data.world

A Platform for Global-Scale Semantic Publishing

http://ow.ly/zwF030g2ryo

Bryon Jacob[0000-0003-0470-9300] bryon@data.world ,

Dave Griffith[0000-0001-9700-0012] dave.griffith@data.world ,

Triet Le[0000-0001-5619-5802] triet.le@data.world

WHAT

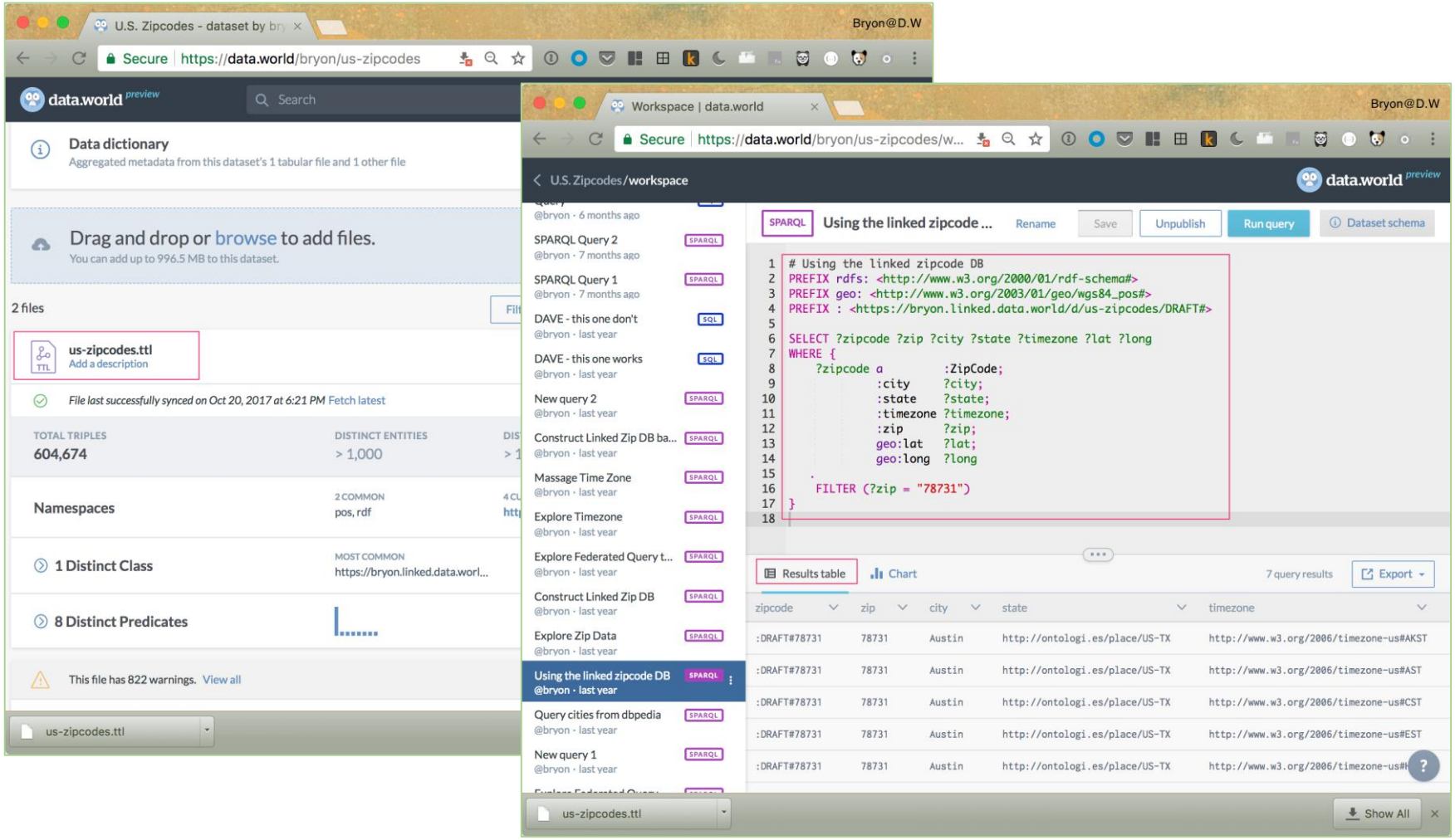
is data.world?

data.world

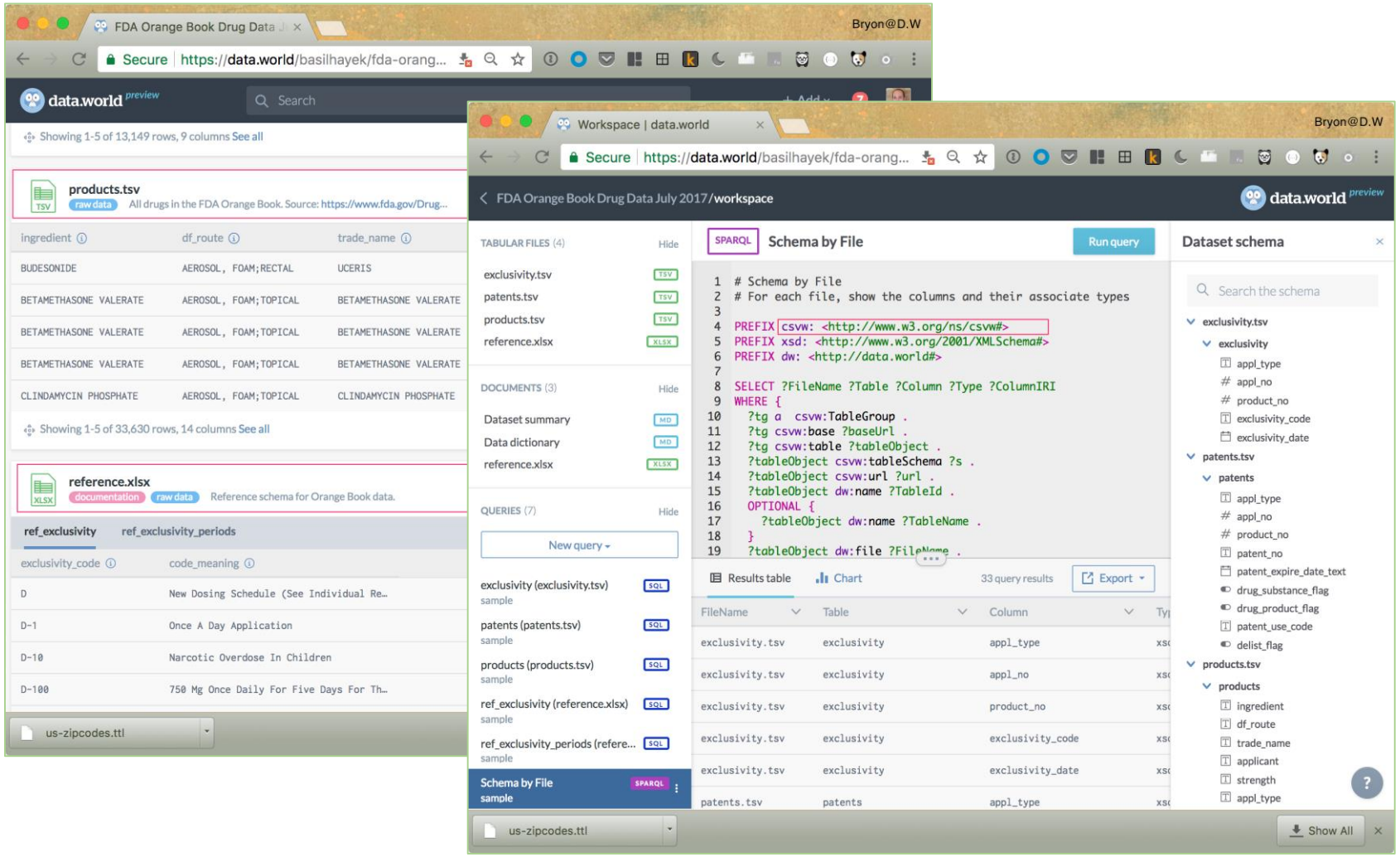
Is a collaborative web platform with

- a user base that are (primarily) not Semantic Web experts
- datasets that are (generally) not published as linked data (mostly tabular/JSON data)

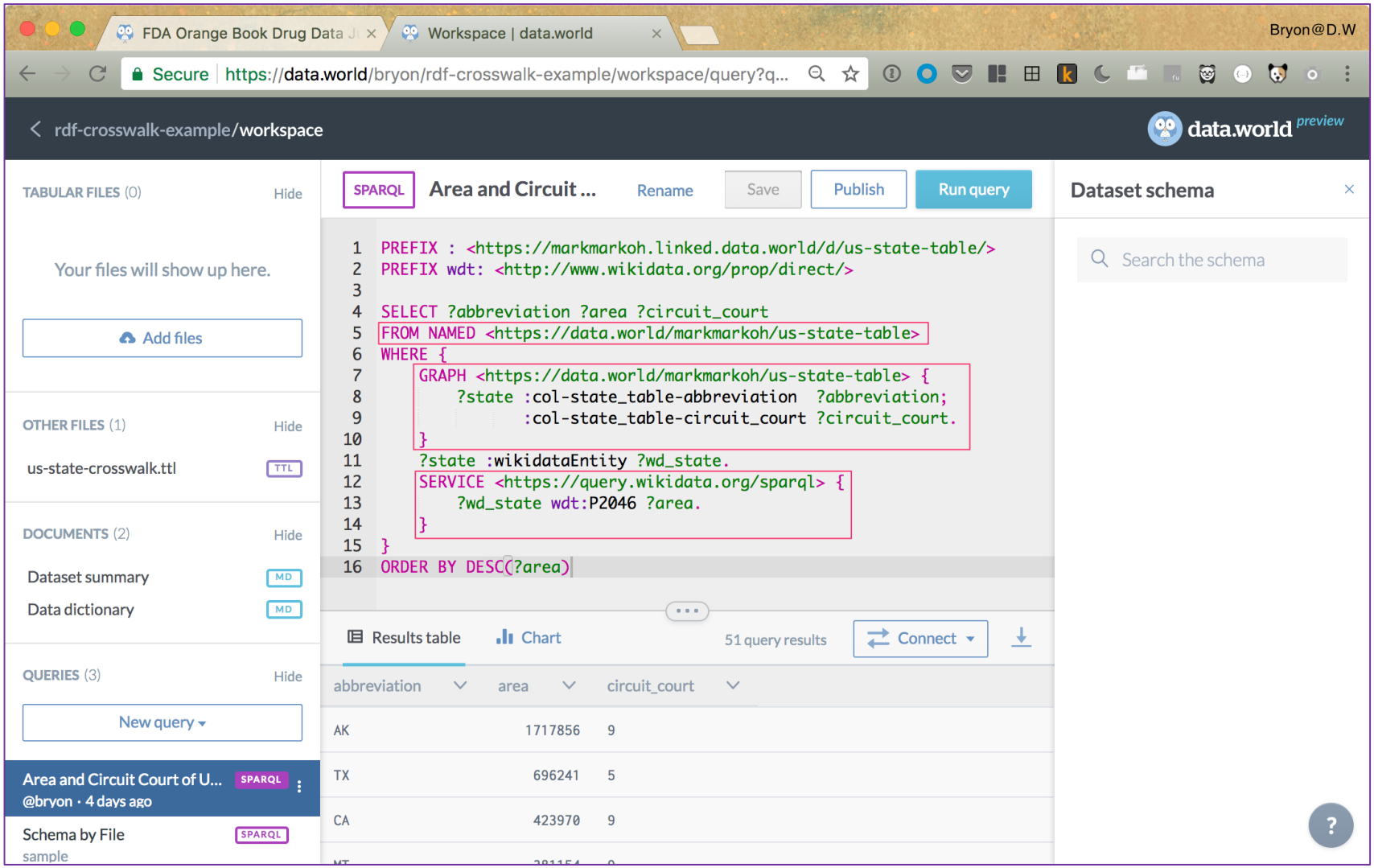
We use web standards to translate that data into RDF



RDF Formats are supported natively, and gets you a SPARQL endpoint to query the data



Tabular formats are also ingested – an RDF model of the data is built using CSVW.



- Each dataset in data.world is accessible as a *named graph* (provided you have access permissions)
- Remote SPARQL endpoints can be federated naturally via SERVICE patterns

WHY

does it exist?

We think more people should care about the Semantic Web.

The real question is *”Why don’t they?”*

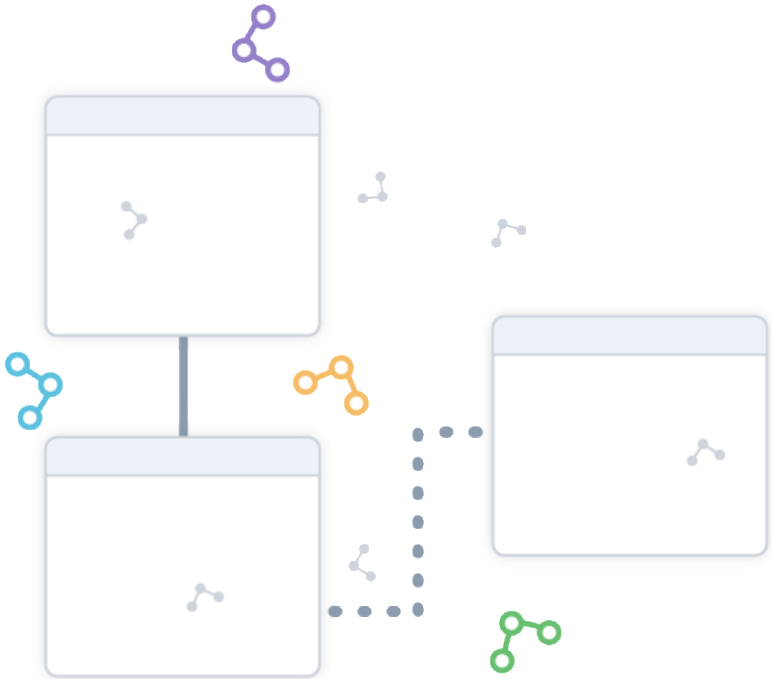
Semantic Web Products – Solutions in search of problems?

In order to make the Semantic Web appeal to mainstream data workers, we need products that focus on their needs.

*We need to find the real-world **problems** that are searching for a **solution** that the Semantic Web can provide.*

N% of Data Scientists' time is spent finding and prepping data.

Is an unsubstantiated claim that is thrown around a lot.



Nonetheless – it is a real problem, data workers do spend a lot of time in discovery and exploratory analysis of data – and the real crime is that work is done redundantly.

Semantic Web offers a powerful set of tools to deal with this:

- a universal structure for data
- an open-world model that adapts to heterogeneous data
- metadata can be added to the data
- SPARQL for federated query

The set of people working on data projects is diverse

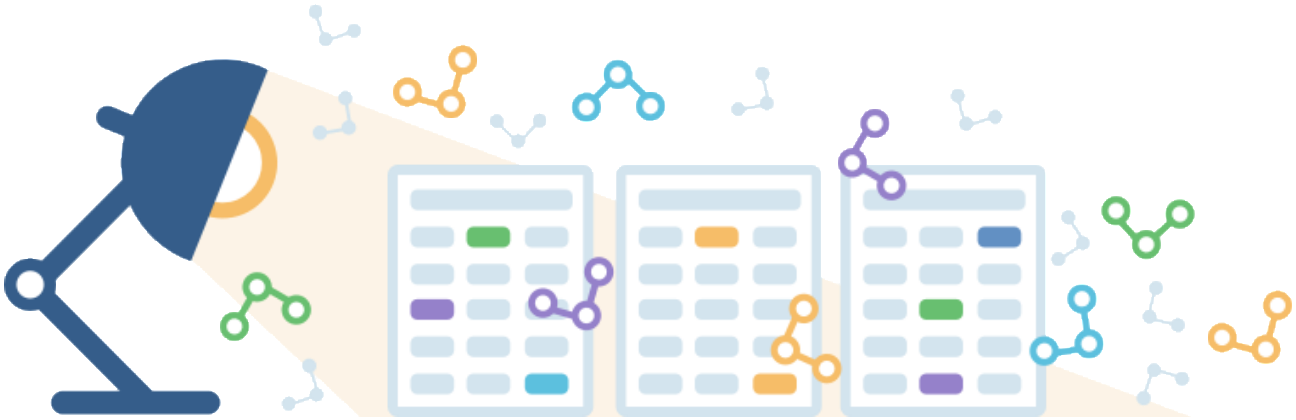
- knowledge engineers who work directly in the creation of ontologies and knowledge bases
- data scientists, machine learning experts, and statisticians who produce models, derivative data works, and visualizations
- analysts, scientists, and students who are accustomed to using spreadsheets or visually-driven analytics platforms
- stakeholders and end-users consuming the high-level conclusions of the work.

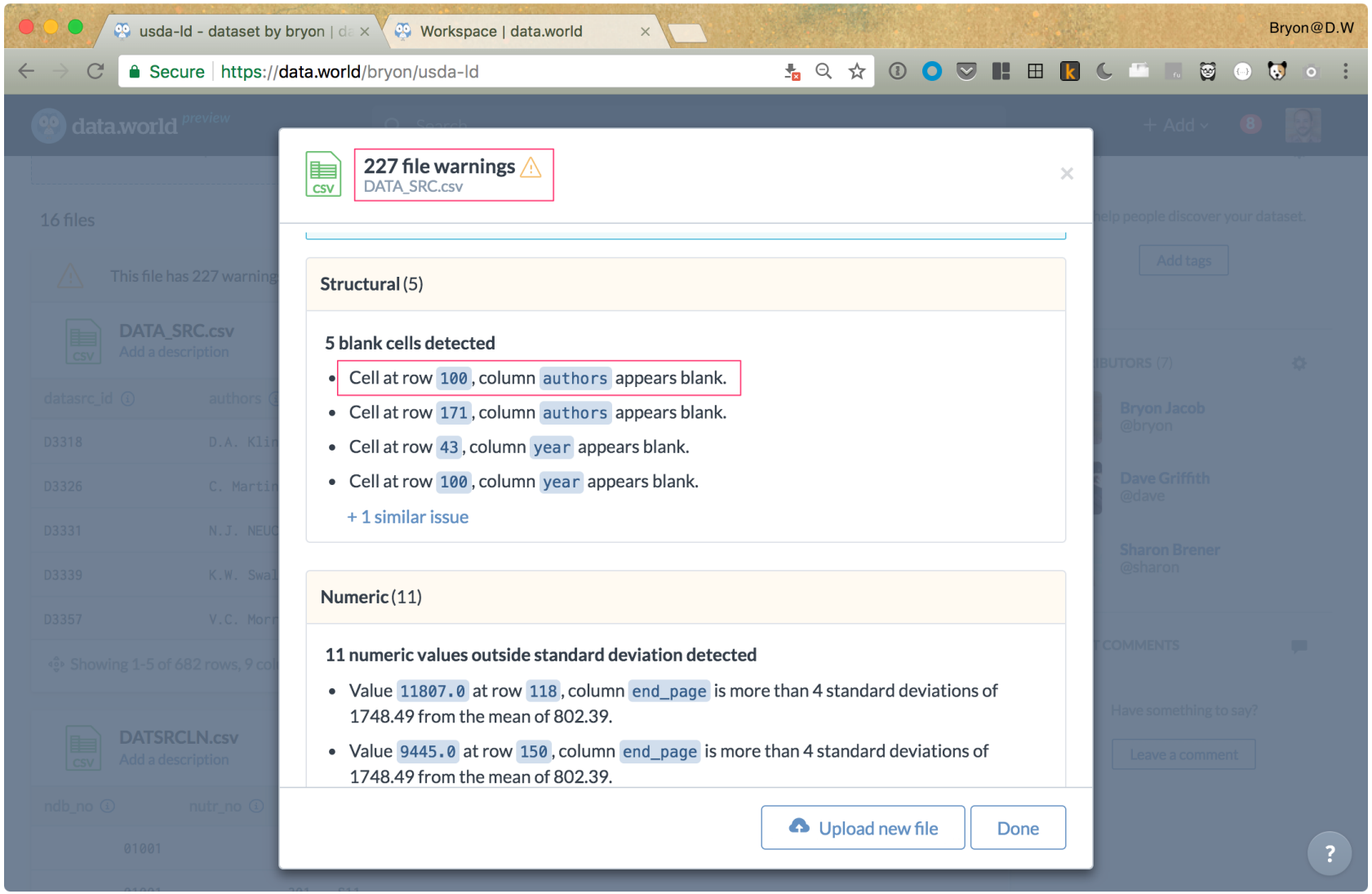
Each of these personas has a process and a tool chain they prefer – we do a lot of user-centered design work that focuses on the elements that support collaboration between them. The unifying factor is the standard data model.

Working on data is not a one-way conversation

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.

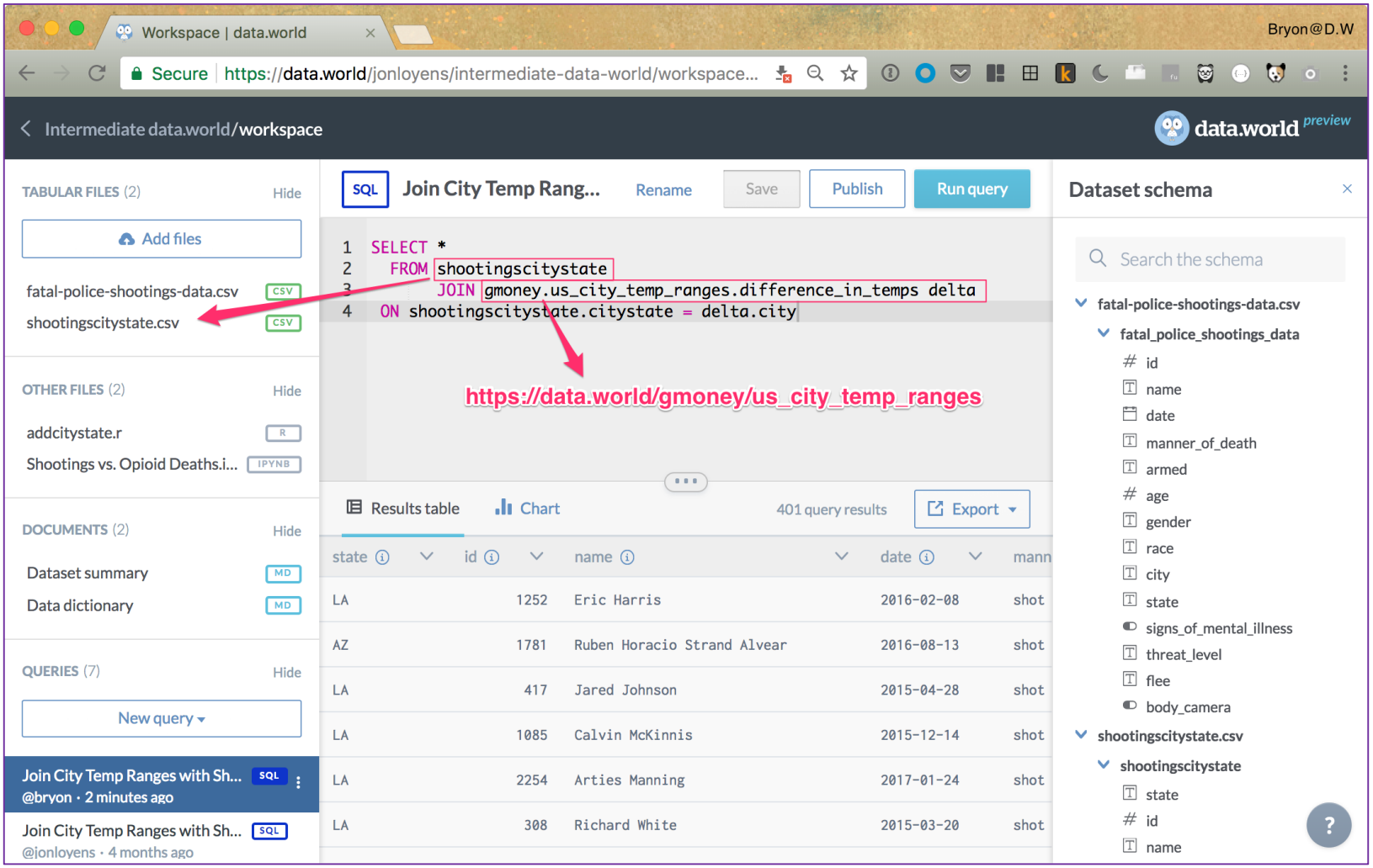
Data can be worked on completely in the open, completely privately, or in groups of collaborators with read/write ACL.





Data inspector powered by RDF & SPARQL

Most data is tabular, many more people know SQL than SPARQL

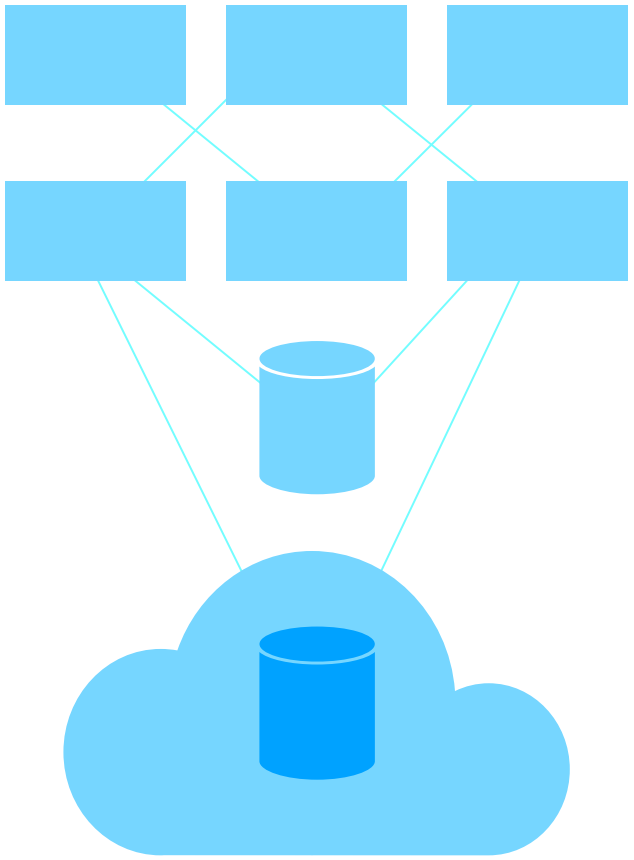


HOW

does it work?

Our query architecture prioritizes scale and query responsiveness over raw performance and update flexibility.

Updates as bulk ingest, with the output of the ingest pipeline an immutable HDT file.

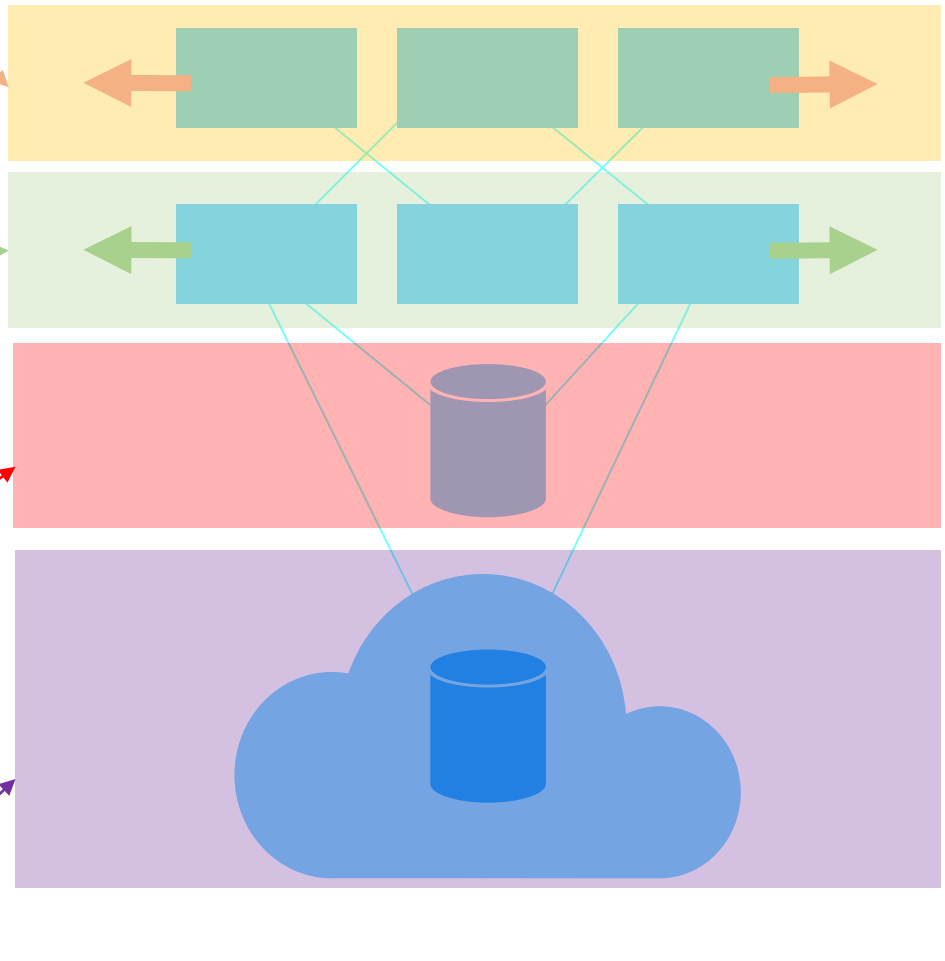


Front-end query servers parse user queries and route them (this is where SQL is transpiled)

Query heads load HDT data on-demand from deep storage and execute SPARQL

HDT and indices are cached in a fast storage layer, leveraging HDT's query-in-place

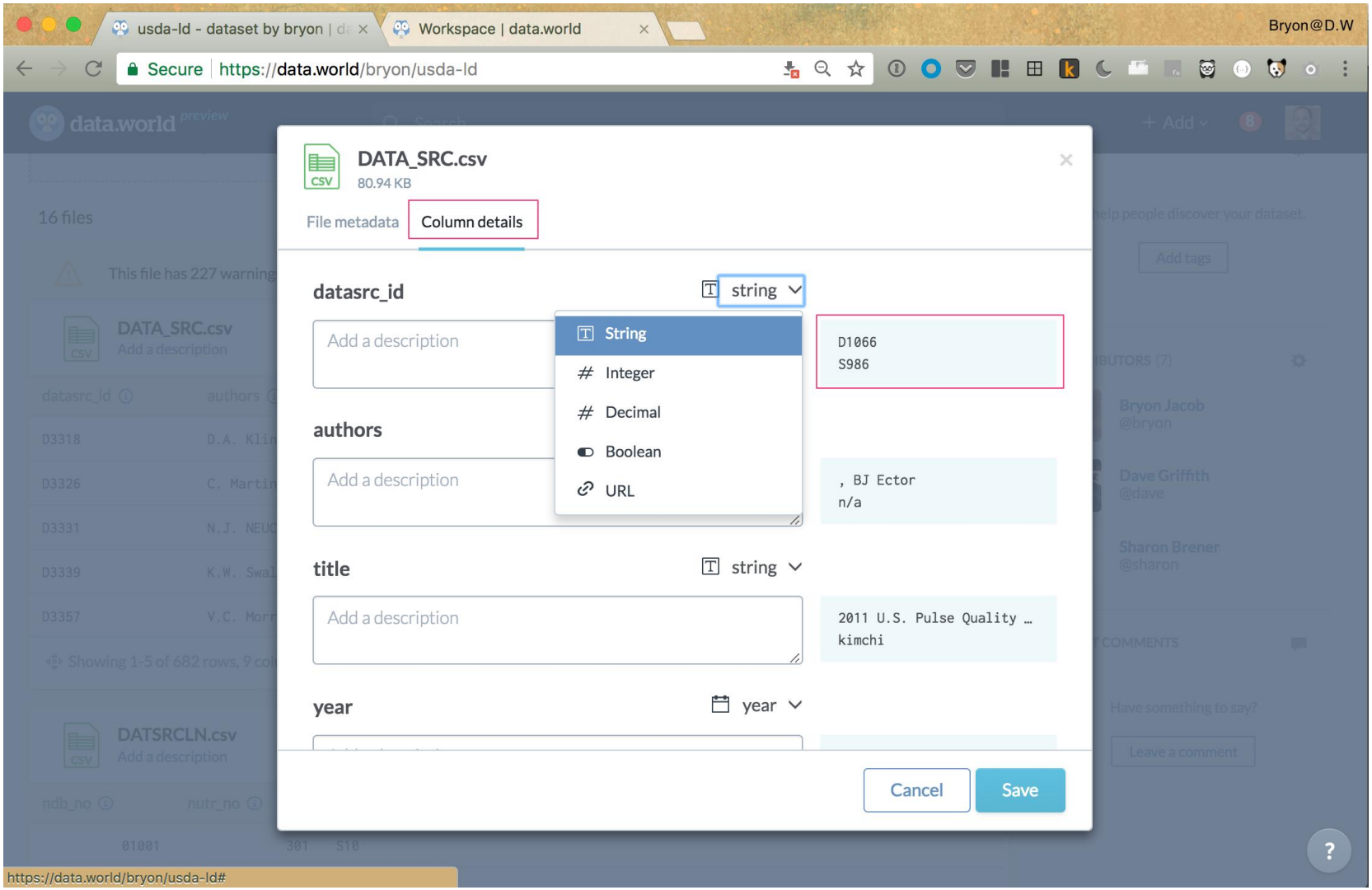
Data is stored at rest in inexpensive, scalable cloud storage



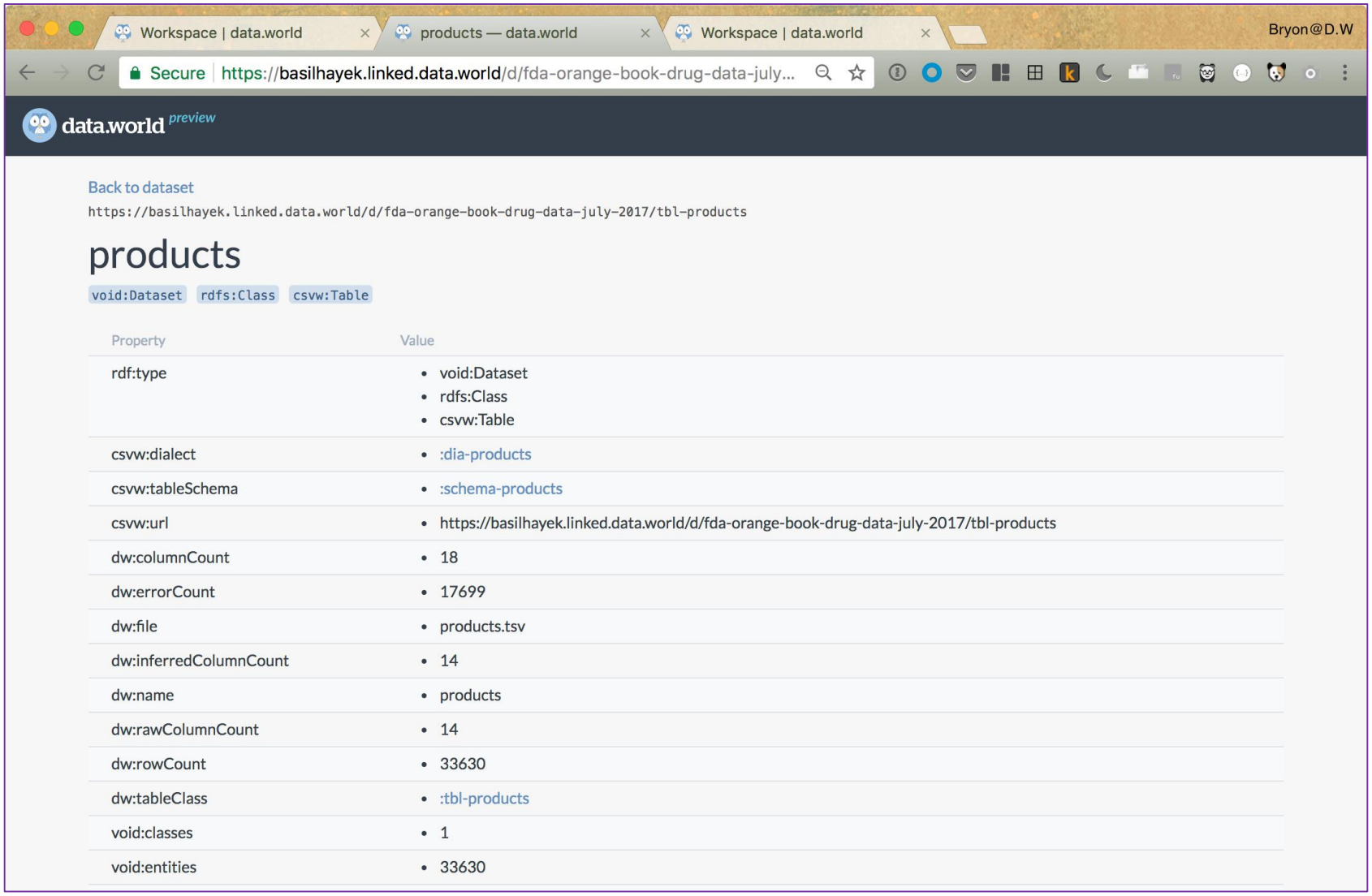
WHAT

are we working on now?

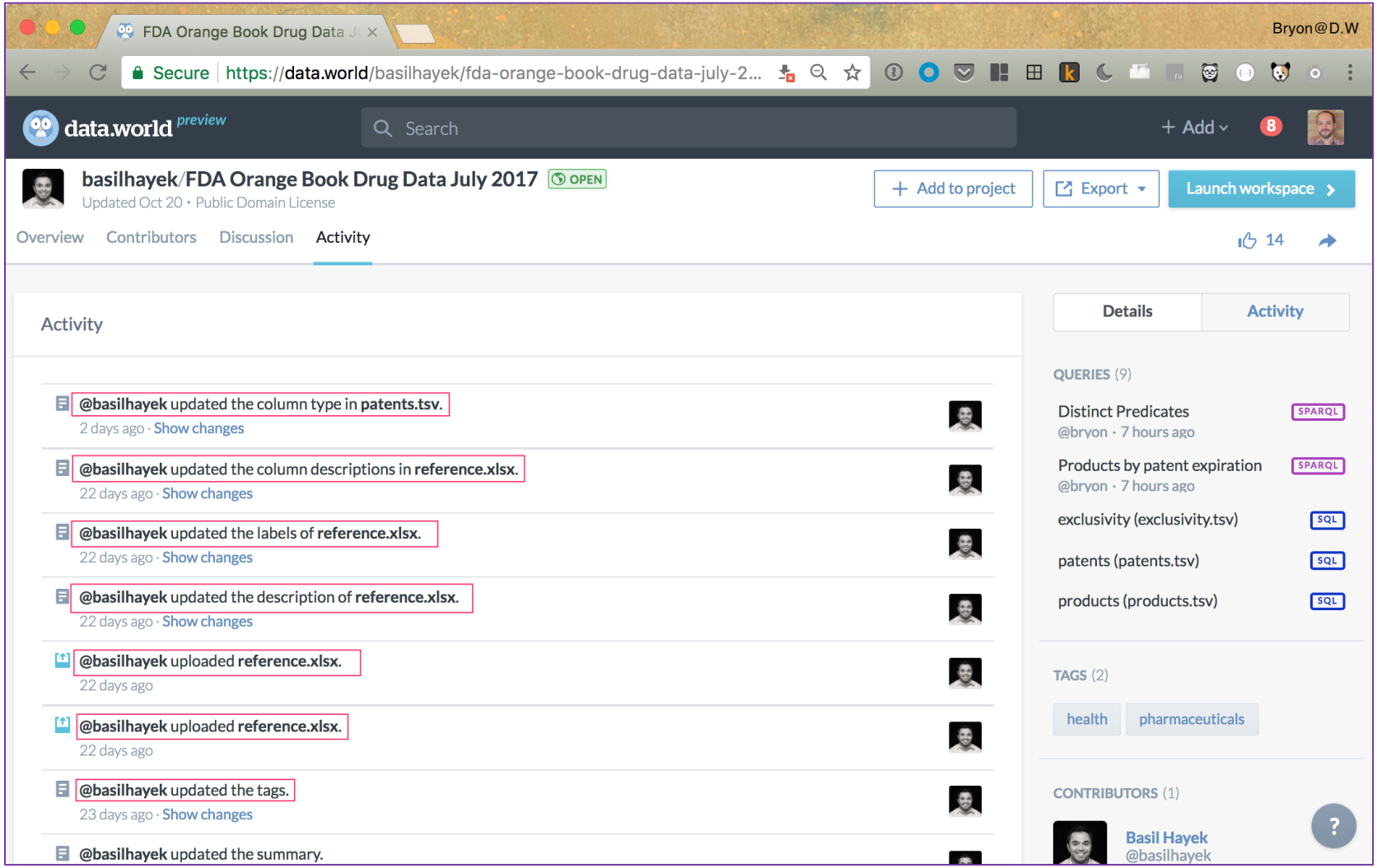
More tools to enrich data with schema, metadata, and reconciliation against standard taxonomies/entities…



Entity browser enhancements – HTML templating for browse interfaces…



More problems to help users solve in: dataset versioning, auditing, governance…



Query Architecture

Our HDT-based query architecture works well for scaling out many exploratory queries, but it’s not optimal for large analytical (non-selective) queries.

We’re pushing the limits of HDT here, and also contemplating a hybrid architecture where large analytical queries are run against columnar formats such as parquet files.

Our hypothesis…

To increase the size and connectivity of the web of Linked Data, we need to increase the connectivity of the network of people working with data.

By using Semantic Web technology to facilitate data work, we can leverage that work to grow the web of Linked Data.

Create an account:

https://data.world/semweb

Use it, and send me feedback:

https://data.world/bryon

bryon@data.world

Thank You!!