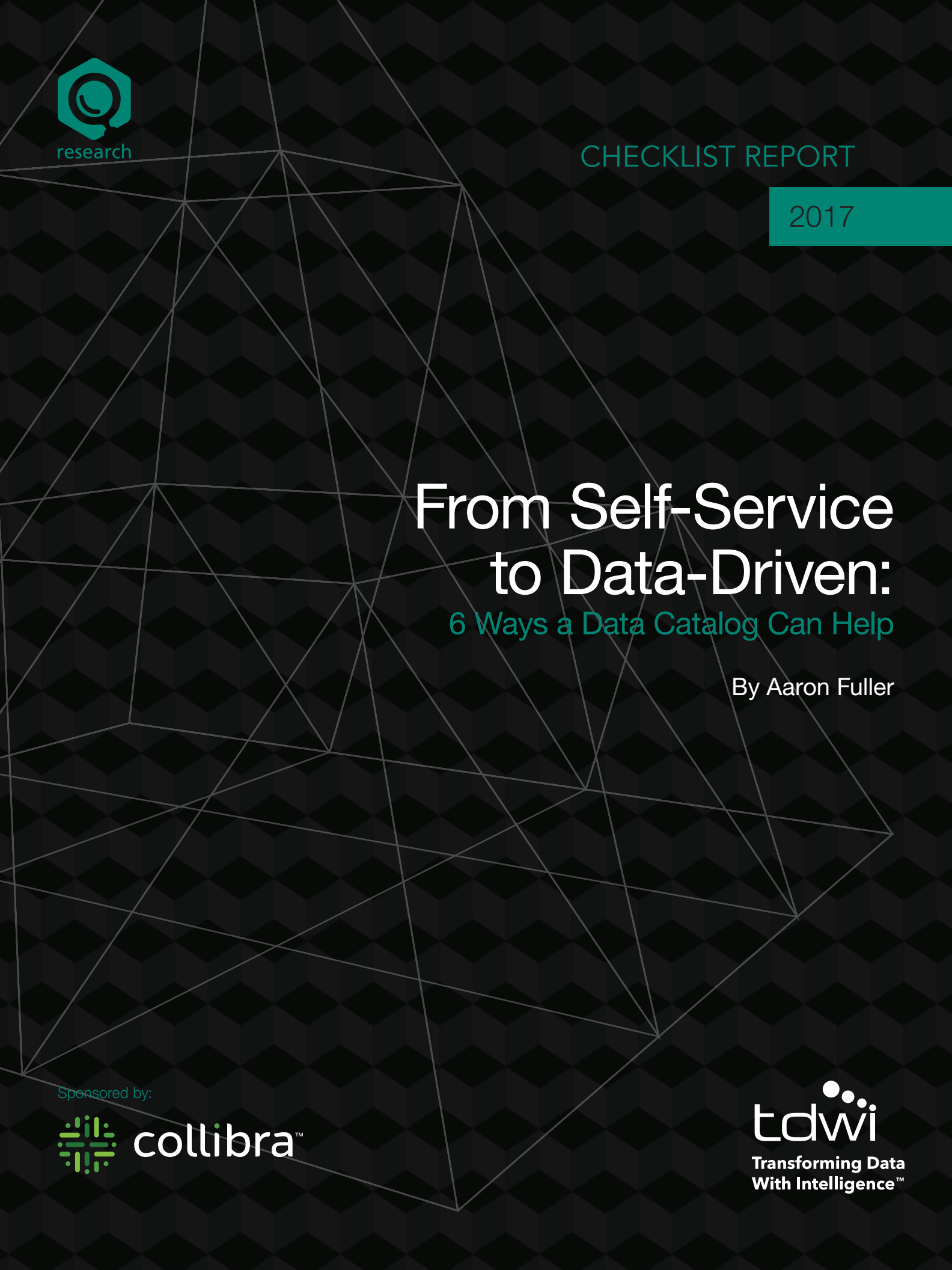

research

CHECKLIST REPORT

2017

From Self-Service to Data-Driven:

6 Ways a Data Catalog Can Help

By Aaron Fuller

Sponsored by:

collibra™

tdwi

Transforming Data With Intelligence™

NOVEMBER 2017

TDWI CHECKLIST REPORT

From Self-Service to Data-Driven:

6 Ways a Data Catalog Can Help

By Aaron Fuller

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

Businesses that built their success on data know something that other organizations do not. They know that to become truly data-driven, they must engage with a broad audience of business users to ensure that all data is easy to find, understand, and trust. However, too often these conversations about data turn technical. IT uses terms such as *metadata management* when engaging with the business about data—and that is part of the problem. Describing the challenge as metadata makes becoming data-driven sound like a primarily technical, IT-driven pursuit. It shouldn't be.

Today, many organizations believe that providing data governance, data management, and BI/analytics teams with a common workspace for capturing a shared understanding of the enterprise's data assets shows promising results in building trust in BI and analytics applications. However, in a 2016 TDWI survey on data preparation, 57 percent of respondents said they were not satisfied with how easily they can find relevant data and understand how to use it appropriately for BI and analytics.[1]

As the importance of self-service solutions for BI, analytics, and data preparation continues to grow, the emphasis is no longer only on centralized, full-time data professionals and their institutional, remembered knowledge of the organization's data. Clearly, these data professionals still have a critical role to play. However, because the bigger goal is to enable the business units to build many or most of their own BI and analytics applications, we also must find a way to support the business' needs with better-documented data assets. This gives rise to the idea of engaging with the business to develop a single point of access for "data" assets (including raw data, reports, metrics, and algorithms) with a consumer-friendly experience. For many organizations, this is a data catalog.

Organizations today need not only full-time architects, analysts, and developers to collect and maintain the wide variety of information that needs to exist in the catalog—they need business data stewards to be the primary driving force in creating and maintaining it. That's why ensuring the data catalog is user-friendly is key to driving broad adoption. Clearly, developing BI and analytics is no longer a one-way street from IT to the business; it goes both ways and everyone must manage the information about data.

The following six tactics will demonstrate why it is critical that every organization implement a data catalog and how your enterprise can create value with this important investment.

[1] For more information, see the 2016 *TDWI Best Practices Report: Improving Data Preparation for Business Analytics*, online at www.tdwi.org/bpreports.

NUMBER ONE

DON'T CALL IT METADATA! MAKE IT BUSINESS-RELEVANT

Metadata management is crucial but the term itself is a turn-off for the broader business audience. When we communicate the need for a managed inventory of all the information we know about our data assets, we will be able to show our business participants that data governance is a business concern, not a technical issue.

Having great metadata means having sustainable wisdom about your data. Too often this wisdom is locked in the minds of the longest-serving, most in-demand subject matter experts in the company and the organization struggles to disperse that knowledge to everyone who needs it. Companies need to move beyond the wisdom of individuals to focus on the wisdom of crowds. They must think about the data audience the same way they think about customers and tailor the data user journey and experience accordingly. For many companies, it's easier said than done.

NUMBER TWO

ENABLE A CONSUMER-FRIENDLY APPROACH: DATA SHOPPING

Data shopping is like shopping for anything else—you need to be informed. It's inevitable that power users in most companies will need to pull together multiple data sources and prepare them for their projects. However, in the 2016 TDWI research report on data science and big data[2], only nine percent of surveyed users felt very satisfied with their access to integrated data from multiple sources. In modern business, users have access to an incredible number of potential data sources. Their ability to be a wise data shopper depends on their ability to easily find the data they need and quickly assess its quality, usage, and value. Without a good data catalog, business data users simply don't know where to begin. This lack of information makes it risky to be a data consumer.

For example, imagine a retail marketing analyst who has been asked to advise her department's directors regarding which specific products should be promoted in an upcoming campaign focused on attracting younger customers. She knows that a wide variety of different data sources will provide valuable input to her project, including internal sales transactions, market-share information, customer and store demographics and geography, product classifications, and so on. If no central repository of knowledge about the company's data exists,

[2] For more information, see the 2016 *TDWI Best Practices Report: Data Science and Big Data*, online at www.tdwi.org/bpreports.

she'll have to begin by hunting and gathering information from the far corners of the organization. She'll talk to subject matter experts (if she can get time on their calendars) or she'll simply explore the wilderness of data assets by gaining access to project directories, databases, and support documentation.

Now imagine that her company has a data catalog. The analyst can begin her work by searching and browsing the catalog for all the company's current data sources that may apply to her effort. When she finds data that looks relevant, she'll be able to see whether it is of high quality and who else has used the data for similar projects. Assuming it meets her criteria, she can check out or request access to the data by adding it to her data shopping basket.

A consumer-oriented approach leads to better support from central data management/IT staff. By taking the perspective of a consumer, data management staff can consider all the possible ways the business would use the catalog. They can determine how best to enable users to browse, search, link, and recommend within the data catalog repository. By keeping this perspective, data management staff can better understand the importance of having a good catalog and ensure the capabilities exist for a variety of user needs. In short, it serves as a point of collaboration between the business and IT.

NUMBER THREE

CREATE THE WORKSPACE FOR BUILDING TRUST IN DATA

Organizations struggle with the abstract nature of data governance. *Governance* is developing trust in data by ensuring that the business receives consistent, compliant, useful, and well-understood data for decision making. Although most companies have tried to improve their data governance situations over the last decade, success rates remain stagnant. In 2008, leading industry analysts estimated a 90 percent failure rate for governance initiatives[3]. Why? Because too many organizations implement governance as a way to control and lock down the data. The organizations that *do* succeed know that they need to think differently about governance—to use it as a way to open up the data so more people can use it to drive efficiencies, uncover insight, and gain a competitive edge.

A data catalog is critical to building and showing the value of governance. By making the creation, expansion, improvement, and maintenance of the data catalog a central responsibility of the data governance program, participants have a shared place to build a tangible asset. Collaboration capabilities of the governance/catalog tools allow all users to keep track of discussions about particular data elements or groups of data over time. This record aids them in in their current efforts and establishes an important historical repository of all the important data collaboration that has occurred.

Data lineage drives trust in the data. A shared understanding of definitions, combined with information about the lineage of data flows, allows your organization to conduct an impact analysis when business processes and systems change. Depending on your data integration and BI and analytics architectures and tools, completing this type of analysis may be a challenge. However, data catalog users will often expect to be able to research the flow of data and its definitions.

NUMBER FOUR

CLARIFY MEANING OF DATA ACROSS BUSINESS AREAS

Misunderstanding data definitions is a huge, costly problem for every organization. There are often good reasons why different business areas need to look at similar data in slightly different ways. However, the general insistence on using shorthand terminology—and the lack of interest from business leaders in ensuring enterprise consistency of definitions—has led to widespread semantic misunderstandings.

Data catalogs help the business to collaborate and clarify these definitions. In a data catalog, business data concepts must have unique identifiers. This requirement forces the business areas to surface their disagreements about terminology, calculations, proper data sourcing, and constraint rules in a constructive way. Collaboration up front saves time and prevents discord, confusion, and data brawls after a report is run.

Returning to our marketing analyst example, consider that she may need to clarify the definitions of several customer demographic data elements that she plans to use in her analysis. If there is a data governance program that promotes the use of a data catalog, she will have a place and process to work with subject matter experts around the company to answer her questions. She might open a new discussion thread related to the specific group of data elements she's interested in. The data owner and steward of the data would see it, and (hopefully) answer her questions quickly. If the analyst uncovered a significant problem with the definitions, she would have a documented process through which she can resolve the data issues.

Without data governance and a data catalog, the marketing analyst would instead need to establish these definitions for her project without the advantage of tapping into the wisdom of the rest of the company. These extra steps could cause delays in her project, or worse, the inability to deliver a complete, trustworthy analysis.

[3] http://www.eiminstitute.org/library/eimi-archives/volume-3-issue-7-july-2009/201chouston-we-have-a-problem201d-2013-why-data-governance-programs-fail

NUMBER FIVE

COLLECT SUBJECTIVE INFORMATION ABOUT DATA QUALITY

In tandem with data lineage, data quality is another natural next step in the catalog. When business data stewards, owners, and SMEs start to look at the definitions of the data elements, they naturally begin to wonder about and investigate the quality of information. After all, if the quality of a data element is not high enough to suit the purposes it is meant for, it doesn't really conform to the definition in the catalog.

The catalog is a good place to gather perceptions about data quality. It is up to business users to decide what levels of quality are acceptable for various types of data. This is where the ability of data catalog software to enable collaboration and provide a flexible repository for what you know about your data really shines. For example, if our marketing analyst can look at the history of discussions and concerns documented in the collaboration features of the data catalog, then she will have a much easier time understanding the potential pitfalls of the data sources and elements she's planning to use in her analysis. Without this ability, she'll spend time trying to figure out which individuals she should reach out to for history of the issues related to this data rather than on the analysis itself.

Objective measures of data quality are associated with the catalog as well. This capability is unlikely to be available for all data elements, but the most often-used and enterprise-critical data should be constantly measured for its quality. Giving data users the ability to evaluate both objective metrics as well as user experiences—in terms they understand—allows them to quickly decide whether the data fits their purpose, what it will mean exactly, and how well it can be trusted.

Our marketing analyst will certainly want to review data quality metrics trends for any of the data she plans to use in her project. If data cataloging, governance collaboration, and data quality monitoring capabilities are all present, she can navigate through her full thought process naturally. First, what are the possible data sources and their elements, and what do they mean? Second, where does the data come from and how was it transformed throughout the data flow? Third, what do the people who are responsible for this data say about its state and what other users might have experiences from which she can learn? Finally, what is the objective quality of the key metrics and attributes for her project? Can she depend on her desired data sources to be of high enough quality to properly support the project?

NUMBER SIX

BUILD TRUST AMONG ALL DATA STAKEHOLDERS THROUGH CAREFUL MANAGEMENT OF DATA ASSETS

Trust is based on understanding. When data consumers see that the company takes its data assets seriously enough to ensure their proper management through investments in tools, processes, and people, it builds trust in the BI and analytics capabilities themselves. Simply put: providing the business with a good data catalog makes data easy to find, easy to understand, and easy to trust. In a recent survey, only 44 percent of organizations indicate trust in data to make important business decisions, and the average C-level executive estimates that one-third of their data is inaccurate[4]. Successful companies will make the right investments to build a higher level of trust.

Trusted data assets are easier to use, allowing power users to focus on developing advanced analytics capabilities. Trusted data allows the user to be more agile and aggressive in implementing higher-level analytics capabilities such as machine learning, predictive analytics, decision automation, and artificial intelligence. The ability to trust existing data assets allows these projects to quickly speed through the iterations of analytics modeling needed to develop strong capabilities.

Data scientists typically spend 80 percent of their time collecting, understanding, and preparing the data they need and 20 percent of their time actually doing analytics modeling. If we can flip that ratio to 20/80—or even bring it down to 50/50—we can provide enormous value. Data scientists are some of the most valuable company resources—a value that isn't derived from data preparation work but from analytics models they produce and the insights they reveal.

Users aren't the only stakeholders in an organization's data management efforts. Customers, partners, and third parties demand that companies ensure the security of confidential data and a reasonably high level of data quality[5]. Data catalogs are not simply making decisions better and faster within organizations; they ensure customers and partners trust the company, too. Long term, that means they are willing to share more data that, in turn, enables the business to drive greater value.

Good decision makers know it is better to have no data than bad data. Experience alone is more trustworthy than experience plus bad information, and experience plus good data is the way to improve decision making. It is our responsibility as data management professionals to implement the trustworthy tools and processes that enable our business decision makers to be bold. The data catalog is a pillar of building this trust.

[4] https://www.edq.com/globalassets/white-papers/2017-global-data-management-benchmark-report.pdf

[5] For more information, see the 2016 *TDWI Best Practices Report: Improving Data Preparation for Business Analytics*, online at www.tdwi.org/bpreports.

ABOUT OUR SPONSOR

collibra™

collibra.com

As the leader in data governance for business users, **Collibra** helps organizations across the world gain competitive advantage by maximizing the value of their data across the enterprise. Collibra is the only platform purpose-built to address the gamut of data stewardship, governance, and management needs of the most complex, data-intensive industries. Our flexible and configurable cloud-based or on-premises solution puts people and processes first— automating data governance and management to quickly and securely deliver trusted data to the business users who need it. Learn more at www.collibra.com.

ABOUT THE AUTHOR

Aaron Fuller, CBIP, is the principal and owner of Superior Data Strategies, LLC. Aaron is an experienced enterprise architect, data architect, data modeler, business analyst, project manager, and Certified Business Intelligence Professional who has developed a keen eye for transforming data into relatable, effective information solutions for businesses. Superior Data Strategies focuses on data warehousing, dimensional data marts, operational data stores, and data integration in a variety of industries. You can reach him at aaron@superiordatastrategies.com and @AaronTheDataGuy on Twitter.

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.