DFIN

WHITE PAPER | A NEW APPROACH TO DATA QUALITY

A New Approach to Data Quality: Preparing for a RegTech, SupTech and AI Future

DFINsolutions.com

Global reporting and compliance have remained document-bound for too long. With recent advances in artificial intelligence (AI) and other forms of machine learning, however, companies are poised to meet their regulatory and compliance requirements in completely new ways.

One sign of the magnitude of change currently underway is that human beings are no longer the primary consumers of regulatory information. The Securities and Exchange Commission (SEC) has said that on any given day as many as 85 percent of the documents visited on the EDGAR filing system are visited by internet bots.

Moreover, on a world-wide basis, the number of countries collecting structured data hits the 50+ mark in 2020, as recently reported by XBRL International in its new interactive directory to all of the global XBRL projects that they are currently aware of around the world.

In its 2019 AI Predictions report, PwC states that 20 percent of U.S. business executives plan to implement artificial intelligence enterprise-wide in their organizations in 2019. It's clear that advances in artificial intelligence, robotic automation processing, enterprise digitization and business intelligence are transforming internal financial reporting as we know it.

A modernized reporting landscape

One of the most striking aspects of the ongoing transformation from documents to machine-readable structured data is the emergence of Inline XBRL—also known as iXBRL—and the SEC's June 28, 2018 decision to mandate its use. iXBRL is an international XBRL standard that is an offshoot of XML. Unlike XBRL, however, iXBRL reconciles data standards and makes traditional XBRL information readable from a web browser.

The technical fine points of the iXBRL standard are complicated, of course. But for users, iXBRL is best understood as a system that allows financial reporting to undergo a transition from documents to a data stream and back again.

The SEC has mandated iXBRL's use, on a phased-in basis, by June 2021. This is a clear sign that U.S. regulators are committed to transforming documents to data and joining the digitization revolution.

There are other signs that XBRL is finally catching on for a variety of different purposes. In January 2019, the U.S. Federal Energy Regulatory Commission (FERC) published a proposal that would swap the Visual FoxPro-based e-forms currently in use to an XBRL-based data collection system. After a brief comment period, FERC announced a formal request for proposals, setting in motion a plan to transition from document-based reporting to standardized XBRL data reporting and analysis.

The FERC project, which involves developing a special FERC taxonomy, is similar to the US-GAAP Taxonomy developed by XBRL U.S. and will make filing easier, more consistent and accurate for filers. For the regulators, it will make information that was difficult to use more searchable and user-friendly. Another goal is to improve the quality and usefulness of reported FERC data.

The current FERC proposal builds on a 2015 pilot project developed by Danny Kermode, assistant director of water and transportation and energy policy advisor at the Washington Utilities and Transportation Commission (UTC) and DFIN to transform the UTC reporting process. As part of this pilot, DFIN helped UTC innovate from document-based collection, to standardized, machine-readable data reporting and validation.

The goal of the original pilot was to allow UTC professionals to spend less time transcribing and validating data and more time analyzing the information being submitted. As part of the pilot, DFIN helped with data point modeling, taxonomy design, XBRL database storage and analysis. The aim of the pilot program – and of the new FERC project – is to collect more accurate data while saving the agency money through productivity gains.

The effort to make information machine-readable is spreading to other corners of government. On Jan. 14, 2019, President Donald Trump signed the Open Public Electronic and Necessary (OPEN) Government Data Act (P.L. 115-435) into law. Under this new law, all government data deemed non-sensitive will be made available in machine-readable data formats. Federal agencies will be expected to build and maintain comprehensive data catalogs.

Every federal agency is now required to designate a chief data officer (CDO) to oversee how data is created and used, and these individuals will be tasked with implementing the various components of the new law. These officers will also be part of a central Chief Data Officer Council.

One major challenge going forward will be how federal agencies establish and equip their CDO functions; they need to recognize, as the law outlines, the CDO function as a distinct and independently established function, apart from traditional information technology leadership. The hope is that CDOs will build communities of practice, and work with members of the council to share best data practices that could be implemented across other agencies, as well.

In the end, CDOs will not be chief information officers under a different name; rather, they will be the sentinels of quality, providing accurate and complete agency data, and, hopefully, they will help shift the prevailing culture to one of data management and data-driven decision-making.

Christian Hoehner, senior director of policy for the Data Coalition, stated, “Better education about how data will be utilized by agencies and the public will help to incentivize agencies to ensure compliance with the law.”

The Federal Data Strategy and President’s Management Agenda show that the government recognizes the value of its vast stores of data. These initiatives will run parallel to the implementation of the OPEN Government Data Act, and in some cases, the initiatives will be intertwined.

Bipartisan congressional efforts and a slew of executive efforts, including the recent executive order on U.S. artificial intelligence, clearly show that open data utilization is a government priority.

The tech transformation

Arguably, there are three pillars of the tech transformation that is underway. The first is RegTech, which has been defined as the use of new technologies to meet regulatory and compliance requirements more effectively. The second pillar is RegTech’s cousin, supervisory technology, or SupTech, which uses new technologies to tackle supervisory requirements. Finally, AI, the third pillar, affords companies machine learning and related techniques, so computers can sort through and analyze copious data and even draw meaningful conclusions from that data.

"The advantage of using machine learning technology is that it can learn what concepts look like and identify them," says Ned Gannon, president of eBrevia, an artificial intelligence-based contract analytics company acquired by DFIN in 2018. "AI can help you pull out data from vast quantities of documents where using human beings alone would be impossible."

Apart from the efficiencies AI affords, regulators are specifically recognizing the promise of RegTech and SupTech. On May 3, 2018, Scott Bauguess, then deputy chief economist and deputy director of the SEC's division of economic and risk analysis, delivered an SEC keynote address in Boston. In his address, titled "The Role of Machine Readability in an AI World," Bauguess specifically referenced both RegTech and SupTech as ways to "lessen the burden of either complying with or supervising a wide range of regulatory requirements in financial markets."

There is, however, a large challenge hindering the successful evolution and convergence of RegTech, SupTech and AI: the "usefulness" of data. In his address, Bauguess explicitly dispels the myth that machine-readable reporting standards ensure high-quality data. The SEC is well aware that whenever a standard reporting element might reasonably be used, but individual companies create custom extensions to the SEC's Generally Accepted Accounting Principles (U.S. GAAP), then it becomes difficult for users to make meaningful comparisons between companies. It is becoming clear that reliability and data quality are key to ensuring AI can be used to provide the insights RegTech and SupTech are designed to deliver.

Until there is true standardization and quality validation within structured reporting, achieving the original SEC vision of "leveling the playing field between companies large and small" for how information is presented and consumed by both institutional and retail investors will remain an elusive goal.

> The Intersection among Regulation, Technology and Data provides for enhanced "usability" and "reliability" of data and promotes machine learning and artificial intelligence.

Quality data through data comparability

For quite some time now, there have been ongoing problems with data comparability within digitized financial documents. The crux of this issue is that public companies, especially large accelerated filers, frequently design and use their own custom axes and concepts even though there are approximately 300 existing axes available in the U.S. GAAP. When a company uses a custom tag or axis, the information reported cannot be compared with information from other companies because those companies are using different measurements. In other words, it is no longer possible to compare apples to apples.

As structured data replaces document-based disclosures, better oversight of audit and quality is needed. Standardized data definitions applied at the creation of financial reporting should reflect the complete lifecycle of data from collection through dissemination. Audited, digital financial statements would facilitate analysis and could minimize errors as documents are transformed into data. With fewer errors, the usability and quality of the digital data under collection would be dramatically improved.

Efforts to automate the semantic mapping of new elements across taxonomies have had disappointing results. One researcher undertook such an effort using natural language processing and machine learning to try and normalize custom tags to their nearest equivalent standard tags. Using SEC filings from 2016, the researcher encountered 285,102 unique tags. To compound the challenges, the language used in naming or describing extension elements can have specific meanings different from general English usage, or the language may even be intentionally vague. While data and period typing could reduce these problems, without more explicit guidance the results were poor.

The good news is that a solution *does* exist. If regulators provide explicit guidance, then the number of custom tags could be dramatically reduced. In fact, the U.S. GAAP and IFRS taxonomies could adopt the concept anchoring model that will be part of the European Single Electronic Format (ESEF) taxonomy from the European Securities and Markets Authority (ESMA). The ESEF taxonomy is derived from IFRS, but extensions are allowed only if filers define an anchor connection to one or more elements from the standard taxonomy. By enforcing the explicit identification of a custom tag with the standard taxonomy, regulators ensure that the resulting custom taxonomies will be more amenable to the normalization necessary for comparisons across taxonomies.

If an anchoring model were voluntarily introduced, there would be benefits to the individual filer. For instance, the enhanced contextual information may make it easier for vendor software to propose alternatives like dimensionalizing rather than creating a whole new extension.

There is, however, a regulatory impediment to supporting this model within the current framework: the SEC would need to introduce a new role that matches the wider/narrower role in ESEF. Without this, vendors that want to offer the ability to anchor custom axes or concepts would be unable to include these relationships in their XBRL submissions without violating an EFM rule against custom arc roles.

"Nevertheless," says DFIN's Truzzolino, "concept anchoring would create a more precise semantic context for any extensions." He is therefore convinced that adopting a concept-anchoring system "would provide many attractive benefits in terms of data usability."

Greater semantic context: FIBO

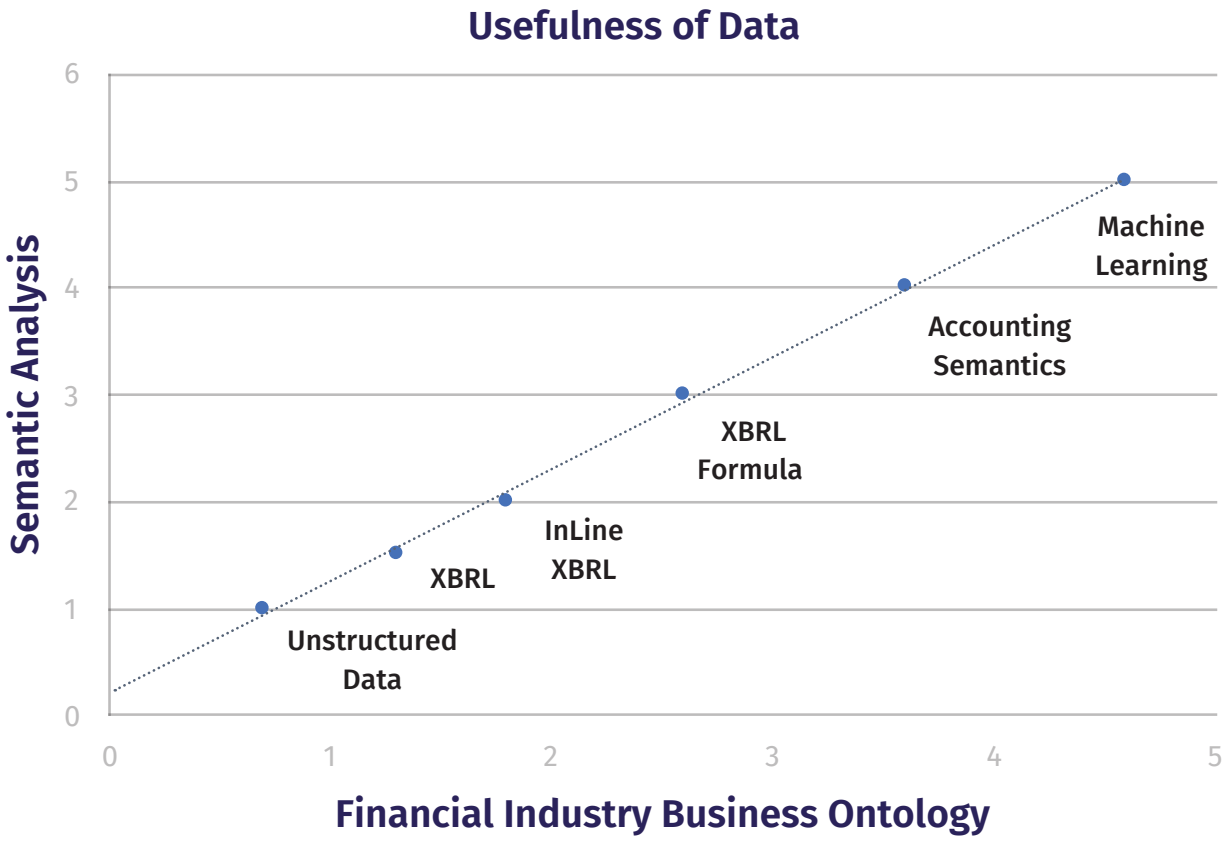
The SEC currently supports additional data models other than XBRL – such as, FpML and FIXML – and provides a common data model for those two standards to use. However, the fact remains that even with improvements to the intra-comparability of XBRL, comparing data across different models is not possible unless filers can convert representations into a common model that can serve as the basis for additional levels of reasoning and analysis.

XBRL taxonomies, like the U.S. GAAP, provide a range of context information for elements, such as rollup hierarchies, data, period and balance types, and references to source accounting codification. However, this is often insufficient to perform semantic analyses that would:

- Help focus or automate selections of tags
- Aid in decisions regarding custom extensions
- Determine the consistency of documentation and naming of custom extensions

Enhancing the semantic context of XBRL tags, whether they are standard or custom, would create internal consistency, improved data quality and simpler conversions of XBRL models to a common ontological model. This common model could then be used for broader analysis and comparison with other data, such as FpML and FIXML.

The Financial Industry Business Ontology (FIBO) standard developed by the Enterprise Data Management Council (EDM) is an example of a common model. This initiative draws together several connected ontologies implemented as RDF triples, which is a statement in "subject/predicate/object" form, that are represented using W3C Web Ontology Language (OWL). According to EDM, FIBO currently consists of 11 core finance industry domains – including securities and equities and loans – in 49 modules and more than 300 ontology files. Proponents of FIBO see this model as the reference point for an emerging intersection of structured data and AI, and there is certainly a case to be made for this approach.



RegTech solutions best support the transformation from documents to data when they can demonstrate and quantify the benefits of new technologies aligned with regulatory and policy changes. An example of this is discussed in the Data Foundation's white paper: "Standard Business Reporting: Open Data to Cut Compliance Costs," where coordinated changes in regulation, technology and data converged for true RegTech transformation. Since implementing SBR, Australia has exceeded its cost-cutting goals. By June 30, 2016, the Australian government and the wider business community saw combined savings of $1.2 billion—with even higher levels of projected savings for the future.

Current approaches to conversion of XBRL to OWL/RDF models seem to focus on simpler taxonomies, and have typically involved some manual steps to complete the mappings. While this may be acceptable for initial modeling of U.S. GAAP and IFRS, it is not feasible for custom extensions that are the mainstay of SEC filings.

One approach is to add more options for the reference linkbase, like FASB currently does for change notes. These options would allow additional semantic characterization of any element and would be available to filers to characterize their extensions. This, in turn, would permit filers to submit reference linkbases as part of their filings to allow additional semantic information to be used in processing the filing to the common model.

How regulators could improve data quality

The U.S. regulatory compliance structure is fragmented by industry and by purpose making RegTech difficult to deploy. But by adopting the right data standards, we can enable new technologies to flourish. Enabled by quality standardized data, RegTech solutions will make regulatory filings more efficient, transparent and useful for everyone who generates, collects, and uses these filings

While full standardization of regulatory data has been acknowledged as the Holy Grail, the problem is that full standardization can be impractical. In a reporting system with thousands of different companies submitting data, the data reported is bound to have inconsistencies.

In response, DFIN proposes that the SEC and other regulators utilize international standards like the XBRL formula model to express data quality rules. It seems apparent that if machine learning technology is taught how to leverage data quality rules, the result would be of considerable value because quality and reliability have been concerns for investors who would like to use XBRL data. It is critical that data be converted to non-proprietary, unambiguous, machine-readable formats, such as XBRL or iXBRL, so that machine learning can be leveraged to improve data quality and usability.

Furthermore, machine learning can only be fed the proper algorithms to standardize data if there is a marketplace of validation rules; these rules can inform the relationships among data and content beyond the current definitions in the XBRL taxonomy.

Regulators would need to provide a method for expressing rules using the XBRL formula model in a way that domain experts could understand. Simultaneously, they would need to permit generation-of-formula implementations that could be run by any compliant

processor – including Arelle, an open source validation engine, as well as DFIN's UBMatrix XBRL Processing Engine, which is proprietary and installed across three continents. Solutions such as these could provide XBRL and XBRL formula validation for very large data sets.

Regardless of the approach taken to strengthen rule development, what matters most is that there is a canonical form, which generally consists in the choice of a specific object in each class, available for each rule – or XBRL formula – that all stakeholders can use as an unambiguous and verifiable declaration of that rule. The IFRS 2017 Formula Linkbase and IFRS Guidance serve as a model of this approach.

DFIN takes AI to the next level with eBrevia

When eBrevia co-founder Ned Gannon was a corporate law associate, he often found himself reviewing and summarizing contracts as part of the due diligence process that preceded any merger or acquisition. "There's got to be a more efficient and accurate way to do this," he recalls thinking.

Gannon and his co-founders, Adam Nguyen and Jake Mundt, began exploring ways to leverage machine learning technology to make due diligence and similar tasks easier and more effective. The eBrevia product, he explains, "focuses on analyzing contracts and extracting unstructured data," but the applications extend to compliance and other areas of data analysis, as well.

"The quality of the data that you're giving to an AI system really is important in terms of how accurate that system is going to be," maintains Gannon. Improving the quality of data is central to eBrevia's mission – and that mission will only strengthen now that eBrevia is part of DFIN.

DFIN's role in technology

Understanding that definitional ontologies and aligned data formats matter for establishing shared meanings within an enterprise and within markets, as well as for regulators to succeed in their missions, DFIN will continue to play an important role in the development of RegTech, SupTech and AI. DFIN is committed to filling the gaps between what companies are analyzing and reporting, and ensuring the data that users access is high quality and standardized.

"We believe that when regulation, technology and data converge, the approach is disruptive as it rests on a few key themes: efficiency, risk minimization and data quality improvement," says Craig Clay, DFIN's president, Global Capital Markets. "Once standardization is achieved, then RegTech, SupTech and AI can begin to realize their incredible promise."

The following are key concepts for smoothly transitioning from static documents to data, and will effectively prime organizations for increased adoption of RegTech, SupTech and AI solutions:

- **Make quality of data paramount.** Data must be reliable so that users are confident in the information generated by technology. The best way to do this is through adopting a single, open data standard for U.S. regulatory agencies.

- **Take steps to enhance the semantic context of XBRL tags.** One model for doing so is FIBO. Another is adding additional options for the reference linkbase.

- **Support enterprise digitization.** Companies can build a competitive information foundation through enterprise digitization; this entails internally aligning common data formats and adopting open standards. For more, see DFIN's paper "How Data Will Determine the Future of RegTech."

- **Use AI to standardize data.** When machine learning is used to fill in the gaps in what companies have reported, users benefit from more standardized data.
- **Encourage vendors using machine learning to create algorithms that standardize data.** In the end, standardization will be achieved by those programs that can harmonize data most effectively. Standardization and data quality are key to a successful RegTech, SupTech and AI future.

DFIN's SaaS solutions support growing trends in enterprise digitization and SEC modernization.

ActiveDisclosure is a cloud-based collaboration, tagging, validation and SEC filing tool. With ActiveDisclosure, you benefit from:

- Advanced tools to simplify your reporting process
- The ability to update hundreds of cells instantly
- Access to DFIN's expert advice and guidance

Venue is a highly secure, virtual data room platform used to confidently share critical information in real time. Venue delivers:

- Rigorous security features to protect your data
- Ongoing compliance advice from industry experts
- A dedicated project manager and support team

eBrevia is an industry-leading, artificial intelligence-based data extraction and contract analytics tool. The platform helps users review documents up to 90 percent faster and more accurately. With eBrevia, you can:

- Find and extract critical information based on your specifications
- Convert scanned documents to searchable text
- Generate summary reports populated with extracted content
- Share documents and control access to them.

About DFIN

Donnelley Financial Solutions (DFIN)

DFIN is a leading global risk and compliance solutions company. We provide domain expertise, enterprise software, and data analytics for every stage of our clients' business and investment lifecycles. Markets fluctuate, regulations evolve, technology advances, and through it all, DFIN delivers confidence with the right solutions in moments that matter.

Learn about DFIN's end-to-end risk and compliance solutions on line at DFINsolutions.com. You can also follow us on Twitter @DFINSolutions or on LinkedIn.

Learn more about DFIN's end-to-end risk and compliance solutions.

Visit DFINsolutions.com | Call us +1 800 823 5304