The Journal of Open Source Software

Carnival: JVM Property graph data unification toolkit

David Birtwell [2], **Heather Williams** [1], **Tom Hutchinson** [1], **Louis Lee** [2], **Hayden Freedman**[3], **and Christian Stoeckert** [1]

1 University of Pennsylvania, Institute for Biomedical Informatics **2** USC Norris Cancer Center, USA **3** University of California Irvine, Department of Informatics

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Mark A. Jensen

Reviewers:

- @kinow
- @KonradHoeffner

Submitted: 07 October 2022
Published: unpublished

Summary

Research activities in data rich areas such as biomedical informatics face many data challenges including harmonizing complex and disparate data, integrating existing knowledge bases into data sets for manual or machine learning analysis, and reproducibility of results. Graphs are a powerful data structure for naturally describing complex data. Information about data provenance can be embedded in the graph itself to aid in quality control and reproducibility. **Carnival** is a semantically driven informatics toolkit that enables the aggregation of data from disparate sources into a unified property graph and provides mechanisms to model and interact with the graph in well-defined ways inspired by the Open Biological and Biomedical Ontology (OBO) Foundry ontologies.

Statement of need

Loading, cleansing, and organizing data can dominate the time spent on a data science project (Press, 2016). This phenomenon is exacerbated in human subjects research at an academic medical institution where data are very complex, reside in disparate repositories with varying levels of accessibility, are coded by separate yet overlapping coding systems, frequently rely on manual data entry, and change over time. Data provenance and reproducibility of results are important factors in human subjects research. It is no easy task to implement a robust consistent data pipeline with clear data provenance that can be rerun when source data change. While there are several mature libraries and toolkits that enable visualization and statistical computation once the analytical data set is generated, there are comparatively fewer data preparation tools.

Existing extract, transform and load (ETL) technologies such as Microsoft SQL Server Integration Services help with data staging. Similarly, data manipulation tools like pandas facilitate transformation of series and matrix data. **Carnival** distinguishes itself by offering a lightweight data caching mechanism coupled with data manipulation services built on a property graph rather than arrays and data frames. Graphs present an alternative to relational data structures that more naturally represent complex and highly relational data and are more adaptive to change. A property graph database is an implementation of the graph structure that represents data as nodes and directed edges (relationships) between the nodes, where nodes and edges can have properties (key/value pairs) associated with them. Carnival's combination of features and graph data representation empowers informaticians and programmers working in complex data domains to build pipelines, utilities, and applications that are comparatively richer in semantics and provenance.

Knowledge bases in Resource Description Framework (RDF) triplestores can be valuable tools to harmonize and enrich complex data. Transforming source relational data into RDF triples reflecting a data model is challenging. While there exist relational-to-RDF mappers such as

Karma (Gupta et al., 2015), the configuration process is labor intensive and the resulting triples may not match a data model particularly one of sufficient complexity.

Carnival was developed to create domain-specific property graph data models, and provide tools to create robust pipelines to import and manage data in that model. There are two main components to Carnival. The primary component is a layer built on top of Apache Tinkerpop that seeks to provide more standardized and semantically driven methods of interacting with a property graph. An additional component is a data caching mechanism that supports the efficient aggregation of data from disparate sources.

Key Features

- a graph modeling framework that ensures graph data remain consistent
- a data caching mechanism to ease the computational burden of data aggregation during the development process and promotes data provenance
- a lightweight graph algorithm framework that facilitates the creation of graph building components with automated provenance tracking

Uses

Production of analytical data sets

Carnival was initially developed to facilitate the production of analytical data sets for human subjects research. The source data repositories included a relational data warehouse accessible by SQL, a REDCap (Harris et al., 2009, 2019) installation accessible by API, and manually curated data files in CSV format. Data pertaining to the set of study subjects was distributed across each of these data sources. Using Carnival, a data pipeline was implemented to pull data from the data sources, instantiate them in a property graph, clean and harmonize them, and produce analytical data sets at required intervals.

Queries over enriched data

A key challenge of human subjects research is to locate patients to recruit to a study, frequently done by searching a research data set containing raw patient data. Potential recruits need to be stratified by attributes, such as age, race, and ethnicity, matched against inclusion criteria, such as the presence of a diagnosis code, and filtered by exclusion criteria, such as a treatment modality. **Carnival** has been used effectively in this area by loading the relevant raw data into a graph, stratifying and categorizing patients by the relevant criteria, then using graph traversals to extract the patients who are potential recruits (David Birtwell, 2019; Freedman et al., 2020).

Integration with OBO Foundry Ontologies

We drew upon ontology modeling in the OBO Foundry as inspiration for the Carnival graph data model. For example, a 'process', is an event that occurs at some time on some material entity. A 'planned process' extends 'process' to include a pre-defined plan, participants, inputs, and outputs. In the Carnival graph, healthcare encounters are modeled as planned processes, where participants include the patient and clinician and the outputs may be diagnoses and medications.

System integrations

Carnival's ability to integrate data from disparate resources into a flexible computational resource enables data driven system integrations. For example, **Carnival** has been used effectively to integrate a custom help desk ticketing system with Monday.com. The help desk ticketing system was developed locally with a back-end relational database. Monday.com is accessible via its API for reads and writes. By modeling the help desk and Monday.com data

as separate graph models, then using a third graph model to integrate the two, a **Carnival** integration application was developed to integrate the two data sets and compute changes that needed to occur based on the state of the data. In this example, **Carnival** was partnered with Micronaut and deployed as a Docker container on Microsoft Azure. The service would build an in-memory graph at regular minute intervals, compute the changes that were required, then call the appropriate web services to execute the logic of the integration.

Additionally, an example repository carnival-demo-biomedical has been provided that demonstrates the integration of Carnival with Micronaut and Docker. This example project models and integrates synthetic electronic health record data from a SQL database and CSV files, reasons over the data to create case and control patient cohorts, and presents an API endpoint to the user to explore the data.

References

David Birtwell, R. P., Heather Williams. (2019). Carnival: A Graph-Based Data Integration and Query Tool to Support Patient Cohort Generation for Clinical Research. *IOS Press Ebooks*, *264*, 35–39. https://doi.org/10.3233/SHTI190178

Freedman, H. G., Williams, H., Miller, M. A., Birtwell, D., Mowery, D. L., & Stoeckert, C. J. (2020). A novel tool for standardizing clinical data in a semantically rich model. *Journal of Biomedical Informatics*, *112*, 100086. https://doi.org/10.1016/j.yjbinx.2020.100086

Gupta, S., Szekely, P., Knoblock, C. A., Goel, A., Taheriyan, M., & Muslea, M. (2015). Karma: A system for mapping structured sources into the semantic web. In E. Simperl, B. Norton, D. Mladenic, E. Della Valle, I. Fundulaki, A. Passant, & R. Troncy (Eds.), *The semantic web: ESWC 2012 satellite events* (pp. 430–434). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-46641-4_40

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., & Duda, S. N. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, *95*, 103208. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Press, G. (2016). Cleaning big data: Most time-consuming, least enjoyable data science task, survey says. *Forbes: ENTERPRISE & CLOUD*. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/