eBISS 2023

State-of-the-Art of Interoperability of Big Data in Healthcare:
Exploring Current Approaches and Advancements

Adina Faye P. Bondoc | Laura Isabella Forero Camacho

adina.faye.bondoc@estudiantat.upc.edu | laura.forero@estudiantat.upc.edu

1. INTRODUCTION

Data integration (DI) is the process of combining and harmonizing data from **multiple heterogeneous sources** to promote its exploitation and facilitate its transparent management. This study explores interoperability solutions with a specific focus on its applications in **healthcare**.

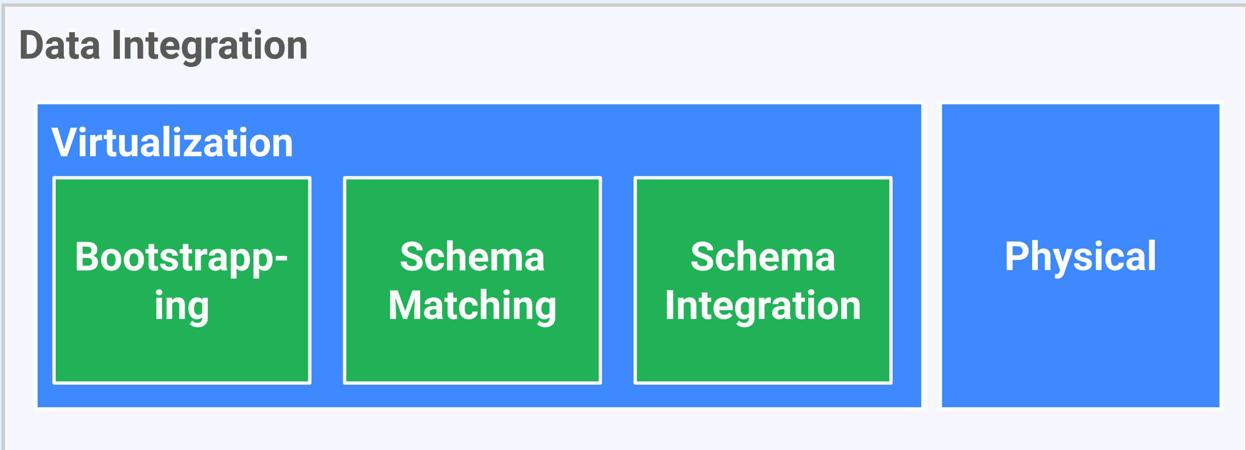

Figure 1. Data Integration System Components

Challenges in healthcare interoperability such as poor data quality, data protection, and compatibility issues will be further explored for each tool presented. Addressing these challenges is vital for maximizing the value of big data.

2. STANDARDS

FAIR DATA PRINCIPLES

(Findable, Accessible, Interoperable, Reusable) framework to improve the infrastructure & reuse of scholarly data

EUROPEAN INTEROPERABILITY FRAMEWORK

guidelines and principles to ensure interoperability of data within the EU

FAST HEALTHCARE INTEROPERABILITY RESOURCES (FHIR)

standardized approach for sharing electronic health records & other health-related data

3. SOLUTIONS

1999

CLIO

a semi-automatic tool developed by the IBM Almaden Research Center facilitating **data integration and exchange**. It was the first to use mappings to exploit relationships between heterogeneous schemas and facilitate data exchange.

2015

OPTIQUE

an advanced virtual data integration system of various data sources, with an **easy-to-use query interface** and leveraging ontology-based data access (**OBDA**) for efficient integration.

2019

ODIN

an automated system developed by the DTIM-UPC group, designed to incrementally integrate diverse data sources into **dataspaces**, and user feedback to generate integrated results and providing **user-friendly querying** mechanisms.

2020

FAIR4HEALTH DATA CURATION TOOL

acts as an Extract-Transform-Load (ETL) tool based on HL7 **FHIR** resources and aligns with the **FAIRification process**.

2020

SQUERALL

Enables virtual access allowing users to query **diverse big data in real-time** without data transformation and efficient distributed query processing, establishing the concept of a **Semantic Data Lake**.

4. ANALYSIS

DI system architectures were compared based on their components:

DI System	Supported Data Sources	Bootstrapping	Schema matching	Schema integration	Query
Clio	Relational and nested schemas	Input	Input	Simple merge, Incremental	Yes
Optique	Relational, Sensor, Stream	Provided and extracted	Input	Simple merge	Yes
ODIN	Structured (e.g., relational) and semi-structured (e.g., JSON)	Provided and extracted	Enhanced LogMap, NextiaDI, Manual	Simple merge, Incremental	Yes
FAIR4Health Data Curation Tool	Relational databases, Files, semi-structured and structured medical data sources	Provided and extracted	Input	Simple merge	-
Squerall	No relational, and relational data sources (Databases, raw files, and structured data)	Input	Input	Simple merge	Yes

Table 1. DI Systems Architecture Comparison

Each system was then assessed based on their ability to address each of the interoperability challenges [8], particularly for healthcare:

	Clio	Optique	ODIN	DCT	Squerall
Enhance data quality	✓	✓ Heuristics - Schema quality	✓ Compliance checks	✓ Data validation phase	✓
Data protection				✓ De-identification & pseudonymization*	
Decoupling between data producers & users	✓	✓ User interface	✓ User interface	✓	✓
Schema-level and data-level conflicts			✓ Matching: LogMap and NextiaJD		
Demand for near real-time analytics		✓ Streaming			✓ Data Lake
Intersystem interfaces & compatibility	✓ Relational	✓ Sensors, Streaming	✓ Structured and semi-structured	✓ Medical domain	✓ Relational and nonrelational db
Scope and scalability	✓ Incremental	✓	✓ Incremental		✓

Table 2. Challenges addressed by each DI System

5. CONCLUSION

Existing DI systems exhibit strengths in improving data quality, bridging the gap between data producers and users, and resolving schema and data conflicts. In particular, **ODIN** and **Squerall** address the most challenges.

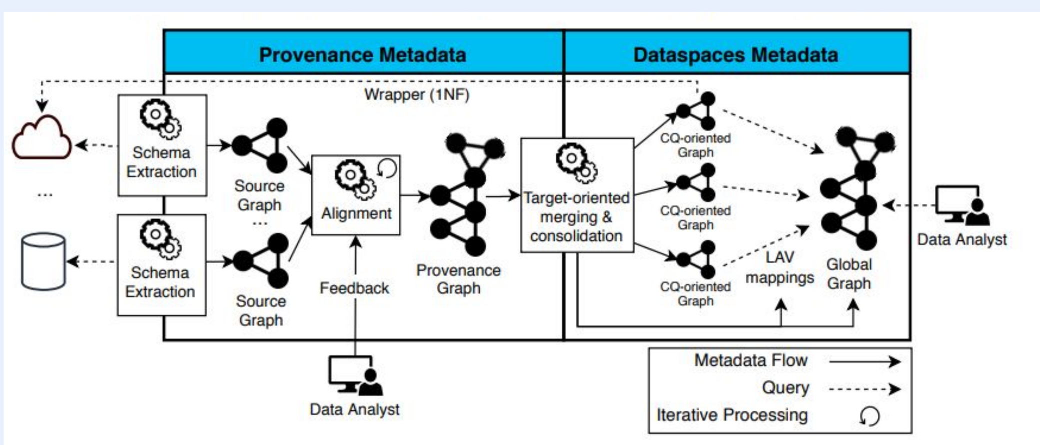

Figure 2. High-level Architecture of Odin [7]

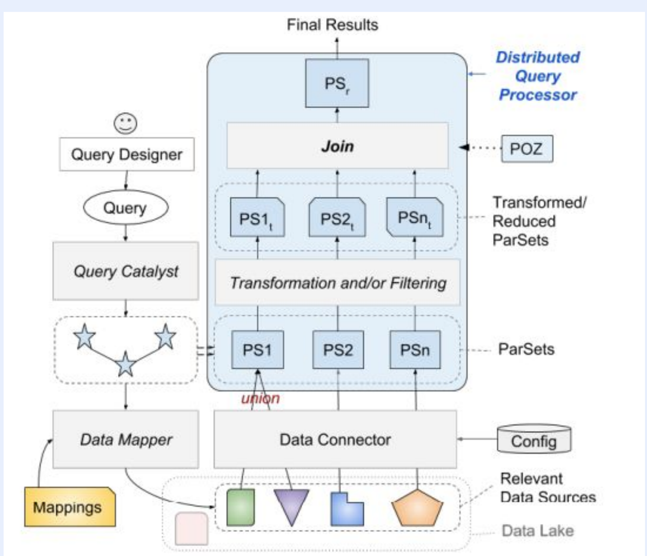

Figure 3. Semantic Data Lake Architecture [6]

However, there are areas that require further attention, such as **data protection and addressing near real-time requirements.** By addressing these challenges, the healthcare industry can achieve improved interoperability for **better patient care and decision making.**

REFERENCES

1. Calvanese, D., Giese, M., Haase, P., Horrocks, I., Hubauer, T., … Zheleznyakov, D. (2013). Optique: OBDA Solution for Big Data. In Advanced Information Systems Engineering (pp. 293–295). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41242-4_48
2. European Commission. Directorate General for Informatics. (2017). New European interoperability framework: promoting seamless services and data flows for European public administrations. Publications Office. https://doi.org/10.2799/78681
3. Fagin, R., Haas, L. M., Hernández, M., Miller, R. J., Popa, L., & Velegrakis, Y. (2009). Clio: Schema Mapping Creation and Data Exchange. In Conceptual Modeling: Foundations and Applications (pp. 198–236). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-02463-4_12
4. HL7. (n.d.). HL7 FHIR Summary. HL7.org. Retrieved June 23, 2023, from https://www.hl7.org/fhir/summary.html
5. Javier Flores, Kashif Rabbani, Sergi Nadal, Cristina Gómez, Oscar Romero, Emmanuel Jamin, Stamatia Dasiopoulou: Incremental Schema Integration for Data Wrangling via Knowledge Graphs. Semantic Web Journal
6. Mami, M.N., Graux, D., Scerri, S., Jabeen, H., Auer, S., & Lehmann, J. (2019). Squerall: Virtual Ontology-Based Access to Heterogeneous and Large Data Sources. International Workshop on the Semantic Web.
7. Nadal, S., Rabbani, K., Romero, O., & Tadesse, S. (2019). ODIN: A Dataspace Management System. International Workshop on the Semantic Web.
8. Scheerlinck, J., Van Eeghem, F., Loutas, N. (2018) Big Data Interoperability Analysis. https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/document/big-data-interoperability-analysis
9. Sinaci, A. A., Núñez-Benjumea, F. J., Gencturk, M., Jauer, M.-L., Deserno, T., Chronaki, C., Cangioli, G., Cavero-Barca, C., Rodríguez-Pérez, J. M., Pérez-Pérez, M. M., Laleci Erturkmen, G. B., Hernández-Pérez, T., Méndez-Rodríguez, E., & Parra-Calderón, C. L. (2020). From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. In Methods of Information in Medicine (Vol. 59, Issue S 01, pp. e21–e32). Georg Thieme Verlag KG. https://doi.org/10.1055/s-0040-1713684
10. Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In Scientific Data (Vol. 3, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/sdata.2016.18