data.world: A Platform for Global-Scale Semantic Publishing

Bryon Jacob[1[0000-0003-0470-9300]] and Jonathan Ortiz[1[0000-0002-1042-0776]]

[1] data.world, 7000 North Mopac Expressway #425, Austin, TX 78731 USA
bryon@data.world jonathan.ortiz@data.world

Abstract. Data scientists, journalists, analysts and academics struggle to find data sources, prepare analyses, collaborate with peers, present their results in context, and find an audience for their conclusions. The data.world project aims to create a semantic-based publication platform for datasets, scalable to hundreds of thousands of heterogeneous users and millions of distinct datasets. With such broad horizontal scale comes a highly varied population of datasets, ranging from multi-kilobyte Excel spreadsheets to multi-gigabyte RDF graphs, and of users, ranging in technological sophistication from journalists and students to highly skilled data scientists. The data.world project leverages semantic web technologies including CSVW, HDT, and dataset metadata ontologies to automate the ingestion, discovery, presentation and linkage of both tabular and native graph datasets. Cloud-based technologies are used to elastically scale the back-end graph storage and query fabric to meet availability and reliability targets with acceptable response latency. This abstract provides an overview of a technical demonstration of data.world, through which data.world's functionality and architecture shall be explained, in addition to the challenges of creating a system of such broad horizontal scale and the role that semantic web technologies have played in powering its development and adoption.

Keywords: linked data, HDT, CSVW.

1 Introduction

Our demo will be driven from our live, production system at https://data.world/. A short slide deck walking through the points we intend to demonstrate, with links to the primary resources, is at https://goo.gl/txcgkD. The demo will begin at data.world's homepage to root ISWC attendees in the company's mission and describe its goals. data.world aims to create a collaborative web platform that can engage a user base consisting primarily of users who are not semantic web experts—trafficking in datasets that are generally not initially semantically linked—and, by using web standards for the automated translation of those tabular data formats into RDF, build a connected network of linked datasets.

An oft-cited stat is that finding, understanding, and preparing data for use can take upwards of eighty percent of the time spent on an analysis project [1]. These projects usually involve multiple data sources in a variety of formats. Semantic web offers a powerful set of tools to deal with this diversity - a universal structure for data and an

open-world model that adapts readily to heterogeneous data acquisition and discovery at web scale. Metadata can be iteratively layered into datasets, by different actors at different times. SPARQL enables structured, federated query between datasets.

Non-semantic solutions abound to address the growing needs of data users in finding data for analysis. However, rather than a data "portal" where there are a small, fixed number of data producers publishing data to data consumers, data.world focuses on collaborations where each actor can play both the producer and consumer roles. Data can be worked on completely in the open, or in private access-controlled datasets. Structured data is converted into RDF and can be queried via SPARQL, but the original data is retrievable as well so all users, not just the semantic web community, can continue working in familiar modes while enjoying the benefits of semantic web technology.

In addition to basing its architecture in semantic web and open standards, data.world also structured itself as a public benefit corporation, and its mission includes improving the "adoption, usability, and proliferation of open data and linked data." As such, data.world's first goal is to connect the people working with data and then leverage those social and collaborative connections to build correlations in the data itself – with the hopes of ultimately achieving a multiplicative "network effect" by connecting those detached data resources in the web of linked data.

2 Connecting diverse user base

The demo will cover data.world's social functionality. In many ways, data.world mirrors other social networks in existence today – one can meet new people, follow other users, talk to them, and collaborate with them. Indeed, data.world endeavors to become the social network for data people just as much as a semantics-based data collaboration platform. Every member has a profile and is encouraged to share content with the community. Instead of status updates, data.world community members post datasets, both public and private. Social currency is provided by likes and follows.

data.world publishes each dataset with a persistent URI, referenceable anywhere, and users upload data files to their datasets, same as they would photos to Facebook. In addition to raw data files, users may upload ontologies, vocabularies, cleaned and derived data files, workflows, scripts, documentation, visualizations, evaluation methods, replication studies, and finalized analysis reports. Users may control access to their datasets by marking them Open – viewable to all – or Private – viewable to contributors they invite. Contributors may have view, edit, and/or ownership rights to a dataset.

Conversations about the datasets take place on data.world within context of the data. By posting a comment to a dataset's discussion tab users discuss dataset accuracy, reusability, related datasets, and news. In addition to discussions, an activity log assists contributors in keeping up-to-date with dataset changes and aids the community in reproducing research and tracking provenance of the data. Users label datasets with titles and tags to assist others in finding their data, and data.world also provides several license types under the Open Data Commons and Creative

Commons to promote open collaboration.

Today, thousands of users—from students and journalists to data scientists and ontologists—collaborate on myriad datasets on data.world. data.world is bridging the gap from the state of the world today to a linked data future by building a social network that connects these diverse data users and their data resources together.

3 Scalable, Heterogeneous Data Publication

The demo will also highlight data.world's data functionality. The main dataset view is designed to give users as much information about the data as quickly as possible by automatically generating data file previews and summary statistics for the user. This information includes sample data, histograms, minima and maxima, averages and standard deviations, skewness, number of distinct entries, number of empty and non-empty, and datatypes like String, Categorical, Numeric, Date, etc.

When a user creates a dataset, data.world generates a unique graph database instance for it. data.world handles each dataset individually, giving each its own SPARQL endpoint, but users can combine graphs easily by referring to any dataset they have access to in data.world as a named graph in any SPARQL query.

When users upload structured data in tabular formats that data.world supports—CSV files and other text-delimited variants, Excel workbooks, relational database tables, JSON Lines, etc.—the data.world ingest pipeline automatically converts those data files into RDF using the CSVW specification, which provides a standardized model for virtualized tables modeled within an RDF graph structure. CSVW tables can use RDF schema and type specification, can be intermixed and queried in conjunction with graph-like structures defined directly in RDF, and can be serialized for transmission or storage in any textual or binary form that RDF can take. Also, when users upload RDF serializations to a dataset, that data is ingested natively into the graph. Finally, datasets can store files of other types as well – not just raw data. Users can save images, PDF documents, visualizations, raw or Markdown text, Jupyter Notebooks, and more alongside their datasets to give vital context to their data.

Although data.world translates tabular data into RDF upon upload, all virtualized tables appear as the tabular sources ordinarily would – as rows and columns of a table. This enables users unfamiliar or uncomfortable with semantic technologies to use their data in familiar ways, including the data.world query tool. Just as users can execute SPARQL queries directly against their datasets, those who have no training in SPARQL may use data.world's SQL-to-SPARQL translation layer. Both SPARQL and SQL queries can federate across all the datasets available to a user on data.world, and SPARQL queries can federate to any RDF data available on the internet with a SERVICE call.

To enable such broad horizontal scale, data.world prioritizes query responsiveness over update flexibility. Updates are handled as bulk ingest, with the output of the ingest pipeline a read-only RDF dataset rendered in the HDT (Header-Dictionary-Triples) file format. This HDT architecture is optimized for the sorts of queries that characterize exploratory semantic web usage and allows us to treat datasets as independent graphs, but loadable together as named graphs for optimized joins.

Finally, getting data out of data.world is as easy as uploading it. Users may extract the raw source files, export a Tabular Data Package, or import into Python, R, Tableau, or Java/JDBC environments.

4 Seamless data preparation workflow

Our demo will cover data projects, a new way of working with data on data.world. Whereas datasets provide a means for users to store canonical references to finalized data artifacts, projects allow users to iteratively join, clean, and analyze disparate datasets they may or may not own or contribute to. This section of the demo will utilize two datasets – a Working in Texas dataset produced by co-author Jonathan Ortiz and a NAICS Codes (2012) dataset produced by data.world user Dean Allemang – and documentation on US Census and IPUMS webpages. The Working in Texas dataset houses CSV collections of demographic info about people who work in Texas categorized by the US Census Bureau's standard industry codes (Census IND codes and INDNAICS codes). The NAICS Codes (2012) dataset contains a SKOS ontology of NAICS codes in their hierarchical structure.

The demo will show how to search for datasets on data.world and link promising results into a personal project workspace, which allows users to create derived data from those sources without affecting the originals. Using SPARQL and the data.world query tool, one can aggregate the Working in Texas data, originally uploaded in CSV, up to a higher-level NAICS entity called a NAICS Sector using Dean Allemang's SKOS ontology of NAICS codes and a crosswalk added to the personal project. This demonstrates how a small network of just three different people, all acting independently, can affect each other's data work. Each open dataset is a community shared resource that can be adapted to support each other's research via the data.world platform in an effort to drive continuous, iterative improvement toward cleaner models and more meaningful data resources.

5 Lessons learned and Future work

Finally, the demo will end with lessons and best practices learned through the development, deployment, and use of the data.world platform. Dataset versioning and provenance is an area of active research and development; our demonstration will cover our current work there. Additionally, our HDT-based query architecture works well for scaling exploratory queries, but it is suboptimal for large analytical (non-selective) queries. We will talk about the work we are doing to leverage a hybrid query architecture to support both simultaneously.

References

1. Forbes, https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/ - 681f0a3c6f63, last accessed 2017/07/25.