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 in the course of his social science research into Cliodynamics. The vocabulary 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 creation of the political violence vocabulary required the evaluation of pre-existing vocabularies for reuse and compatibility. The original US political violence dataset was stored in a spreadsheet and an initial vocabulary was extracted from that. A process was designed for the semi-automated harvesting of political violence data from online corpora of historical documents such as newspaper archives. The vocabulary was refined to support dynamic interface generation by a vocabulary-neutral data harvesting tool. 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.

The main practical outcome of this work to date is a prototype political violence data harvesting tool-chain. This will enable us to quickly collect the UK and Ireland political violence dataset and perform experimental evaluations on the collection process and tool chain.

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]. Typically, the processes by which the structured data is extracted from archives are largely manual with limited tool support. The schemas which structure the extracted data are rarely specified formally and it is frequently the case that each project designs its own schema. This renders inter-operability, aggregation and reuse of data difficult. When the collected data-sets are made available, it is frequently the case that simple, file-based formats such as CSV or Excel are the means by which they are distributed. Or, even worse, the data is locked in layout-centric formats such as PDF. Where these data-sets are maintained over time, they generally rely upon labour-intensive, ad-hoc processes to manage and update the data. Where resources are available for IT support, this typically consists of a relatively rigid web-based application on top of a relational database, which is expensive to maintain yet does not provide much in the way of information interoperability.

ENRICH'13, August 1, 2013, Dublin, Ireland.

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. LOD technologies have a number of features which address many of the basic challenges of the domain:

1. RDF is based upon a graph and triple-based format, rather than being table-based. Thus it can efficiently store heterogeneous, sparse, data.
2. LOD is based upon the re-use of existing vocabularies and allows them to be easily integrated into new schemas. Thus, schemas can be designed which are specific to a particular data-set, while still facilitating inter-operability with other data-sets. Data sets can also take advantage of facilities provided by existing published datasets to enrich their data – for example, rather than defining one's own location vocabulary, the Geonames vocabulary can be used which provides links into their extensive information base.
3. In LOD approaches, tightly-defined formal schemas and are optional and can be overlaid upon existing datasets. This makes it much easier to design a schema incrementally.
4. RDFS and OWL are semantically rich schema-languages which provide support for inference. This enables the creation of multi-layered data-sets which provide interpretative and theoretical structure on top of the underlying raw data.
5. Although LOD is a relatively new approach, it leverages a generation of semantic web-research. Thus a significant number of software tools exist which support the creation and publication of LDO datasets on the web.

However, despite the clear attractiveness of LOD approaches, take up in the social sciences has been limited to date. There are several reasons for this:

1. Social scientists are rarely technologists and developments in computer science research permeate only slowly in that community.
2. The advantages provided by RDF and semantic technologies come at a certain cost – designing schemas in such a way that they can take advantage of the technology is a difficult task that requires considerable experience and expertise that is not readily available.
3. Due to their provenance in computer science research, most LOD tools are based on the assumption that the data sets will be designed and populated by specialist knowledge engineers and consumed by non-expert users. Thus, there are few publication tools that are usable by non-expert end-users,
4. The open nature of the LOD philosophy supports flexibility in schema design – there are many ways to specify any given structure. However, this flexibility is a double-edged sword, it provides the possibility of mixing together different, incompatible constraints. Furthermore, it makes it easy for errors in specification to permeate through the data-set's structure.

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. This is an important requirement as it allows the often tedious process of manual data harvesting to be performed by those who are less expert in the specifics of the domain.
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. For example, class properties that have defined domains and ranges allow DaCura to generate widgets that are more finely tuned to the data.

In designing our schema, we adopted a philosophical stance whereby 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. Rather than defining every event as a specialization of the most general concept of event possible (a very abstract thing indeed), we chose an event concept that was general enough to describe all of the types of events that we might conceivably encounter in the data and tailored our schema and choice of vocabulary accordingly.

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

The vocabulary design process is necessarily iterative however there were five distinct activities involved – 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. Most often this feature is used to combine several more specific models into a broader framework such as reusing the W3C's time ontology [4] to model the time of occurrence of an event within our political violence event vocabulary. Reuse at the level of individual classes or properties (i.e. terms) can also be done in RDF. This requires even closer analysis of vocabularies that are candidates for reuse. If a term is to be reused, then there are typically two choices to the designer – to import the term directly or to create an equivalent local term with the ability to declare an owl:sameAs mapping at a later stage.

Another important vocabulary design consideration is that 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. When this factor is combined with the fact that RDF is still an emerging technology and hence many of the applications are research projects with limited appeal or longevity, it complicates the reuse choices – should proliferation on the web of data be a consideration for adoption? What if this conflicts with the technical demands of the vocabulary design, such as the ability to express concepts succinctly, or ease of querying?

In recent years the Linked Data community [5] has resolved some of these concerns by focusing 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 Generalization to UK and Ireland Dataset

After the formulation of a vocabulary based on the USPV dataset, this was then analyzed to ensure that it was suitable for the compilation of the United Kingdom and Ireland Political Violence (UKIPV) dataset. Historical knowledge of the period 1785-2007 was used to determine the suitability of the vocabulary for the 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.

In the USPV, 'land' is a motivation used to describe only one incident of political violence. Conflicts about rent and control of land are covered by the 'economic' motivation. Due to the prevalence of such conflict in the United Kingdom and Ireland, we decided to separate it out from other economic conflict. In a similar manner, while it is useful to distinguish between conflict between whites and African-Americans and other ethnic conflict in the United States (coded as "race and ethnic", respectively), due to the prevalence of the former, this distinction is less important for the UKIPV dataset.

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. Thus, it forms a parallel process to that of converting the schema, or data structure into an RDF vocabulary. This semantic uplift process focuses on instances or the individual event data.

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. This leads to two conclusions – designing a vocabulary without an example dataset is prone to error and that it is important to automate the semantic uplift process early in the vocabulary design activity. In our case the semantic uplift process was written as a PHP script that processed a CSV (comma separated value) representation of the spreadsheet.

One final vocabulary consideration was driven by the semantic uplift process – would the final vocabulary support true lossless conversion of the original dataset? This impacts the vocabulary because real data-sets contain inconsistencies and errors that will not easily convert into a strongly typed vocabulary. If these errors are artifacts to be maintained for posterity or if the vocabulary creators do not know the value of the inconsistencies, then it may be necessary to add unstructured representations of the relevant fields into the vocabulary.

2.1.5 Creation of Links to 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.

This introduces vocabulary design considerations as to how best to achieve these interlinks. 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, browsingoriented 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 [6].

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. This enables applications to follow the links from one dataset to another. This approach is facilitated by a vocabulary design that includes these externally defined instance types as the objects of properties within the location concept. It is our belief that this approach facilitates the most direct integration of the datasets and hence it has been followed whenever possible within the political violence vocabulary. 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 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. For example it was seen that the original US political violence dataset includes events that were subsequently marked as duplicates or unsuitable for inclusion in the analysis as although initial research was promising they did not ultimately meet the exact requirements for inclusion in the analysis dataset.

2. Establishing the possible actors or roles in the data collection process. This could also have an impact on the vocabulary, for example in recording the provenance of event data records.

3. Specializing the process to consider the requirements placed on it by the knowledge that the UK and Ireland political violence dataset would be harvested from the London Times online archive and the types of workflows that it supported.

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. This required a cost-benefit analysis of the likely implementation effort required to achieve the desired automation especially with respect to the skills and knowledge already present in our research group. For example Natural Language Processing (NLP) technology can obviously be applied to processing some electronic versions of historical documents but since limited expertise was available to us in this area we decided to defer this topic until we had established a baseline dataset and recruited a suitable collaborator.

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.

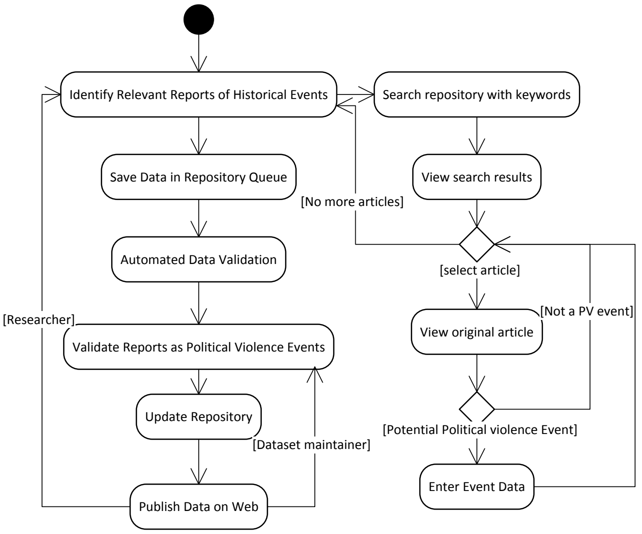
Identify Relevant Reports of Historical Events
Search repository with keywords
Save Data in Repository Queue
View search results
[No more articles]
Automated Data Validation
[select article]
[Researcher]
[Not a PV event]
View original article
Validate Reports as Political Violence Events
Update Repository
[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 Developing Harvesting Tool

A major goal of our research is to automate or semi-automate the event data harvesting process. This will be achieved by providing new tool support for the right-hand side of our data harvesting process (fig. 1) and integrating the new tool with our linked data curation platform, DaCura. [3].

2.3.1 Corpus Archive Analysis

The initial phase of tool development was to analyze the repository in order to determine requirements and gold standards. The structure of the site was inspected in order to assess how feasible the construction of overlaid tools for data capture would be, and to determine the requirements for building these tools.

Two periods (the year 1831 and the first six months of 1982) were chosen for gold standard testing. A list of 94 political violence reports was compiled manually for 1831. The CAIN database

(CITE), a database of all deaths during the conflict in Northern Ireland, lists 38 political violence events for the first six months of 1982. Of these, 30 were reported in the repository, and these were used as a gold standard for testing.

2.3.2 Identification of Search Terms

A variety of search terms were then tested against the gold standards established above in order to identify search terms which would be effective in identifying candidate reports for volunteer analysis. Effective search terms would provide good precision and recall on testing against the gold standard, while returning a number of search results per year which volunteers would be able to process in an acceptably short period of time. Furthermore, they should provide consistent samples across time periods. Two search terms were eventually chosen, one for early time periods and one for later time periods, in order to reflect changes in the language used in the repository.

2.3.3 Integration with DaCura Platform Workflow

Most of the lifecycle steps on the left-hand side of our data harvesting process (fig. 1) are already directly or partially supported by the DaCura platform. Thus it forms a basis for our overall tool chain. However it had to be integrated with the new data harvesting client that supports the right-hand side of the overall data harvesting process diagram. It is possible to create a loose coupling between the two parts of the system since the DaCura platform already includes data lifecycle management support. Thus once the data is in the DaCura system it can continue to be revised, updated and maintained. This means the data harvesting client can focus on supporting event report data creation and submission to the DaCura system. When a researcher creates new event report data, the data harvesting client generates a dataset update call to the DaCura server interface. This standard interface is implemented as a REST (representational state transfer) web interface. It holds the event report data payload natively as RDF with a minimum of control or header information such as the credentials of the researcher requesting the event report data creation.

2.4 Developing the Online Repository

Providing a way for non-specialist users to access the dataset was a key consideration. If the information it contains can only be accessed by linked data experts, then its usefulness for social science purposes will be very limited. In order to allow social scientists to access the data without requiring linked data expertise, we built an online repository for political violence datasets. This repository contains schema documentation, access to the SPARQL endpoint, and documentation for the UKIPV data collection volunteers. We also built a web-based interface for the dataset [19], which allows users to search, order, and visualize the dataset or parts of it, and export sections which are relevant to their research or interests.

2.5 A POLITICAL VIOLENCE VOCABULARY

This section details the structure and terms of the political violence historical event vocabulary at its current stage of development. It also discusses the potential interlinks between political violence datasets and linked data datasets inherent in the vocabulary design.

Existing vocabularies that influenced the design of this vocabulary were LODE: Linking Open Descriptions of Events [7], the Time Ontology in OWL [4], the Open Annotation Data Model [8], the Dublin Core Metadata Element Set [9], Geonames[1], the vCard Ontology [10] and the ontology meta-data structure defined by the Live OWL documentation environment [11].

2.6 Vocabulary Structure

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. However it is still a useful visualization of the overall vocabulary structure. There are three main classes defined (upper left corner): the historical Event and its two sub-classes (sub-sets) the Report and the Political Violence Event. Events are sub-divided in this fashion because it is necessary to distinguish between reports of violence and events that are classified by a domain expert as meeting all the criteria for inclusion in an analysis of political violence events.

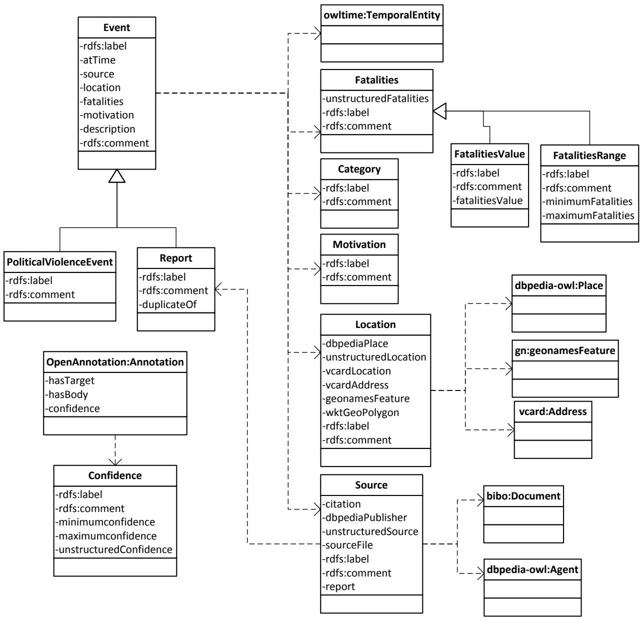

Figure 2. UML class diagram illustrating structure of the political violence vocabulary.

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 vocabulary to enable researchers or other consumers of the data-set to annotate individual dataset elements. One special property is defined in the PV vocabulary to enable this, the confidence property of an annotation. This enables researchers to mark individual event properties as uncertain, something that occurs frequently in collection of event data from historical sources. Thus it is important that this feature can be represented in our model.

One other aspect of the vocabulary not shown above is the large number of instances (allowed values) defined for the event category and motivation fields. These standard terms, e.g. riot, from political violence research provide a means to constrain the allowed values of each field and hence support automation.

[1] http://www.geonames.org/

2.7 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.

2.7.1 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. Assassination and terrorism describe political violence which is perpetrated by small groups – on another small group for assassination, and on a large group for terrorism. Lynching and riot describe political violence perpetrated by large groups – on a small group for lynching, and on a large group for riot. Following previous work, small groups refers to groups of 12 or fewer individuals, and large groups to any group of more than 12.

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.

2.7.2 Motivations

The motivation class describes the reasons political violence event occurred. Events may have multiple motivations if they have numerous or complex causes. Motivation definitions are given below. The motivations volunteers are advised to use in categorization are bolded.

Table 1. Motivation terms in the political violence vocabulary

Criminal	Violence with no political motivation - performed for financial gain or personal motivation.
Economic	Violence centred on economic issues.
Education	Violence which occurs in/around schools and colleges.
Ethnic	Violence between national groups or narrow ethnic groups.
Extralegal	Violence which occurs as a punishment for violation of legal, moral, or cultural codes.
Family	Rampages with a large proportion of killed family members/close acquaintances.
Indian	Conflicts involving American Indians.
Insane	Rampages committed by mentally ill people.
Labour	Violence between employees and their employers, or in the context of industrial action.
Land	Violence which occurs in defence of individual land rights and ownership.
Military	Violence which occurs between soldiers, or which involves off-duty soldiers.
Other	Violence which does not fit into any of the other categories. The purpose of the violence can be ascertained, but it does not fit into one of the categories listed.
Personal	Violence involving family/faction fighting, personal grudges, etc
Political	Violence between political factions or over political events, including rebellions
Prison	Violence which occurs in prison.
Race	Conflict between broad ethnic groups.
Revenge	Violence motivated by revenge for perceived slights.
Religion	Violence which occurs between religious groups, or which is motivated by religious belief.
Section	Violence between pro-/anti-slavery in the immediate run-up to the American Civil War.
Sex	Rampages inspired by 'sexual frustration'.
Shopping	Violence which occurs as a result of shopping - e.g. sales-related frenzies.
Work	Violence which occurs in/around the workplace.

2.8 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 [18], (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. For example the Location class includes the property “dbpediaPlace” which enables instances of the Place concept defined in the DBpedia ontology to be directly referenced. Thus a political violence event that takes place in the US state of Ohio can be linked to the DBpedia concept of the US state of Ohio. This means that the DBpedia (and hence Wikipedia) data on Ohio can easily be queried to gain additional context on Ohio-based events, for example a picture or historical synopsis of the state in a web-based visualization of both datasets.

2.9 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.

DaCura contains a widget-generation component which provides a simple web-based user-interface tool which allows users to input new instance data that will conform to DaCura's constraints, without requiring any knowledge of RDF on the part of the user. These widgets are automatically generated from the RDF schema. However, in order to allow the widget to be fine-tuned to the specific dataset, there are a number of additional requirements on the schema design. For example, by defining ranges and domains of each property in the schema, user-input elements can be generated which can both provide more convenient user-input elements and better validation of user input. The tool also allows very specific 'pre-boiled' data input elements to be associated with particular properties in any given dataset.

Thus, as a final stage in our vocabulary design, we added RDFS ranges and domains to each property in the schema and further defined mappings from the more complex properties to specific user input elements that covered the sub-set of instances that we would encounter in the data-harvesting task. The design of this widget-generation component will be addressed in a forthcoming paper.

3. RELATED WORK

In this section we briefly discuss sources of RDF vocabulary design advice, prior work on representing events in RDF and examine a contemporary example of related work on a historical corpus annotation tool.

3.1 Sources of Vocabulary Design Advice

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 [12]. However Dodds and Davis [13] give a concrete set of examples for designs that are based on Linked Data use cases and were influential on this paper. Another source of vocabulary design conundrums is the W3C linked open data mailing list [14]. The W3C Government Linked Data (GLD) Working Group is also a source of best practice guidelines such as their recent informative note [15].

3.2 Representing Events in RDF

Shaw et al. [3] 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 on what they characterise as the factual aspects of events, i.e. "What happened, Where did it happen, When did it happen, and Who was involved.". 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 [16] 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. However we did adopt the "atTime" property defined in this vocabulary to give us a common basis for mapping between events described with our political violence vocabulary and the LODE ontology.

3.3 Corpus Annotation

The process of event data harvesting is closely related to research on cultural heritage corpus annotation. For example, the event data includes a reference to the original newspaper article and the harvested event data is similar to structured annotation meta-data.

Recently Ferro et al. have demonstrated their FAST-CAT (Flexible Annotation Semantic Tool - Content Annotation Tool) system [17]. This tool, in common with our approach, is a generic corpus annotation framework based on an RDF knowledge model. Their work takes place in the context of the CULTURA project [1] that provides adaptive and personalized access to online historical collections. FAST-CAT has been integrated into the CULTURA environment to provide users with an additional means of interacting with the portal. This parallels our work on documenting and exposing prior Cliodynamics research on political violence events in a way that facilities new researchers or members of the public engaging with the historical record, analyzing it, linking it with other linked data data-sets and re-using or adding value to the research in as yet unanticipated ways. Our work specifically enables the application of the US political violence dataset collection methodology to the creation of a new UK and Ireland political violence dataset and the ongoing improvement and maintenance of both data-sets into the future.

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.

After the creation of the ontology, we were able to import the original USPV dataset into an RDF triplestore (graph data store) which describes the events in accordance with the vocabulary described here. This dataset will soon been published to the web of data and the web interface is already available [19]. The publication of this data will allow social scientists to access extensive datasets recording political violence, and to build on this research.

The development of a tool-supported data-harvesting and curation process is potentially of use on a more general level. Social scientists who need to generate structured data from human-readable sources such as newspaper archives may find our approach provides them with a significantly improved way of obtaining such data.

One area with potential for improved accuracy is in the precision and recall of the chosen search terms. The development of our gold standards was time-intensive, particularly for 1831, due to low accuracy OCR in the source archive, the large quantity of material to be analyzed, and the lack of tool support. This meant that tests could only be performed on a small subset of potential results, which may not be large enough to fully account for linguistic and stylistic changes. Analysis of a larger corpus may provide an insight into whether alternative search terms might improve precision and/or recall.

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. A political violence event may potentially be recorded in numerous reports, which will need to be associated. The domain expert moderator queue will support historians performing this task, automatically suggesting reports which are likely to be associated with a particular event due to metadata similarities and allowing them to easily associate related reports with the relevant event.

5. ACKNOWLEDGEMENTS

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

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] Hobbs, J.R. and Feng, P. (eds.). 2006. Time Ontology in OWL. W3C Working Draft, 27 September 2006. Retrieved June 7, 2013, from W3C: http://www.w3.org/TR/owl-time

[5] 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.

[6] 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).

[7] 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

[8] 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/

[9] Powell, A., Nilsson, M., Naeve, A., Johnston, P., and Baker, T. 2007. DCMI Abstract Model. DCMI Recommendation, 2007. Retrieved June 9, 2013 from Dublin Core Metadata Initiative: http://dublincore.org/documents/2007/06/04/abstract-model/

[10] Iannella, R., McKinney, J. 2013. vCard Ontology For describing People and Organisations. W3C First Public Working Draft 2 May 2013. Retrieved June 9, 2013 from W3C: http://www.w3.org/TR/vcard-rdf/

[11] Peroni, S., Shotton, D., Vitali, F. 2013. Tools for the automatic generation of ontology documentation: a task-based evaluation. To appear in *International Journal on Semantic Web and Information Systems, 9* (1).

[12] 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.)

[13] 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/

[14] public-lod@w3c.org, retrieved June 6, 2013, from archive available at http://lists.w3.org/Archives/Public/public-lod/

[15] Hyland, B., Villazón-Terrazas, B. and Atemezing, G. (eds.). 2013. Best Practices for Publishing Linked Data. W3C Note, 6 June 2013. Retrieved June 6, 2013, from W3C: https://dvcs.w3.org/hg/gld/raw-file/default/bp/index.html

[16] Scherp A., Franz T., Saathoff, C., and Staab, S. 2009. F—A Model of Events based on the Foundational Ontology DOLCE+ Ultra Light. In *5th International Conference on Knowledge Capture* (K-CAP’09, Redondo Beach, California, USA, 2009).

[17] Ferro, N., Munnelly, G., Hampson, C., and Conlan, O. 2013. Fostering Interaction with Cultural Heritage Material via Annotations: The FAST-CAT Way. In *Proceedings of the 9th Italian Research Conference on Digital Libraries* (IRCDL2013, Via Ludovico Ariosto, Rome, Italy, 31 Jan-1 Feb 2013).

[18] 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.

[19] Political Violence Datasets 2013, Knowledge and Data Engineering Group, Trinity College Dublin, Retrieved June 6, 2013 from http://tcdfame.cs.tcd.ie/dacura/PV