Transforming Data into Meaning: Standards-based Capabilities to Bring Disparate Data Sources Together

By Darren Selsky, MS, MHA

Data integration is a costly activity, but one that is used by healthcare organizations on a daily basis to unite information from diverse sources, such as lab, pharmacy, and radiology. This data is typically centralized in electronic health records (EHRs) or other niche systems for clinical and operational review. Current approaches naturally lead to the question: Is it possible to bring multiple data sources together in a less costly, fully automated way and still satisfactorily respond to the full breadth of business needs?

The answer to that question would represent a breakthrough in several industries. In healthcare alone, current spending on health information technology (HIT) integration exceeds $2 billion annually, with growth expected to reach $2.7 billion by 2018.[1] Factors such as rising healthcare costs, strong government support and initiatives, and a growing need to integrate healthcare systems have increased the demand for healthcare IT integration. Unfortunately, most organizations that deal with lots of data are using resource intensive, often manual methods and approaches to integrate data from heterogeneous sources that are becoming obsolete.

Part of the reason for this is that typical data sources are complex. Data must be tended to by technical staff. In large data warehouses, technical staff often work in islands of interoperability. For example, one department's niche Oracle database (with tables designed 20 years ago) might not play well with another department's niche Microsoft SQL Server database.

Extract, Transform, and Load

Data is not easily aggregated. This is why so much time and effort goes into an old school process known as "extract, transform, and load" (ETL). From a business perspective, how often have healthcare leaders identified a new business opportunity only to be told that the data to support that opportunity is not obtainable, or that it will take weeks or months to identify, or that the means to access the data is too complicated and costly to be feasible? In some situations, operations come to a stop because key technical people in an organization who are familiar with a legacy database leave the organization. Maintenance and modifications have to be put on hold. Time and resources go into finding candidates that can learn proprietary ETL scripts or consultants to transform the existing data into supported formats. Yet addressing all the above problems still does not solve the root problem of deriving sufficient business value from disconnected data sets.

The dominant form of data storage for many years in healthcare has been based on relational databases. Because of that, business intelligence (BI), search functions, and data analytics have required those difficult ETL operations.

The problem with current ETL methods is that they take time and resources and include hidden costs that are not apparent at the beginning of a project. In an article published by the Data Warehousing Institute, David Linthicum states: "When it comes to the cost of a BI deployment, it's not the software that will get you; it's the miscellany—the miscellaneous integration work, in particular."[2]

In the implementation phase of any BI exercise where disparate sources of data need to be integrated, it is estimated that 80 percent of the cost of that BI project is wrapped up in data integration, compared to the analytics component at 20 percent.[3] Data integration hassles are legendary, including bringing together all relevant data from various operational systems not designed to feed BI systems. In addition to integration and conversion costs, there are ongoing costs as well.

"When you look at ongoing costs, though, the roles reverse, making data integration 20 percent of the costs versus reporting and analytics," Linthicum writes.[4]

Why so expensive? On a commercial scale, data integration is difficult and complex. It's perhaps one of the most difficult jobs in the world of BI. However, it's also critical. Indeed, the ability to bring in the right data on a timely basis is directly related to the value that the BI system will bring to the business—more so than the types of analytics it's looking to drive. It's the old "garbage-in-garbage-out" concept.

The future of healthcare relies on the improved flow of health information across the entire patient care continuum. This means a shared information strategy linking disparate systems across the healthcare continuum, inclusive of enterprise EHRs, niche departmental EHRs, practice management systems, and external manufacture device registries—while still maintaining patient privacy and security standards. Such a realization would not only enhance the clinician and patient experience but also enable faster treatment and better care coordination for patients.

For this to occur, a solution would have to be built on a participatory platform, where all organizations share a vision to create an interoperable information space. While this utopia has support through the "meaningful use" EHR Incentive Program standards, such as the Fast Healthcare Interoperability Resources, there are bridge solutions built upon sophisticated artificial intelligence platforms systems that enable semantic integration across federated platforms, regardless of database type, to bring efficiency to clinical decision-making.[5]

We May Boldly Go Where No One has Gone Before

At the 2014 Semantic Technology and Business Conference in San Francisco, CA, a big step was taken towards tackling the problem of semantic interoperability. It came in the form of The Yosemite Manifesto,[6] which recommended using the World Wide Web Consortium's Resource Description Framework (RDF) standard model for data interchange as a universal healthcare exchange language, describing RDF—one of the core technologies of the Semantic Web—as the best available candidate for the job.

Technology recently released out of the University of Texas' (UT) Department of Computer Science bridges the gap of current barriers to interoperability through the use of semantics and RDF. The technology leverages these standards to integrate and search data across disparate databases. This technology is already in use by many healthcare organizations for population health analysis, pharmaceutical vendors for pre-clinical discovery, and healthcare professional organizations to integrate "in situ" data from member organizations for large scale cross-organizational best practice analysis.

The UT technology integrates data from multiple disparate sources and then maps these data sources to standard ontologies (i.e., CPT, SNOMED, LOINC, RXNORM, etc.) for federated search.[7] It also provides semantic search capabilities and, in particular, what is known as faceted search. Faceted search included the ability to semantically parse search terms to get the best results. For example, the software will take the terms "child kidney cancer" and process this using the terms "pediatric renal oncology."

In addition, a SQL-like language called SPARQL for query processing has developed a method that dismisses the notion of performance degradation of graph-based queries on top of relational databases. In a nutshell, the UT technology has figured out how to run SPARQL queries on top of relational data as fast as SQL queries alone. Which means that a user would see the same level of search performance when looking at databases spread across an organization—internal and external—as it would if all the data were in one central repository.

By enriching data with semantics improvements, business intelligence advancements are realized as follows:

- Search: Enable search across multiple, heterogeneous data sources
- Analytics: Enables data analytics in real time between previously unmapped data sets
- Speed to Market: Reduction in time, capital, and human resources associated with data mapping

Semantic technology gets at the root problem of data integration and search across disparate databases. Mapping strategies and medical terminology management will play a key role in moving data from setting to setting and use to use, from informing patient care to influencing national policy decisions. While not an end in itself, data normalization through semantics moves the industry closer to the interoperability level needed for better information sharing including reporting, enhanced quality, and more robust analytics to support patient care.[8]

Notes

[1] MarketsandMarkets.com. "Healthcare Integration Market by Products (Interface Engine, Medical Device Integration, Media Integration), Services (Implementation, Maintenance, Training), Applications (Hospitals, Radiology, Laboratory, Clinics, HIE) - Global Forecast to 2018." March 2014. www.marketsandmarkets.com/Market-Reports/healthcare-it-integration-market-228536178.html.

[2] Linthicum, David. "The True Cost of Integration in the World of BI." The Data Warehousing Institute. August, 20, 2013. https://tdwi.org/articles/2013/08/20/true-cost-of-integration.aspx.

[3] Ibid.

[4] Ibid.

[5] Health Level Seven. "Fast Healthcare Interoperability Resources (FHIR)." www.hl7.org/fhir/.

[6] Guess, A.R. "Working On Taking 'RDF as the Universal Healthcare Exchange Language' from Proposal to Policy at SemTechBiz." Dataversity. June 5, 2015. www.dataversity.net/working-on-taking-rdf-as-the-universal-healthcare-exchange-language-from-proposal-to-policy-at-semtechbiz/.

[7] Whetzel, P.L. et al. "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications." *Nucleic Acids Research* 39, June 14, 2011. www.ncbi.nlm.nih.gov/pubmed/21672956.

[8] Levy, Brian. "Health Care's Semantics Challenge." *For the Record* 26, no. 5 (May 2014): 26.

Darren Selsky (dselsky@capsenta.com) is the senior vice president of business development at Capsenta, Inc.

Article citation:
Selsky, Darren. "Transforming Data into Meaning: Standards-based Capabilities to Bring Disparate Data Sources Together" *Journal of AHIMA* 87, no.3 (March 2016): 38-39.