research

CHECKLIST REPORT

2018

Six Strategies for Balancing Compliance with Data Value

By Philip Russom

Sponsored by

technologies®

Transforming Data With Intelligence™

MARCH 2018

TDWI CHECKLIST REPORT

Six Strategies for Balancing Compliance with Data Value

By Philip Russom

TABLE OF CONTENTS

tdwi

Transforming Data With Intelligence™

555 S. Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org

FOREWORD

The world of data is evolving at an accelerating pace. On the upside, most of the recent changes can lead to business opportunities. For example, data from new sources—such as social media, customer channels, and the Internet of Things (IoT)—can be applied to improve operational efficiency, business decisions, competitiveness, profitability, and growth. On the downside, customers and prospects grow increasingly concerned about data privacy and security as data captures increasing volumes of detailed information about their behavior, finances, health, and sensitive demographics. In response to these concerns, governments are introducing more legislation that ensures protection for citizens, residents, and other parties. Even in light of these changes and challenges, businesses can seize new data-driven opportunities, but only if they recognize sensitive data and handle it responsibly.

To keep pace with evolving data and privacy concerns, the Parliament of the European Union (EU) has written and passed a new General Data Protection Regulation (GDPR), which resembles prior ones that covered the handling of data about EU residents, largely within the boundaries of the EU. Now the regulatory scope is extended to encompass corporations and other information technology user organizations worldwide that handle data about EU residents. Organizations that handle EU data and fail to comply with the new GDPR can be subject to fines as high as 4 percent of their annual revenues.

The GDPR is looming large on the horizon, going into effect on May 25, 2018. Many organizations worldwide are now revising their data governance programs for GDPR compliance.

Organizations experienced with "legacy regulations"—e.g., HIPAA in the U.S., the DPRs of Canada and the U.K.—are well positioned to incrementally extend existing governance programs. Organizations that have complied with prior EU data regulations are even better positioned. Obviously, organizations new to the challenges of data governance must work hard and fast to prepare for GDPR compliance. At the same time, even the best data governance programs can age and require updates and modernization to remain relevant. Regardless of your organization's level of maturity and readiness relative to compliance, now is the time to act.

Focus of this report. One way to prepare for any new compliance regulation is to review and revise your organization's data governance policies and business processes around data's capture, management, and usage. However, those policies and processes can be hamstrung when the underlying user best practices and tool technologies for data management do not form a solid foundation. Successful compliance demands a sound data management foundation.

Therefore, this report will focus on how targeted improvements to specific data management best practices and technology can contribute significantly to your success with GDPR compliance, as well as data governance and data-driven programs in general. After all, winning at data governance is similar to winning in competitive sports: if you and your team do not have solid fundamentals, you will probably not play a competitive game.

Note that the EU's GDPR is definitely a direct driver for compliance modernization as well as an indirect driver for data management modernization. However, there are other drivers and goals for data management best practices and tool updates, such as competing on analytics, leveraging new big data, supporting self-service use of new data sources and analytics capabilities, capturing IoT data, assembling complete customer views, adopting new data platforms to enable new forms of data and analytics, and embracing the digital enterprise concept. As we'll see, the data management best practices and tools that contribute to success with data compliance also contribute to success in other data-driven enterprise areas.

Data value. Modernizing data management and its governance is demanding in terms of investments in time, resources, technologies, and changes to policies and procedures. However, in the long run it is more of an opportunity than a problem—if business and technology people are willing to also modernize how they think about data and its use. We all need to think more about *data value* and act accordingly.

Data value is a business-oriented way of thinking focused on the business potential of data. It complements the all-too-common obsession with data's technical requirements. Data value assumes data is a valuable business asset and therefore should be leveraged accordingly. It assumes that work in data management should align with and serve the data requirements of high-value business goals—growth, operational efficiency, and competitive advantage—not just technical requirements. It assumes that your asset portfolio of data should be protected and grown, as well as made fit for multiple purposes and easier to use, so that data becomes even more valuable for even more businesspeople and departments. Furthermore, thinking in modern terms about data value can accelerate modern programs for big data and the digital enterprise.

A number of best practices and tool technologies have contributed to data value for years—but in silos, usually with a technical bent. For example, think of best practices and data management tool functions for data quality, stewardship, curation, metadata management, cataloging, lineage, data integration, analytics, and reporting. Each leverages, improves, controls, or presents data

that has business potential—though with little coordination with others. It's time to take a more holistic approach; data value needs all these and more, but through modern tools that do a better job coordinating multiple value-adding data management best practices. Data success requires a strategic view of data value.

Data intelligence. Finally, it's not enough for data to have value. Data's value should also be documented and quantified centrally to provide data intelligence for both business and technical people who work with data. *Data intelligence* is information about data, and it assumes familiar forms such as metadata, lineage, and quality metrics. Data intelligence assumes emerging forms as well, such as data cataloging, glossaries, and machine learning for the automation of data management. Without data intelligence there can be no data value.

Data intelligence also involves analytics for change impact, quality tracking, and usage trends. When data intelligence is centralized and shared holistically, users can find relevant data and prepare it for production use more easily, quickly, and accurately, all within compliance rules. Data intelligence must be extended via crowdsourcing, where users review and score the accuracy, quality, risk, trust, relevance to specific use cases, and compliance requirements of data sources, databases, reports, and analyses.

This report will now drill into a list of data management strategies that provide a solid foundation for successful compliance, ranging from internal data governance policies to GDPR and other external regulations. The report will also stress how data value and data intelligence make compliance (and other data-driven activities) more agile, accurate, credible, and auditable.

STRATEGY ONE

MANAGE THE ENTERPRISE DATA PORTFOLIO IN A HOLISTIC AND BALANCED MANNER

New ways of thinking about data and its value are inherently broad, reaching across every unit of a business and beyond to customers, partners, and regulatory bodies. Although data is still managed locally in its applications and databases, there's a need to also see all that data holistically, as a collection of diverse and related assets within a single portfolio. Without holistic data portfolio management, data's value is limited to siloed use cases and data intelligence is possessed by only a select few.

Compliance assurance may be compromised, too. Because data travels across the modern enterprise and its ecosystem, data may be compliant in one use case but not in others. Without a holistic view and management, how can you quantify your current state of enterprise compliance? How do you track and maintain compliance? Enterprise data portfolio management may require a balance of holistic business goals and compliance realities.

TDWI has seen organizations successfully turn to enterprisewide data governance programs as a way to get a holistic view of data. Such data governance boards or committees are often two programs in one:

- A business program that creates enterprisewide policies for the compliant use of data
- A technical program that develops enterprisewide data standards that help data travel across more systems for broader business value

Note that business and technical programs balance each other, but the business side leads so that technical standards and other data management work support business goals. This way, compliance is assured and the value of data is enhanced by its availability to a wider range of users and departments. Depending on execution, data intelligence and data quality may also be enhanced by the enforcement of data standards.

Data governance boards face another balancing act. On the one hand, compliant data usage tends to limit access to sensitive data for some users. On the other hand, data standards for sharing data tend to give more users access to more data. Successful user organizations are balancing these competing goals by defining user roles, assigning users to roles, and writing policies that define the access privileges of each role. In a similar practice, the data governance board writes policies that define categories of sensitive data and parameters for the access and use of each category.

In a related area, TDWI has also seen an increase in the use of data masking, where data analysts or data scientists can access sensitive data sets for analytics as long as personally identifiable information (PII) is masked. As another example, there's a strong trend toward amassing multiterabyte data sets in Hadoop, data lakes, cloud storage, and large data warehouses to enable broad and holistic data exploration and discovery-oriented analytics, sometimes in a self-service fashion. User roles, masking, and other security measures are required to prevent compliance infractions with these types of large data collections. GDPR has given new focus to this area, requiring "pseudonymization" of personal data.

Finally, let's look at another balancing act. The quest for data value can lead organizations to hungrily acquire and ingest any data that comes their way under the assumption that greater data volume equates to greater data value. TDWI regularly sees this situation in firms that are pursuing complete customer views. This can be a problem relative to the GDPR because merely capturing data about EU residents constitutes a commitment to compliance.

Data governance policies should establish controls for new data acquisition to limit exposure risk for EU data. An emerging new best practice is to catalog data as it comes in so you have better

intelligence about the lineage, domain, location, and use of sensitive data—and hence you can more accurately recognize, track, and assure compliance, then prove compliance to auditors, your managers, and EU officials. In other words, data can be a source of risk. Getting an idea of data value and meaning at the point of acquisition is rapidly becoming a best practice.

STRATEGY TWO

EXPAND METADATA MANAGEMENT INTO NEW BEST PRACTICES

Metadata management has been with us since the dawn of database management systems, and remarkably it is still highly relevant today in the age of big data, unstructured data, and IoT. This is because metadata is still the preferred language for naming and modeling data as well as accessing data platforms for data inserts, updates, and queries. Likewise, data professionals have significant metadata skills and the majority of data-driven tools support or require metadata in some form. Furthermore, metadata and its management continue to evolve to accommodate more technical best practices and business use cases.

Metadata comes in multiple forms, each contributing to data value, intelligence, and compliance:

Technical metadata continues to be foundational to tool interfaces, queries, and data models as well as interpreting data, collaborating via data, and managing data portfolios. Other value-adding uses of metadata have also become common in recent years, such as using a metadata repository as an inventory of data.

Today, users regularly browse, query, and search metadata repositories. Without touching actual data, viewing metadata gives users an inventory of data to be managed, developed, and governed. At the metadata level, stewards and curators can be assigned to data sets they foster and improve for broader business use and value. Much of the information that goes into data intelligence is documented via metadata. Based on that intelligence, a variety of user types can quickly find PII, EU residents, customer data, and other data domains of interest.

Operational metadata is similar to log data in that it records information about data server events such as access by users and tools, the names of users or processes (when available), the types of data functions executed, and date and time stamps for the beginning and end of the event. This information, when analyzed, is useful for auditing, change detection, capacity planning, and chargeback accounting. However, it is also a kind of data intelligence that assists with compliance because it can reveal unauthorized access, improper usage of data, and security breaches.

Business metadata is a value-adding semantic layer built atop technical metadata. Business metadata translates the technical and arcane names of data structures seen in technical metadata into user-friendly names that nontechnical users can understand and use autonomously. In fact, business metadata is an absolute requirement for increasingly popular self-service data practices such as self-service data access, exploration, data prep, visualization, and analytics. By improving the accessibility and usability of data for many classes of users, business metadata delivers tremendous data value for the business.

Business glossaries are similar to business metadata (and often built atop a metadata repository), but have far more advanced tool functionality. A business glossary enables diverse user types who work together to define a common vocabulary of business terms, precise definitions for each term, links to data associated with each term, and provenance for each term (who created it, business processes it applies to, departments that use it, etc.). Because terms are used by multiple people and departments, creating or updating a term or its definition must be a team effort, meaning that the glossary tool must support collaborative and workflow-approval processes in a user-friendly interface.

Business glossaries underpin data governance, and they are key to business/technology communication and collaboration. A well-defined and populated business glossary provides a map of the business—organization, structure, policies, and processes. Business glossaries are seeing brisk adoption because having a common vocabulary for business entities and their associated data greatly enhances data value, intelligence, and compliance as well as user productivity, accuracy, and consistency.

Intelligent metadata is where a metadata management tool includes special analytics or automation. For example, leading tools can deduce metadata structures for data and sources that are devoid of metadata, as are many big data, unstructured, and IoT data sets; metadata thus created can be suggested to a developer or applied automatically without human intervention. Hence, intelligent metadata is a necessary component of successful big data programs.

As another example, machine learning algorithms in some tools can automatically map from sources through transformation logic to targets. Other algorithms may scan metadata as it comes in to spot sensitive data or domains of interest. The intelligence built into modern metadata management helps multiple user types be agile and productive with development, stewardship, compliance, and other data-driven tasks.

STRATEGY THREE

CATALOG DATA MULTIPLE WAYS FOR A MULTILAYERED BUT HOLISTIC VIEW OF YOUR ENTERPRISE DATA PORTFOLIO

The data catalog is where all the practices under discussion come together—a catalog enhances data's value, intelligence, and compliance through a multilayered but holistic view of an enterprise data portfolio, and that view is easily used autonomously by a wide range of users.

Depend on a catalog for a holistic view of your enterprise data portfolio. A mature catalog can visualize the whole portfolio or subsets of it to help users understand the data available and its relations to other data. Ideally, a data catalog integrates with metadata repositories and business glossaries for consistent terminology and definitions across the holistic view. This is vital for agility, rapid business decision making, and the ability to find the right data, right now.

Enrich data value by cataloging data by numerous characteristics. For example, users may categorize data by source, lineage, transformations, compliance sensitivity, privacy issues, PII, EU relevance, data domain, and so on. Users can tag, document, and annotate every data element or multielement structure. The multifaceted characterization enables users to scan data by numerous criteria. Cataloging can apply to existing data assets, thereby extending their value and usable life cycles. One of the characteristics of data is that its value depends on business circumstances; catalog richly, because you don't know what you will want to know.

Provide multiple access methods to catalog data. Ideally, users can browse, query, and/or search the catalog as is appropriate to the user's talents, the current use case, or the data of interest. Flexible access increases the likelihood of finding the data most relevant to a user's needs.

Extend data science to a wider range of users. The high ease of use, multiple methods, and visual interface of a data catalog makes data exploration, discovery, and analytics accessible to everyone, from the technical data scientist to the business analyst and nontechnical information worker.

Keep data scientists and other technical data professionals focused on their own work. These sophisticated users regularly devote time to helping businesspeople find the data they need for their jobs. When businesspeople can work autonomously with a data catalog, it frees up sophisticated users' time so they can focus on advanced analytics and true data science. A data catalog must be centralized and easy to use, bringing valuable data to the widest range of users possible. Enabling self-service via a data catalog enables expert support for hard problems.

Apply the catalog to multiple use cases. The catalog can help technical people design data models, deploy interfaces, profile data, monitor data quality, and curate data. Conversely, the catalog's ease of use helps users with self-service data access, exploration, discovery, visualization, and some forms of analytics. For any user, the catalog presents a holistic inventory of data that can guide governance and compliance, not just development and data management.

Tap the catalog's tremendous data intelligence. The average catalog is loaded with information about data and the entities it represents, all in one place, easily accessed by anyone.

Foster crowdsourced data intelligence. Information about data assets is enriched by employees who have direct experience with the data. The catalog should publish their scores and annotations immediately so that people across the enterprise benefit from their insights and discoveries.

STRATEGY FOUR

TRACK AND ANALYZE DATA LINEAGE ACROSS DATA'S WHOLE LIFE CYCLE

The most common question asked in business intelligence is: Where did the data in this report come from? Related questions include: How did you aggregate and transform this data? How old is this data? What is the level of quality and value of this data? How accurate are the sources? Does this report and its data respect policies for compliance, privacy, and data standards?

Problems with data trust. If you cannot answer these questions with credibility, data-consuming users will not trust the data and will not use it. Furthermore, a lack of trust in data can delay or prevent business decision making. This is a problem for you as a data professional when your bonus calculation is based on user adoption of your data products. Furthermore, this is a severe problem for your enterprise when the people asking the questions are regulators, lawyers, law enforcement officers, or other entities that have audit privileges (as investors and partners sometimes do). Without documented answers, delivered quickly, an audit can easily burn up valuable enterprise resources, including the time of both business and technical people, plus lots of money when lawyers and auditors are required to sort it out. Don't forget: your failure to prove regulatory compliance could lead to a fine or loss of licensing (GDPR can levy up to 4 percent of revenue).

Data lineage as the solution. The best way to prepare for and solve these problems is to implement *data lineage*. This is a special form of data intelligence that recognizes, tracks, and analyzes the introduction and movement of data sets and data elements into

and across your enterprise data portfolio. Lineage information is indispensable for answering the common questions mentioned above as well as for the agility and accuracy demanded of formal audits. Data lineage information can also guide improvements in data quality and process management.

End-to-end data lineage should be your goal. In some tools and methods, data lineage is little more than a record of data sources. However, modern and mature approaches aspire to *end-to-end lineage*, where data elements are tracked in detail as they move from source to target, including stepping stones along the way for data aggregation, transformation, and quality. Lineage also records dates, users, tools, and event types (e.g., access, load, copy). This detailed data intelligence enables highly credible auditing that can apply to compliance, security, and process improvements.

Impact analysis based on lineage information. Lineage intelligence—especially when it is end-to-end—enables deep impact analytics. For example, analysis may reveal that noncompliance risk is high where a process involves multiple sensitive data elements; the process can be adjusted or monitored to keep it compliant. A good impact algorithm can predict the consequences of changing a data-driven business or technology process, even with processes that span multiple systems. Impact analysis also contributes to better understandings of multistep processes so they can be optimized for speed, accuracy, data value, and general governance.

Tool requirements for end-to-end data lineage. Providing a holistic view of data elements, their relations, and their use across an enterprise data portfolio is one of the strongest benefits of data lineage. However, this is also its greatest technology challenge, given the large number of data items and IT systems in the average enterprise today.

It takes special tooling to scan and map a data portfolio while tracing data elements from origin to destination, understanding changes, and closing gaps. That tooling must interface with hundreds of applications, tools, databases, and other platforms, plus support COBOL, Java, SQL, ETL transformations, and many languages. Furthermore, the tool should have both a user interface that is friendly for business users and deep functionality for technical developers. It visualizes the business, applications, and data lineage. Ideally, the tool should be more than data lineage; for the sake of holistic scope, it should also support related functions for data value and intelligence, including metadata management, data cataloging, and impact analysis. Finally, given the size and complexity of today's big data, automation is a critical component of data lineage analysis; the tool should be able to link information sources and enrich lineage using existing data, metadata, and application understanding.

STRATEGY FIVE

MANAGE DATA'S QUALITY TO ENHANCE ITS VALUE, INTELLIGENCE, AND COMPLIANCE

We say *data quality* as if it were one monolithic practice, but it actually consists of many techniques and best practices, each with its own specific way of adding value to data. For instance, the most common data quality method is data standardization, which ranges from name-and-address cleansing to anomaly remediation. Similarly, data set deduplication includes multiple householding and match-and-merge techniques.

Other data quality best practices include data validation, verification, internationalization, and augmentation (with geocoding or third-party demographics). Data quality has many attributes, including accuracy, relevance, timeliness, and usability.

***Value* has diverse meanings in a data quality context.**

For example, engineering methods for data quality were originally inspired by value-adding concepts in manufacturing, where raw material is preprocessed to make it fit for a particular product. In data management, we preprocess raw data to make it fit for operations, reporting, analytics, marketing, and financials. In essence, we manage a data supply chain. That's a technical kind of value, which needs to be managed and directed for the greatest business value from data. For this reason, data stewardship is almost always included in a successful data quality program.

A data steward is typically a businessperson who understands how the state of data affects the success of his/her department or business process. The steward works closely with a data professional to prioritize data quality and other data management work so that the steward's business gets the greatest value from its data. Stewardship is so effective at aligning data quality work to business goals that it has become common in other data management disciplines, especially data integration, data modeling, report creation, and analytics development.

Data profiling is another data management discipline practiced in a data quality program. This is a technical exercise in which a data professional develops metrics and statistics that quantify the state of a data set or element. The result is a form of data demographics. Once a profiling solution is developed, it is scheduled to run regularly to provide monitoring that guides the continuous improvement of data.

The information that results from profiling and monitoring is a form of data intelligence.

It is important to share that intelligence in ways that make sense to both technical developers and data-consuming businesspeople.

In addition, metrics and statistics are numeric, therefore they can guide machine learning and other algorithms or rules to automate and accelerate many data-driven activities.

Data quality enforces compliance with internal data standards, not external regulations.

For example, the quality of data is usually one of the things that a data governance board governs. Data quality metrics, statistics, and other numeric quantifications tell governors whether critical data sets in the enterprise data portfolio are compliant with the board's guidelines for quality. Data governance boards care about the quality of data across the holistic data portfolio because they know it increases data's usability, shareability, trustworthiness, and accuracy for greater overall business value. The governance board also knows that the awareness of data quality supports confident business decision making.

STRATEGY SIX

GIVE SOME DATA DOMAINS SPECIAL MANAGEMENT AND TOOLS FOR GREATER VALUE WHILE RESPECTING COMPLIANCE

Your approach to data value, intelligence, and compliance will vary across diverse data domains. This includes general data domains at the data set or entity level (customers, products, financials) and granular domains at the field level (email address, street address, city, state, credit card number, URL, company name, etc.).

Finding an approach appropriate to a given domain can be challenging, and success will depend on balancing requirements for external regulations and internal standards with your need to actively grow and enhance high-value data domains. Balancing cost and value—so that the investment in data generates returns at least equivalent to other business investments—is a critical business skill. It underpins the rise of the chief data officer.

The customer data domain is a case in point. Many firms are pursuing a 360-degree view of each customer and prospect because they know that knowledge is power when it comes to gaining, retaining, and growing accounts. In other words, the customer data domain is high value, and the data's value to the business increases with each new data element quantifying a customer characteristic.

On the flip side, many customer characteristics such as PII, race, gender, salary, criminal record, and health status are inherently sensitive. A number of national regulations limit how certain data elements can be copied or made public. Most customer-oriented firms have also published a data privacy policy, which is tantamount to a contract with the customer. For companies and other organizations hoping to expand into European markets, difficulty managing the customer data domain is further exacerbated by the GDPR.

Despite conflicting goals, TDWI has seen user organizations in this situation successfully strike a balance between monetizing critical data domains and respecting legislated regulations and internal policies. Success comes from people and process measures, such as careful planning and practical governance policies. Given the complexity of the situation and the data itself, significant software tooling for data value, intelligence, and compliance are also critical components of the solution.

Other data domains need special handling and appropriate tool functions:

Geocodes. This domain is best curated by humans (either fully or partially) to assure that a relevant spot is coded (the entrance of a building instead of its center) and that related geocodes are grouped and labeled accurately (a factory may need codes for its delivery entrance, visitor center, emergency exit, etc.).

Product data. A product of any complexity will consist of multiple parts and subassemblies. These are regularly modeled as a hierarchical bill of material, and product data's value and intelligence are reduced when hierarchies are flattened.

Partner data. This domain is critical for supply-chain-oriented industries, such as retail, manufacturing, and logistics. The data enables B2B operations and the analysis of partners' performance, supply quality, and financials. Much of this data is coded in EDI, XML, and JSON, thereby requiring special tools that can read, catalog, and index the arcane structures of these document types.

Reference data. Customer masters and other collections of reference data have many handling requirements intended to foster high data value. For consistency and facile sharing, reference data is collected, standardized, improved, and curated in one central repository before being distributed holistically across the data portfolio.

The single central repository is also a control point for compliance and governance as well as the classic single version of the truth. Ideally, the tools that manage the repository should enhance this domain's data intelligence by providing or integrating with high-quality metadata, cataloging, and end-to-end lineage information.

ABOUT OUR SPONSOR

asg technologies®

ASG Technologies brings peace of mind to every enterprise with information access, management, and control within legacy and leading-edge environments. Many *Fortune* 500 companies trust ASG to optimize their IT investments. With our Enterprise Data Intelligence solution, we give companies deep insights into critical business data. Achieve regulatory governance and compliance requirements with end-to-end data lineage analysis, which has allowed customers to see more than 200 percent savings on time and effort. Our solution provides quick, reliable data intelligence to drive your business forward.

ABOUT TDWI CHECKLIST REPORTS

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, analytics, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.

ABOUT THE AUTHOR

Philip Russom, Ph.D., is senior director of TDWI Research for data management and is a well-known figure in data warehousing, integration, and quality, having published over 550 research reports, magazine articles, opinion columns, and speeches over a 20-year period. Before joining TDWI in 2005, Russom was an industry analyst covering BI at Forrester Research and Giga Information Group. He also ran his own business as an independent industry analyst and consultant, was a contributing editor with leading IT magazines, and was a product manager at database vendors. You can reach him at prussom@tdwi.org, @prussom on Twitter, and on LinkedIn at linkedin.com/in/philiprussom.

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on analytics and data management issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data management solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.