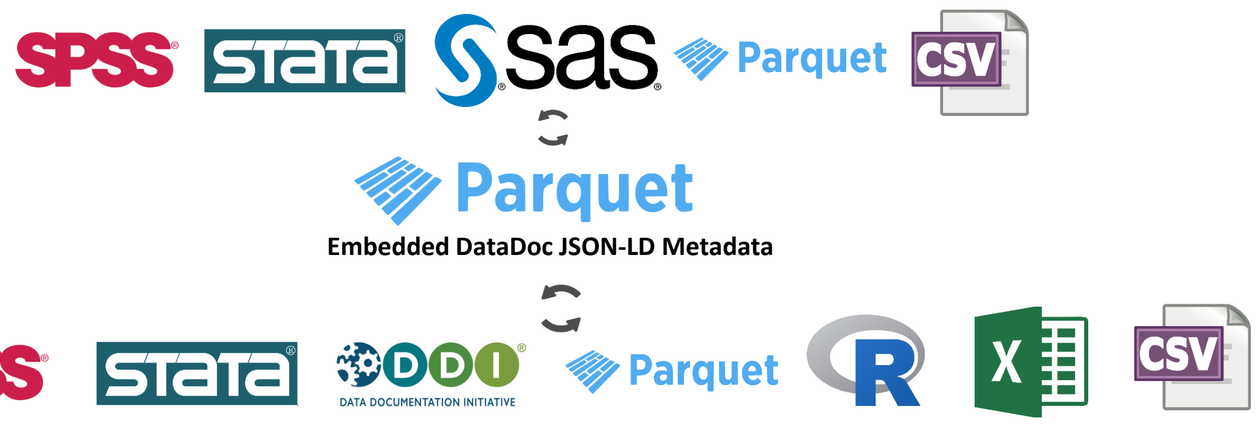
Integrating Parquet and JSON-LD for Embedded Smart Metadata in Statistical Data Storage

Dan Smith *Colectica, Minneapolis, MN, United States*

Traditional statistical data file formats from SAS, SPSS, and Stata have traditionally embedded structural metadata to detail additional variable information, code lists, missing values, and other types of annotations. While useful within their respective statistical packages, these formats suffer from limitations in terms of interoperability, transparency, online processing efficiency, metadata connections, open dissemination, and long-term archival suitability. This project introduces a novel **data** and **doc**umentation methodology, **DataDoc**, which combines the open Parquet file format with embedded RDF metadata that seeks to address the limitations of proprietary statistical file formats, while also offering efficiency, archival, and FAIR benefits. This work is motivated by the pressing need to overcome the deficiencies of the traditional formats, aligning with the evolving landscape of modern data analysis tools and FAIR sharing guidelines.



Smartly converting between file formats

Apache Parquet is an open source columnar file format

- Very fast
- Ubiquitous support in cloud-scale data query engines tools such as Apache DataFusion, Google BigQuery, Azure Synapse, Amazon Athena or DuckDB while utilizing inexpensive block storage
- Supports embedded metadata
- Built in support for compression and missing (empty) values
- DataDoc uses additional flag columns for tagged and defined missing values during conversion
- DataDoc adds additional metadata present in statistics data but not natively supported in parquet.

Statistical Data File Feature Inventory

Feature	Parquet	SPSS	Stata	SAS
Native Data Types				
Boolean	Yes			
Int32	Yes			
Int64	Yes			
Int96	Yes			
Float	Yes		Yes	
Double	Yes	Yes	Yes	Yes
Byte Array (or Char Array)	Yes	Yes	Yes	Yes
Logical Data Types				
String	Yes	Yes	Yes	Yes
Enum	Yes			
UUID	Yes			
Signed Integer 8	Yes		Yes	
Signed Integer 16	Yes		Yes	
Signed Integer 32	Yes	Yes		Yes
Signed Integer 64	Yes		Yes	
Unsigned Integer 8	Yes			
Unsigned Integer 16	Yes			
Unsigned Integer 32	Yes	Yes		Yes
Unsigned Integer 64	Yes			
Decimal	Yes			Yes
Date	Yes	Yes		Yes
Time (Milliseconds)	Yes	Yes		Yes
Time (Microseconds, Nanoseconds)	Yes			
DateTime (Milliseconds)	Yes	Yes		Yes
DateTime (Microseconds, Nanoseconds)	Yes			
Interval of Time (Milliseconds)	Yes	Yes		Yes
Interval of Time (Microseconds, Nanoseconds)	Yes			
Textual Metadata				
Dataset Title		Yes	Yes	Yes
Dataset Title (Multilingual)			Yes	
Dataset Description		Yes		
Dataset Timestamp		Yes	Yes	
Variable Label		Yes	Yes	Yes
Variable Label (Multilingual)			Yes	
CodeList Value Labels		Yes	Yes	Yes
CodeList Value Labels (Multilingual)			Yes	
CodeList Value Labels with ranges				Yes
Missing Values				
System missing	Yes	Yes	Yes	Yes
Predefined range		Yes	Yes	
User defined range		Yes		
User defined range plus another number		Yes		
One, two, or three specific missing values		Yes		
Tagged missing				Yes

Motivation

The motivation for this project arises from the need to **support a superset of features** provided by the traditional statistical data file formats of SAS, SPSS, and Stata. While these software formats have traditionally been the backbone of data analysis and documentation, they suffer from significant limitations, including poor interoperability across applications, limited transparency, inefficiency in on-demand processing, limited metadata options, and inadequate support for open dissemination and long-term archival. The current landscape of data analysis tools necessitates a more adaptable, efficient, and transparent storage format that can seamlessly integrate with modern platforms and support FAIR sharing guidelines. The research team desired to develop a unified methodology for the **storage**, **real-time analysis**, **publishing**, and **archival** of statistical data that retains **a superset of all the features** inherent to each of the proprietary formats. Simultaneously, it sought to introduce additional capabilities related to identification, metadata management and openness. There were three main initial goals, compile a comprehensive inventory of all features supported by the three proprietary data formats, select a storage format for the primary data, and identify an optimal format for preserving metadata information.

Vocabulary and JSON-LD

The DataDoc vocabulary consists of several parts in addition to the internal Parquet schema: a tabular dataset and column description, missing value definitions, code list value labels, and additional documentation. JSON-LD was chosen for the metadata storage since it **allows non-RDF aware applications** to make use of the additional structural metadata. A specific **JSON-LD framing** is provided by the vocabulary to ensure that the Json representation is consistent across all DataDoc usages and **provides developers with straightforward access** to the additional definitions. DataDoc supports **the superset of all missing value schemes** that are used with the proprietary statistical file formats. In the case where a dataset column can contain both values and the system missing value, Parquet can denote in its self-describing schema that a column can contain null values. When additional data values are designated to represent missing values, separate flag columns are created within the dataset. The flag columns are marked nullable and exclusively contain data values that signify a missing values.

Incorporating multiple metadata standards

In addition to its own definitions, pieces of several other well-known ontologies or systems are incorporated into DataDoc.

- The tabular dataset and column descriptions make use of a subset of the **W3C tabular metadata** standard.
- Code lists for values and missing values are represented by a portion of the **SKOS** vocabulary with several extensions for ranged values.
- The **Dublin Core** vocabulary is used for descriptive citation information.
- Several terms from the **DDI Lifecycle** metadata standard
- Several terms from Apache Iceberg's format metadata
- UNF data fingerprints

Find out more about Parquet with DataDoc

- **https://github.com/datadocumentation**
- **https://datadocumentation.org**

DataDoc Tools

- **Colectica Datasets**
- **Python pandas integration**