



Data Governance Practices for Cyber Risk Management

By Evan Wheeler

Ask a cyber risk professional about data governance practices, and they will likely tell you tales of classification schemes, access controls, and encryption ... but we often overlook the importance of data quality, integrity, and usability that are core tenets of a robust data governance process. Especially as risk professionals, we rely so heavily on data that is sourced both throughout our organization and from industry partners. No matter how sophisticated your analysis methodology is, garbage data in often means garbage results coming out.

Data is the lifeblood of the business. Don't delay implementing basic data governance

Start-ups, especially, can be caught in the push-pull of needing to be light and nimble to get ahead of the entrenched firms and at the same time having high client or regulatory demands that are data heavy. Often investments in data governance structures are deferred or entirely overlooked by start-ups.

However, it is a reality that **data is truly the lifeblood of the business**, and it can be a costly mistake to delay implementing some basic data governance practices early on.

About the Author

Evan Wheeler is VP of Risk Management at Fintech firm NVDR, Inc., and a FAIR Institute Advisory Board Member. He has spoken at several FAIR Conferences, taught risk at the graduate level at UCLA and other universities and is the author of *Security Risk Management* (Elsevier).

To put it into perspective, the Chair of a company's Board Risk Committee was once asked in a town hall meeting if he worried about the risks of an ongoing migration of their core systems into the cloud, and he responded that he was far more worried that most of the strategic decisions made in the company are based on spreadsheets with no quality controls or validation. Let us not forget that data quality needs to have equal priority to security and privacy when it comes to data governance.

The Risk team should be an example to the rest of the company for robust data governance


Risk teams are in the ideal position not only to shine a light on the data governance gaps, and how those can severely hinder well informed decision-making, but they should also be an example to the rest of the company for what it looks like to implement robust data governance practices by applying these

principles to their own risk programs first. Think about every KRI on your reports to senior leaders – do you know where the supporting data is sourced from, are automated validation checks in place to flag anomalies, and would you know in advance if the data sources were changed or decommissioned?

Let's take a simple example, assume that you're tracking the percentage of company-owned laptops that are encrypted and the endpoint security tool's dashboard reports that agents are deployed to 1,800 laptops and that all 1,800 laptops have full disk encryption enabled. Seems like good news, but wait, how do you know that the company only has 1,800 active laptops? Has this been compared to an asset inventory? What about the decommissioned laptops sitting in a closet waiting to be wiped or destroyed? In this case, measuring control coverage is closely related to the concept of data completeness, which is just one element of data quality.

In some ways this is a good news / bad news story. The good news is that you have automated a bunch of the data inputs into your risk program – take a minute to savor this success. Now put the champagne down and let it sink in that you are entirely dependent on all these data sources, data pipelines, and reporting tools which you likely don't directly control. If data is missing, inaccurate, or simply misinterpreted, then your program could be causing the organization to make some really poor decisions. Anyone who has experienced the embarrassment of having to report back to a risk committee or board that the last month's numbers were materially "off" will do anything to avoid being in that position a second time.

Launch a data governance program



Start with a working group focused on the data the risk management team uses

If a data governance program already exists within your organization, then make sure that your risk management team has a seat at the table and is advocating for more than just more access controls and encryption. Ideally, the data governance program would set rules and standards for data-related matters across the enterprise. It is helpful to establish policies,

standards, and reusable models for how to manage the full lifecycle of data – from gathering to destruction.

If a data governance program/initiative doesn't already exist in your organization, start by creating a working group focused first on the data that the risk management program uses as input, then expand out to other areas of the organization.

Consider these objectives as a starting point for your working group:

- Include representatives from both Business and Technology groups
- Share and coordinate ongoing initiatives related to reporting and data infrastructure
- Clearly define each data element, and set expectations for the types of processes and controls necessary to accurately collect and report on that data

- Formalize controls to restrict access to modify the data, from extraction to final reporting
- Implement change controls to log any manipulations applied to the data, and processes to reconcile reported data back to its source

Whether you're a tiny startup or a well-established player, you will find that there are more data and reporting related initiatives happening than any one person can be aware of, and probably as many different approaches to the same problems. So, the initial goal of the working group is surface these initiatives and address any overlap or conflicts. Then start getting consensus on standards.

Create a data catalog

There are many important aspects of a data governance program, from master data modelling to data warehouse designs, and it is easy to take on too much at first. Let's assume you're looking at the metrics that the risk program reports. When getting started, consider these steps:

1. **Labeling your data is crucial** (i.e., establish a data catalog with clear definitions for each data element)
2. **Create clear paths for data to flow through** (i.e., know where the data is sourced from and how it is transformed along the way)
3. **Be able to govern changes in how data is calculated** (i.e., managing changes to the methodologies for calculating and reporting metrics)

Start by establishing a data catalog or data dictionary and identify your most critical audiences for data. For example, is there a quarterly risk metrics report that goes to the Board? What data elements are in that report, and are there clear definitions for each metric? Where do they source from? What is the QA process for the data being reported? If something changed upstream in how the data was being generated, who would communicate that to stakeholders? Does the owner of the source data even know that this is being used in a regular report to the Board? etc.

Essentially the data catalog becomes the Rosetta Stone of requirements for:

- **Data quality** - start with the master data sets/sources
- **Security & privacy** - start with inventory of most sensitive data elements (like personal information or intellectual property)

Once the data catalog is established, start identifying who owns each data element and where the data is stored, but equally important where it is used and by whom. Some metadata you might want to capture in your data catalog includes:


- Data element name
- Description
- Owner or custodian
- Source data repository or lineage of upstream systems
- How the data is gathered (i.e., user input, external data feed, calculation, etc.)
- Refresh frequency
- Where the data is used (i.e., audience)
- Data classification designation
- Business continuity priority (i.e., priority to restore and/or tolerance for data loss)
- Retention requirements (i.e., how long it must be retained, how readily accessible it needs to be when archived, and whether it should be automatically purged at the end of the retention period)

This probably looks a lot like the metadata that you track for your KRIs/KPIs, and that is no accident. For metrics that use this data, you would also want to capture some additional details:

- **Purpose** – what is the metric intended to measure?
- **Business logic** – what is the “recipe” behind the scenes to generate the metric?
- **Thresholds** – what thresholds have been set, and how are threshold breaches escalated?

If you’re struggling to identify the relevant data sources, consider which reports drive decisions within your organization. Identify those reports and then trace the data lineage back to its source. Eventually data governance needs to happen at the ingestion point to be effective, but you can often work backwards from the reports to the master data sources.

Ensure data quality



It is critical to implement quality controls at each stage of the data sourcing flow.

Oversight of data quality is a logical starting point, but ultimately, we care about *information quality* for decision-making. Even with the highest quality data, we can lose information quality if the data isn’t relevant or is misinterpreted.

Data governance controls should cover:

- **Data** – controls that help to understand the accuracy, completeness and timeliness of raw data used for analysis
- **Analysis** – controls that help to ensure that data are correctly interpreted and turned into accurate insights
- **Reporting** – controls that provide useful analysis results to stakeholders in time to support decision-making

When we think about analysis and reporting of data, we may forget that there is often “code” in the background that is joining, querying, summarizing, transforming, and maybe calculating values based on some raw data that we will use in our analysis. It is critical to implement quality controls at each stage of the data sourcing flow. There are two aspects to ensuring the quality of information created from raw data:

1. **Quality assurance** - was the data gathered correctly?
2. **User acceptance** - does the information answer the question?


If the provided information is in the form of a dataset, tests should be run and documented to provide confidence that the result meets the requestor’s requirements. Exception tests and regression tests should be run regularly to minimize the likelihood of errors being introduced during any transformation:

- **Exception tests** look for prohibited data relationships and set thresholds for expected data types or ranges (e.g., you have ten times as many laptops as employees, or a laptop is associated with more than one employee).
- **Regression tests** examine data assets before/after code changes.

Any tests that were useful during QA testing of the data would great candidates to be automated and applied to dataset updates before they’re ingested into your systems.

Along the way you may consider “coding around” an issue in upstream data quality (i.e., to fix bad data during your transformation process), but be very cautious. It almost always pays off in the long run to fix the issues at the data source rather than applying a band-aid that will likely blow up in the future.

Data quality supports risk management



Ever FAIR factor is prone to data quality issues, depending on source and use of the data.

Clearly the discipline of risk analysis is highly dependent on data, and the most mature programs strive to be even more data-driven and evidence-based. But this comes with an obligation to scrutinize the data and implement controls to quickly identify anomalies or inconsistencies in the data feeds.

In fact, poor quality data isn’t even necessarily a problem if we can recognize it, account for the lower confidence in our modeling, and reflect it in our reporting.

Thinking about data through the lens of Factor Analysis of Information Risk (FAIR™) lens, every factor in the ontology is prone to data quality issues, and the considerations will vary depending on the source of the data and how it is being used. Let’s touch on a few examples.

Loss Magnitude – creating reusable loss tables is a staple of a risk analysis function, but over time we can easily lose the traceability for the source of these ranges and the quality may degrade.

For example, assume you have a range of potential PCI violation fines as \$5,000 to \$100,000 a month, depending on factors like the size of your business and the length and degree of your non-compliance. You've been happily applying this range to your scenarios involving breaches of credit card data, but wait, this data was gathered from an accounting firm's report back in 2017 when you first kicked off your risk program, and hasn't been updated since. So many questions come to mind: 1) was the source credible in the first place, 2) was this based on data of similar firms to yours, 3) has anything significantly changed in the enforcement landscape that these ranges need to be updated, etc. It is so critical that you capture the source of the data being used in your loss tables and maintain a basic change log. You'll need to consider what is an appropriate trigger or frequency to prompt updating the loss tables, and how to update your previously scenario assessments if the range changes significantly.

Another input to your loss tables may be your own internal incident data. This may or may not be captured in a system over which you have direct control, and likely isn't managed by a process that you own. What if the incident management team (or another risk team) changes the definition for what is classified as an incident, or decides to stop tracking incidents with an impact below a certain threshold – would you know and be able to account for this in your loss tables?

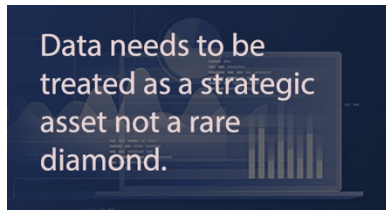
Loss Event Frequency – by far, the largest volume and the most frequently sourced data is related to this branch of the FAIR ontology. By now the potential issues are likely clear, but let's discuss a few examples to reinforce the points about data quality.

For example, you might be leveraging a threat profile for a hacker group that was developed by another team – you would want to apply the same scrutiny that we discussed previously for the industry reports feeding the loss tables. Take time to validate that the data being provided is credible and also that it is relevant to your risk scenario. Don't assume that someone who is analyzing threat groups will be familiar with your use case for a threat actor's motivation or capability. Explicitly ask them to rate their confidence level. Think about how this will be maintained and updated.

Let's assume that you're measuring the effectiveness of a control – this could be through observation of its operation during normal conditions or through simulated testing or sampling. This could involve sourcing data from multiple different security tools, system logs, and asset inventories just to measure the effectiveness of one control. All of these sources need to be in sync and cross-checked for integrity. Too often we're just so thrilled to get access to the data in the first place, we don't follow a thorough vetting process before incorporating it into our assessments. Clearly the degree of QA testing and ongoing validation recommended in this paper isn't scalable manually, and automation is key.

A lot of time is spent discussing the reliability of subject-matter expert sourced data, and it is easy to overlook the issues that can arise from the many other data sources (automated and not) that we rely on for our assessments and metrics.

Security and privacy in data management



So often in cybersecurity circles you'll hear references to the "crown jewels" that need to be identified and protected. The concepts of *need to know* and *least privilege* are foundational principles for most cybersecurity programs.

Unfortunately, this leads to a mentality of locking it away in a vault that is very difficult to access – in fact the harder it is to access the data, the safer we feel.

Let's face it, data isn't a rare crown jewel that you bring out only on special occasions, it is the lifeblood of our organizations. Even the ransomware criminals have realized this and are capitalizing on it. Data is flowing through our systems every millisecond of every day, and in and out of our environments just as often. We need a new analogy. **Data needs to be treated as a strategic asset**, not a rare diamond. Of course, we want to control access to sensitive datasets and limit exposure, but we also need to shift our culture to be thinking about how we enable data sharing in a responsible manner. Instead of the need-to-know and least-privilege principles, let's recognize that **data exists to be shared**. We should assume data will be shared as a default. So how do we enable this without exposing the organization to unnecessary risk?

Technical solutions will include many techniques of data segmentation, de-identification, tokenization, etc. that need to be part of the fundamental design of our operational processes and technologies. But even more fundamentally, the risk management function should be an advocate for thoughtful data governance. As already discussed, this goes far beyond the cybersecurity or privacy considerations (although these are central to an effective data governance structure) by also considering the many aspects of data quality, integrity, and usability.

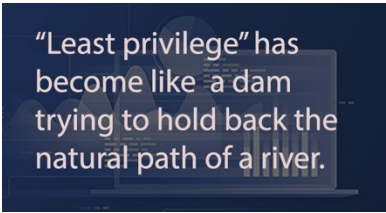
Let's start by describing what the risk management program can design to help the rest of the organization succeed and make their jobs easier – by anticipating friction points and reducing them.

No matter what industry you're in or in which countries you operate, there are multiple laws and regulations related to data protection and appropriate use. In 2018, global companies implemented many new privacy controls to comply with GDPR, and in 2020 companies with employees or clients living in California had another set of requirements with which to comply. Some of these overlap, but many of them either define the scope of what is considered *personal information* differently or have different requirements. The evolving privacy laws are

also demanding that consumers have more transparency into the data we're collecting, how we're using it, and in some cases gives them a right of deletion. A data governance program should be preaching the advantages of designing our systems with some fundamental privacy concepts in mind and anticipate that these systems will need to adapt as the laws evolve.

Privacy by design principles:

- Let's collect as little personal information as possible
- Let's find ways to make data available for business analysis and research by de-identifying
- Let's create safe and contained environments to work with the sensitive datasets
- Let's make it easy to do the right thing ...
 - Make handling requirements easy to remember and follow
 - Define roles and access to match job functions
 - Some information (like passwords) should go beyond the baseline controls and should be treated as "zero knowledge"



"Least privilege" has become like a dam trying to hold back the natural path of a river.

We know the threats are real ... so why challenge the concepts of "need to know" and "least privilege" in such a hostile environment? Because the implementations of these principles have become like dams trying to hold back the natural path of a river to build a town. Let's just build the town next to the river instead and use the river to our advantage.

If we assume data will freely flow throughout the organization, what measures can we build to de-identify the datasets without reducing the value to the business. Does your product team want to know the age groups of clients who are accessing your site most often? Don't give them access to a database with every client's name and full date of birth ... have a dataset already created (and even provide access through an API) to provide demographics such as age ranges with a client GUID (global unique identifier) that they can freely analyze and share internally. You never need to expose that full date of birth to the product or marketing teams.

As we're designing systems and processes, we can greatly reduce our potential exposure by following these simple guidelines:

- Justify any personal data collection
- Use segmentation, obfuscation, and anonymization to lower the sensitivity of datasets
- Mask or redact sensitive data when practical
- Minimize storage of paper documents - digitize when practical
- Maintain audit logs to track access to sensitive data
- Don't keep data longer than it is useful

Going back to our concept of privacy, emerging laws will prohibit even collecting personal information from consumers without a clear business use defined. So, we can't just grab or keep data "just in case".

Data aggregation considerations:

It is common to just classify the sensitivity of each individual data element on its own, but it is also important to think about the varying sensitivity of aggregated datasets. A well-known study¹ showed that it's possible to personally identify 87% of the U.S. population based on just three data points:

- Five-digit ZIP code
- Gender
- Date of birth

So, whenever we are evaluating the sensitivity of data or classifying it, we need to consider the changing sensitivity of data based on the context of the dataset – not just a static label on each individual data element. This can make anonymizing data sources challenging, but it is worth the upfront investment.

Data governance use cases

There is no end to the use cases for when and how to apply these data governance principles. Let's consider just a few.

When you're designing a new product or service, ask data questions such as:

- Which internal databases have this information?
- Who has access to it?
- Have we clearly defined what we call a 'customer' or a 'household'?
- Are the structures of billing data already defined?
- What is the quality of the source data?
- Which business metrics will be supported by this data?

If your project involves migrating data, converting data, replacing an existing technology tool, or integrating a new data source, then you should follow these guidelines:

- What is the QA process for the data?
 - *Can you get a 100% data copy for testing, if not what is the sample size or how similar is the test data to production?*
 - *What volume of data is required for adequate testing? A couple hours of data, a few days, months, or years?*
- Compare feature parity - plan to address each gap
- There must be both roll-back and fallback plans

¹ <https://dataprivacylab.org/projects/identifiability/paper1.pdf>

- For third-party services, how will you validate their documentation?
 - *An API might claim to provide a certain calculation or composition of data. Can we pull data from the source and test this? Don't just take them at their word.*
- Have all downstream data receivers and reports been identified in the design or technical specification?
 - *Have all changes in data calculations, presentation, or meaning been communicated in advance to applicable clients and our own support staff?*
- Has the existing solution been analyzed for custom rules, hard-coded behavior, and other exceptions?
 - *Does the technical specification address all these cases, and will each be tested?*
- What is the smoke-testing plan post-implementation?
- Post-implementation, what monitoring and validation will be performed?
 - *Extra monitoring controls or delays on processing might be needed to allow time for finding and resolving errors.*
 - *Can it be run in parallel for some time period?*
- What metrics or reporting will track data quality pre- or post-implementation?
- When data is imported or transferred, how will you track data that has to be excluded or fails quality checks?
 - *What's the procedure to handle/resolve these issues?*
- Are any client accounts, transactions, communications, or other activities being suppressed or on hold in legacy system that need to be considered?
 - *What is the plan to address these pre- and post-implementation?*
- How long will legacy data source or system be retained or kept online in case we need to investigate issues?

Data transformations are pervasive in every organization, and just because they might be automated, we can't assume they can be trusted. Several things can go wrong in the transformation of data:

- There could be an error (or bug) in the transformation itself
- The raw data from the upstream system might be corrupted
- The transformation process or the meaning of the end result may be misunderstood

Each of these examples hopefully gives you some ideas for risk scenarios to consider in your organization, whether they cover a new product, a migration project, or just assessing an established process.

Risk team's leadership role in oversight

The Risk team can play a central role in sponsoring a data governance program. In many organizations, the Risk team is well positioned to initially set the standards for:

- Understanding the usage of data in the business
- Setting and enforcing policies and standards

- Creating clear and unambiguous definitions of data, and reconciling conflicting definitions
- Defining roles and responsibilities, and sponsor training
- Monitoring data quality and sponsoring root-cause investigations when problems arise



Start by applying these standards to the data that the Risk team analyzes and reports on regularly and set the example for the rest of the organization. As the value of the program becomes clear to the business, the Risk team can help identify critical business processes and top risks that should be high priorities for data governance. Depending on the structure of your risk

program, the Risk team or an Internal Audit function could also validate the effectiveness of data governance controls.

There is a tendency for business leaders to view data governance as a one-time project or initiative, but it is crucial to help them to understand that it is an ongoing need that will evolve over time – in essence it is a program to support more informed decision-making. What could be more important?

We should be thinking about the value of data (or more precisely the value of “information” that the data supports) in terms of its role in decision-making, and we should retire the paradigm of the crown jewels.

A good data governance program can shape a firm’s decision-making process, but it isn’t a small undertaking. It’s a cultural shift that requires both business and technology sides of the organization to come together to define data elements and the rules that will govern this data across the enterprise. Do this early and the business will reap the rewards.

Copyright 2021 FAIR Institute. All rights reserved. No part of this paper may be reproduced or transmitted in any form or by any means, electronic or mechanical, without the written permission of the copyright holder. For permission requests, please [contact the FAIR Institute](#).