4. Comparison with the *factoid* model and the CIDOC CRM

Let us note that an essential distinction was introduced into the *symogih.org* project's model between those statements which model "facts" as they where (*states of affairs* - for example the fact that Galileo Galilei taught in Padua), and those which reproduce the content of a document literally, so to speak, with each source providing different points of view both on the date and circumstances of such an event and on the interpretation thereof. This is 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; knowledge of this model is fairly widespread in the field of digital history. 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" [11].

In other terms, the *factoids* tend to model the content of the sources, while the "information" defined by the *symogih.org* project tries to describe the "reality" of the past; the "facts." In order to account for this essential distinction for historical research, "contents" have been introduced into the *symogih.org* VRE since 2010 as sub-class of the *KnowledgeUnit* class (Figure 1). These are built so as to be analogous to the "information" units, i.e. by using the generic model instanced in the form of content types, but with a substantially different epistemological level: the "contents" (much like *factoids*) model the assertions of the source, including the full range of uncertainties, contradictions and ambiguities it may hold, while "information" units model the assertions made by the historian having applied the critical method with the aim of establishing "states of affairs". As an expression of the content of the source, "contents" or *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*[19], and by then proceeding to semantic annotation in connection to a shared reference base [16-6].

The *factoid* structure is comparable to the one of "information" in the *symogih.org* project: an assertion is qualified by a type; we then indicate which objects it links by using roles, for which the semantics are specified using a type defined by the researchers [11]. It is not the structure, therefore, but the epistemological value of the data that marks the difference between the two models: in order to go from one to the other level of knowledge 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, the "facts" as they where.

This process of aggregation and changing the epistemological value of the data is generally required in order to meet the needs of the majority of historical research agendas. Indeed, as a general rule, when data is submitted for processing and analysis it is necessary to have at our disposal information that is consistent and non-redundant or contradictory in terms of the same state of affairs, avoiding repetitions and distortions that can bias the results. For example, we cannot compare the careers of a population of university teachers if the data available does not contain unique information units 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 states of affairs and to indicate, as far as possible, their degree of probability in relation to the sources available.

The complex process of aggregation of the contents in the sources, producing historical "states of affairs", must also be modelled and documented. The ontology of the *symogih.org* project offers a simplified approach to this procedure, by specifying at the information sourcing level (the *Sourcing* class in Figure 1), the method used to produce it, as well as the level of reliability of the source in question. This procedure enables other historians to measure the value and reliability of the data produced. One or several sources, or even instances of the "content" already created in the information system, may be linked to one and the same information unit. We may also indicate the precise wording of the objects as they appear in the archival sources.

However, this method is not entirely satisfactory, as it does not allow us to specify in a distinct and precise fashion the reliability and the degree of veracity of the content of each distinct source or *factoid* mobilized within the process of integration that produces the information. This significant limitation depends on, among other things, an ontological "shortcut" which was unintentionally introduced into the abstract model of the *symogih.org* project, and which appears obvious when compared with similarly structured ontologies. The model presents the same cognitive structure as the *Descriptive Ontology for Linguistic and Cognitive Engineering* (DOLCE), a high-level ontology designed as a means of studying the essential structures of natural language as an understanding of reality [17] [20]. In particular, the category of *endurants* (entities which subsist with the same essence over time, such as physical objects, concepts or human beings) are equivalent to the *Object* class in the *symogih.org* project, while *perdurants* in DOLCE (entities which develop over time and may be modified from one instant to another, such as events or processes), correspond to knowledge units: they express the relation subsisting between objects at a given moment in time, whether temporary or extended [7]. The same cognitive perspective is found in the conceptual model of the CIDOC CRM, which was created to enable interoperability between data produced in the field of preservation of cultural assets (standardized by the ISO in 2006) [14]: the *Persistent Item* and *Temporal Entity* classes are respectively equivalent to objects and information units in *symogih.org*, but with an important distinction in terms of epistemology.

A *Knowledge unit*, a piece of "information" from the *symogih.org* project encapsulates both the *perdurant* as such, the event as a state of affairs, and the historian's source-based assertion describing it. Of course, thanks to the *Sourcing* class, multiple sources may be associated with a piece of information, and for each of these sources reliability parameters, and the method adopted to produce the knowledge may be indicated. But within this process, each piece of information is conceived as a whole, with all its roles, as one and the same knowledge unit or assertion. In the CIDOC CRM, by contrast, sourcing is virtually carried out for each property (or for each role, to use the equivalent term in *symogih.org* onto-

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

[20] Cf. http://www.cidoc-crm.org/ - http://www.cidoc-crm.org/collaborations - http://www.loa.istc.cnr.it/dolce/overview.html

logy) by producing instances of the *E13 Attribute Assignement* class which, as a sub-class of *E7 Activity,* comes with a wide range of properties enabling users to specify, thanks to suitable typing, the manner in which the knowledge was produced. Sourcing is applied to roles/properties, which enables users to provide, for each component of the information, a specific source and a detailed description of its reliability and the method applied in terms of historical criticism[21].

In the epistemological structure of the CIDOC CRM, the *Temporal entity* class represents a "fact", a *state of affairs* as such (for example a birth or a teaching activity), rather than a mention thereof in a source or an assertion regarding it. A temporal entity is therefore never directly sourced, as in principle it is not sourceable, it exists as such: it is in terms of the association of objects that participate in this event brought about by properties, as well as via the definition of the class to which it belongs, that we may define its identity. The assertion of veracity that arises from the application of the historical method is therefore displaced from the temporal entity as a whole (*symogih.org*) towards its properties and the related "attribute assignements" (CIDOC CRM), which enables a much more detailed and relevant form of sourcing. Consequently, conflicting points of view of various sources regarding the same property can be recorded: "The CRM has been designed to accommodate alternative opinions and incomplete information, and therefore all properties should be implemented as optional and repeatable for their domain and range"[22]. Using this modelling method we could tackle, much like a *detective*, the case of a homicide in which the murderer's identity is unclear, by linking the same role ("being a murderer") with several persons with the potential to be guilty, while coupling the discussion to the sourcing level for each one of these alternatives.

Concerning its epistemological value or level, the CIDOC CRM's temporal entity is equivalent to an "information" in the *symogih.org* ontology (once the assertion element is removed), rather than to "content", as it does not involve modelling what the source says but the event itself. It does not therefore seem legitimate to model the *factoid* as a sub-class of the *Temporal Entity*, as has been proposed by the authors of the article mentioned above [11], since a *factoid* expresses the assertion of a source, not the "fact" itself. In the context of CIDOC CRM, we could model a *factoid* (as well as a "content" in *symogih.org*), considered as a graph gathering all its properties, as an equivalent class to *E89 Propositional Object* , or, if we sought to also encapsulate its formulation in the form of a text, as an instance of the *E73 Information Object* class[23].

5. OntoME (Ontology management environment) and the *Data for History* consortium

These considerations of the alignment, ongoing since 2014, between the ontology of the *symogih.org* project and that of CIDOC CRM have enabled us to highlight the specificities and originalities of the collaborative modelling developed, and at the same time to understand the limitations pointed out in the previous pages [7]. Furthermore, the act of publishing RDF data modelled using an original ontology (one not explicitly aligned with a recognized standard) does not guarantee the interoperability of this data, at least not in the sense of the *Interoperability* article in the FAIR principles. In order to surpass these limitations, an active partnership with the *Special interest group* (SIG) that maintains the CIDOC CRM was launched in 2016, and the LARHRA is now actively participating in the development of this conceptual model, notably with regard to a dedicated extension to create a *Model for social phenomena* (CRMsoc)[24]. A process of alignment between the *symogih.org* ontology with the classes and properties in the CIDOC CRM is underway. The approach adopted within the *symogih.org* project, which enables a collaborative and progressive production of model instances to match the various fields of research, is now integrated into the more robust modelling method used by the CIDOC CRM: the notion of sub-classes enables users to build specialization trees with property inheritance, and to formally defined guarantee coherence between the overall ontology. This interlocking of various levels of abstraction is an essential principle in enabling data issuing from different informa-

[21] CIDOC CRM (version 6.2.1, October 2015): http://www.cidoc-crm.org/Version/version-6.2.1. A CIDOC CRM extension under development, CRMinf, understood as a formal ontology whose purpose is to integrate "metadata about argumentation and inference making in descriptive and empirical sciences" allows to further detail this process of knowledge production, but its presentation is beyond the scope of this paper: http://www.cidoc-crm.org/crminf/

[22] *Definition of the CIDOC Conceptual Reference Model*, Version 6.2.1, October 2015, p. xiii.

[23] http://www.cidoc-crm.org/Entity/e73-information-object/version-6.2.1

[24] http://www.cidoc-crm.org/crmsoc/fm_releases

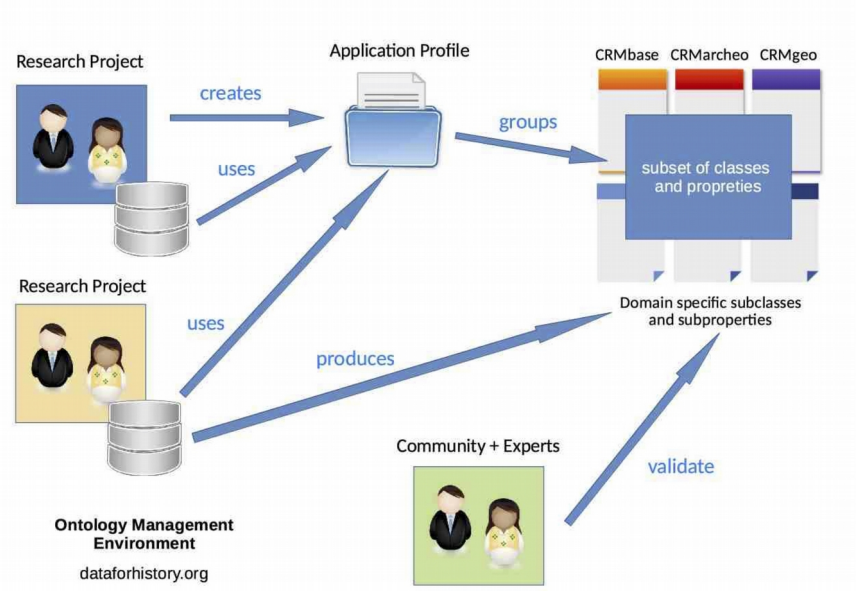

Fig. 2. OntoME (Ontology Management Environment) Use Cases

tion systems, designed with different levels of specialization, to be rendered interoperable.

Finally, while we may observe that several projects and infrastructures have adopted the CIDOC CRM as a conceptual framework, and that its use is becoming more and more widespread[25], this adoption is often in the form of developments to local extensions, or interpretations of the model which do not converge with those of other projects, or which are perhaps never even published. This situation represents a major obstacle to the interoperability of the data. It seemed therefore appropriate to launch an initiative to federate these efforts, notably in the field of geo-historical data, and at the same time to make available to the research community the ten years' worth of experience in collaborative modelling of historical data achieved by the *symogih.org* project.

This ambition has 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[26], it seemed opportune (given the limitations they then had in terms of collaboration between projects), to establish a new online application enabling users to align the data models of existing information systems with that of the CIDOC CRM, in the form of an ontology management environment named OntoME[27]. The development of this platform, which is already operational and accessible on the web, is still ongoing, and the various use cases presented here will be implemented gradually. Once the base framework is completed, the code will be switched to *open source*.

As Figure 2 shows, OntoME enables the application of several use cases, adapted to the needs of various projects. First and foremost, OntoME is a CIDOC CRM learning space, enabling users to get to grips with the ontology more rapidly than they would by reading the standard, thanks to the display of inherited properties for each class and to an integrated overview over all the CRM family of extensions (FRBRoo, CRMgeo, CRMsci, etc.)[28]. A system of filters enables them to analyse and compare a single limited set of ontologies, and more specifically the versions of each one selected.

A research project could go on and, on the one hand, create sub-sets of CIDOC CRM classes and properties, as well as those of its extensions, by defining application profiles that can then be exported via an API to act as a model for the production of data in a distributed information system. If, on the other hand, the project already has its own database,

[25] https://doc.biblissima.fr/ontologie-biblissima - https://masa.hypotheses.org/500

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

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

[28] http://www.cidoc-crm.org/collaborations

it could import the existing model and align it with the CIDOC CRM and its extension, then export the alignment generated and use it to wrangle the data, before publishing them in RDF. Scheduled data exporting, or rewriting in real time, would enable users to continue generating data in the project information system, based on its original model, all while transposing the available data into an interoperable format.

A project could also produce more specialized classes and properties, specifically dedicated to the field of study, while also entering them into the more abstract tree of classes of the CIDOC CRM. In order to do so, the project's members have access to a dedicated namespace of which they are the sole master, in which it is possible to also add descriptions and additional explanations regarding the existing classes and properties; this can be done in several languages in order to make the ontology more intelligible to the users of the distributed information system.

These new classes and properties will enable users to produce data that will be interoperable at a higher level of abstraction, thanks to their being registered within a specialization tree. The new classes may also be combined with the classes and properties of the CIDOC CRM under application profiles that will make up a coherent model. The process is carried out publicly: the members of the community, or the experts, can evaluate and discuss the new classes and properties, or the application profiles, in a dynamic that will enable them to progressively improve the quality of the model and to supplement it with more specific research fields, while also accommodating the various research agendas of historians or integrating the standards adopted by heritage institutions. The elements of the new models could be freely added to the application profiles of the various projects, by virtually creating a model that is both multiform and coherent, built progressively by the community. We may also note that there are plans to enable the importing of other standards into OntoME, so as to enable the alignment of existing classes with these models where desired.

Finally, since the CIDOC CRM was created to enable interoperability for the museum field, and the field of Cultural Heritage more generally, its ontological commitment is somewhat different from that which pertains to the production of data for the sake of historical research. It therefore seemed opportune to create, within OntoME, a specific namespace: the "CIDOC CRM Extension for Historical Data Management"[29], which gathers high-level classes and properties missing from the CIDOC CRM, or which specify its characteristics so as to be more easily intelligible and adapted to historical research. This "experimental" namespace, based on the experience of the *symogih.org* project, will be subject to discussion with the domain experts and will enable them to clarify, in its successive versions, those concepts which are sometimes defined ambiguously or which are missing from the CIDOC CRM. For example, a misunderstanding often appears about the definition of a geographical place due to the purely abstract definition of class E53 (Place), while the geographical location in its more "real" dimension is modelled in class E26 (Physical Feature). This led to the introduction of class histC8 Geographical Place, in the new namespace dedicated to historical research, enabling clarification of the concept[30].

The collaborative dimension of ontological design enabled by OntoME invites managers of VRE to participate in the process with a view to creating a wide, distributed information system (Figure 3) enabling the implementation of the FAIR principles. In the centre of the figure are existing or future IT systems which contain data instances and are directly administered by research projects or heritage institutions. The ontology specific to each field of research or cultural heritage preservation, extended towards a global model, will be managed in spaces per project in OntoME based on the method presented above. The level of specialization in the ontology will be flexible and defined for each project, whereas data instances are produced locally within different information systems using application profiles.

As regards shared vocabularies, the other pillar of interoperability, an alignment with the IdRef authority files from ABES, or other plarforms providing stable identifiers in form of URIs, e.g. Wikidata, will be implemented by using shared *gazetteers* and *thesauri* in order to align not only the model, but also, as far as possible, the instances contained in distributed information systems, e.g. concepts, places, persons etc. The use of *Opentheso*, developed by Miled Rousset[31], as an intermediate layer between the silos of the various projects and public authoritiy files is currently in its trial phase. This should enable researchers to gain access to richer vocabularies which are more specific to the field and time periods being studied, and at the same time to align them (in

[29] https://ontome.dataforhistory.org/namespace/11#summary

[30] I will dedicate more developments to this topic in further publications.

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

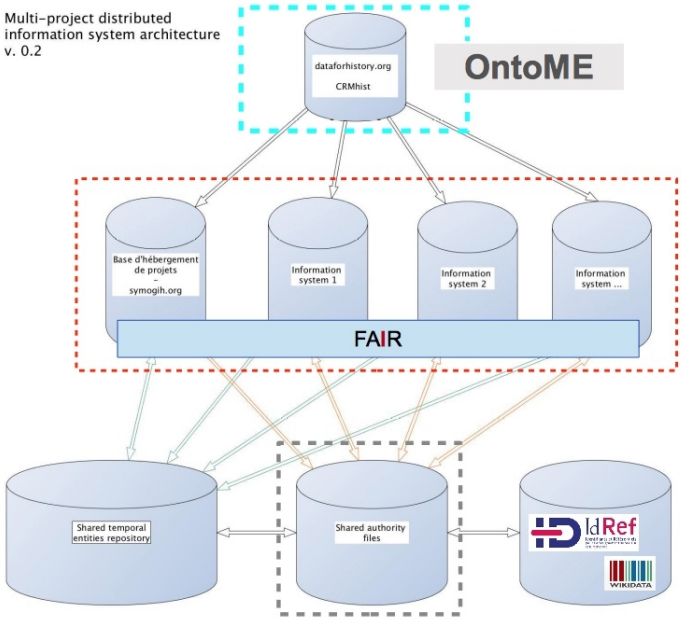

Fig. 3. Architecture of a geo-historical distributed information system

an ergonomic manner and at a higher level of abstraction), with the IdRef authorities and with other authority controls (VIAF, Wikidata, Library of the Congress, etc.). Finally, the projects' data will be made available to other researchers and the public in the form of LOD, either directly *via* a local API for each project, or by harvesting the data made available and storing them in a shared location in the form of a *triplestore* or an indexed search engine, such as Huma-Num's Nakala platform[32].

This vision, initiated by the Digital history research team of the LARHRA, is driven and shared by a community that is growing progressively under the umbrella of the *Data for History* consortium, a network of interested projects. Ahead of its formation, two French workshops were organised on the topic, one held in Lyon in June 2016 and one in Brest in March 2017. The consortium was then officially constituted in a workshop in Lyon, in November 2017; in attendance were around thirty historians, art historians, archeologists and information science specialists from six European countries. The ABES IdRef team, as well as the team from the Archives de France (SIAF) expressed their interest and support for this initiative, which will help bring together the work of those producing metadata from documents preserved in archive deposits and libraries with those of historians, using an *open data* approach. The second *Data for History* workshop was held in May 2018, again in Lyon, alongside the meeting of the CIDOC CRM SIG. A *Data for History* panel on the interoperability of data was presented to the EADH conference in Galway in December 2018[33], followed by the annual meeting of the consortium held in April 2019 in Leipzig. The consortium operates a public forum and a mailing list, both of which are open to anyone upon request.

5. 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 the answer to this question is certainly positive, but that, at the same time, it requires the roll-out of a process that can only work under certain conditions. I shall summarize these in three points.

[32] https://www.nakala.fr/index.html.en

[33] https://eadh2018eadh.wordpress.com/

Firstly, it will be necessary to adopt one or several ontologies to suit the historical research model. The comparative analysis of the ontologies of the *symogih.org* project, the *factoid* data model and the CIDOC CRM shows that the act of characterizing the essential epistemological difference between the modelling of source assertions (*factoids* and "content") and the reconstitution of the historical world ("states of affairs" and "information") provides a robust solution to the issue of ambiguity and contradictions in sources, in accordance with the critical principles on which the historical discipline is based. At the outcome of this analysis, the CIDOC CRM appears to be a highly suitable conceptual framework for promoting the interoperability of geohistorical data, but at the same time it must be added to and adapted so as to take into account the specificities of historical research, and be extended to other domains and research fields by adding sub-classes and sub-properties which are essential to the work of historians.

Secondly, this approach requires the implementation of an infrastructure to match the vision of elaborating a common fine-grained and adaptive understanding of information, one which is easy for projects to use. OntoME, coupled with Opentheso and the IdRefs/Wikidata, constitute the first pillars of an infrastructure whose endeavour is to develop projects within an ecosystem of VREs and distributed information systems connecting to communal services.

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 foundations of a network. We must hope that the holders of projects and platforms, notably those which are very broad in scope, will have the foresight to get on board with this dynamic so that it can truly flourish.

References

[1] Akoka, J. et Comyn-Wattiau I. (2001). Conception des bases de données relationnelles, Vuibert, Paris.

[2] Alerini, J., Lamassé, S. (2011). Données et statistiques. L'avenir du travail en ligne pour l'historien. *Les historiens et l'informatique.* Un métier à réinventer, Rome, École française de Rome, p.171-187.

[3] Allen, J. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26, p. 832-843.

[4] Audibert L. (2009). *Bases de données : de la modélisation au SQL*, Ellipses, Paris.

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

[6] Beretta F. (2016). Pour une annotation sémantique des textes: le projet *symogih.org* et la *Text encoding initiative*. *Bruniana & Campanelliana,* vol. 22, n°. 2, p. 453-465.

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

[8] Beretta F. et Butez C. (2013). Un SIG collaboratif pour la recherche historique. *Géomatique Expert*, n. 91, p.30-35 / n.92, p.48-54.

[9] Beretta F, Letricot R. (2017). Le portail XML du projet symogih.org : un projet d'édition numérique collaborative de sources et d'informations historiques », *Humanités numériques et construction des savoirs*, London, ISTE Editions, p.125-143.

[10] Beretta F., Vernus P. (2012). Le projet SyMoGIH et la modélisation de l'information : une opération scientifique au service de l'histoire. *Les Carnets du LARHRA*, n.1, p.81-107.

[11] Bradley J., Pasin M. (2015). Factoid-based prosopography and computer ontologies: Towards an integrated approach. *Literary and Linguistic Computing*, vol. 30, issue 1, p.86–97.

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

[13] Dedieu J.-P. (2004). 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, vol.5, 2004, p.101-114.

[14] Doerr M. (2003). The CIDOC CRM. An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine* vol. 24, number 3, p.75-92.

[15] Domingue J., Fensel D., Hendler J. A., eds (2011). *Handbook of semantic web technologies*. Vol. 1. Foundation and technologies (Berlin / Heidelberg, Springer.

[16] Eide Ø. (2014-2015). Ontologies, Data Modeling, and Tei. *Journal of the Text encoding initiative*, vol. 8.

[17] Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A. (2003). *WonderWeb Deliverable D18 Ontology Library* (final), Trento, Laboratory For Applied Ontology.

[18] Meroño-Peñuela, A., Ashkpour A., van Erp M., Mandemakers K., Breure L., Scharnhorst A., Schlobach S., van Harmelen F. (2015). Semantic Technologies for Historical Research: A Survey. *Semantic Web*, 6, p. 539-564.

[19] Segaran T., Evans C., Taylor J. (2009). *Programming the Semantic Web.* Beijing e.a., O'Reilly.

[20] Wilkinson M.D., Dumontier M., Aalbersberg I.J., Appleton G., Axton M., Baak A., Blomberg N., et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3 (March 15, 2016): 160018.