CRITICAL PATH INSTITUTE

jscdm-130-alarcon

ORIGINAL RESEARCH

Leveraging Biolink as a FAIR "Rosetta Stone" Between Clinical Semantic Models Provides Emergent Interoperability

Pablo Alarcón-Moreno[1], Ian Braun[2], Emily Hartley[2], Daniel Olson[2], Ninguama Benin[2], Ronald Cornet[3], Mark D. Wilkinson[1] and Ramona L. Walls[2]

Interoperability between clinical datasets is challenging due to, in part, the number of data models and vocabularies in use and the variety of implementations. Here we describe the first steps in an ongoing effort to achieve interoperability between two clinical datasets currently being constructed within independent international projects. Both are utilizing the FAIR Principles but have constructed their data models independently and have selected different ontologies. In this initial exploratory experiment, we examined the degree to which a mapping of both models into an independent schema, Biolink, can increase interoperability. Mapping was achieved by categorizing the key nodes in both data models as "types" of concepts in the Biolink schema. We found that with this very thin mapping in place, and without changing either model, queries could be constructed that extracted data from both datasets, demonstrating that at least some degree of interoperability had been achieved. Our results support the use of FAIR-compliant data representations, which are, by nature, more interoperable than legacy clinical data representations, even when the models have not been coordinated upfront.

Keywords: Common Data Element (C19984); Interoperability (C142381); Data Integration; FAIR; SDTM

Introduction

The achievement of personalized and precision medicine demands not only massive-multi-modal analysis of medical records, but also detailed international healthcare data to build accurate predictive models properly stratified to different sub-populations. Data integration is particularly crucial for rare diseases, where data are scarce, extremely heterogeneous, and disseminated in many repositories around the globe. In this study, we tested how well the use of a high-level semantic model, Biolink,[1] was able to support integration of data from two independent sources: Critical Path Institute's (C-Path) Rare Disease Cures Accelerator Data and Analytics Platform (RDCA-DAP) and European Joint Programme on Rare Diseases (EJP RD)– whose both sources used Findable, Accessible, Interoperable, and Reusable (FAIR) Data Principles to represent their data, but each used an independently developed data model.[2]

Background

Healthcare data, like most scientific data, is spread over many formats and repositories. This alone makes them challenging to integrate, but the data also have features, such as being highly privacy-sensitive, that dramatically inhibit integration. Within life and health sciences, interoperability projects that focused on machine actionability began in the late 1990s, and some came to fruition in the early 2000s. Some approaches pushed the interoperability problem on to the data owner, trying to enforce harmonization at source, such as caBIG and DiXA.[3–5] caBIG created an interface standard for the cancer genetics/genomics community that all participating organizations should implement, whereas DiXA, for the biodiversity community, pursued a query language that all sites should respond to. With respect to more general-purpose approaches, myGrid and BioMoby were both Web-services-based interoperability projects that used semantic annotations of Web-service interfaces,

[1] Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, and Centro de Biotecnología y Genómica de Plantas (CBGP, UPM-INIA), Universidad Politécnica de Madrid (UPM) – Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), ES

[2] Data Collaboration Center, Critical Path Institute, 1730 E. River Rd., Tucson, AZ 85718, US

[3] Department of Medical Informatics, Amsterdam Public Health Research Institute, Amsterdam University Medical Centers, University of Amsterdam, Meibergdreef 9, Amsterdam, NL

Corresponding authors: Pablo Alarcón Moreno (palmoreno@inia.es); Ian Braun (ibraun@c-path.org)