Building the Seshat Ontology for a Global History Databank

Rob Brennan[1], Kevin Feeney[1], Gavin Mendel-Gleason[1], Bojan Bozic[1], Peter Turchin[2], Harvey Whitehouse[3], Pieter Francois[3,4], Thomas E. Currie[5], Stephanie Grohmann[3]

[1] KDEG and ADAPT, School of Computer Science and Statistics, Trinity College Dublin
{rob.brennan, kevin.feeney, mendelgg, bozicb}@scss.tcd.ie

[2] Department of Ecology and Evolutionary Biology, University of Connecticut
peter.turchin@uconn.edu

[3] Institute of Cognitive and Evolutionary Anthropology, University of Oxford
{harvey.whitehouse, pieter.francois, stephanie.grohmann}@anthro.ox.ac.uk

[4] History Group, School of Humanities, University of Hertfordshire

[5] Centre for Ecology & Conservation, Biosciences, University of Exeter
T.Currie@exeter.ac.uk

Abstract. This paper describes OWL ontology re-engineering from the wiki-based social science codebook (thesaurus) developed by the Seshat: Global History Databank. The ontology describes human history as a set of over 1500 time series variables and supports variable uncertainty, temporal scoping, annotations and bibliographic references. The ontology was developed to transition from traditional social science data collection and storage techniques to an RDF-based approach. RDF supports automated generation of high usability data entry and validation tools, data quality management, incorporation of facts from the web of data and management of the data curation lifecycle.

This ontology re-engineering exercise identified several pitfalls in modelling social science codebooks with semantic web technologies; provided insights into the practical application of OWL to complex, real-world modelling challenges; and has enabled the construction of new, RDF-based tools to support the large-scale Seshat data curation effort. The Seshat ontology is an exemplar of a set of ontology design patterns for modelling unncertainty or temporal bounds in standard RDF. Thus the paper provides guidance for deploying RDF in the social sciences. Within Seshat, OWL-based data quality management will assure the data is suitable for statistical analysis. Publication of Seshat as high-quality, linked open data will enable other researchers to build on it.

Keywords: Ontology Engineering, Ontology Design Patterns, Cliodynamics

1 Introduction

The success of linked data has seen semantic web technology widely deployed. However in many domains such as social sciences, despite a strong tradition of quan-

adfa, p. 1, 2011.

titative research, linked data has made little headway. This stems partially from a lack of social sciences research ICT infrastructure but also from the challenges of describing human systems with all their uncertainties and disagreements in formal models.

Here we describe re-engineering an OWL ontology from the structured natural language codebook (thesaurus) developed by the international Seshat: Global History Databank initiative[1] [1]. This evolving codebook consists of approximately 1500 variables used to study human cultural evolution at a global scale from the earliest societies to the modern day. Each variable forms a time series and represents a single fact about a human society such as identifying the capital city, the capital's population or the presence of infrastructure such as grain storage sites. The variables are grouped – measures of social complexity, warfare, ritual, agriculture, economy and so on. However the historical and archaeological record is incomplete, uncertain and disagreed upon by experts. All these aspects, along with annotations need to be recorded. An example variable definition in the codebook is: "**Polity territory** in squared kilometers". An instance of this variable, showing uncertainty and temporal scoping of values is "**Polity territory** 5,300,000: 120bce-75bce; 6,100,000:75bce-30ce **".**

Current data collection in Seshat uses a wiki based on the natural language codebook. This is unsustainable as data quality assurance is impossible and better tools are required to manage the collection, curation and analysis of the dataset. In addition it is desired to publish the dataset as linked data to enable other scholars to build upon the Seshat work. The new tools will be RDF-based using the Dacura data curation platform developed at Trinity College Dublin[2] as part of the ALIGNED H2020 project[3].

This paper investigates the research question: what is a suitable structure in RDF to represent the Seshat codebook that will support data quality assurance? Our technical approach is to develop an OWL ontology describing the codebook based on a set of design patterns for Seshat variables that capture the requirements for variable uncertainty, temporal scoping, annotations and provenance while producing a compact, strongly typed data model that is suitable for quality assurance in a very large dataset.

The contributions of this paper are: an identification of challenges for converting social science codebooks to RDF, a description of the Seshat ontology, new ontology design patterns for uncertainty and temporal scoping, a case study of the Seshat ontology deployed in a data curation system and finally the lessons learned.

The paper structure is: §2 background on Seshat, §3 ontology re-engineering challenges, §4 the Seshat ontology and design patterns §5 deployment of the ontology in the RDF-based data collection infrastructure, §6 lessons learned for social sciences ontology development, §7 surveys related work and §8 is conclusions & future work.

2 Background – Seshat: The Global History Databank

The study of past human societies is currently impeded by the fact that existing historical and archaeological data is distributed over a vast and disparate array of databases,

[1] http://seshatdatabank.info/
[2] http://dacura.scss.tcd.ie
[3] http://www.aligned-project.eu

archives, publications, and the notes and minds of individual scholars. The scope and diversity of accumulated knowledge makes it impossible for individual scholars, or even small teams, to engage with the entirety of this data. The aim of 'Seshat: The Global History Databank' is therefore to systematically organize this knowledge and make it accessible for empirical analysis, by compiling a vast repository of structured data on theoretically relevant variables from the past 10.000 years of human history [1]. In this way, it becomes possible to test rival hypotheses and predictions concerning the 'Big Questions' of the human past, for example the evolution of social complexity[4], the deep roots of technologically advanced areas[5], or the role axial age religions play in explaining social inequality[6].

Seshat data is currently manually entered either by domain experts (historians, archaeologists and anthropologists), or by research assistants whose work is subsequently reviewed and validated by domain experts. The aim is to move to quality assured data collection facilitated by customized software that can automatically import data from existing web resources such as DBpedia. A central requirement for the Seshat information architecture is a flexible and agile system that allows for the continuous development of the Codebook (which structures the data), the adaptation of variables to different research interests and theoretical approaches, and the participation of a large number of additional researchers and teams.

The databank's information structure comprises of a range of units of analysis, including polities, NGAs (i.e. 'Natural Geographic Areas'), cities and interest groups [2]. These are associated with temporally-scoped variables to allow for a combination of temporal and spatial analyses. Each variable currently consists of a value, typically marking a specific feature "absent/present/unknown/uncoded", and indicating levels of inference, uncertainty or scholarly disagreement about this feature. In addition to the values, which are used for statistical analysis, variables contain explanatory text as well as references to secondary literature. Where it is not possible to code variables due to missing or incomplete source data, variables are sometimes coded by inference (for example, if it cannot be ascertained if a given feature was present for a certain time period, but it is known to be present in the time periods immediately before and after, the feature would be coded 'inferred present'). By linking descriptions of past societies to both sources and coded data amenable to statistical analysis, the databank thus combines the strengths of traditional humanistic and scientific approaches.

In the initial stages of the project, the database was implemented in a Wiki, however, as the number of coded variables has been rapidly growing, it was decided to move the Seshat data to an RDF-based triplestore. Based on the Dacura data curation platform, this will facilitate all steps of the Seshat research process, from data gathering, validation, storage, querying and exporting down to analysis and visualization.

---

[4] 'Ritual, Community, and Conflict' research project funded by the ESRC/UK (http://www.esrc.ac.uk/research/our-research/ritual-community-and-conflict/)

[5] 'The Deep Roots of the Modern World: Investigating the Cultural Evolution of Economic Growth and Political Stability', funded by the Tricoastal Foundation/US (http://seshatdatabank.info/seshat-projects/deep-roots-economic-growth/).

[6] 'Axial-Age Religions and the Z-curve of Human Egalitarianism', funded by the John Templeton Foundation; (http://seshatdatabank.info/seshat-projects/axial-age-egalitarianism/)

3 Seshat Codebook to Ontology Re-Engineering Challenges

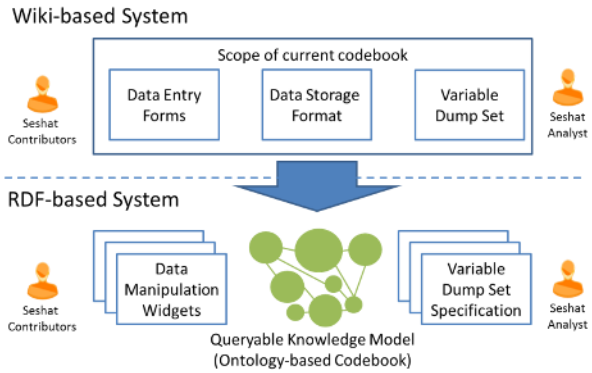

Fig. 1. The Seshat Codebook Re-Engineering Vision

The purpose of creating the Seshat ontology was not simply to translate or uplift an existing dataset to RDF for publication as linked data. Instead we wished to use the ontology at the heart of a set of RDF-based tools that would produce a step change in the data collection and curation capabilities of the Seshat consortium by improving data quality, productivity and agility (fig 1). The primary goal of the formal OWL model is to enable data quality management as even uncertain facts can be omitted, mis-typed, duplicated, inconsistent and so on. This creates a huge data cleaning overhead before statistical processing in the pre-OWL system. Later we hope to extend the utility of DL reasoning to support inference, fact reuse and other advanced features.

The characteristics of the Seshat codebook that made this re-engineering process challenging were as follows:

1. **The codebook was specified in semi-formal structured natural language** designed for human consumption. While a common approach in social sciences it is not often studied in ontology engineering, e.g the methodology for ontology re-engineering from non-ontological resources [3] doesn't consider it.
2. **The ontology must not depend on custom reasoning or triple-stores**. Rather than moving beyond RDF triples to specify qualified relations or temporal scoping it must be possible to use standard, state of the art, scalable triple-stores.
3. **The ontology must be expressive enough to support data quality validation**. The flexibility of wiki-based collection means that the data collected needed extensive cleanup before analysis. The ontology must eliminate this workload.
4. **Every historical fact (Seshat variable value) recorded was potentially subject to uncertainty**. The historical and archeological record often does not permit definite statements of the sort normally recorded by RDF triples.
5. **Each Seshat variable assertion is temporally scoped**. This is because historical facts are typically only true for a certain period of time.
6. **Each temporal scoping was potentially subject to uncertainty**. Many historical dates are unknown or only have known ranges of values.

7. **Time-series variables must support human-readable annotations in addition to data-values.** Seshat is primarily data-oriented but the data collection and expert verification process depends upon the availability of a flexible annotation scheme.
8. **Efficiency of representation for storage and query**. The Seshat dataset is going to be very large. Hence it is desirable to create a tight data model.
9. **Seshat variables do not represent a full model of the domain**. Each Seshat variable is a time series that is conceptually linked to other variables in the codebook based on social science concerns. However there are many missing relations between variables or unifying concepts that only reside in the minds of the domain experts that constructed the codebook and perform analysis on the collected data.
10. **Dataset will be sparse, sampling rates not fixed**. History does not provide sufficient data to populate a classical data cube, there are too many gaps and it is necessary to record data when available rather than imposing a rigid sampling scheme.
11. **Hierarchical structures present in the codebook are often arbitrary**. The hierarchical patterns used to organize variables within the Seshat codebook serve purposes such as navigation, templating or grouping of items for data entry.
12. **Data provenance important but cannot overload infrastructure**. In addition in the RDF-based data curation platform will use provenance to record activities, agents and entities within the platform.
13. **Representing time from circa 10,000BC to the present day**. Typical IT applications and date-time formats do not deal with >4 character BC dates well.

The next section describes our solutions in the Seshat Ontology for each challenge.

4 The Seshat Ontology

In this section we introduce the Seshat ontology[7], describe the development process and describe the key design patterns deployed in the ontology.

4.1 Overview

The Seshat codebook is primarily aimed at collecting geo-temporally scoped time series variable values describing two main units of analysis – the Polity, representing an independent historical human culture or society and the natural geographical region (NGA) which is a unit of data collection or analysis defined spatially by a polygon drawn on a map. In the RDF-based approach we use three named graphs to represent the dataset: V, the data value graph which is described by the Seshat ontology; A, the annotation graph (based on Open Annotation) where textual annotations of data values are held and P, the provenance graph (challenge 12, §3) where W3C PROV statements are recorded that describe the annotation and variable value lifecycles as they travel through the data curation system (fig. 1). The Seshat ontology extends the set of units of analysis by creating a hierarchical structure of entity classes as seen in fig.1. Each of these entities has a set of Seshat variables associated with it. Each variable value for an entity is associated with geographical and temporal scoping information.

[7] http://www.aligned-project.eu/ontologies/seshat

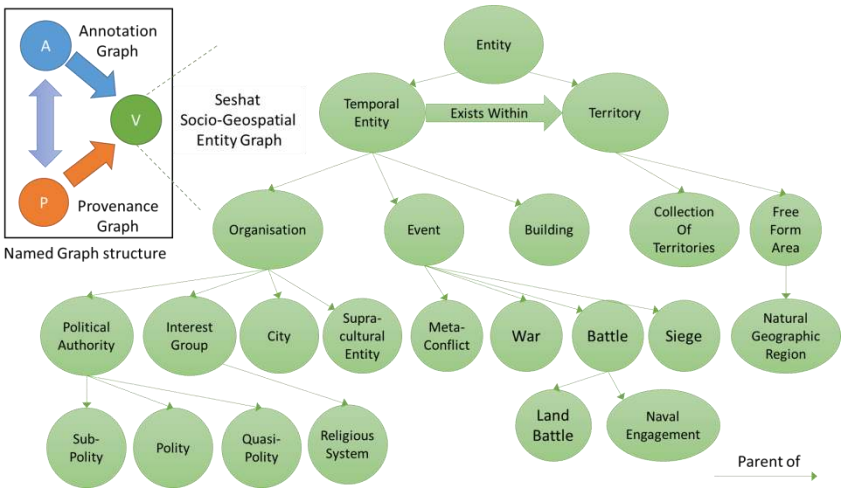

Fig. 2. Seshat Named Graph Structure and Seshat Ontology Geo-temporally Scoped Entities

In order to model the additional context required by the qualified nature of a Seshat variable, each is modelled as an OWL class and a property pointing from the appropriate Seshat entity to that class (challenge 2, §3). In order to keep the data model compact a large number of data pattern upper classes are defined for each variable. By exploiting multiple inheritance and OWL DL's complete class definitions it is possible to overload the class definition to provide a toolbox of assertions which can be automatically classified and constrained by an appropriate OWL reasoner (challenge 8, §3). Each value type associated with a variable is either an XSD datatype or a custom OWL class definition, often with a declared set of allowed values. At the variable definition level in the Seshat ontology it is possible to associate a unit of measure with data values.

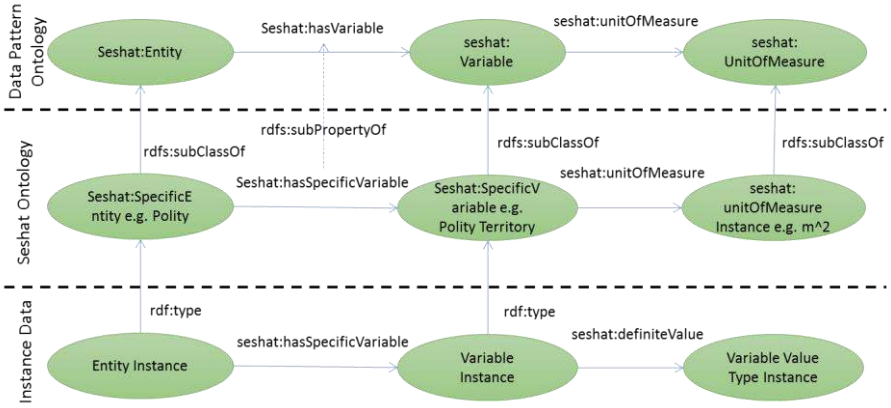

Fig. 3. Seshat Ontology Variable Structure - Modelled as a Qualified Relation

4.2 Development Methodology

The ontology has been developed at Trinity over the last 18 months. No formal ontology engineering process has been followed exactly. We used an iterative development model where the domain was explored in group sessions and requirements established. Then individual knowledge engineers worked on surveying the literature and generating solutions for specific aspects of the model. Then new versions of the combined model were developed. Then hand-coding of instance data was done to evaluate the consequences of designs. The ontology was primarily written in turtle in a syntax-highlighting text editor. Using Protégé for editing has several drawbacks – turtle comments on development are silently dropped, the import of a file often reduces properties to annotations if Protégé cannot understand them, additional meta-data and comments were generated. RDF validation has been periodically performed with the rdf2rdf[8] command line tool. More recently the ontology has been validated by the Dacura Quality Service [4], a custom OWL/RDFS reasoner that can check an ontology for a wider range of logical, typographical and syntactic errors. In addition the ontology has been used for testing the Dacura data curation tools being developed for Seshat. The ontology was split into an upper part containing basic patterns and a lower part containing the ontology of the Seshat codebook based on those patterns.

Close collaboration with the domain experts that developed the codebook was necessary. Several workshops have been held to understand their modelling concerns and describe our approach. Developing a common understanding and hence appropriate model of data unreliability and uncertainty was the most conceptually challenging topic. Three separate sources of uncertainty were identified: (1) within the codebook there was a syntax defined for variable bags of values or ranges (2) some apparently boolean variables were assigned enumerated values of "uncoded, present, inferred present, absent, inferred absent, unknown", and (3) the codebook syntax allowed multiple experts to disagree on a value. It was discovered that the use of "inferred" and "uncoded/unknown" tags in the dataset instances went wider than the variable definitions of the codebook and hence these represented generic patterns that needed to be available for all variables, not just those specified as an enum. Modelling of values, bags and ranges was straightforward (§4.3). The concept of an "inferred" value was added as an attribute for any value to indicate a human researcher had gone beyond the direct evidence to infer a value. Both unknown and uncoded were collapsed into one concept that of epistemic incompleteness – a statement of the limits of human knowledge about the past, given the expertise of the person asserting it (in the Seshat wiki a research assistant would put uncoded and an expert unknown but our PROV logs could distinguish these cases).

4.3 Design Patterns

In this section we use description logic and commentary to describe how each ontology re-engineering challenge is overcome by using the basic patterns of the Seshat

[8] http://www.l3s.de/~minack/rdf2rdf/

ontology. In the following description logic we define $\uplus$ as the disjoint union operator where $A \uplus B \equiv A \sqcup B$ where $A \sqcap B \sqsubseteq \bot$.

Representing Uncertain Time

Two main references were used as a basis for representing time - the W3C draft Time Ontology in OWL (henceforth owltime) and the W3C PROV-O ontology. Owltime is attractive since it makes explicit the granularity of representation, for example in cases where the historical record only records a year but no month or day, whereas PROV-O uses a simpler structure for time whereby activities are directly linked to an xsd:datetime value using the prov:hasBeginning and prov:hasEnd properties. In contrast owltime uses 4 intermediate nodes for each time value in an interval. Neither specification has any support for uncertainty in time assertions or non-Gregorian calendars (although Cox [5] has recently extended owltime to handle this).

Our approach, based on triple efficiency concerns, has been to re-use the expressive owltime:DateTimeDescription directly linked to a qualified variable object via the atDatetime, hasEnd and hasBeginning properties in the PROV-O pattern. i.e.

$$\begin{aligned} Instant &\equiv (= 1\ \forall atDateTime^{-}.DateTimeDescription) \\ Interval &\equiv (= 1\ \forall hasEnd^{-}.DateTimeDescription) \sqcup (\\ &\quad = 1\ \forall hasBeginning^{-}.DateTimeDescription) \end{aligned}$$

We have then extended the definition of an InstantValue to be either an Instant or UncertainInstant, which is defined as a thing having two or more assertions of the atDateTime property:

$$\begin{aligned} InstantValue \equiv Instant \sqcup UncertainInstant \text{ where } Instant \\ \sqcap UncertainInstant \sqsubseteq \bot \\ UncertainInstant \equiv (\geq 2\ \forall atDateTime^{-}.DateTimeDescription) \end{aligned}$$

Then we generalized an IntervalValue to be either an Interval or an UncertainInterval which is defined as the disjoint union of the three types of temporal uncertainty:

$$\begin{aligned} IntervalValue &\equiv Interval \uplus UncertainInterval \\ UncertainInterval & \\ \equiv UncertainEndInterval &\uplus UncertainBeginInterval \uplus UncertainBothInterval \\ UncertainEndInterval &\equiv (\geq 2\ \forall hasEnd^{-}.DateTimeDescription) \sqcup (\\ &\quad = 1\ \forall hasBeginning^{-}.DateTimeDescription) \\ UncertainBeginInterval &\equiv (= 1\ \forall hasEnd^{-}.DateTimeDescription) \sqcup (\\ &\quad \geq 2\ \forall hasBeginning^{-}.DateTimeDescription) \\ UncertainBothInterval &\equiv (\geq 2\ \forall hasEnd^{-}.DateTimeDescription) \sqcup (\\ &\quad \geq 2\ \forall hasBeginning^{-}.DateTimeDescription) \end{aligned}$$

This gives a flexible and compact notation (challenge 8, §3) for defining certain or uncertain temporal scopes (challenge 6, §3). We currently use Gregorian dates, which we project back in time using the common interpretation of ISO 8601 that allows for greater than 4 digit dates if preceded by a minus sign (challenge 13, §3).

Representing Uncertain Data Values

A key feature of Seshat is that many uncertain facts must be recorded (challenge 4, §3). We deal with this through the intermediate qualification node in a Seshat variable value. From this we define four properties: definiteValue, valuesFrom, maxValue and minValue. This enables a given variable to have a single value, a bag or a range:

$$DefiniteValue \equiv (= 1\ \forall definiteValue^{-}.\top)$$
$$BagOfValues \equiv (\geq 1\ \forall valuesFrom^{-}.\top)$$
$$RangeMaxValueRestriction \equiv (= 1\ \forall maxValue^{-}.\top)$$
$$RangeMinValueRestriction \equiv (= 1\ \forall minValue^{-}.\top)$$
$$Range \equiv RangeMaxValueRestruction\ \sqcap RangeMinValueRestriction$$

One special type of value in the Seshat codebook is one that is inferred from the historical record by the person entering the data, rather than by reference to a historical source. This is modelled as a new type but it is always a form of definite value:

$$InferredValue \equiv Inferred\ \sqcap DefiniteValue$$

When a Value is present it is always a member of the disjoint union of definite values, bags or ranges:

$$Value \equiv DefiniteValue \uplus BagOfValues \uplus Range$$

However in addition to these types of uncertainty it is important for Seshat data collectors to be able to express the presence of epistemic incompleteness, i.e. that a search has been performed and that, to the extent of the current author's knowledge, the data value is not present in the historical record. In this case we set the variable to UnknownValue which carries these semantics and record the author in the PROV graph. This leads to the full definition of an UncertainVariable in Seshat:

$$UncertainVariable\ \equiv Value \uplus UnknownValue$$

In fact due to OWL's inability to create properties that have a range of both datatypes and objects it is necessary for us to create 4 additional properties named definiteDataValue, dataValuesFrom, maxDataValue and minDataValue and parallel class definitions (DefiniteDataValue etc.) to the above to allow variables to have data or object properties. The base range for data values is rdfs:Literal rather than owl:Thing.

Temporal Constraints.

The final pattern needed is the ability to express temporal constraints as part of the qualification of a Seshat variable (challenge 5, §3). To do this we build upon our uncertain representation of time above to add scoping properties to the variable qualification class. Hence we first define the TemporalScoping as the disjoint union of the temporal types:

$$TemporalScoping$$
$$\equiv Instant \uplus Interval \uplus UncertainInstant \uplus UncertainInterval$$

Then we construct a TemporalScopedVariable as the intersection of uncertainvariables and things with a defined temporal scoping.

$$TemporalScopedVariable \equiv UncertainVariable \sqcap TemporalScoping$$

Finally we have our Seshat variable qualifier base class the UncertainTemporalVariable which can pick and mix both certain and uncertain temporal scoping and values:

$$UncertainTemporalVariable \equiv UncertainVariable \sqcup TemporalScopedVariable$$

Again it is necessary to have a parallel definition of an UncertainTemporalDataVariable for variables that refer directly to xsd:datatypes instead of OWL classes. These parallel definitions are all available in the online version of the Seshat ontology.

Example Seshat Datatype Variable Definition

To illustrate the use of the previous sections we define here an example Seshat datatype variable based on xsd:dateTime. In order to enable quality analysis and constraint checking we need to make this as strongly typed as possible. This means that all our data accessor properties must be restricted to using a single datatype (xsd:dateTime in this example) and the base type of UncertainTemporalVariable. We do this by declaring the 4 restriction classes (one for each data accessor property) and the intersection of these with our base type:

$$DateDataValueRestriction \equiv (= 1\ \forall definiteDataValue^{-}.XsdDateTime)$$
$$DateBagOfDataValuesRestriction \equiv (\geq 1\ \forall dataValuesFrom^{-}.XsdDateTime)$$
$$DateRangeMinDataValueRestriction \equiv (= 1\ \forall minDataValue^{-}.XsdDateTime)$$
$$DateRangeMaxDataValueRestriction \equiv (= 1\ \forall maxDataValue^{-}.XsdDateTime)$$
$$UncertainDateTimeVariable \equiv UncertainTemporalDataVariable \sqcap DateDataValueRestriction \sqcap DateBagOfDataValuesRestriction \sqcap DateRangeMinDataValueRestriction \sqcap DateRangeMaxDataValueRestriction$$

This is a full, usable Seshat variable and we would follow the same pattern if we had defined a custom OWL Class to hold our variable value. In practice we have defined all the common xsd:datatypes in this way as part of our base ontology and when a specific Seshat variable is based on a specific datatype we declare a sub-property in the Seshat ontology to declare specific annotation properties (rdfs:comment, rdfs:name) and meta-properties such as the units of measure for that variable.

5 Application and Use Case

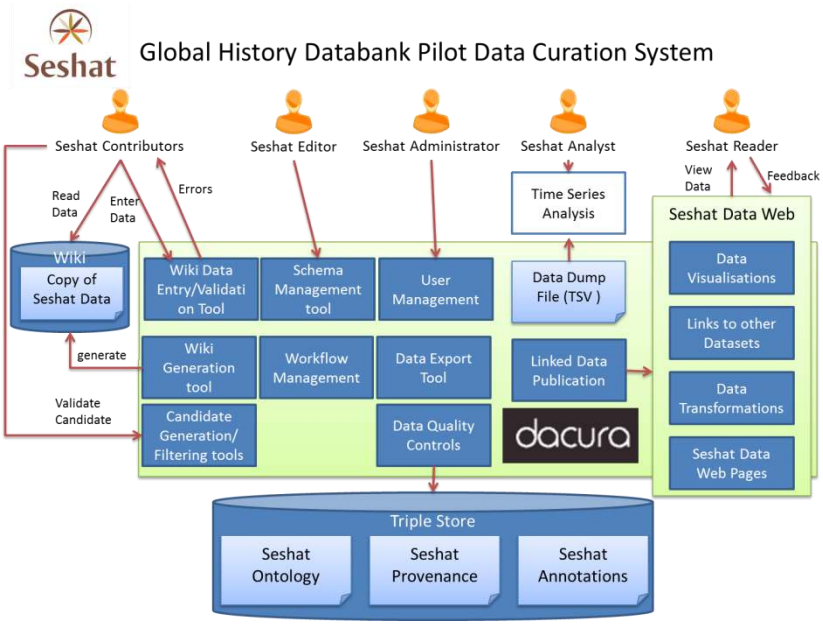

Fig. 4. Seshat Ontology Deployment in Data Curation System

The Seshat ontology is deployed in the pilot Seshat data curation system[9] based on the Dacura platform developed within the H2020 ALIGNED project. This platform allows Seshat users to enter data, manage the structure and quality of the entered data and output it in standard formats. In the pilot system, four of the components from fig. 4 are used: (1) The wiki data entry/validation tools (top left in figure); (2) The schema management tools; (3) The data quality controls (lower middle of figure) which perform schema and data integrity checks; and (4) the data export tool which can transform Seshat data into the TSV dumps required by statistical analysts. The Seshat ontology in this system is used by all our tools and enables more structured information to be captured than the original Seshat wiki, data validation at the point of entry and triple-store data integrity enforcement by the Dacura Quality Service.

6 Lessons Learned

The exercise of re-engineering the Seshat codebook into an OWL DL ontology has provided us with valuable experiences in the areas of social science codebook translation, data uplift to RDF, OWL modelling and Linked Data publishing. Each of these is summarized in table 1 and further discussed below.

[9] For a video demonstration see https://www.youtube.com/watch?v=OqNtpSClczU

Table 1. Lessons Learned

Area/Issue	Resolution/Impact	OWL Adv[1]
1. Codebook Translation		
1.1 Implicit data patterns in codebook	Required manual design of new data patterns	Y
1.2 Implicit semantics of blank values	Explicit modelling of epistemic incompleteness	Y
1.3 Lack of data-typing	Defined variables as xsd:floats, ints or unsigned ints	P
1.4 Domain model incomplete	Attached OWL classes to variable definitions	P
1.5 Atomic concepts evolve	Require patterns for composite and inferred variables	Y
1.6 Support mandatory annotations	Model at the variable definition level	P
1.7 Measurement unit definitions	Model in variable definition, link to units ontology	Y
2. OWL Modelling		
2.1 RDFS insufficient for data quality	Moved to OWL to express constraints	Y
2.2 Minimizing number of properties creates complex OWL restrictions	Knowledge model complexity increases faster than an interface specification as properties are reused	P
2.3 OWL data/object property split	Parallel definitions for owl:Thing and rdf:Literal	N
2.4 Compact data representation	OWL disjoint unions to access a palette of properties	Y
2.5 OWL Restriction classes verbose	Automated generation of OWL from design patterns	N
2.6 Intermediate logical classes needed	Additional classes defined, hide from users	N
2.7 Constraints for xsd:datatypes	OWL restrictions provide excellent property reuse	Y
3. Linked Data		
3.1 Open Annotation Inconsistent	OA imports 64 vocabularies, hard to work with as OWL (see also [6])	-
3.2 Time vocabulary	Compromised between owltime and W3C PROV-O	-
3.3 GeoSPARQL	Badly named specification, not clear is an ontology	-
4. Uplift/Import of Wiki		
4.1 Seshat coding sheet variations	Need flexible uplift mappings	N
4.2 OWL model drift from codebook	The more complex the knowledge model, the harder the uplift and dump as TSV	P
4.3 Modelling inter-entity relations	Important to provide support for text-based links as well as true relations	Y

[1] Was OWL an advantage for resolving this issue, especially wrt the wiki: Y = yes, P = partial, N = no

The overwhelming experience of developing the Seshat ontology from the wiki-based codebook is that taking a semantic web approach will add a lot of value. However given the emphasis on fixing the data quality issue in the wiki it has proved necessary to move to OWL for the ontology rather than using a linked data/RDFS approach. This is because the demands of data validation and the imprecision of what Gómez-Pérez terms "Frankenstein" linked data ontologies were ultimately incompatible. In general the process has helped the domain experts too as they have had to clarify and make explicit the semantics embedded in the codebook. The biggest hurdles in terms of OWL modelling have been the lack of support for a property top that spans both object and datatypes. This has created a doubling-up of the data patterns required. In terms of the future, by moving to a natively RDF-based system it is hoped to be able to automate the exploitation of the vast quantity of structured data produced

by the semantic web community and of course this would not be possible in a manual approach based on the wiki without a lot of brittle, custom development.

7 Related Work

The major influences on this work have been Dodds and Davis' catalogue of design patterns [6], especially the modelling patterns section, the W3C PROV ontology [7] and Open Annotation [8]. In terms of ontology engineering process, the many works of Gómez-Pérez, e.g. [2], have been influential. Our treatment of uncertainty is inspired by the work of the W3C Uncertainty Reasoning for the World Wide Web group [9]. The works of Horrocks, Patel-Schneider and their collaborators, e.g. [10], have been vital in shaping our understanding of OWL DL. Finally the survey of Zaveri et al. [11] has been instrumental in guiding the development of a Seshat ontology that is suitable for data quality assurance.

There have been many initiatives that tackle the challenge of representing historical data using semantic web technology. One important standard is CIDOC CRM [12] published by ISO. It has the broad remit of defining an ontology for cultural heritage information. In contrast to Seshat, its primary role is to serve as a basis for mediation between local representations of cultural heritage resources such as museum collections. Hence the term definitions and subsumption hierarchy are incomplete, there is no full grounding of datatypes, for example as xsd:datatypes but instead the lowest level is abstract types such as string. The RDFS-based ontology definition the standard includes is not the primary reference but a derived one. Nonetheless the FP7 ARIADNE infrastructure project[10] has made progress with using it as a basis for linked data publication and interworking between collections. There is great potential for future collaboration with the Seshat consortium in terms of data sharing.

DBpedia [13] of course contains many historical facts that are of interest to Seshat and it is hoped that by leveraging the work already done there it will be possible to quickly import candidate data for Seshat, to be then curated by the Seshat research assistants and domain experts. Nontheless the current DBpedia data is not in a format suitable for processing as time series and does not comply with the conceptual models underlying the Seshat codebook so mapping techniques will have to be employed. Through the ALIGNED project we are collaborating with the AKSW group at the University of Leipzig and it is planned to establish a virtuous circle whereby DBpedia extracts crowd-sourced facts from Wikipedia, Seshat uses those facts as input to their historical time-series, the Seshat team curates and refines the facts and publishes them as high quality linked data which in turn is available to DBpedia+, the new multi-source, improved quality version of DBpedia in development by the DBpedia community. This integration will be trialed in year 3 of ALIGNED (2017).

There are also a large number of other curated RDF datasets describing historical locations and facts such as Pleiades[11] that focuses on ancient names, places and locations. Nonetheless these datasets are typically based on controlled vocabularies rather

[10] http://www.ariadne-infrastructure.eu/
[11] http://pleiades.stoa.org/home

than formal semantic data models and RDF is provided as a dump that transforms the internal representation. This gap presents an opportunity for Seshat as a provider of high quality native linked data with strong consistency assurances. Once again it is hoped that Seshat will work with these other dataset publishers in the future.

Finally there are a wide range of historical time series data collection efforts in the social sciences that are not RDF-based or publishing linked data. Most of these have much more limited scope than Seshat. For example Sabloff's datasets describing the limited geographic region of Mongolia throughout time [14] or the Database of Religious History [15] that has similar geo-temporal scope to Seshat but deals only with religion rather than all aspects of human cultural evolution.

8 Conclusions and Future Work

Our ambition for the Seshat ontology goes beyond constraining, structuring and classifying the uncertain and sparse (although voluminous) historical time series data that forms the basis of the Seshat: Global History Databank. In future work we will enrich the knowledge model by adding semantic relationships between Seshat time-series variables to support domain knowledge-based quality assurance. This will enable, for example, the identification of inconsistent statements about a historical society's military metal technology and the metals used for agricultural tools.

The current ontology reflects the modelling foci in the original Seshat codebook and several areas would benefit from generalization or extension. Two high priority areas are (1) the creation of richer models of the politico-geographical relationships between historical societies as this will add greater flexibility to the model and (2) adding support for inferred variable values in addition to collected values as this will reduce data collection effort and improve consistency. Similarly the ontology will be extended for publication as linked data. For example, creating interlinks between Seshat and the web of data or mapping Seshat to common linked data vocabularies like GeoSPARQL to make it more easily consumed.

In addition to data validation and quality assurance, a key use of the ontology within Seshat is the generation of customised, dataset-specific, high usability user interfaces for data entry, import, interlinking, validation and domain expert-based curation. This requires the development of form generation tools for presenting ontology elements and widgets that streamline data entry and constrain the entered data to be syntactically and semantically correct. As this form generation technology develops it may produce new design patterns for the structure of the Seshat ontology.

Acknowledgements: This work was supported by a John Templeton Foundation grant, "Axial-Age Religions and the Z-Curve of Human Egalitarianism," a Tricoastal Foundation grant, "The Deep Roots of the Modern World: The Cultural Evolution of Economic Growth and Political Stability," an ESRC Large Grant, "Ritual, Community, and Conflict" (REF RES-060-25-0085), European Union Horizon 2020 research and innovation programme (grant agreement No 644055 [ALIGNED, www.aligned-project.eu]) and the ADAPT Centre for Digital Content Technology, SFI Research Centres Programme (Grant 13/RC/2106) co-funded by the European

Regional Development Fund. We gratefully acknowledge the contributions of our team of research assistants, post-doctoral researchers, consultants, and experts. Additionally, we have received invaluable assistance from our collaborators. Please see the Seshat website for a full list of private donors, partners, experts, and consultants and their respective areas of expertise.

References

1. Turchin, P., Brennan, R., Currie, T., Feeney, K., Francois, P., Hoyer, D., et al.: Seshat: The Global History Databank. Cliodynamics: The Journal of Quantitative History and Cultural Evolution, 6: 77-107, (2015).
2. Francois, P., Manning, J., Whitehouse, H., Brennan, R., Currie, T., Feeney, K., Turchin, P.: A Macroscope for Global History. Seshat Global History Databank: a methodological overview. (Submitted)
3. Villazón-Terrazas, B., Gómez-Pérez, A.: Reusing and Re-engineering Non-ontological Resources for Building Ontologies, in Ontology Engineering in a Networked World, pp 107-145, Springer Berlin Heidelberg, 2012. http://dx.doi.org/10.1007/978-3-642-24794-1_6
4. Mendel-Gleason, G., Feeney, K., Brennan, R.:Ontology Consistency and Instance Checking for Real World Linked Data, Proceedings of the 2nd Workshop on Linked Data Quality co-located with 12th Extended Semantic Web Conference (ESWC 2015), Portorož, Slovenia, June 1, (2015).
5. Cox, S. J. D.:Time Ontology Extended for Non-Gregorian Calendar Applications, (to appear) Semantic Web Journal (2015) http://www.semantic-web-journal.net/content/time-ontology-extended-non-gregorian-calendar-applications-0
6. Feeney, K., Mendel-Gleason, G., Brennan, R.: Linked data schemata: fixing unsound foundations, submission to Semantic Web Journal, (2015) http://www.semantic-web-journal.net/content/linked-data-schemata-fixing-unsound-foundations
7. Lebo, T., Sahoo, S., McGuinness, D. (eds.): PROV-O: The PROV Ontology, W3C Recommendation 30 April (2013)
8. Sanderson, R., Ciccarese, P., Van de Sompel, H. (eds.): Open Annotation Data Model, Community Draft, 08 February (2013). http://www.openannotation.org/spec/core/
9. W3C Uncertainty Reasoning for the World Wide Web XG, UncertaintyOntology, (2005) http://www.w3.org/2005/Incubator/urw3/wiki/UncertaintyOntology.html
10. Grau, B., C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., and Sattler, U.: OWL 2: The next step for OWL. J. Web Semantics. 6, 4, 309-322, (2008). DOI=http://dx.doi.org/10.1016/j.websem.2008.05.001
11. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J. and Auer, S..: Quality assessment for linked data: a survey. Semantic Web. Vol. Preprint, no. Preprint, pp. 1-31, (2015) DOI: 10.3233/SW-150175
12. ISO 21127:2014 Information and documentation -- A reference ontology for the interchange of cultural heritage information, 2nd Edition, ISO, (2014)
13. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia – A Crystallization Point for the Web of Data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Issue 7, Pages 154–165, (2009).
14. Sabloff, P. L.W. (Ed.): Mapping Mongolia:Situating Mongolia in the World from Geologic Time to the Present, Univ. of Pennsylvania Press, ISBN 978-1-934536-18-6, (2011)
15. Slingerland, E., and Sullivan, B.: Durkheim with Data: The Database of Religious History (DRH), Journal of the American Academy of Religion (in press), (2015).