Publishing Social Sciences Datasets as Linked Data: a Political Violence Case Study

Rob Brennan
Knowledge and Data Engineering Group
Trinity College Dublin, Ireland
rob.brennan@cs.tcd.ie

Kevin C. Feeney
Knowledge and Data Engineering Group
Trinity College Dublin, Ireland
kevin.feeney@cs.tcd.ie

Odhrán Gavin
Knowledge and Data Engineering Group
Trinity College Dublin, Ireland
gavinod@cs.tcd.ie

ABSTRACT

This paper discusses the design, application and generalization of a Linked Data vocabulary to describe historical events of political violence. The vocabulary was designed to capture the United States political violence 1795-2010 dataset created by Prof. Peter Turchin and has been generalized to support a semi-automated data collection process suitable for the creation of a complimentary dataset of political violence events in the UK and Ireland. Both datasets will be published as managed linked data that is inter-connected with other web-based datasets such as DBpedia, a computer-readable version of Wikipedia. The lifecycle of the datasets will be actively managed with tool support for further harvesting, evolution and consistency checking. The harvesting tool, data harvesting process, political violence vocabulary and US political violence dataset were connected to our existing linked data management platform, DaCura.This political violence vocabulary described herein has been validated by application to a real-world dataset and publication use-cases. Our data harvesting process is potentially applicable to a wide range of social science or historical research activities that focus on generating structured data-sets or annotations of human-readable corpora. The publication of the US political violence dataset as linked data is a contribution towards the emerging fields of Digital Humanities and Linked Science. This paper describes a new linked data vocabulary for political violence events, provides insights into the processes of creating a new vocabulary for social science datasets. It also illustrates the potential benefits of publishing social science or other cultural heritage datasets as linked data.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User / Machine Systems – *Human information processing*. H.2.1 [**Database Management**]: Logical Design – *Data models, Schema and sub-schema*. H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Retrieval models, Selection process*.

General Terms

Management, Documentation, Design, Experimentation, Human Factors, Standardization.

Keywords

Linked Data, Vocabularies, RDF, Schema Design, Cliodynamics, Data Curation.

1. INTRODUCTION

The collection and curation of structured data-sets from unstructured and semi-structured sources is a common requirement for research both in social sciences and more general cultural heritage projects [1]. Linked Open Data (LOD) approaches to online data publishing are based upon RDF and semantic-web technologies such as RDFS, OWL and SPARQL. These should, in theory, be a very attractive solution for harvesting, curating and publishing structured social science or humanities data-sets.

In this paper, we describe a case-study of an approach to migrating a social-science dataset to an LOD platform. The dataset in question is the United States political violence 1795-2010 dataset created by Prof. Peter Turchin in the course of his research into Cliodynamics [2]. The dataset was originally distributed as an Excel spreadsheet, consisting of 1828 event records, each of which had several properties associated with it. This process formed a test-case of the DaCura system which we have been developing in Trinity College Dublin [3]. That system is designed to provide easy-to-use tool support for non-expert users to allow them to easily harvest data from web-based sources into an RDF based triple-store. It furthermore provides support for the management of that data-set over time with a focus on supporting constrained schema evolution.

The focus on this paper is on the process by which we designed the LOD schema from the original dataset spreadsheet. The schema is represented as an RDFS (RDF Schema) vocabulary. In designing this schema we had the following goals:

1. Re-use, wherever possible, existing LOD vocabularies to represent the events and their properties in the data set.

2. Provide support within the schema for the process by which the data is collected and not just the final data format. Thus, for example, a requirement is that we can capture candidate events in our dataset which may need to be approved for inclusion in the final dataset by a domain expert.

3. Design the schema in such a way that it would integrate well with our DaCura platform. DaCura provides several features such as the ability to generate simple web-based widgets to represent dataset instances. To take full advantage of this facility certain properties must be present in the schema.

In designing our schema, we attempted to describe entities in a general and extensible way while minimizing the overall complexity of the schemata upon which we were relying. Rather than trying to define everything in an entirely general way, we attempted to steer a pragmatic middle-ground between generality and specificity and only introduced more general schema in situations where we could envisage future situations in which we might take advantage of this generality.

2. SYSTEM DESIGN

This section discusses the development of the political violence vocabulary, a formal process for harvesting political violence events from a historical corpus, our harvesting tool and finally the online repository for political violence datasets.

2.1 Political Violence Vocabulary Design

There were five distinct activities involved in the vocabulary design: survey of other vocabularies; examination of the original US dataset; consideration of the requirements for the UK and Ireland dataset; the semantic uplift process and creating interlinks to other linked data datasets. Each of these activities is discussed below.

2.1.1 Survey of Other RDF Vocabularies

One of the key features of vocabularies based on RDF (Resource Description Framework) is that they can easily be combined to produce larger models. RDF-based systems do not depend on the existence of a single, canonical ontology into which every vocabulary or specialized ontology must fit. This frees vocabulary designers to create domain or application specific designs but it also creates a proliferation of overlapping vocabularies published on the web. In recent years the Linked Data community [4] has focused on reuse of a few well-known vocabularies such as the Dublin Core metadata for describing documents. This has the beneficial outcome of reducing the requirements for applications that consume linked data, as terms defined by these common vocabularies appear again and again in datasets published on the web.

2.1.2 Evaluate and Analyze the Example Dataset

The United States Political Violence (USPV) dataset was initially compiled in order to assist research into the dynamics of political instability in the United States [2]. It was compiled from a number of sources and was published as a spreadsheet consisting of 1,828 reports of incidents of violence, recording date, category, motivation, fatalities, location, source, a description of the event, and research-specific coding. In conjunction with the appendix to [2], historical research was undertaken in order to formulate precise definitions of the types of political violence events in the dataset, as described by the category and motivation fields. Our vocabulary was designed to ensure that all information contained within the published dataset could be captured without loss.

Two features of the dataset particularly informed design choices in the vocabulary. The presence of duplicate reports in the dataset led to the decision to differentiate between reports and events. The presence of reports marked with question marks to indicate uncertainty, led us to decide to include the capability to report levels of uncertainty about reports of political violence.

2.1.3 Generalisation to UK and Ireland Dataset

Historical knowledge of the period 1785-2007 was used to determine the suitability of the vocabulary, based on the USPV dataset, for the United Kingdom and Ireland Political Violence (UKIPV) dataset. In most cases, vocabulary terms used to describe political violence in the United States were also appropriate to describe political violence events in the United Kingdom and Ireland. However, due to historical differences between the two regions, a small number of terms describing motivations required changes in order to capture the characteristics of political violence for the UKIPV dataset more accurately.

2.1.4 Semantic Uplift

We define semantic uplift as the process of converting non-RDF data, for example the original US political violence spreadsheet, into an RDF-based knowledge representation such as a set of RDF triples describing the individual events according to the Political Violence vocabulary. Semantic uplift is often ignored in favor of focusing on schema modeling tasks. However it has an important impact on the vocabulary design process. Converting events into RDF exercises the vocabulary and exposes flaws or weaknesses. In our case the semantic uplift process was written as a PHP script that processed a CSV (comma separated value) representation of the spreadsheet.

2.1.5 Creating Links to other Linked Data Datasets

One of the major motivations for publishing the political violence datasets as (RDF-based) linked data is to enable combination of the data with other datasets already available on the web. In theory once the dataset is published as RDF on the web it is available to all RDF-consuming applications. However this can place onerous requirements on those applications if a new vocabulary is used and no interlinks are created between the political violence dataset and already existing datasets. In general this means that generic, browsing-oriented applications are able to display the data but that more sophisticated use cases such as mash-ups of the data are less likely.

At the dataset consumption level, enabling discovery is a topic addressed by several ongoing research efforts such as the Data Hub / CKAN by the Open Knowledge foundation and the Sindice semantic web index by DERI [5]. At the vocabulary level it is possible to reuse common vocabularies such as Dublin Core that are often used in linked data datasets. At the dataset level it is possible to include interlinks to instances in other datasets. For example when recording the location of an event as the US state of Ohio it may be preferable to record this as the instance of that concept defined by the Dbpedia or Geonames datasets. Thus a “dbpediaLocation” property is defined in the PV vocabulary which enables us to directly embed references to instances of the DBpedia concept “Place”.

2.2 A Data Harvesting Process

The manual process of extracting US political violence events from the historical record was described by Turchin in his analysis of that data-set [2]. However for this work it was necessary to formalize and document the harvesting process model with six goals in mind:

1. Establishing the requirements placed by the collection process on the political violence vocabulary in terms of what concepts need to be modeled.

2. Establishing the possible actors or roles in the data collection process.

3. Specializing the process to consider the requirements placed on it by the UKIPV dataset sources.

4. Reviewing the process with respect to the possible activities where automation could both be beneficial and could leverage the advantages of having a formal vocabulary describing the data being extracted.

5. Linking the data collection process to our previous work on DaCura, a managed linked data curation platform [3].

6. Determining the experimental process by which we would gather data to validate the utility of our tool support for data collection, validation, publication and management of the datasets.

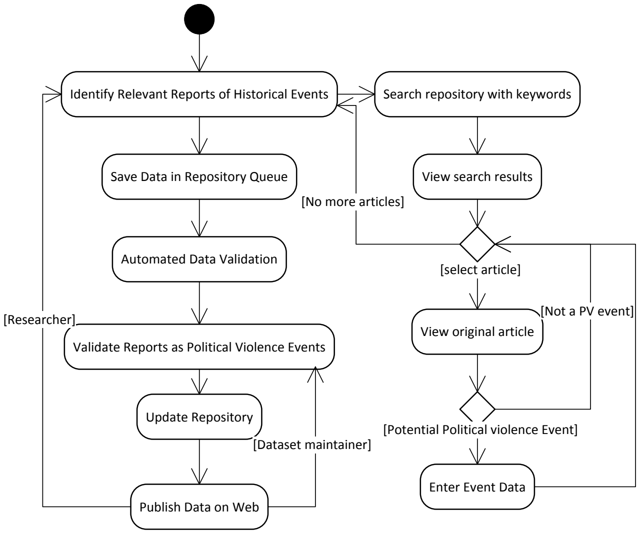
Identify Relevant Reports of Historical Events
Search repository with keywords
Save Data in Repository Queue
View search results
[No more articles]
Automated Data Validation
[select article]
[Researcher]
View original article
[Not a PV event]
Validate Reports as Political Violence Events
Update Repository
[Potential Political violence Event]
[Dataset maintainer]
Enter Event Data
Publish Data on Web

Figure 1. The political violence data harvesting process.

Figure 1 illustrates the data harvesting process as a UML activity diagram. The left hand side of the figure focuses on the overall data lifecycle. In this lifecycle, events are identified in the repository by a researcher, then data is validated as conformant to the vocabulary by the DaCura platform, then a dataset maintainer examines the report data to validate whether or not it should be recorded as a new political violence event (or a duplicate, or out of scope, etc) and finally the data is updated as linked data on the web. This is an iterative process that can continue as long as there is event data to be found or maintained.

The right-hand side of the figure shows details of the event report extraction from the online newspaper repository. First a set of search keywords are used to retrieve a set of articles that are candidates for event reports. The researcher then views each article in turn to evaluate it against the requirements for inclusion in the dataset as a report. If it is to be included then data about the article and the underlying event as reported is recorded. This event report data is then placed into the overall process flow on the left-hand side of the figure.

2.3 Political Violence Vocabulary

The approximate structure of the political violence vocabulary is represented as a UML class diagram in figure 2. This is an approximation because the RDF semantics do not exactly align with the object oriented modeling assumptions of UML. There are three main classes defined: the historical Event and its two sub-classes, the Report and the Political Violence Event.

In addition to all the classes used to model the properties of events (on the right of the figure and discussed further below) the vocabulary makes use of the Open Annotation Data model [7] vocabulary to enable researchers or other consumers of the data-set to annotate individual dataset elements.

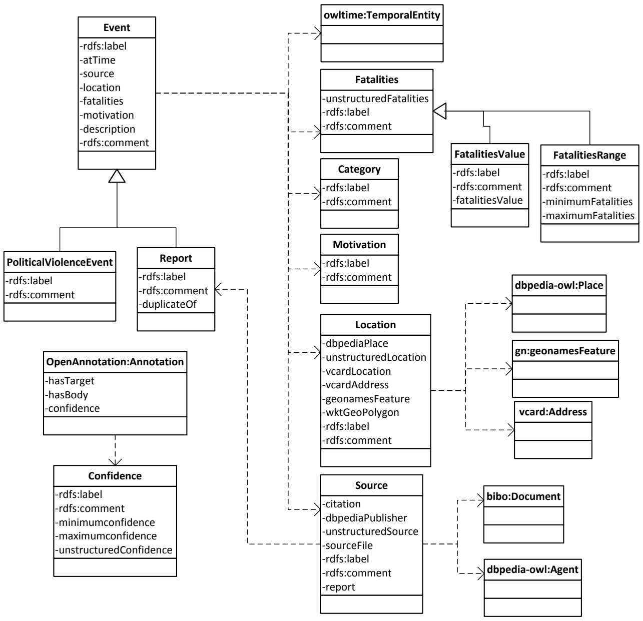

Figure 2. UML class diagram illustrating vocabulary structure.

2.3.1 Vocabulary Terms

The basic building-block of the dataset is our concept of an event, which is defined as any individual historical event. Based on the dataset and requirements, events were further subdivided into two classes, political violence events and reports. A report refers to a source's record of an event, e.g. a newspaper article. A political violence event refers to the event itself. In general, political violence events are referred to by one or more reports. This division reproduces both the existence of duplicate records of events in the original USPV spreadsheet, and the occurrence of multiple reports of individual historical events in the historical source material for the UKIPV dataset.

2.3.2 Categories

The category class identifies what form the political violence event takes. In the USPV and work based on it [2], most events are categorized into one of four categories – assassination, terrorism, lynching, and riot – based on the number of perpetrators and victims. There are a number of other categories which describe less commonly-occurring political violence events. The most common of these is rampage, which refers to events such as school and workplace shootings. The remaining categories describe uncommon events or are excluded from the analysis, and are included to fully capture the USPV dataset.

2.3.3 Motivations

The motivation class describes the reasons political violence event occurred. Events may have multiple motivations if they have numerous or complex causes.

2.4 Links to other DataSets

The value of datasets is expanded if it is possible to easily combine them with other datasets already published on the web. Hence this vocabulary contains multiple connection points to three important linked data datasets: (1) DBpedia, the RDF version of Wikipedia [8], (2) Geonames, a geographical database accessible through RDF and (3) vCard a vocabulary for representing people and organizations in RDF that is reused in many open datasets. These links are created by creating properties in the PV vocabulary that reference the other datasets.

2.5 Integration with DaCura software

The DaCura system is designed to improve the manageability of RDF datasets over time by imposing a set of constraints on RDF schemas and updates to RDF datasets above and beyond those that are mandated by RDF and RDFS standards themselves. For example, it requires that properties must have labels specified and requires that classes cannot be removed from a schema if there are instances of those classes in the dataset. It also defines naming conventions for RDF URLs. The combined effect of these constraints is to allow schema and dataset evolution while maintaining the consistency of the dataset over time.

3. RELATED WORK

Vocabulary and ontology design is an evolving subject area as the actual deployment of Semantic Web technologies and Linked Data is immature. The focus of theoretical and practical design concerns have rarely overlapped. A major venue for this debate is the annual Workshop on Ontology Patterns [9]. However Dodds and Davis [10] give a concrete set of examples for designs that are based on Linked Data use cases and were influential on this paper.

Shaw et al. [6] provide an overview of current ontologies for representing events in RDF and show the common attributes of event representations and how the differing modelling approaches tackle each aspect. In addition they provide a “Linked Open Data Event Model” (LODE) that encapsulates the common attributes in other representations but concentrates. This is a laudable and useful outcome but it was found to be lacking for our application to political violence datasets in two main respects. First, it assumes that these factual aspects represent some form of “consensus reality” whereas in harvesting data from the London Times archive it is often found that newspaper reports over time can be inconsistent or contain incorrect factual assertions. Second, it uses the DOLCE+DnS Ultralite [11] upper ontology for several property value types and we didn’t want to be constrained to using such an abstract and complex description of our dataset because of the resultant complexity in querying the dataset.

4. CONCLUSION AND FUTURE WORK

In this paper we have examined the process of generating a vocabulary to support extraction of political violence event data from online historical sources. The ontology is flexible enough to capture the original US political violence dataset while still supporting the needs of the proposed UK and Ireland political violence dataset. It is potentially suitable for collecting political violence event data from other sources. Using this vocabulary, we have created a set of tools which allow for harvesting and collation of political violence events. These tools will be used to construct the UK and Ireland political violence dataset. They will also underpin the experimental process examining the utility of tool support for collecting and managing linked data datasets.

Future work will involve extending the functionality of the data extraction toolset. Currently, candidate political violence reports are selected via a small set of searches chosen to offer acceptable and consistent precision and recall. We intend to provide users with the facility to suggest potentially useful search terms after data retrieval, in order to improve the precision and/or recall of the results. Another planned feature is to implement a domain expert (historian or social scientist) moderator queue.

5. ACKNOWLEDGEMENTS

The authors wish to thank Prof. Declan O’Sullivan and Prof. Peter Turchin for their encouragement and feedback on this work.

6. REFERENCES

[1] Hampson, C., Agosti, M., Orio, N., Bailey, E., Lawless, S., Conlan, O., and Wade, V. 2012. The CULTURA Project: Supporting Next Generation Interaction with Digital Cultural Heritage Collections. In progress in Cultural Heritage Preservation - 4th International Conference (EuroMed 2012). 668-675. Lecture Notes in Computer Science (LNCS) 7616, Springer, Heidelberg, Germany.

[2] Turchin, P. 2012. Dynamics of political instability in the United States, 1780–2010. Journal of Peace Research, 49(4). 577-591. DOI:10.1177/0022343312442078

[3] Tai, W., Feeney, K., Brennan, R., and O'Sullivan, D. 2012. Manageable Dataset Curation for Linked Data. 18th International Conference on Knowledge Engineering and Knowledge Management, (EKAW, 8 - 12 October, Galway, Ireland, 2012).

[4] Bizer, C., Heath, T., and Berners-Lee, T. 2009. Linked Data - The Story So Far, International Journal on Semantic Web and Information Systems (IJSWIS), 5 (3). 1–22.

[5] Käfer, T., Umbrich, J., Hogan, A. and Polleres, A. 2012. Towards a Dynamic Linked Data Observatory. WWW2012 Workshop: Linked Data on the Web (LDOW2012, Lyon, France, 16 April, 2012).

[6] Shaw, R., Troncy R., and Hardman L. 2009. LODE: Linking Open Descriptions of Events. In Gómez-Pérez A., Yong, Y., and Ying, D. (eds.), Proceedings of the 4th Asian Conference on The Semantic Web (ASWC '09), Springer-Verlag, Berlin, Heidelberg, 153-167. DOI=http://dx.doi.org/10.1007/978-3-642-10871-6_11

[7] Sanderson, R., Ciccarese, P., and Van de Sompel, H. (eds.) 2013. Open Annotation Data Model. Community Draft, 08 February 2013. Retrieved June 9, 2013 from W3C: http://www.openannotation.org/spec/core/20130208/

[8] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. 2009. DBpedia - A crystallization point for the Web of Data. Web Semantics. Science, Services and Agents on the World Wide Web, 7(3), 154-165.

[9] Blomqvist, E., Gangemi, A., Hammar, K., and Suárez-Figueroa, M.C.(Eds.) 2012. Proceedings of the 3rd Workshop on Ontology Patterns (11th International Semantic Web Conference 2012 (ISWC 2012), Boston, USA, November 12, 2012.)

[10] Dodds, L., and Davis, I. 2012. Linked Data Patterns, A pattern catalogue for modelling, publishing, and consuming Linked Data, 2012-05-31, Retrieved June 6, 2013, from: http://patterns.dataincubator.org/book/

[11] Powell, A., Nilsson, M., Naeve, A., Johnston, P., and Baker, T. 2007. DCMI Abstract Model. DCMI Recommendation, 2007.