Ontology-based Data Organization for the Enslaved Project

Cogan Shimizu and Pascal Hitzler
Kansas State University

1. Introduction
Recent, and not so recent, advances in technology have changed the face of academia. The World Wide Web allows for the sharing of information at an unprecedented level. Data storage is the cheapest it has ever been. These have spurred a renaissance in many domains, allowing data collection and archiving to scale to previously unimagined levels.

Unfortunately, though, these new advances do not come without their disadvantages. Such incredible *access* to information has become *inundation*; anyone may publish their data in any format. The sheer heterogeneity of data and metadata formats, provenance, licensing, makes it difficult for anyone to adequately discover the data they require (Hitzler and Shimizu 2018). These problems are faced by any data-intensive science, but is perhaps doubly felt in history, where the drive to meticulously identify, record, and trace both source and provenance is paramount.

But it is certainly not only researchers that are affected. Researchers, corporations, organizations, and the average citizen (e.g. a curious family member investigating their genealogy) alike want or need to be able to discover information, to be able to pass it on, to ingest and reuse.

In order to make it easier to discover and thus use or reuse research data, we are thus presented with a number of important questions.

1. How can a researcher discover data?
2. Once discovered, how can a researcher make use of it?
3. How do two different researchers working on two different perspectives of the same thing share their data?
4. How does a generation of researchers share their findings and data with the next?
5. How can collaborators with highly heterogeneous sources fuse their data?

It is to be noted, that data discovery and reuse are not fully driven by the data itself, but also how that data is *modelled*. That is, the data is accompanied by sufficient *metadata* that describes how the data is organized, what the data describes, does the data utilize existing vocabularies, and other such documentation. Recently, to address this need, there has been a movement towards so-called FAIR data, that is ``Findable, Accessible, Interoperable, and Reusable.''

Computer Science research in recent decades has gravitated towards a particular approach for representing such metadata, commonly known under the term ``Ontologies''. Ontologies have been described as ``explicit shared specifications of conceptualizations,'' and as such they indeed seem like a natural fit for such a role (Gruber 1993), with respect to interoperability (Hitzler, et al. 2012) and reusability. They offer a way to create human accessible organization for large amounts of complex data and act as a vehicle for the sharing and reuse of knowledge. The research field known as ``Semantic Web''

within Computer Science and adjacent disciplines has driven the development of corresponding standards (Hitzler, Krötzsch and Rudolph 2010), both for ontologies (the Web Ontology Language OWL (Hitzler, et al. 2012)) and also for (some) actual content data (the Resource Description Framework RDF (Cyganiak, Wood and Lanthaler 2014)). In a newer development, the term ``Linked Data'' has been used to describe such data and accompanying metadata, and most recently, yet another term, ``Knowledge Graph'' (and ``Knowledge Graph Schema'' for what used to be called ontologies) is increasingly used. These together with common shared ontologies (e.g. Schema.org, Dublin Core, Vocabulary for Interlinked Datasets) allow for indices to be built, aiding in discovery. The use of established standards generally makes access and reuse easier due to robust tooling infrastructure.

Yet, despite all these advances which already simplify data discovery and reuse, due to the sheer amounts of data generated every day, management of such data is still a very significant drain on resources for both the data provider and the data user, and thus corresponding research continues to investigate in order to improve methods and tools for data management. One particular aspect of this research pertains to the question how ontologies (i.e., the metadata or schema) should best be organized such that data integration, reuse, and maintenance become easier. Our research on this topic, which we advance and apply in the Enslaved project, is particularly based on the premise that one effective way to obtain ontologies that emphasize interoperability and reusability is to develop them with a modular structure and that sufficiently modularized ontologies are designed in such a way that researchers may easily adapt the ontology to their particular use-case, while still maintaining integration and relationships with previous versions of the ontology or, even, another researcher's version of the same ontology.
In this chapter, we will explain the particular approach we employ, and provide examples from the Enslaved project. In Section 2, we define the key concepts behind modular ontologies. Section 3 discusses the methodology used to construct a modular ontology for the Enslaved project. Sections 4 and 5 present examples from the Enslaved ontology and its documentation, respectively. In Section 6, we conclude.

2. Patterns, Modules, and Modular Ontologies

At their most general, *patterns* are simply an observed invariance, e.g. a pattern of usage, or a pattern of structure, which is conceived as recurring. In the case of ontology engineering, we have an analog: ontology design patterns (ODP). These are *conceptual* patterns that recur across different data models. Modelling with ODPs has established itself as an ontology engineering paradigm (Hitzler, Gangemi, et al., Ontology Engineering with Ontology Design Patterns -- Foundations and Applications 2016) which continues to be refined.

For instance, the conceptualization of an Event is a frequently occurring pattern (Krisnadhi and Hitzler, A Core Pattern for Events 2016). Figure 1 displays a schema diagram for an Event pattern. A schema diagram is a graphical representation of how concepts are related to each other via properties. It should be noted

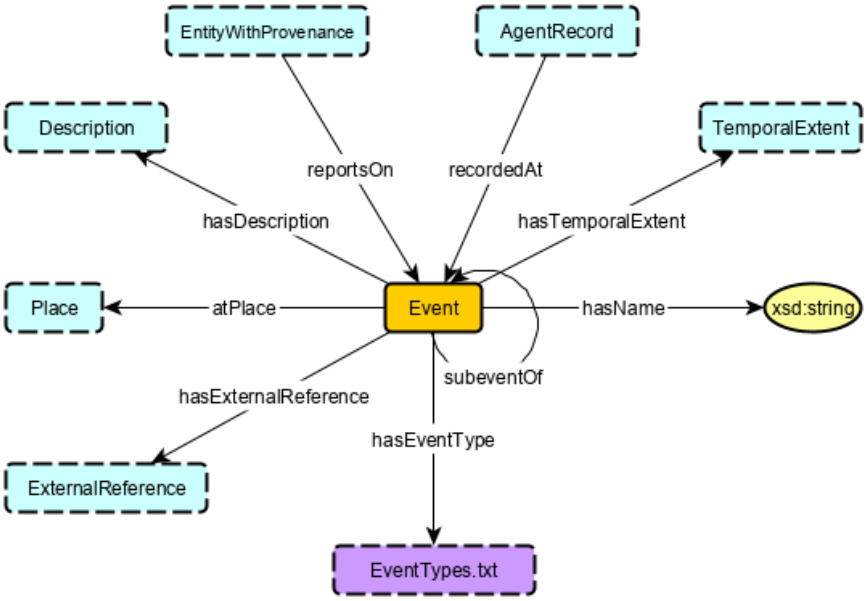

Figure 2 This is the Event module in the Enslaved Ontology

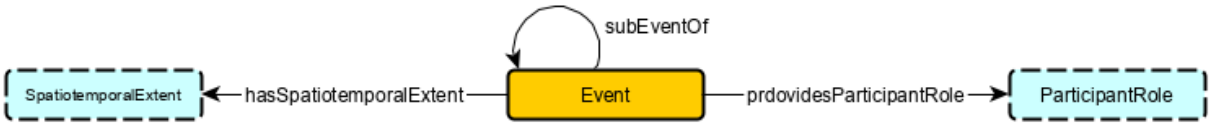

Figure 1 This is the Event pattern

that schema diagrams aim to be an intuitive representation of the formally encoded knowledge. The rounded rectangles are the concepts. The coloring of a node in the diagram is used by us to indicate where a concept ``belongs'' and how it should be used. For example, the node Event is the ``core concept'' of the pattern. The other nodes, colored light blue with dashed borders, indicate that they are ``external'' to the pattern. That is, the pattern acknowledges that they are outside of its scope -- it may be more complex than the pattern needs, or may be used to indicate that additional, domain specific modelling is needed. This is because patterns are designed to be sufficiently general as to apply to many different cases as possible.

From the schema diagram of this particular pattern, it can be read off that an event has a spatio-temporal extent (i.e., it occurs in space and time), and that it provides roles for participants (i.e., there is something or somebody participating in the event). Furthermore, events can be subevents of other events. Of course, this is a very simple model, and it is easy to come up with events which do not fit this pattern, and in particular there is usually much more to an event than time, space, participation, and sub-events. A good pattern is not made to capture all aspects of our complex reality. Rather, it is a simple model which can serve as a core organization principle for data, and which will be suitable in many -- though certainly not all or even most -- use cases.

In the Enslaved Ontology, as we discuss in later sections, Events play a key role, since they also play a key role for the corresponding historic data, mainly because recording is an event in itself, and corresponding

historic records were often produced on the occasion of a historic event. In this sense, we strive to make our models correspond as well as possible to the native data organization principles followed by the experts in a field. However, a general pattern such as that depicted in Figure 1 is usually too generic and minimalistic for a concrete use case. Thus, it is necessary to create a corresponding *module* from it by adapting the pattern to the specific domain and use-case in mind. Figure 2 shows the Event module of the Enslaved ontology (V1.0). As can be seen, there are numerous additions to the original pattern. Further, it was deemed helpful to disconnect the spatial and temporal aspects of the event into separate relationships (a discussion of this is outside the scope of this paper). The purple box, in this case, indicates a relationship to a controlled vocabulary of *event types*.

Now, when engineering an entire ontology, using patterns and turning them into modules, the end result is a so-called Modular Ontology. Modular ontologies are ontologies which are made in such a way that the original information pertaining to the patterning and modular structure is retained via metadata in the form of Ontology Design Pattern Representation Language (OPLa) annotations (Hitzler, Gangemi, et al., Towards a Simple but Useful Ontology Design Pattern Representation 2017). These annotations facilitate the maintenance and future evolvability of the ontology as a whole. For example, when a new type of data source is encountered or as more (or less) complex representations are needed for a concept, the annotations can guide the developer in ensuring all the axioms in a module are updated.

Before we move on to more detailed explanations of our modeling approach, we want to emphasize that the schema diagrams do not by themselves constitute (part of) the ontology or pattern. The ontologies or patterns are rather described in a formal language, usually the Web Ontology Language OWL as standardized by the World Wide Web Consortium (Hitzler, et al. 2012). It would be out of scope for this chapter, though, to discuss this in more depth.

3. Modular Ontology Modelling Approach

We follow a modular ontology modeling approach, which is an extension of ontology design pattern-based modelling (Blomqvist and Sandkuhl, Patterns in Ontology Engineering: Classification of Ontology Patterns 2005) (Gangemi 2005) (Hitzler, Gangemi, et al., Ontology Engineering with Ontology Design Patterns -- Foundations and Applications 2016). As previously mentioned, it is designed to yield a high-quality ontology, emphasizing reusability, as well as interoperability with future expansions, both in terms of scope and in terms of granularity, and with other ontologies in the domain. This modular ontology modeling approach and its rationale have been described in (Krisnadhi and Hitzler, Modeling with Ontology Design Patterns: Chess Games as a Worked Example 2016), and it is closely related to the eXtreme Design approach (Blomqvist, Hammar and Presutti, Engineering Ontologies with Patterns -- The eXtreme Design Methodology 2016). Following this methodology as laid out in (Krisnadhi and Hitzler, Modeling with Ontology Design Patterns: Chess Games as a Worked Example 2016) and further detailed in (Krisnadhi and Hitzler, A Tutorial on Modular Ontology Modeling with Ontology Design patterns: The Cooking Recipes Ontology 2018), we took the following subsequent steps to construct the Enslaved Ontology.

3.1 Step 1: Define use case or scope of use cases.

African enslavement was fundamental to the making of Europe, Africa, the Americas, and Middle East and parts of the Asian subcontinent. Its enduring legacy continues to shape the moral questions of humanity in modern times. In the past decade, there has been a steady growth of interest through film, television, and historical fiction. Indeed, there has also been constantly growing scholarly interest. At the same time,

however, it is a worthy goal to further scholarly output, as well as to push results to the general public. The Enslaved Project aims to shed light on questions such as:

- How can we more effectively answer important moral questions?
- How can we make those questions part of a broader public discourse?
- What sources are available?
- How can we give broad access to them?
- How in the decades to come will scholars answer questions about black bondage and its legacies when much valuable source material is deteriorating due to inattention, siloed scholarly activities, and underfunded archives?

With recent advances in technology, there are more options in gathering, saving, and representing information about enslaved Africans, their descendants, and those who asserted ownership over them throughout the world. The cutting edge of the digital humanities and social sciences is seeing to the identification, digitization, analysis, and public availability of such resources. As a result, a growing number of collections of original digital manuscript documents, digitized material culture, and databases, that organize and make sense of records of enslavement, are free and readily accessible for scholarly and public consumption. Over time, projects and teams have generated a wide variety of databases with different foci and specializations represented in myriad of formats. Thus, although the data is available through these data silos, scholars, students, and the interested public are presented with a number of challenges:

- Most of these databases focus on the individuals of the slave trade, but data is often limited to the focus of the project. Further, the task of disambiguating (or merging) individuals across multiple datasets is nearly impossible given the current, siloed nature of all databases about slavery and the enslaved;
- There is no central, universally recognized clearinghouse for slave data. As such, it is difficult to find projects and databases;
- Individual projects and databases are isolated, preventing federated and cross project searching, browsing, and quantitative analysis;
- There are no best practices for digital data creation collectively agreed upon by the scholarly community;
- Many projects are in danger of going offline and disappearing;
- Important data is often lost or remains locked away in scholars' files, completely inaccessible to other scholars, students, descent communities, and the general public;
- Project participants rarely get scholarly credit for the work that goes into creating and releasing digital data;
- Humanists may often have little incentive to deposit datasets.

To address these challenges, the Enslaved project is set to pioneer a new model for collaborative humanities scholarship. The technical goal of Enslaved is thus to establish what we call the Enslaved Hub to provide one-stop querying and inspection capabilities for integrated historic data on the slave trade, originating from heterogeneous contributors and their data sources, thereby allowing students, researchers and the general public to understand and reconstruct the lives of individuals who were part of the historical slave trade. To address the underlying data integration issues, Enslaved opted to construct a knowledge graph, expressed in RDF, with an underlying schema in the form of an OWL ontology.

3.2 Step 2: Collect competency questions while looking at possible data sources and scoping the problem, i.e., decide on what should be modeled now, and what should be left for a possible later extension.

Competency questions were collected in a variety of ways. First, search suggestions were solicited from partner projects, seeking expert input from historians actively engaged in slavery studies and databasing. The partners discussed potential audiences for the Enslaved project and how the different audiences would have their own usage characteristics, expectations, and needs.

Specific search suggestions focused on categories such as names, events, relationships, and place and time constraints. Secondly, project partners drafted competency questions by formulating queries about the data in natural language. Below we show a representative selection of the competency questions.

- Public 5: List the enslaved people in Reed County, NC, in the second half of the eighteenth century.
- Public 12: Who were the godparents of my great-great grandmother, Beatriz of the Ambaca nation, baptized at Sao Jose church in Rio de Janeiro on April 12, 1840.
- K12 1: Who did Thomas Jefferson enslave at Monticello?
- K12 9: How many enslaved children lived in Boston when Phillis Wheatley lived there?
- Pro 4: What were the gender ratios of enslaved people identified as being of XXXX ethnicity?
- Pro 6: In what records does the enslaved person named XXXX appear? What were XXXX's professions? What places did he live? Who were his/her children and children's children? Who did he marry?
- Pro 9: I am researching an enslaved person named Mohammed who was a new arrival from West Africa in Charleston in 1776. Is there data about what slave ship he might have been on?
- Pro 20: What ever happened to Bernarda Angola, a Free African who ran away from her mistress Maria dos Santos Pereira, in June 1845?

The historians thus designed a suitable scope based on the competency questions and the availability of data sources. From the potential data sources to be incorporated into the Enslaved Hub, eight were selected, namely African Origins[1], Voyages: The Trans-Atlantic Slave Trade Database[2], Slave Societies Digital Archive[3], Dictionary of Caribbean and Afro-Latin American Biography, Dictionary of African Biography and African American National Biography,[4] Freedom Narratives,[5] Legacies of British Slave-ownership,[6] The Liberated Africans Project,[7] and Slave Biographies[8] as they provided a range of different kinds of data and seemed representative as a starting point.

3.3 Step 3: Identify key notions from the data and the use case and identify which pattern should be used for each. Construct a set of modules from these.

[1] http://www.african-origins.org/
[2] http://www.slavevoyages.org/
[3] http://www.vanderbilt.edu/esss/
[4] https://hutchinscenter.fas.harvard.edu/AANB
[5] http://freedomnarratives.org/
[6] http://www.ucl.ac.uk/lbs/
[7] http://liberatedafricans.org
[8] http://slavebiographies.org/

The list of key notions was quickly finalized during a modeling meeting where historians, data experts, and ontology engineers first drafted the modules and the overall ontology. Due to the ontology's primary focus on historic data, time, place, and provenance play important roles. The content focus of the ontology is largely on persons and key biographical data, such as their name, age, sex, occupation, status (e.g. enslaved or freed), race, ethnolinguistic or geographical origin, whether or not (and the extent to which) they participated in events, and relationships to other persons or organizations such as family relations or ownership relations. The occurrence and records of events play a central role in the data, as historic records usually originate from specific events (e.g. mass baptisms, ledgers). In order to allow this data to connect to outside databases, it seemed necessary to model a generic way for providing descriptions of data, external references, and specific research projects contributing data.

Whilst identifying ontology design patterns as a basis for corresponding modules, some were choices were easily made, such as using the core of PROV-O (Lebo, Sahoo and McGuinness 2013) for provenance; using agent instead of person in order to encompass organizations when needed; an agent roles pattern (Krisnadhi, The Role Patterns 2016) for participation in events and for inter-agent relationships. On the other hand, some of the person data, such as sex, occupation, enslaved/freed status, race, and origin seemed to be best represented with controlled vocabularies, in order to more closely match the data, as well as reduce ambiguity. An interesting exception was age, as the data presented it both numerically and in age categories. Due to the inherent complexity of names, due to ethnolinguistic origins, ambiguity of data, and temporality, we opted to simply utilize a name stub (Krisnadhi and Hitzler, The Stub Metapattern 2017), that is, a relatively simple placeholder module to be more robustly modelled at a later time. Further, a comprehensive treatment of historic places was clearly out of scope for the initial phase of the project, and so we took a limited approach compatible with external efforts, such as historic gazetteers.[9]

As the focus is on historic data, as reported by various (and possibly conflicting) sources, it was also obvious that the modelling, and the resulting knowledge graph, would contain possibly conflicting data, rather than anything resembling a base truth. From a historian's perspective, this is of course obvious, but required careful consideration in our approach to the knowledge graph schema design. To overcome this hurdle, we acknowledge the possibly conflicting nature of reports by referring to historic data about agents in the form of records, rather than directly about any ontological individual.

3.4 Step 4: Put the modules together and add axioms which involve several modules.

In the interest of brevity and space, we defer this discussion to Section 4 where we briefly describe two of the core modules of the Enslaved Ontology.

3.5 Step 5: Create OWL files.

The OWL files for the Enslaved Ontology were generated using Protégé (Musen 2015). The requisite metadata information pertaining to patterns and modules were added using the tool presented in (Shimizu, Hirt and Hitzler 2018), which guides ontology developers in adding Ontology Design Pattern Representation Language (OPLa) annotations (Hitzler, Gangemi, et al., Towards a Simple but Useful Ontology Design Pattern Representation 2017).

4. Examples from the Enslaved Ontology

[9] http://whgazetteer.org/

This section will present two of the core patterns and modules of the Enslaved ontology. These descriptions are adapted from the technical documentation for the Enslaved Ontology found at https://docs.enslaved.org.

4.1 The AgentRecord Pattern

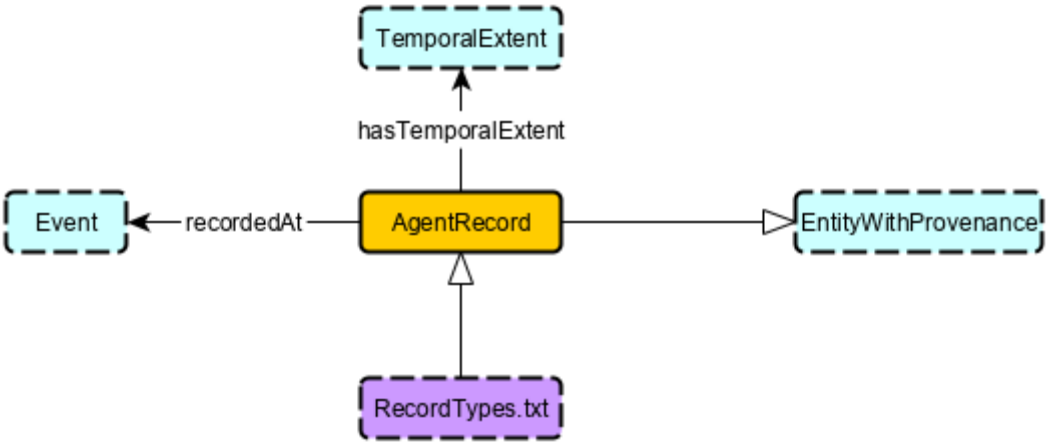

Figure 3:Schema Diagram for the AgentRecord module. Orange boxes indicate classes. Dashed blue boxes indicate classes which are part of another module. White-headed arrows indicate subclass relationships.

Each piece of information about an agent is organized as a record for that agent. Generally, each agent will typically have many agent records associated with it, and the collection of them constitutes everything that we know about that particular agent. Each agent record by itself usually contains one piece of information about the agent, plus provenance information about this piece of information. AgentRecord has a number of sub-modules which are described separately; the intended usage is that every agent record is of one of the types given by these sub-modules. PersonRecord is a sub-module of AgentRecord, for records that pertain to persons only, and not to agents in general, such as racial information. For example, a SexRecord, a submodule of PersonRecord, captures exactly one ``piece'' of information, about the recorded sex of the person, from exactly one source, the provenance of the information.

4.2 The EntityWithProvenance Pattern

This module provides the ability to talk about the source document that an agent record is generated from. It also allows for provenance chains that describe the transformation of the data by different activities.

The intended use, which is mirrored by the axiomatization, is that any agent record is at the same time an entity with provenance, which is directly based on *exactly one* other entity with provenance. In addition, entities with provenance may carry the document type of the original source, and indeed at most one,

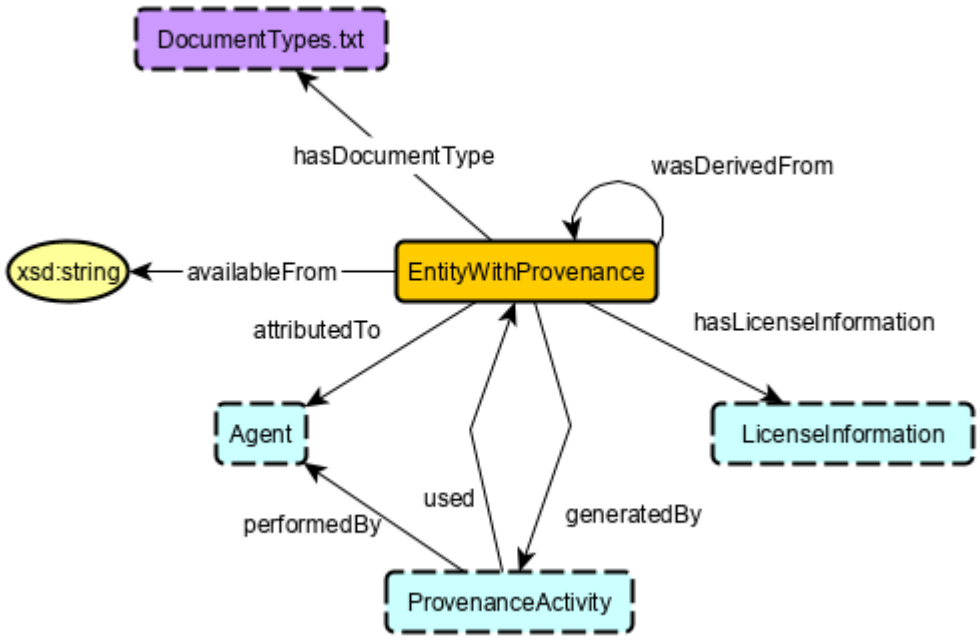

Figure 4: Schema Diagram for the Provenance module, color and shape usage is the same as in the previous diagrams. The dashed purple boxes indicate relationships to controlled vocabularies. Yellow ovals represent datatypes.

which is governed by a controlled vocabulary: this document type refers to the type of document of the historic source document from which the current information has originally been retrieved. Furthermore, an entity with provenance may carry information from where the document may be available, e.g., what library it is hosted at or in which database it can be found in, as well as reference URIs which point at concrete online resources. These Entities may furthermore carry license information, also taken from a controlled vocabulary.

Parts of this module are heavily borrowed from PROV-O (Lebo, Sahoo and McGuinness 2013).

5. Conclusions

Building a data model that is capable of adapting to a wide, and growing, variety of heterogenous contributors and their sources, while supporting various use case scenarios can be a complicated process. To meet these design and usage characteristics, as well as supporting FAIR data practices, we built a modular ontology to model the data, the Enslaved Ontology.

This chapter documented the modular ontology design methodology, as it pertained to the construction of the Enslaved Ontology, perhaps allowing others to follow in our footsteps. For more information and background on the Enslaved Ontology, see https://docs.enslaved.org.

Acknowledgement. This work was supported by The Andrew W. Mellon Foundation through the Enslaved project.

References

Blomqvist, Eva, and Kur Sandkuhl. 2005. "Patterns in Ontology Engineering: Classification of Ontology Patterns." *7th Interational Conference on Enterprise Information Systems.* 413--416.

Blomqvist, Eva, Karl Hammar, and Valentina Presutti. 2016. "Engineering Ontologies with Patterns -- The eXtreme Design Methodology." In *Ontology Engineering with Ontology Design Patterns -- Foundations and Applications*, by Pascal Hitzler, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnadhi and Valentina Presutti, 23-50. IOS Press.

Bonatti, Piero Andrea, Stefan Decker, Axel Polleres, and Valentina Presutti. 2018. "Knowledge Graphs: New Directions for Knowledge Representation on the Semantic Web." *Dagstuhl Reports* 29-111.

Cyganiak, Richard, David Wood, and Markus Lanthaler. 2014. *RDF 1.1 Concepts and Abstract Syntax.* W3C Standard.

Gangemi, Aldo. 2005. "Ontology Design Patterns for Semantic Web Content." *4th International Semantic Web Conference.* Springer. 262--276.

Gruber, Thomas R. 1993. "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition* 199-220.

Hitzler, P., M. Krotzsch, B. Parsia, P.F. Patel-Schneider, and S. Rudolph. 2012. *OWL 2 Web Ontology Language: Primer (Second Edition).* http://www.w3.org/TR/owl2-primer/.

Hitzler, Pascal, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnadhi, and Valentina Presutti. 2016. *Ontology Engineering with Ontology Design Patterns -- Foundations and Applications.* IOS Press.

Hitzler, Pascal, Aldo Gangemi, Krzysztof Janowicz, Adila Krisndahi, and Valentina Presutti. 2017. "Towards a Simple but Useful Ontology Design Pattern Representation." *8th Workshop on Ontology Design and Patterns.* CEUR-WS.org.

Hitzler, Pascal, and Cogan Shimizu. 2018. "Modular Ontologies as a Bridge Between Human Conceptualization and Data." *International Conference on Conceptual Structures.* Springer. 3-6.

Hitzler, Pascal, Markus Krötzsch, and Sebastian Rudolph. 2010. *Foundations of Semantic Web Technologies.* chapman and Hall/CRC Press.

Krisnadhi, Adila. 2016. "The Role Patterns." In *Ontology Engineering with Ontology Design Patterns -- Foundations and Applications*, by Pascal Hitzler, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnadhi and Valentina Presutti, 313-319. IOS Press.

Krisnadhi, Adila, and Pascal Hitzler. 2016. "A Core Pattern for Events." *Workshop on Ontology and Semantic Web Patterns.* IOS Press. 29--37.

Krisnadhi, Adila, and Pascal Hitzler. 2018. *A Tutorial on Modular Ontology Modeling with Ontology Design patterns: The Cooking Recipes Ontology.* Technical Report, CoRR.

Krisnadhi, Adila, and Pascal Hitzler. 2016. "Modeling with Ontology Design Patterns: Chess Games as a Worked Example." In *Ontology Engineering with Ontology Design Patterns -- Foundations and Applications*, by Pascal Hitzler, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnadhi and Valentina Presutti, 3--21. IOS Press.

Krisnadhi, Adila, and Pascal Hitzler. 2017. "The Stub Metapattern." In *Advances in Ontology Design Patterns*, by Karl Hammar, Pascal Hitlzer, Agnieszka Lawrynowics, Adila Krisnadhi, Andrea Nuzzolese and Monika Solanki, 3964. IOS Press / AKA Verlag.

Lebo, Timothy, Satya Sahoo, and Deborah McGuinness. 2013. *PROV-O: The PROV Ontology.* W3C Recommendation.

Musen, Mark A. 2015. "The Protege Project: a look back and a look forward." *AI Matters* 4-12.

Shimizu, Cogan, Quinn Hirt, and Pascal Hitzler. 2018. "A Protege Plug-in for Annotating OWL Ontologies." *The Semantic Web: ESWC 2018 Satellite Events.* Springer. 23--27.