HAL
archives-ouvertes.fr

Automated Monitoring of Data Quality in Linked Data Systems

Kevin Feeney, Rajan Verma, Max Brunner, Andre Stern, Odhran Gavin, Declan O'Sullivan, Rob Brennan

▶ **To cite this version:**

Kevin Feeney, Rajan Verma, Max Brunner, Andre Stern, Odhran Gavin, et al.. Automated Monitoring of Data Quality in Linked Data Systems. 2nd International Workshop on Computational History and Data-Driven Humanities (CHDDH), May 2016, Dublin, Ireland. IFIP Advances in Information and Communication Technology, AICT-482, pp.121-123, 2016, Computational History and Data-Driven Humanities. <hal-01616350>

HAL Id: hal-01616350
https://hal.inria.fr/hal-01616350

Submitted on 13 Oct 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution| 4.0 International License

Automated Monitoring of Data Quality in Linked Data Systems

Kevin Feeney, Rajan Verma, Max Brunner, Andre Stern, Odhran Gavin(✉), Declan O'Sullivan, and Rob Brennan

Knowledge and Data Engineering Group, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland
{feeneykl, vermarl, gavino, declan.osullivan, rbrenna}@tcd.ie,
emailparaomax@gmail.com, andrestern88@gmail.com

Abstract. This paper describes the Dacura system's ability to monitor data quality. This is evaluated in an experiment where a dataset of historical political violence is collected, enriched, interlinked, and published. The results of the experiment demonstrate that automated quality measures enable the construction of publication pipelines which allow datasets to evolve rapidly without loss of quality.

Keywords: Linked data quality metrics · Data curation · Visualization · Semantic web

1 Introduction

Ensuring high-quality data is a difficult task. Most large-scale data sources contain a significant amount of inconsistent data, due to differences in encoding and available information, and human error. The Dacura system [3] developed at TCD is a data curation system designed to support the harvesting, assessment, management and publication of high-quality Linked Open Data. We wish to support an internationally distributed community of humanities and social science scholars collaborating on the Seshat Global History Databank project [1], which aims to encode time-series datasets describing the most important features of all human societies since Neolithic times. The scope of the project – over 100 researchers and approximately $10 million in total funding, divided across multiple autonomous collaborating projects, with a 10 year time-frame – is such that the data-collection process is necessarily incremental.

The goal of our work is to harness the input human experts to efficiently transform the wealth data into high-quality datasets and to provide visualisations, analysis and modelling, data-export and a variety of other tools based upon that data. The system must be dynamic because a requirement of the research program is to iteratively publish datasets which cover specific regions and time-slices and subsets of the Seshat variables and to evolve the datasets so that they improve progressively over time as their coverage is extended.

© IFIP International Federation for Information Processing 2016
Published by Springer International Publishing AG 2016. All Rights Reserved
B. Bozic et al. (Eds.): CHDDH 2016, IFIP AICT 482, pp. 121–123, 2016.
DOI: 10.1007/978-3-319-46224-0

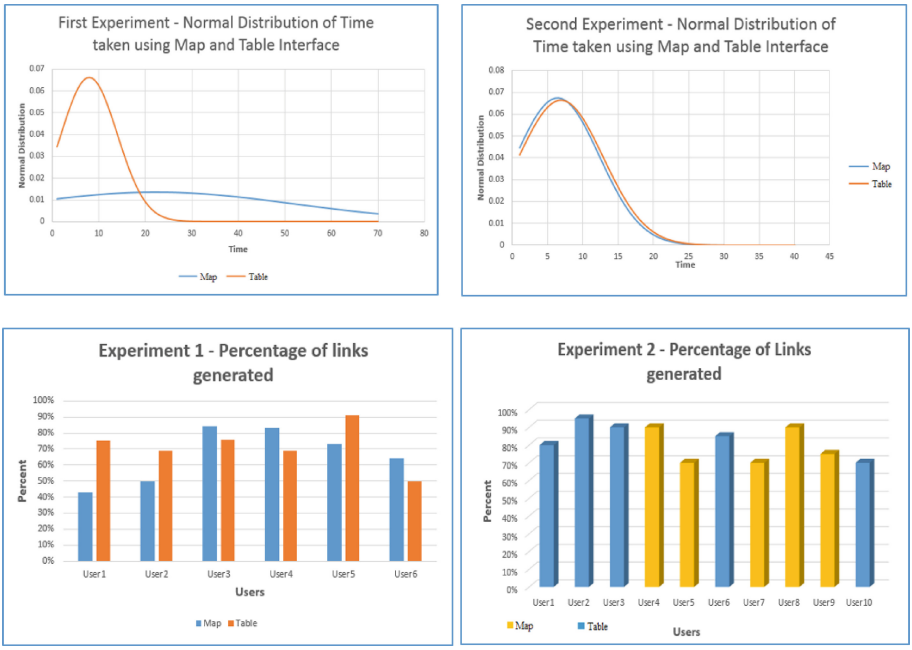

Fig. 1. Distribution of results by time and user

2 Experiments

The goal of the experimental work described here was to use the Dacura system to create and improve a linked dataset of historical political violence events. The source of data was a database of 1599 events that took place between the years of 1784 and 2010, collected by Peter Turchin [2]. This database had been manually compiled and was of unknown quality. We applied an uplift script to import it into the Dacura system. We then used Dacura to assess the quality of the data and to improve this quality wherever possible using a pool of non-expert volunteers. The goal was to use this labour as efficiently as possible to improve the quality of the dataset.

The first step where Dacura's quality assessment features were used was immediately after uplift. The RDF generated by this step was tested for conformance with the Political Violence schema. This process immediately identified a problem with the dataset which was traced back to a bug in the uplift script. When this was fixed and the basic Dacura quality control check rerun, the dataset passed the quality checks. However, the accuracy of the data itself remained an unknown quantity.

The next stage in the data-quality improvement process was to load the dataset through our visualization application. Upon cursory observation, a number of inconsistencies were found. The locations of the events had been encoded as simple text strings in the original database and the structure of these strings varied, meaning that in many cases they could not be mapped to structured linked data location properties. Furthermore, the way that these properties were encoded varied over the course of the

data-collection process, which made their mapping to real locations less reliable. The next challenge was to modify the dataset to improve these location properties to such an extent that they would be sufficiently accurate for analytics - we aimed to produce at least 90 % accuracies.

In order to achieve this improvement, we combined a selection of Dacura's automated and manual tools. We used Dacura's dataset interlinking tool to generate links between our dataset and DBpedia, based on the location in the original database. Where it proved impossible to match our strings to DBpedia location, we used Google's autocomplete API to suggest location names. We then ran a series of experiments where we got volunteers to go through the records in the dataset, using the automated interlinking tools that Dacura provided, to assist them in choosing the correct location. To evaluate these experiments we went through each of the corrected location properties manually to identify whether it was, in fact, the correct location information for that particular event. The goal of these experiments was to assess whether we could use relatively unskilled volunteers with Dacura's tools to produce a high-quality dataset from the inconsistent database without requiring expert input. We ran several iterations of this experiment, improving the tools in each iteration, in order to ensure that User Interface issues were not impacting upon experimental results.

The process of improving the accuracy of the data by non-expert volunteer users was measured through two experiments – the first involved 6 users, the second 10 users. Each user was presented with a series of events which had locations that could not be linked automatically to DBpedia locations. Two different Dacura inter-linking tools, one map-based, one table-based, were used by each user in both experiments (Fig. 1).

The most important conclusion is that we now know that by devoting approximately 2 h and forty minutes of non-expert human effort, we can reduce the amount of data that has to be examined for approval by an expert by approximately 80 %.

3 Conclusion and Future Work

In this paper we described how the Dacura framework has been used to improve and extend a dataset of political violence events. Dacura's tools were used to allow non-expert users to correct location errors in approximately 80 % of cases, requiring, on average, 6 s per record. Future work will see us extend the quality assessment aspects of the framework to non-location properties and other domains.

References

1. Turchin, P.: The SESHAT databank project: the 2014 report. Cliodynamics: J. Quant. Hist. Cult. Evol. **5**(1) (2014)
2. Turchin, P.: Dynamics of political instability in the United States, 1780–2010. J. Peace Res. **49**(4), 577–591 (2012). doi:10.1177/0022343312442078
3. Feeney, K., O'Sullivan, D., Tai, W., Brennan, R.: Improving curated web-data quality with structured harvesting and assessment. Int. J. Semant. Web Inf. Syst. **10**(2), 35–62 (2015)