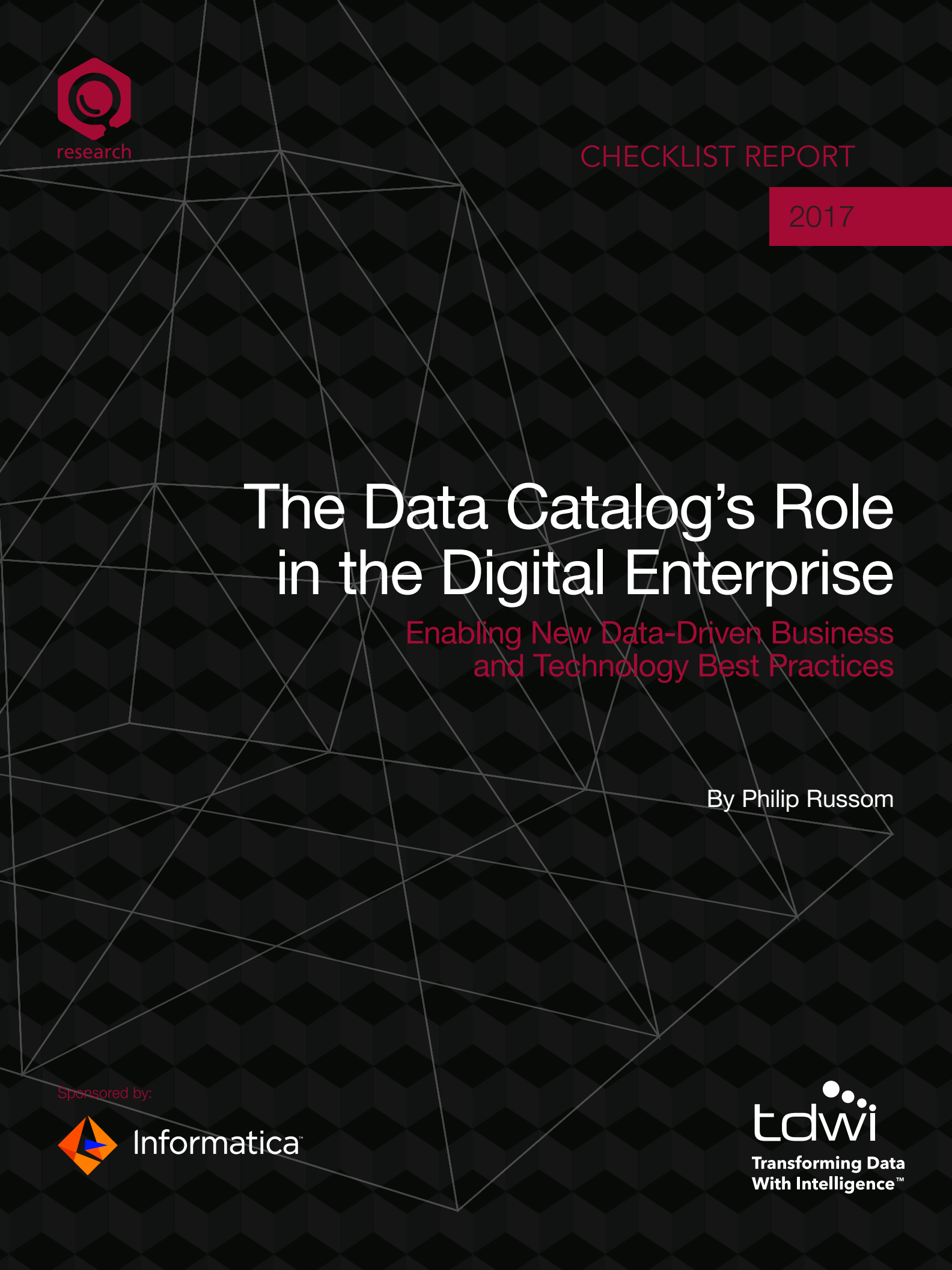

research

CHECKLIST REPORT

2017

The Data Catalog's Role in the Digital Enterprise

Enabling New Data-Driven Business and Technology Best Practices

By Philip Russom

Sponsored by:

Informatica

tdwi

Transforming Data With Intelligence™

NOVEMBER 2017

TDWI CHECKLIST REPORT

The Data Catalog's Role in the Digital Enterprise

Enabling Data-Driven Business and Technology Best Practices

By Philip Russom

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

TABLE OF CONTENTS

2 **FOREWORD**

3 **NUMBER ONE**
Rely on modern data cataloging for a unified, enterprise-scope view of data

3 **NUMBER TWO**
Modernize metadata management with a new data cataloging facility

4 **NUMBER THREE**
Build a data catalog that documents all data in diverse ways

4 **NUMBER FOUR**
Demand tool intelligence to cope with the growing number of data sources, targets, and structures

5 **NUMBER FIVE**
Allow some classes of users to access data via the catalog in a self-service manner

6 **NUMBER SIX**
Deploy an enterprise data catalog that fosters collaboration among technical and business users

6 **NUMBER SEVEN**
Depend on data catalog functionality to automate processes for data governance, stewardship, and curation

7 **ABOUT OUR SPONSOR**

8 **ABOUT THE AUTHOR**

8 **ABOUT TDWI RESEARCH**

8 **ABOUT TDWI CHECKLIST REPORTS**

FOREWORD

When we design and develop data management solutions, one of the first and most important steps is to catalog the data that will be captured, managed, analyzed, and shared. This is true whether the solution is a simple modernization of data management tools and practices or a dramatic paradigm shift in the business use of data that will transform the organization into a digital enterprise. A data catalog and its processes are critical for the technical success and business innovation of a modern data management solution.

The process of cataloging data. To catalog data, we usually start by naming the components of a data structure, including the names of databases, tables, rows, columns, and keys. We also describe each component in terms of data type, attributes, acceptable parameters, and lineage. These best practices are well established but should be extended to support modern practices such as cataloging by quality metrics, profiling statistics, data domain, trust level, and governance sensitivity. Cataloging must also support new data types, as seen in big data, multistructured data, social media, and the Internet of Things (IoT).

The contents of the data catalog. The descriptions of data components and types are usually expressed as metadata. Because metadata is the language of cataloging, a data catalog must manage all forms of metadata, including technical, business, and operational (or usage) metadata. In many ways, a data catalog is the modern approach to metadata management, yet there's much more to it than just metadata. The catalog and its platform must also support other forms of semantics, including master and reference data, plus the standard and proprietary semantics of tools for reporting, analytics, and integration. In addition, due to today's collaborative teams, the data catalog should ideally enable and manage discussion threads and annotations that users enter to further describe data and to share their discoveries.

Metadata and other valuable information about data assets are primarily managed per tool, application, department, or business process. A truly modern data catalog seeks to collect such information broadly from most enterprise systems, then centralize and unify collected information assets so they may be shared among many teams, solutions, tools, and business functions. The resulting unified enterprise view of data enables enterprise-scope governance, technical standards, visibility, sharing, exploration, and analytics.

Use cases for the data catalog. An enterprise-scale data catalog gives a productivity boost to technical professionals who depend on the catalog's metadata and other semantics when they design data models, deploy interfaces, profile data, monitor data's quality, and curate data. For business users, the business metadata managed by the catalog's infrastructure (and other user-friendly semantics) can enable several new data-driven self-service practices that business people are demanding. These include self-service data access, exploration, prep, visualization, and some forms of analytics. For both business and technical users, a catalog provides an inventory of data that can be browsed, searched, and queried to get "the lay of the land" or to guide governance and collaboration, not just data management and development.

Note that these use cases stretch across business operations, analytics, development, compliance, integration, and system administration. This shows the tremendous value that a modern enterprise-scope catalog of information can add to just about every nook and cranny of a digital enterprise.

The importance of intelligent automation for data cataloging. From these examples you can see that a fully operational data catalog documents vast numbers of data sources, targets, and structures. The data catalog size increases as organizations introduce new sources and structures of data, commonly from new customer channels, machinery, vehicles, consumer devices, and IoT. To survive the glut of new data, systems, and structures to be cataloged, users need cataloging software with intelligent functions that can onboard new systems and catalog data automatically with little or no human intervention. Modern tools are doing more of this via artificial intelligence (AI) and machine learning (ML) algorithms, as well as old school business rules and application logic, as described later in this report.

This TDWI report will examine the many components and functions of a modern enterprise data cataloging facility. The report is organized as a checklist of seven recommendations, each focused on a high-priority practice, tool function, or issue in data cataloging, and will help technical and business users understand the potential of modern data cataloging as well as how and where to apply it.

NUMBER ONE

RELY ON MODERN DATA CATALOGING FOR A UNIFIED, ENTERPRISE-SCOPE VIEW OF DATA

The word "catalog" comes from the Greek and literally means "complete list." In that spirit, a data catalog is not complete unless it aspires to enterprise scope. Although your catalog may never represent 100 percent of enterprise data assets, it still presents a single, unified view that enables growing practices that depend on enterprisewide visibility and access, such as the following:

Viewing the data catalog as a data inventory. This information can provide a valuable "big picture" that can assist many cross-platform data disciplines, including database administration, data integration and synchronization, enterprise data architecture, data warehouse architectures, and data governance.

Visibility for enterprise data governance. The data inventory seen in a data catalog helps governors, stewards, and curators determine which data sets need policy-based controls, structural improvement, and additional security. The catalog may also present information that relates to the governance of technical standards, such as the quality of data and semantics, plus the structure of data models, record formats, and field data types.

Browsing or searching catalog contents. This helps data analysts and data scientists discover data sources and data sets that are appropriate to a new analytics project. It also helps business users with elementary tech skills learn about customers, products, financials, partners, and other business entities that are critical to the success of their departments.

Sharing all information about data. To attain enterprise scope, a data catalog should support many approaches to documenting data assets and their characteristics. Metadata is critical for this purpose. However, other valuable information about data assets can also be amassed in a data catalog, then disseminated in a controlled way to a wide range of users, systems, and teams. Imagine a unified and comprehensive view of data assets as seen via metadata, master data, profiles, quality stats, categories, data domains, lineage, annotations, and entity relationships, whether crowdsourced, developed by technical experts, or generated via AI/ML.

Catalog infrastructure that is hardened for enterprise scale. For a data catalog to serve an enterprise, it needs an underlying infrastructure that can scale to the volumes of big data, perform quickly for the large number of users and applications that will access it in real time, and handle unstructured data and many container types (such as messages, result sets, files, and documents). Other enterprise requirements for catalog infrastructure include high availability, feature-rich administrative tools, and multiple security approaches (from user authorization to data encryption and masking).

NUMBER TWO

MODERNIZE METADATA MANAGEMENT WITH A NEW DATA CATALOGING FACILITY

Metadata management modernization is typically the highest priority of a data catalog because its takes metadata management from its backwater silos to a centralized cross-platform facility that is feature-rich and comprehensive. Imagine metadata extracted from all sources—whether on premises or in the cloud, whether within the enterprise, on the Internet, or at partnering firms. Metadata thus amassed is then improved and shared across an enterprise and beyond for unprecedented consistency, productivity, trust, and governance.

Multiple metadata types. The catalog should support all forms of metadata because each enables important use cases:

- **Technical metadata** documents data's structures, components, and data types. This is a foundation for data extraction and load, other computerized processes, and highly technical interfaces.
- **Business metadata** describes data in user-friendly terms that people with basic tech skills can understand. It enables new practices, such as self-service data access, exploration, prep, and visualization.
- **Operational (or usage) metadata** records access to data by users and applications. These records can be analyzed to understand compliance, security, capacity, and chargeback accounting issues relative to data access and data usage.

Deducing and injecting metadata. Many forms of big data, Web data, and IoT data lack metadata. Users should look for tools that can scan data to deduce its structures and develop metadata from that information. This helps data explorers and developers be more productive as they search, query, profile, and develop. Depending on the platforms and file types involved, users might also look for tools that can "inject" metadata into files (e.g., Hadoop Avro files), data documents (XML), and containers (JSON) to make them more usable. Ideally, these advanced functions should be automated by AI/ML, rules, tool logic, and services.

Data enrichment via metadata. Note that different kinds of data sets need different amounts of metadata for documentation and enrichment. For example, master data (highly governed) may have multiple dimensions of classification and associated metadata, whereas sensor data (lightly governed) may have relatively little associated metadata.

Extensive metadata connectivity. To achieve enterprise scope, metadata connectivity must be comprehensive, supporting many brands of relational database management systems (RDBMSs) and NoSQL databases, Hadoop distributions, mainframe and other legacy systems, reporting and analytics tools, file systems (HDFS, S3), and popular packaged applications (e.g., Salesforce, SAP, Oracle, Microsoft). Any of these may be deployed on premises or in the cloud.

The infrastructure of a good data catalog will include prebuilt scanners for all these sources and targets to collect metadata from databases, data warehouses, applications, cloud storage, BI tools, Hadoop, NoSQL, and other sources. Import aside, a data catalog must also export metadata to most of these platforms as well as to human-readable files and spreadsheets.

Metadata indexing. The metadata recorded in a data catalog should be indexed in multiple ways to enable multiple use cases, namely quick updates to the catalog, analytic correlations, semantic searches, and high-performance queries.

NUMBER THREE

BUILD A DATA CATALOG THAT DOCUMENTS ALL DATA IN DIVERSE WAYS

Categorization functions (both manual and automatic) should be applied to all data recorded in a data catalog, regardless of the data's sources, structures, and containers. Ideally, data sets should also be classified and indexed according to the data domains and business entities represented within them. When a catalog is this thorough, many types of users and tools can explore data more broadly, but with better guidance, and hence more easily and accurately discover the ideal data they need for specific use cases. Attaining this level of comprehensive cataloging requires a number of capabilities, as follows.

Unstructured data cataloging. This usually involves the unstructured data found in the files of popular productivity tools, namely Microsoft Excel, Word, and PowerPoint, plus Adobe Acrobat. The catalog should categorize and scan each file to infer and classify the semantic types, domains, and entities represented in each. Through this functionality, organizations can finally get business value and greater reuse from unstructured data while correlating this data with structured data.

Big data cataloging. This may involve data from Web applications, IoT, vehicles, sensors, and social media. Because these are often captured and processed on Hadoop, the catalog infrastructure should support Hadoop clusters and the Apache tool ecosystem. That infrastructure must also support the files, documents, and containers in which big data is often packaged (JSON, XML, Avro, messages, events). Furthermore, cataloging must scale to the large data volumes common with big data, and it must provide automation for quickly onboarding new data sources, as is common with IoT and other big data environments.

Semantic search. Many data catalogs and similar collections (e.g., business glossaries and metadata repositories) allow users to query and browse information about data assets. Semantic search complements browsing and querying by making data exploration as simple as Google. The search facility could index common names of data structures, domains, and business entities, making it easy to find data that references these. In addition, the index should also record and graph analytic correlations to help users understand relationships across multiple platforms, sources, and latencies.

Data domains. A good catalog will classify sources, data sets, and containers by the domains referenced within them. This includes general data domains at the data set level (customers, products, financials) and granular domains at the field level (email address, street address, city, state, credit card number, URL, company name, etc.). The cataloging tool should include a library of common domains as well as provide a tool for the creation of user-defined domains. The tool should also enable the development of rules, logic, and reference tables that control the recognition of domains and their classification. All this should be automated via rule-based or AI/ML-based domain inference.

NUMBER FOUR

DEMAND TOOL INTELLIGENCE TO COPE WITH THE GROWING NUMBER OF DATA SOURCES, TARGETS, AND STRUCTURES

Artificial intelligence and machine learning have recently expanded from analytics tools into tools for data management (DM). AI and ML give DM tools intelligence so they can automate many of the tasks a DM professional must perform repeatedly. Software automation powered by AI/ML helps data developers and others be more productive and cope with rising numbers of data sources, targets, interfaces, and domains that need to be cataloged across an enterprise.

Metadata management. With IoT and other new sources that are notoriously devoid of metadata, a modern tool can parse data and deduce credible metadata. The tool can suggest a metadata structure to a data developer for approval or log that structure in the catalog without human intervention.

Recommendations to users. AI/ML can watch user interactions with data and learn from them. As new users browse and search the catalog, AI/ML can suggest data assets and data pathways with which prior users succeeded. When the cataloging platform also supports publish/subscribe mechanisms, AI/ML can recommend subscriptions.

Data mappings. Time-consuming source-to-target mappings can now be performed by algorithms. The algorithms' accuracy and breadth increase as they watch successful users map manually. Automated mappings increase the productivity of data developers, data scientists, and data-savvy business users.

Data domains. ML algorithms and other tool logic can recognize and catalog data sources and structures that are part of particular domains. This helps users browse or search the catalog for domains of high interest, such as the customer, product, and financial domains. Advanced algorithms can even detect domains and domain relationships across data sets.

Data lineage. Ideally, cataloging tools should automatically record data's lineage as data is handled via metadata and other functions of the cataloging platform. Lineage should be detailed down to the entity and attribute level, with links among business terms within the glossary. With these details, data lineage enables fast and deep insights into data provenance and impact analysis.

Sensitive data. AI/ML algorithms can recognize and catalog data components that are potentially sensitive in terms of privacy and compliance. Similar algorithms can monitor user activity and alert users when they attempt to access data for which they do not have access rights.

Data anomaly detection. AI/ML has the potential to spot and react to data defects, such as outliers, nonstandard data, and data quality issues. Some tools go beyond detection and automatically remediate data quality issues based on governance rules and policies.

Analytic correlations. Algorithms can make analytic and other correlations among diverse data sets when they discover duplicate, similar, or related data sets, data domains, and entities. This, in turn, enriches users' queries, exploration, and analyses.

Upcoming use cases for AI/ML automation. In the near future, catalog-based AI/ML will also contribute to data security, governance, capacity planning, and system performance.

NUMBER FIVE

ALLOW SOME CLASSES OF USERS TO ACCESS DATA VIA THE CATALOG IN A SELF-SERVICE MANNER

A growing number of business users with basic tech skills want to work with data hands-on. They need to work autonomously—in a self-service manner—with little or no time-consuming support from IT or data management personnel, especially when they explore and profile data.

Modern self-service best practices require two things:

- **An easy-to-use graphical user interface (GUI)** is a critical success factor. Both technical and business people need tools at their skill level. An easy, role-based GUI can tailor functions to specific but diverse user types. Without ease of use, the roles of nontechnical users are limited.
- **Business metadata** is the foundation for all data-driven self-service practices and nontechnical users will not succeed without it. The collection of business metadata may be organized as a metadata repository, business glossary, or data catalog.

Once these requirements are satisfied, several emerging self-service practices are possible.

Self-service data access. With a tool's intuitive click-and-drag GUI, just about any user who knows the basics of data can quickly and independently view information about data managed by a catalog. Imagine self-service users browsing a broad catalog of data to find just the right source or data set for a specific report or analysis. They may be stewards looking for data quality issues. For the greatest success with self-service data access, the catalog infrastructure should support browsing, querying, and semantic search.

Data exploration and discovery. In organizations that are introducing new sources of data (from IoT, social media, big data), a growing constituency of end users is demanding more and better functions for data exploration and discovery. They need to access and explore new data, to understand its content, structure, and valuable uses, and they need to discover the new facts and insights that often come from new data.

Self-service data prep. During or after data exploration, an end user typically wants to develop a data set that represents an insight or some data integration and quality work to be done. To enable business users (not just technical ones), a new practice and tool type called "data prep" has arisen. It is a carefully selected subset of data integration, quality, and query functions, coupled with business metadata and high ease of use. The limited but effective functionality of data prep makes constructing a data set practical and productive for many user types.

Self-service data audit. Imagine users accessing a data catalog in a self-service way to answer common questions, such as: What's the lineage of this data set? How was it transformed? What is its level of quality and standardization? What are its trust and value levels? The user can get answers to these questions immediately without assistance when data sets and sources have been rated by other catalog users.

Self-service development. Most self-service toolsets were designed for business users. However, technical users tap them, too. For example, data scientists, data analysts, and data developers may use self-service tools when they need a quick and simple view

into data, perhaps before stepping back for more formal work. As another example, self-service can enable fast-paced technical practices, such as agile or lean development. In these methods, a prototype data set required early in a project can be created quickly and easily via self-service data access, exploration, and data prep.

NUMBER SIX

DEPLOY AN ENTERPRISE DATA CATALOG THAT FOSTERS COLLABORATION AMONG TECHNICAL AND BUSINESS USERS

Several collaborative and crowdsourcing use cases can be enabled by the functionality of a data catalog. Collaboration via information about data is important because it stimulates data-driven innovation and assures the alignment of data management work to business needs and goals.

Collaborative data development. Medium-to-large enterprises are trending toward dozens of data professionals, specializing in multiple data disciplines (integration, quality, modeling, analytics, etc.). This horde of data pros may be organized across multiple teams, each augmented by numerous consultants. When so many data management professionals work together, they need a central data catalog extended with collaborative features so they can communicate, coordinate efforts, learn from each other, adhere to data standards, be governed and managed, and share data and development artifacts.

Collaborative assessments of data and sources. All data assets—both old and new, regardless of source—need considerable collaboration to determine business value, compliance issues, classifications, and technical requirements. Over time, data sets and sources need collaborative review to foster continuous improvement and greater business relevance. A data catalog can greatly facilitate these processes by supporting collaborative features, such as annotations, discussion threads, workflows, and mechanisms where users can rate a data asset's quality, trust, usability, and other factors.

Collaborative revisions to business metadata and other catalog collections. Most catalogs include a business glossary, business metadata repository, or equivalent collection. Business users (especially stewards and curators) can collaborate via this function to manage the life cycles of business terms, reference data, and customer definitions.

Collaborative data stewardship. Most data stewards are business people who study data from their line of business to understand data's structure and content to determine where the greatest needs for data quality or data model improvement are, or to determine which systems are the best sources for a view of a new customer behavior. Business managers-turned-data-stewards are ideal for making these determinations because they understand how the condition of data affects business processes. For stewards, accessing and finding the right data themselves assures that the selected data aligns with their business needs. A technical user can then apply the steward's catalog annotations to the development of a data management solution.

Collaborative remediation processing. As an example, a brand manager may explore product data (perhaps in a mail-order catalog) and discover entries that are nonstandard. A marketer may see incomplete and redundant records in "leads and prospects" data just purchased. (The data catalog's tool should enable users to annotate errant records. The catalog then sends an alert to a technical user, who corrects the data.) A business user might make repairs to data if the catalog infrastructure provides appropriate self-service functionality.

NUMBER SEVEN

DEPEND ON DATA CATALOG FUNCTIONALITY TO AUTOMATE PROCESSES FOR DATA GOVERNANCE, STEWARDSHIP, AND CURATION

Data governance (DG) is usually manifested as a committee or board of both business and technology people. They establish a process for creating and enforcing policies concerning the capture, storage, access, repurposing, and end-user use of data, whether from traditional sources or new big data sources. Note that a mature and comprehensive data governance program serves two goals:

- **Business compliance for regulations, data privacy, and security.** This goal is about the control of data and its use, with a focus on reducing business liability and risk relative to data management.
- **Technical standards for data and data management solutions.** This involves the communal creation of enterprisewide standards for data quality metrics, data models, exchange formats, metadata, semantics, and data-driven development processes.

Although DG is very much a people-driven process, it increasingly depends on software automation to make governance more collaborative (as in Number Six), automatic (by monitoring data's condition and users' adherence to policies), effective (by raising compliance rates), and scalable (to the end-to-end governance of potentially hundreds of data sets and sources across an entire enterprise). An enterprise-scope data catalog and its integrated data-management infrastructure can provide much of the automation that a growing data governance program needs to ascend to the next level, as in these use cases:

The catalog as an inventory of governable data. For many governors, stewards, and curators, the first step is to create an inventory of data to be controlled, monitored, and improved. An enterprise-scope data catalog provides such an inventory.

The catalog as a collaboration hub for governance. The collaboration best practices described in the previous section of this report can apply to data governance, especially collaborative data stewardship and collaborative remediation processing.

Quality metrics recorded by a catalog. Stewards, governors, curators, and other users can report and analyze the information gathered by data monitoring via a dashboard built into a data quality tool or catalog platform. A comprehensive data quality dashboard is indispensable because its graphs and charts provide at-a-glance views of key data quality metrics and breached thresholds for data standards, thereby enabling users to quickly pinpoint and address issues.

Metadata-driven governance. Technical and business metadata, as managed by a data catalog and its platform, have proved to be instrumental in cataloging enterprise data as a first step in establishing a DG program. Operational metadata can track data lineage and data access that apply directly to understanding the movement of sensitive data as well as policing access to it.

Intelligent catalog automation in the service of DG. The cause of DG is furthered by some AI/ML algorithms and other tool logic that automate data cataloging functions, namely those for data anomaly detection, data lineage (as an audit trail), and the automatic classification of data domains prone to privacy and compliance regulations.

Technology integrated with the data catalog. Integrating a catalog with a data quality tool is a common first step in this direction because it is an achievable goal and delivers a tangible return on the effort. Catalogs also integrate with tools and repositories for other catalogs, metadata management, data integration, and a long list of tools for analytics, reporting, and data management. An emerging best practice is to integrate a data catalog with a data governance application; this greatly extends the reach of DG automation in all directions, including direct interaction with enterprise IT systems. Integrating a DG application with the data catalog provides a closed loop, spanning governance policy definition, management, and adherence validation.

ABOUT OUR SPONSOR

informatica.com

Informatica is the only enterprise cloud data management leader that accelerates data-driven digital transformation. Informatica enables companies to unleash the power of data to fuel innovation, become more agile, and realize new growth opportunities, resulting in intelligent market disruptions. With over 7,000 customers worldwide, Informatica is the trusted leader in enterprise cloud data management.

Businesses must transform to stay relevant and data holds the answers. We're prepared to help you intelligently lead—in any sector, category, or niche. Because our focus is 100% on everything data, we offer the versatility needed to succeed. We invite you to explore all that Informatica has to offer—and unleash the power of data to drive your next intelligent disruption.

With full access to data, IT roles shift dramatically to become more strategic, more essential, to become partners in leading the business. With Informatica, you can use data to develop new business models and capture growth opportunities. We enable you to unleash the power of data to intelligently disrupt your industry.

The Informatica Intelligent Data Platform is the industry's most complete and modular solution, built on a microservices architecture, to help companies unleash the power and value of all data across the hybrid enterprise. The AI-driven platform spans on-premises, cloud and big data anywhere—ensuring data is trusted, secure, governed, accessible, timely, relevant and actionable. This enables the world's most progressive companies to deliver data-driven digital transformation outcomes.

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 data management 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. His Ph.D. is from Yale. 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.

ABOUT TDWI CHECKLIST REPORTS

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, 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.