Alternative converters from tabular data to RDF

This page points to external listings as well as providing a place to describe different approaches for converting existing tabular data (spreadsheets and CSV files, mostly) to RDF.

References gleaned from the public-semweb-lifesci@w3.org list:

Rafael Richards: Healthcare data published by the CDC unfortunately comes as nearly 200 separate spreadsheets:

http://www.cdc.gov/nchs/hus/contents2011.htm#chartbookfigures
The only thing I am aware of that is designed to keep large numbers (potentially hundreds) of spreadsheets continuously integrated and in sync across an enterprise, each independently curated, is Anzo by Cambridge Semantics. Most of the other tools I am aware of do not do real-time updating of the RDF model from the CSV model, and are one-off conversions, so if you have more than one spreadsheet to update, it will be time consuming.
For one-off conversion Google Refine is quite easy to get started. It has a great deal of data cleaning facilities for noisy or illogical data. With its RDF extension you have *automated* data reconciliation with outside linked data sources of your choice as DBpedia. This is a feature I have not seen with any other conversion tool. It does not do visualization, but there are plenty of desktop applications that do this very well.

Just be sure to install the RDF extension to Google Refine: http://refine.deri.ie/
This will give you the capability to reconcile and interlink your spreadsheet data against external SPARQL endpoints or RDF dumps, to search the web for related RDF datasets, and export your data as RDF. On export you can define the shape of the RDF graph using your own vocabulary or import existing ones.

---

editor's note: OpenRefine is apparently continuing from its predecessor Google Refine ; Eliza from Weill Cornell wrote and has documented a Google Refine plugin for VIVO.

Lee Feigenbaum: As you say, Anzo (in particular Anzo for Excel) is designed for enterprises to curate large numbers of spreadsheets, map them to ontologies & to existing RDF instance data, and maintain them as changes are made to the spreadsheets or to the data in the spreadsheets. It can be used for CSV-style "tabular" spreadsheets and also for arbitrarily "human-oriented" spreadsheets. It can be used both in interactive modes (where people are opening up and interacting with spreadsheets) and also in automated batch modes.

Anzo stores the RDF data from spreadsheets in an RDF database. Anzo includes both authenticated and unauthenticated SPARQL endpoints for this data; Anzo can also directly publish the data as Linked Data. Finally, Anzo gives you several ways to export RDF data from the database.

Anzo is available in several editions:

- Anzo Express Starter -- includes Anzo for Excel as above for limited #s of users; freely available
- Anzo Express -- includes Anzo for Excel and Anzo on the Web, a user-friendly browser-based dashboard tool for visualization, searching, and analyzing RDF data
- Anzo Enterprise -- includes the above in addition to tools to connect to data in relational databases, to integrate unstructured data from documents, web pages, etc., to run rules and reasoning and work flow processes, various server-side and client-side APIs, etc.

We also make Anzo available for free for academic use.

Jim McCusker: Tim Lebo's csv2rdf4lod is designed for repeatability and scalability. It was developed to handle transforming the data from data.gov into RDF, and has been set up to automatically convert thousands of datasets. We even have one project that regularly updates conversion configurations from github and be converted and loaded into a triple store automatically via cron. Further, it supports dataset versioning, where you can keep multiple versions of data around without URI collisions. It also supports re-using conversion configurations for multiple files that share a common format.

https://github.com/timrdf/csv2rdf4lod-automation/wiki

Tim Lebo maintains a Alternative Tabular to RDF Converters page on Github.

Eric Gombocz: IO Informatics' Knowledge Explorer. Professional Edition, also provides an automated way to facilitate import and updating a triplestore backend of your choice via monitored folders which will map and import incoming spreadsheets to RDF. You can set up multiple monitored folders with different data mappings, and this will run as background processes to continuously update one or multiple connected triplestores (or different graphs in a single triplestore.

The Knowledge Explorer also provide scripting within the import mapping, application of thesauri and other mechanisms for data transformation to clean, consolidate and harmonize data during the import.

You can find out more about this tool here: http://www.io-informatics.com/products/sentient-KE.html

---

Peter Ansell: Michel Dumontier's php-lib library is what Bio2RDF has been using for converting TSV, CSV files (and other file formats) to RDF [1]. It contains some aspects that are Bio2RDF specific, namely its support for prefixed URIs, but any Pull Requests on GitHub would be appreciated to generalise that. OSX has PHP installed by default as far as I know so you can use it on the command line without any other dependencies.

You can find examples of scripts using php-lib in the bio2rdf-scripts repository on GitHub [2]. A fairly simple example would be the HGNC converter, which is Tab separated, but quite similar [3].

[1] https://github.com/micheldumontier/php-lib
[2] https://github.com/bio2rdf/bio2rdf-scripts
[3] https://github.com/bio2rdf/bio2rdf-scripts/blob/master/hgnc/hgnc.php#L129

Katy Wolstencroft: Our tool, RightField (http://www.rightfield.org.uk), allows you to embed ontology term selection into spreadsheets, and to extract these selections as RDF. It is designed more for assisting in the data collection process (i.e. when users fill in a spreadsheet that has been marked-up using RightField, they are automatically collecting semantically enriched data).

Our paper from last year's eScience conference describes the RDF extraction in more detail:
Wolstencroft, Katherine; Owen, Stuart; Goble, Carole; Nguyen, Quyen; Krebs, Olga; Muller, Wolfgang; , "RightField: Semantic enrichment of Systems Biology data using spreadsheets," *E-Science (e-Science), 2012 IEEE 8th International Conference on* , vol., no., pp.1-8, 8-12 Oct. 2012
doi: 10.1109/eScience.2012.6404412