The Semantic Data Lake in Healthcare

We are entering the age of precision medicine and learning healthcare systems. These principles depend on robust evidence generation frameworks that synthesize genetic, clinical, behavioral and non-traditional data (e.g. patient generated data, environmental, socio-economic, and behavioral data) and knowledge bases and open information from public domain (linked open data). Frameworks that account for reproducibility of research findings, empower patients through engaging them with their own health information and analytic use of their data, provide precise contextual information about meaning and quality of the underlying data, and support all aspects of regulatory compliance while enabling broader collaborations through data and information sharing. This will require informatics platforms that go beyond depositing data in repositories or data warehouses provided by traditional big-data platforms.

We introduce a semantic approach to underpin evidence generation in healthcare. The Semantic Data Lake extends state of the art in big-data management to account for large scale integration of data, metadata, knowledge, and linked open data to support analytics across the spectrum of applications from precision medicine to accountable and learning healthcare systems.