Structured Data To RDF II

Deliverable D4.3.2

Version Final

Authors: W.R. Van Hage[1], T. Ploeger[1], J.E. Hoeksema[1]
Affiliation: (1) SynerScope B.V.

Building structured event indexes of large volumes of financial and economic data for decision making
ICT 316404

Grant Agreement No.	316404
Project Acronym	NEWSREADER
Project Full Title	Building structured event indexes of large volumes of financial and economic data for decision making.
Funding Scheme	FP7-ICT-2011-8
Project Website	http://www.newsreader-project.eu/
Project Coordinator	Prof. dr. Piek T.J.M. Vossen VU University Amsterdam Tel. + 31 (0) 20 5986466 Fax. + 31 (0) 20 5986500 Email: piek.vossen@vu.nl
Document Number	Deliverable D4.3.2
Status & Version	Final
Contractual Date of Delivery	September 2014
Actual Date of Delivery	September 15, 2014
Type	Report
Security (distribution level)	Public
Number of Pages	16
WP Contributing to the Deliverable	WP4
WP Responsible	SynerScope B.V.
EC Project Officer	Susan Fraser
Authors: W.R. Van Hage[1], T. Ploeger[1], J.E. Hoeksema[1]	
Affiliation: (1) SynerScope B.V.	
Keywords: structured data, rdf, conversion	
Abstract: In this deliverable we describe the second conversion of four data sets to RDF for use within the NewsReader project. These data sets are intended to supplement the event indexes extracted from news articles. We describe TechCrunch, CrunchBase, the World Bank Indicators, and Yahoo! Finance. We show the changes with respect to the first conversion process.	

Table of Revisions

Version	Date	Description and reason	By	Affected sections
0.1	15 Aug 2014	Deliverable skeleton	Thomas Ploeger	All
0.2	30 Aug 2014	Initial content for all sections	Thomas Ploeger, Jesper Hoeksema	All
0.3	9 Sep 2014	Processed VUA feedback on all sections	Thomas Ploeger	All
Final	15 Sep 2014	Internal review	Thomas Ploeger, Jesper Hoeksema	All

Executive Summary

The NewsReader project aims to support decision making by building structured event indexes of large volumes of news articles and financial data. To provide additional context, it is necessary to supplement the event indexes with additional data sets.

TechCrunch is a set of news articles about tech startups. CrunchBase is a database of startups, people, and financial organizations that serves as the structured data companion to TechCrunch. The World Bank Indicators are statistical indicators of development (such as GDP or number of hospitals) for countries world wide. Yahoo! Finance provides historical stock prices.

To be stored in the NewsReader KnowledgeStore, these datasets were converted to RDF in Deliverable 4.3.1. Based on feedback and third party changes to the original datasets, we performed a second conversion of each data set to RDF. We present the changes with respect to the earlier conversion process in this deliverable.

Contents

List of Figures

1 Introduction

The NewsReader project aims to support decision making by building structured event indexes of large volumes of news articles and financial data. In this deliverable we describe the second conversion of four existing structured data sets to the Resource Description Framework (RDF). These data sets are intended to supplement the event indexes with additional context.

The reader is assumed to have at least a basic understanding of RDF (including vocabularies, ontologies, and named graphs), JSON, CSV, and the NewsReader project in general.

The four datasets and their purpose within the NewsReader project are described below. These data sets need to be converted to RDF because this is the format the NewsReader KnowledgeStore (see Deliverable 6.1) is designed for.

TechCrunch[1] A news website about information technology companies.

CrunchBase[2] A database of technology companies, people, and investors. Together with TechCrunch, this dataset will be used in the evaluation (Deliverable 8.2.1) of the first versions of the decision support systems (Deliverable 7.3.1).

World Bank Development Indicators[3] Per-country statistical indicators of development and quality-of-life. This dataset will be used to supplement the event indexes with developmental context.

Yahoo! Finance[4] Historical prices of individual stocks as well as stock market indexes. This dataset will be used to supplement the event indexes with financial context.

The reasons for selecting specifically these data sets over other similar data sets are described in Deliverable 1.1: Definition of Data Sources. In that deliverable, the selection criteria are explained in detail, together with several example usage scenarios for the data sets.

In deliverable D4.3.1, we described the first version of the conversion process. This deliverable describes the second version of the conversion process. In Section 2 of D4.3.1 we described the data sets in more detail. We gave an overview of available methods for converting structured data to RDF in Section 3 of D4.3.1. Section 4 of D4.3.1 contains the details of the actual conversion process for each individual data set. We do not reproduce these sections in this deliverable to avoid redundancy.

D4.3.1 concluded with an overview of fixes and improvements that were planned for Deliverable 4.3.2. In this deliverable we describe the actual changes made to the conversion process.

[1] `http://www.techcrunch.com`
[2] `http://www.crunchbase.com`
[3] `http://www.worldbank.org`
[4] `http://finance.yahoo.com`

The rest of this deliverable is structured as follows. In Section 2 we describe the changes that were planned in D4.3.1. In Section 3 we describe the actual changes made to the conversion process. We conclude this deliverable in Section 4.

2 Changes Planned in D4.3.1

In Deliverable 4.3.1, we described four data sets that were converted to RDF for use within the NewsReader project: TechCrunch, CrunchBase, the World Bank Indicators, and Yahoo! Finance. For each data set, we described what kind of data it contains, how that data is structured, how the data was acquired, and what purpose it will serve within the project.

We gave an overview of different approaches for converting existing structured data to RDF: Writing a custom script, using an off-the-shelf tool, or taking advantage of an existing RDF version of the data. We have shown how we used these methods to convert our data sets to RDF and what the result looks like.

When D4.3.1 was written, the following changes and improvements (grouped by data set) were planned for this deliverable:

TechCrunch

1. Do not use blank nodes as article identifiers, but create proper URIs.
2. Investigate how the data set can be continuously updated with newly published articles.

CrunchBase

1. Create detailed specification of CrunchBase vocabulary.
2. Do not use blank nodes as date identifiers, but mint proper URIs.
3. Investigate how the data set can be continuously updated with newly added entities.
4. Do not use asserted typing to not confuse reasoning software. Rather, specify a proper ontology (see point 1) and use inferred typing.
5. Do not place provenance triples and provenance metadata triples (e.g. triples about persons involved in the conversion) in the same named graph.
6. Add a validity context to triples where appropriate (e.g. an e-mail address can only be valid in a certain time period).

World Bank Indicators

1. Investigate whether it is necessary to re-write the World bank Indicator RDF for use within NewsReader or whether it can be used as-is.

Yahoo! Finance

1. Investigate which stock symbols need to be downloaded for use within NewsReader and research which vocabularies are appropriate.

3 Second Version of Conversion

Based on the planned changes described above (in Section 2), feedback from the NewsReader consortium, and changing circumstances as a result of third party actions, we have (had to) make several changes to the conversion process of most datasets. We describe these changes per dataset in the following sections.

3.1 TechCrunch

Only two minor changes were made to the conversion process for the TechCrunch dataset. As planned in D4.3.1, we no longer use blank nodes to identify articles. We now use the MD5 hash of the article's original URI as a unique identifier, which is prepended with the string "http://techcrunch-rdf.org/articles/" to create a proper URI.

We have investigated other options for keeping the set of articles up-to-date, but for the current and projected use of the articles within the project the most economical solution seems to be incremental (e.g. monthly or weekly) scrapes of the TechCrunch website. These scrapes can then be fed to the conversion script for conversion and added to the existing set of converted articles.

The new conversion script can be found in the NewsReader BitBucket code repository together with the old conversion script. The input and output data remain the same, so they are not reproduced for D4.3.2.

3.2 CrunchBase

On April 22nd, 2014, CrunchBase released CrunchBase 2.0, a "more beautiful, easier-to-use, and more powerful version of the database that the tech world has come to depend on." As a consequence of this change, their old API interface, used in Deliverable D4.3.1, was deprecated and will no longer serve any updated information.

A new API was announced shortly thereafter, and was released on June 24th, 2014. Unlike the old API, this new version restricts the number of requests that can be performed per day, essentially making the collection of the entire dataset through the API prohibitively expensive in terms of time.

Fortunately, CrunchBase also provides static dumps of their data in the form of Excel files. These dumps do not contain the full extent of the data that could be retrieved using the old API, but they do contain the subset of the data that is the most interesting: Companies that actually raised money from investors.

In addition to re-wiring the conversion scripts to use the data dumps instead of the API, we took into account all of the feedback from the NewsReader consortium regarding

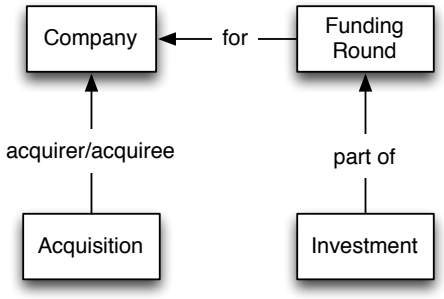

Figure 1: Diagram of relationships between CrunchBase entities in CrunchBase static export.

the modelling of CrunchBase in RDF. The new conversion scripts can be found in the NewsReader BitBucket code repository.

Because the changes are quite significant, we describe them in detail in the following subsections.

3.2.1 Input Data

The input data from CrunchBase contains of a dump of their data in the form of an Excel workbook file. This file contains a separate sheet for each type of entity in CrunchBase (Companies, Funding Rounds, Investments and Acquisitions). Additional sheets (e.g. licence, analysis) exist, but we do not use these.

Each sheet contains a flat table with data, one record per row. Data from linked records (e.g. information about the company funded during a funding round) is often replicated in several sheets (in this case both in the Companies sheet, the Funding Rounds sheet and the Investments sheet). The relations between the different sheets are shown in Figure 1.

3.2.2 Conversion Process

The data from CrunchBase was converted to RDF using a Ruby Script. This approach was chosen because it allowed us to do a number of transformations more complex than would be possible with off-the-shelf mapping tools like D2RQ or Sparqlify. These tools operate on the basis of a simple one-to-one mapping between rows in a CrunchBase dump sheet and RDF entities, while we need to generate more complex graph patterns such as interconnected events with multiple actors.

At a high level, the conversion scripts work by iterating over the rows in CSV files derived from the Excel workbook, creating a number of RDF statements for each row. A separate named graph is used to contain the data derived from each CSV file, and hence from each sheet in the Excel workbook. These named graphs allow us to add statements describing the provenance of the triples contained in the graph. These statements, which

are triples themselves, for example state which file the RDF triples in a certain graph were derived from, which person performed the conversion, and when the conversion took place.

The conversion from XLSX to CSV was done using Excel. Some third-party Excel-compatible tools (such as Apple Numbers) do not support Excel sheets with over 65500 rows. Therefore using any software other than recent versions of Excel might result in data loss, because some of the sheets from CrunchBase contain more rows than the 65500 row threshold.

In order to reduce redundancy in the data, we have used a simple hashing scheme to generate URIs for entities that could be used in multiple places, such as addresses and timestamps. Previously, every occurrence of an address or timestamp would receive a randomly generated distinct ID even if the same underlying address or timestamp was the same as an existing one. This also allows easier querying for events that occurred at the same time or place.

We have used the following vocabularies to model entities, provenance, events, addresses and times:

SEM The Simple Event Model is used for event-related triples, such as specifying that something is an event, that it has certain actors participating in that event, where the event took place, and when it took place.

PROV-O The Provenance Ontology is used for statements about the provenance of triples in named graphs (as explained above), specifically which sources they were derived from and who was responsible for the conversion.

FOAF Friend-of-a-Friend is used for statements about addresses and contact information.

OWL Time OWL Time is used for representing instances and durations of time. DC Dublin Core for certain metadata properties.

vCard vCard was used for defining detailed addresses.

The new conversion scripts (together with a sample of input and output data) can be found in the NewsReader BitBucket code repository together with the old conversion scripts. In addition, we have also created a 'mini' ontology that describes the relationships of some of our self-defined predicates to existing vocabularies. This ontology can also be found in the code repository.

3.3 World Bank Indicators

No changes were made to the World Bank Indicators, as they are provided by a third party and are ready to be used in their current state.

3.4 Yahoo! Finance

We have discovered an existing Linked Data Wrapper[5] for the Yahoo! Finance API. This wrapper returns historical stock prices in the RDF Data Cube vocabulary for modeling statistical observations and the SDMX vocabulary for statistical codes, similar to the World Bank Indicators. This wrapper was developed by OntologyCentral[6].

Initially, this wrapper only returned historical stock prices in a daily granularity, which would make retrieval of years of historical stock prices for the NewsReader project quite inefficient.

We have been in contact with the developers of the wrapper to ask them if they would consider offering the stock prices in a yearly granularity. The developers indicated that this would not require much effort, and they implemented this feature for us.

Once the consortium decides for which stock symbols we want to retrieve the historical stock prices, we can use this linked data wrapper to easily retrieve them in RDF format.

4 Conclusion

In this deliverable, we have described the RDF conversion of four data sets that needed to be converted to RDF for use within the NewsReader project: TechCrunch, CrunchBase, the World Bank Indicators, and Yahoo! Finance.

For each data set, we have described how the conversion process was changed from the previous conversion process as described in D4.3.1. The majority of changes were for the CrunchBase dataset. The other 3 datasets remain largely unaffected.

References

[5] http://yahoofinancewrap.appspot.com/
[6] http://ontologycentral.com/