

# Metadata for BI and Analytics Solutions



## Executive Summary:

Whereas data is potential information, metadata is a statement about that potential information. There are several distinct categories of metadata that make different kinds of statements -- descriptive, administrative, structural, markup languages, and use metadata -- with numerous subcategories. A metadata schema, also known as an element set or data dictionary, provides a structured framework for metadata. Within the field of business intelligence (BI), metadata is essential for extracting, transforming, and loading data into a data warehouse, and for analyzing data once it's there. In BI, metadata can be leveraged to support governance, risk management and compliance (GRC), launch automation, enable chargeback, facilitate upgrades and migrations, organize content for the purpose of monitoring, and provide insight into the adoption of BI tools. Most organizations benefit from third-party metadata services solutions, like 360Suite, that centralize and process metadata inputs and transform them into data for business. What distinguishes 360Suite from a data warehouse is that it extracts only relevant data -- metadata related to BI -- into a data mart, which provides authorized users with easy, fast, reliable, and secure access to metadata that answers their business intelligence questions. As the volume of data grows exponentially, it becomes increasingly difficult to discover and understand the *potential* in potential information. Metadata services represent a new source of information to which other services are able to connect. This, in turn, generates new metadata and creates a greater need for metadata on metadata -- BI on BI -- to make information available for machine learning, artificial intelligence, and business analytics.

<b>What is Metadata?</b>	<b>4</b>
<b>Categories of Metadata</b>	<b>5</b>
Descriptive Metadata	5
Administrative Metadata	6
Subcategory: Technical Metadata	6
Subcategory: Provenance Metadata	6
Subcategory: Preservation Metadata	6
Subcategory: Rights Metadata	7
Structural Metadata	7
Markup Languages	7
Use Metadata	8
Subcategory: Data Exhaust	8
Subcategory: Paradata	8
<b>Metadata Schemas</b>	<b>9</b>
Metadata in Business Analytics	10
The Importance of BI Metadata	11
Governance, Risk Management, and Compliance	11
Automation	12
Chargeback	12
Upgrades and migrations	13
Monitoring	14
Adoption	14
A Word on Non-usage	14
<b>BI Metadata Collection</b>	<b>15</b>
<b>Third-party Metadata Services</b>	<b>16</b>
<b>Conclusion</b>	<b>17</b>

## What is Metadata?

Metadata is all around us and most of the time we aren't even aware of it. So what exactly is metadata? The word meta (Greek for "beyond") has been embraced by popular culture to mean "about the thing itself." For example, metafiction is a self-conscious literary style that alludes to the artificiality or literariness of a work. In other words, it's "fiction about fiction." Similarly, metalanguage is words and symbols for talking about language, or "language about language." Following this logic, metadata translates to "data on data," and that's indeed a popular definition. But what does it really mean?

It's impossible to define metadata without first assigning a meaning to data. For the purposes of this paper, data refers to objects (i.e., facts) that can be processed to yield useful information. In other words, data is *potential information*. Information scientist and professor Jeffrey Pomerantz, author of the book *Metadata*, published by The MIT Press, defines metadata as "a statement about a potentially informative object."<sup>1</sup> Perhaps the simplest way to define metadata is with an example.

Consider a telephone call. In most cases, the spoken words are the data -- potential information (assuming you can understand the language spoken and the meaning of the speaker). Metadata are statements about that data: the time of the call, duration, completion status, source number, identification numbers associated with the exchange (set of equipment that connects telephone lines), and route by which a call entered and left the exchange.

A telephone call is a good example because it highlights just how revealing metadata can be. By providing statements about a potentially informative object -- in this case a telephone call -- metadata can often expose the information exchanged in the call (the data) to a surprising degree. For example, in a 2016 study of telephone metadata involving 800 volunteers, Stanford University computer scientists were able to infer private information, such as health details, at the

---

<sup>1</sup> Pomerantz, Jeffrey. *Metadata*. The MIT Press, 2015, 26.

individual level from metadata alone.<sup>2</sup> That's why many people were outraged when classified information leaked by Edward Snowden brought to light the extent to which the U.S. National Security Agency (NSA) was collecting telephone metadata in bulk from citizens and allies. They understood that, what constituted metadata to most people (phone numbers, call durations, relays, etc.), was actually data for the NSA. This prompted Congress to pass the USA Freedom Act in 2015, which banned the bulk collection of telephone metadata (but allowed for a more targeted approach) by the U.S. intelligence community.

Despite this simple example, metadata isn't simple at all. For one thing, metadata is itself a type of data -- potential information. So what's the difference between data and metadata? The line can be fuzzy, but it generally depends on the purpose of the information. Is it to provide content or context? Another aspect that complicates metadata is the number of different metadata categories and schema. Add to that the abstract nature of metadata and it's no surprise that when two people are discussing metadata, they may or may not be talking about the same thing.

## Categories of Metadata

The National Information Standards Organization (NISO) publication, "Metadata: What Is Metadata, and What Is It For?" refers to four distinct categories of metadata: descriptive, administrative, structural, and markup languages.<sup>3</sup> Jeffrey Pomerantz adds a fifth: use metadata.

## Descriptive Metadata

Descriptive metadata is the simplest kind of metadata. It provides descriptive information about the characteristics or attributes of a resource to aid in "data discovery" -- finding or understanding it. For example, data catalogs rely on descriptive metadata to locate individual items.

---

<sup>2</sup> Mayer, Jonathan, Patrick Mutchler, and John C. Mitchell. "Evaluating the privacy properties of telephone metadata." *Proceedings of the National Academy of Sciences of the United States of America*. May 16, 2016. <https://doi.org/10.1073/pnas.1508081113>

<sup>3</sup> Riley, Jenn. "Understanding Metadata: What Is Metadata and What Is It For?" *National Information Standards Organization, 2017*.

## Administrative Metadata

Administrative metadata is information that relates to the creation of a resource and the management of that resource throughout its life cycle. Because there are so many different types of resources, administrative metadata is a big category with numerous subcategories.

### Subcategory: Technical Metadata

Technical metadata is a type of administrative metadata that provides information about the characteristics of a resource. It's often captured automatically by software when digital files are created or modified. For example, computer files contain technical metadata related to format, size, location, etc.

### Subcategory: Provenance Metadata

Provenance metadata is a type of administrative metadata that provides lifecycle context that a user might need to evaluate the history of a resource, including authenticity and credibility. A provenance metadata record is created, often automatically, whenever a digital resource is created or modified to describe entities and processes involved in producing or delivering or otherwise influencing it.<sup>4</sup>

### Subcategory: Preservation Metadata

Preservation metadata is a type of administrative metadata that provides information to support the preservation of digital resources. The goal of preservation metadata is to ensure that a digital object continues to exist, that it can be used, and that the original is distinguishable from subsequent versions. For example, lifecycle management (e.g., upgrades, migrations rely on preservation metadata to identify object formats, promote compatibility, and avoid regressions).

---

<sup>4</sup> "W3C Provenance Incubator Group Wiki," W3C.org, last modified September 14, 2011, [https://www.w3.org/2005/Incubator/prov/wiki/W3C\\_Provenance\\_Incubator\\_Group\\_Wiki](https://www.w3.org/2005/Incubator/prov/wiki/W3C_Provenance_Incubator_Group_Wiki).

## Subcategory: Rights Metadata

Rights metadata is a type of administrative metadata that deals intellectual property rights and security. The goal of rights metadata is to record copyright-related information, including the right to access, digitize, collect, or provide access to digital works.<sup>5</sup> In the context of BI, rights metadata includes information about permissions and security (i.e., who has permission to access what).

## Structural Metadata

Structural metadata provides information about the properties and organization of a resource, and the relationship between objects. The aim of structural metadata is to assist with navigating content and assembling new content from a variety of smaller parts.<sup>6</sup> For example, structural metadata makes it possible to quickly sort documents by type (e.g., invoice, purchase order, inventory report, etc.) based on structure. Note that some information scientists consider structural metadata to be a type of administrative data, while others treat it a separate category.

## Markup Languages

A markup language is “a system (such as HTML or SGML) for marking or tagging a document that indicates its logical structure (such as paragraphs) and gives instructions for its layout on the page especially for electronic transmission and display.”<sup>7</sup> By inserting tags in content to denote notable features, markup languages mix metadata and content together. In the context of BI, markup languages make it possible to tag specific objects (e.g., SSNs as sensitive) so that data catalogs are more usable for humans and machines.

---

<sup>5</sup> Marcia Lei Zeng and Jian Qin, *Metadata* (New York: Neal-Schuman, 2008), 64.

<sup>6</sup> Michael Andrews, “Structural Metadata: Key to Structured Content,” StoryNeedle.com, October 11, 2017, <https://storyneedle.com/structural-metadata-key-to-structured-content/>.

<sup>7</sup> *Merriam-Webster*, s.v. “markup language,” accessed January 15, 2019, <https://www.merriam-webster.com/dictionary/markup%20language>.

## Use Metadata

Use metadata provides information about the actions taken by the user of a resource. Traditionally, this kind of information has been classified as data, not metadata. But when use data is treated as a statement about a potentially informative object, it becomes metadata. Like other types of metadata, use metadata can reveal a startling amount of information about individuals and their networks. But, unlike other types of metadata, use metadata is not produced deliberately. Instead, it's produced incidentally as a result of other processes. In his book *Metadata*, Jeffrey Pomerantz highlights two significant emerging subcategories of use metadata.

### Subcategory: Data Exhaust

Data exhaust is a term that describes use metadata produced incidentally as a result of certain activities. For example, the act of opening a document generates metadata that represents a record of that activity.

### Subcategory: Paradata

Paradata refers to auditing data, produced when someone triggers an action (e.g., checks out or checks in a document) and provides information (e.g., how, where, why, and by whom the document was used).

# Metadata Schemas

A schema is “a structured framework or plan.”<sup>8</sup> Therefore, a metadata schema is a structured framework for metadata. In other words, it’s a simple language with basic rules about what kinds of statements can be made about the data. A single metadata schema may be designed to include multiple categories of metadata.

**A metadata schema provides a formal structure designed to identify the knowledge structure of a given discipline and to link that structure to the information of the discipline through the creation of an information system that will assist the identification, discovery and use of information within that discipline.**

*Association for Library Services and Technical Collections  
Committee on Cataloging: Description & Access  
Task Force on Metadata  
Final Report*

If a schema is a language, then the words are called elements. Every metadata schema has its own set of elements (aka semantics) specific to the type of statement it was designed to make. That’s why a metadata schema is sometimes called an *element set* or a *data dictionary*.

When developing or selecting a metadata schema, interoperability is an important consideration. NISO defines interoperability as “the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality.” The interoperability of metadata schemas determines how usable the metadata is by BI solutions and artificial intelligence.

---

<sup>8</sup> *Merriam-Webster*, s.v. “schema,” accessed January 15, 2019, <https://www.merriam-webster.com/dictionary/schema>

An incomplete list of metadata schemas follows (in alphabetical order):

- [Categories for the Descriptions of Works of Art \(CDWA\)](#): Metadata schema for the description of art, architecture, and other cultural works
- [Creative Commons Rights Expression Language \(CC REL\)](#): Metadata schema for machine-readable expression of copyright licensing terms and related information.
- [Dublin Core](#): Metadata schemas focused on networked resources
- [Exchangeable image file format \(Exif\)](#): Metadata schema that provides a tag structure for embedded metadata within digital image files
- [Metadata Encoding and Transmission Standard \(METS\)](#): Metadata schema for encoding descriptive, administrative, and structural metadata regarding objects within a digital library
- [Metadata Object Description Schema \(MODS\)](#): Metadata schema for a bibliographic element set that may be used for a variety of purposes, especially library applications
- [Preservation Metadata Implementation Strategies \(PREMIS\)](#): Metadata schema developed by the Library of Congress to be a core set of elements for the preservation of digital objects
- [VRA Core](#): Metadata schema used to describe works of visual culture, including objects or events such as paintings, drawings, sculpture, architecture, photographs, book art, decorative art, and performance art, as well as the images that document them

## Metadata in Business Analytics

Business analytics depends on ETL tools to EXTRACT data from a data source, TRANSFORM it as necessary, and LOAD it into a data warehouse -- a centralized location that brings data together and optimizes it for analysis and reporting. The ETL process relies on existing metadata to find the desired data, organize it, and move it to the right location. The ETL process also generates new metadata related to the transformation.

Metadata remains essential after data has been loaded into a data warehouse. Every table, column, and row has associated metadata (e.g., format, security, modification date, etc.), which makes it possible to perform business analytics activities like data integration, data transformation, online analytical processing (OLAP), and data mining.<sup>9</sup>

---

<sup>9</sup> Rahman, Nayem, Jessica Marz and Shameem Akhter. "An ETL Metadata Model for Data Warehousing." *Journal of Computing and Information Technology* 20, no. 2 (2012). 95-111. <https://pdfs.semanticscholar.org/5013/bc35ad83319aaac456884f3e994e77ed2ce6.pdf>.

Business intelligence (BI) tools query data warehouses to generate reports for business analysis, a process which generates its own metadata, including:

- The connection used to query the data warehouse;
- Identification of the person who launched the query;
- Document(s) in which the data was used;
- Security associated with the data; and
- The instance/subscription/publication creation date.

Every step in the process -- from ETL to database (DB) to business analytics -- relies on metadata and every action creates new metadata, including paradata and exhaust metadata. As mentioned previously, paradata refers to audit metadata -- metadata focused on the use of data<sup>10</sup> -- and exhaust metadata (aka data exhaust) refers to the trail of data left by actions, such as search activities. Both are fundamental to maximizing the potential of metadata for business intelligence.

## The Importance of BI Metadata

In the context of business intelligence, metadata can be leveraged to support governance, risk management and compliance (GRC), launch automation, enable chargeback, facilitate upgrades and migrations, organize content for the purpose of monitoring, and provide insight into the adoption of BI tools.

## Governance, Risk Management, and Compliance

Metadata assists with many GRC tasks. Consider tagging, which attaches metadata to an object. When tagging is used indicate that objects are subject to regulations (e.g., [SOX](#), [GDPR](#), HIPAA, etc.), metadata facilitates regulatory compliance by making it possible to monitor the use of sensitive content. Metadata also facilitates [account recertification](#), segregation of duties, and security by locating permission loopholes, identifying who has access to what, and determining whether or not individuals are taking advantage of that access. Finally, metadata makes it possible to track the creation, modification, and deletion of objects and users, user activity, IP addresses, and much more.

---

<sup>10</sup> "What is Paradata," IGI-global.com, accessed January 15, 2019, <https://www.igi-global.com/dictionary/the-importance-of-being-honest/56385>.

## Automation

Metadata, in the form of tags/flags, can guide automation. For example, objects can be flagged in a way that corresponds with a schedule or trigger event (e.g., #PromoteThisWeek). Automation ensures that a job runs on a predefined schedule (e.g, promote every Wednesday at 11 PM) and metadata ensures that it acts on the appropriate objects. Depending on whether or not the metadata continues to serve a purpose, flags can be retained or removed automatically as part of the process.

Metadata can also be used to launch more complex automation, like auto-cleaning to keep BI content from growing out of control. In this case, metadata serves two purposes. First, it provides insight into BI usage. Second, it guides automation in the form of tags/flags. When an object crosses a trigger threshold (e.g., not used in X days), it can be flagged automatically for archiving (e.g, #Archive). As in the previous example, automation ensures that the job runs according to a predefined schedule, and metadata ensures that it acts on the appropriate objects.

Finally, metadata makes it possible automate reporting related to security, account recertification, tagged content, and more.

Learn more about [SAP BusinessObjects automation](#).

## Chargeback

In this paper, chargeback refers to the process of allocating the cost of resources to the departments that use them. When it comes to IT, organizations may attempt to determine what departments used what percentage of IT resources in order to perform cost-benefit analyses and justify investments. BI tools -- a subset of IT resources -- are particularly hard to chargeback because of their complexity. That's when metadata comes in handy. Metadata makes it possible to analyze BI usage, for example, to determine that 5% of business analytics maintenance resources were associated with a particular document, and that Department A represented 70% of the activity on that document. Similarly, metadata makes it possible to determine that Department A consumed 20% of the total memory associated with BI tools.

No matter how deep an organization drills down, the purpose of analyzing metadata for chargeback is to understand the total cost of ownership (TCO) per group of users or business unit, to support technology business management (TBM), and to forecast future BI investments.

Learn more about [SAP BusinessObject cross charging](#).

## Upgrades and migrations

Metadata makes it possible to understand BI usage, including:

- What documents were used in a specified time period
- When a document was last used
- What documents are used most frequently
- What documents are duplicates
- What variables and objects were used in a specified time period (aka document stripping)
- What objects are sensitive

Whereas data obtained from user surveys is subjective, metadata is objective. Analyzing metadata to understand usage makes it possible to archive superfluous content, which shortens the time frame and lowers the cost of upgrades and migrations.

As mentioned previously, metadata in the form of tags/flags can drive automation. In preparation for an upgrade or migration, unused content can be flagged for automatic archiving and used content can be tagged for automatic promotion.

Following a migration or upgrade, metadata can also support regression testing initiatives. Although, in most cases, regression testing focuses on data, testing for regressions in metadata can identify changes to structure and security.

## Monitoring

Metadata makes it possible to filter BI content based on a metadata description (e.g., sensitivity, class, security, activity, etc.) in order to monitor it. This is particularly useful in the following scenarios:

- When organizations outsource BI activities -- Metadata related to document actions (e.g., creation, deletion, modification, etc.), log-on time per user, check-in/check-out activities, etc. allows organizations to build reports containing metrics that keep track of outsourced BI.
- When organizations must monitor service-level agreements (SLA) on multiple BI solutions (e.g., Business Objects, Tableau, Power BI, etc.) -- Metadata related to platform availability, server performance, successful/failed schedules, etc. allows BI managers and business leaders to compare BI tools and determine if they are complying with SLAs.

## Adoption

Metadata makes it possible to measure the adoption of BI tools. BI adoption is an important metric for BI managers, who aim to optimize resources and need to know the following:

- What percentage of the documents they maintain are being used, and by whom?
- Which of the BI tools they allocated are being used, and by whom? (This is particularly important since most organizations invest in multiple BI tools.)

Beyond measuring adoption, metadata also helps BI managers understand connections, identify sources, recognize when documents are being used as data dumps, locate ungoverned data, and more.

## A Word on Non-usage

The flip side of adoption is non-usage. Non-usage is not reflected in metadata, but in the *absence* of metadata. It represents what is leftover after usage is identified with metadata. Knowing what *isn't* used is just as important to IT and business as knowing what *is* used.

For example, comparing usage and non-usage highlights trends and gives insight into BI reporting and the platform lifecycle. It informs decisions around training, investments, cleanup, etc. That's why non-usage is a key idea throughout business analytics.

- GRC: Non-usage can identify users who never connect and identify sensitive content that has never been used.
- Automation: Non-usage can trigger automated actions or reporting.
- Upgrade Migration: Non-usage can pinpoint objects and documents that can be archived prior to an upgrade or migration.
- Adoption: Non-usage is the flip side of adoption and identifies objects and documents that are not being used.

Note that both usage and non-usage are trimensional; they are determined based on metadata (object creation date, actions on object) -- or lack thereof -- with the addition of a time dimension.

Because metadata can become bulky, there are limits on the quantity of and length of time that metadata is retained. When metadata disappears, so does the opportunity to analyze usage and non-usage . That's why it's so important to continuously track usage and non-usage, and to take advantage of metadata service solutions, like 360Suite, that capture and retain auditing metadata.

## BI Metadata Collection

Metadata management solutions encompass metadata registries, repositories, and development and production services.<sup>11</sup> Metadata services typically harvest metadata from the ETL, DB, and BI tools, store them in a data warehouse, and organize them with a semantic layer.

Data warehouses offers several advantages. First, they facilitate information retrieval. Second, they minimize the impact on live production systems by loading data during off-peak hours, often in delta mode. Third, they make it possible for BI managers to perform metadata impact analysis -- to understand what documents

---

<sup>11</sup> Zeng, *Metadata*, 212.

are impacted by changes to an object. Fourth, they enables BI managers to trace the lineage of objects from ETL to reporting.

## Third-party Metadata Services

Most organizations benefit from third-party metadata services solutions, like 360Suite, that centralize and process metadata inputs and transforms them into data for business. Interestingly, metadata services generate their own metadata related to adoption.

Data catalogs, sometimes called “searchable business glossaries,” are often the most visible part of metadata service solutions. By virtue of being self-service, data catalogs make it possible for IT and business users to perform data discovery, obtain data lineage, add comments to objects, and tag content.

Metadata management solutions, like 360Suite, assist with information retrieval, tagging, documentation, and even triggering actions with metadata. What distinguishes 360Suite from a data warehouse is that it extracts only relevant data -- metadata related to BI -- into a data mart.

The narrow focus makes it easier to control access and provides authorized users with easy, fast, reliable, and secure access to metadata that answers their business intelligence questions.

360Suite also features pre-built, self-service reports, based on simple and complex queries, that assist with data exploration and make metadata usable to IT (data analysts and information scientists) and business users. Metadata searches offers many advantages, including the ability to:

- Compare snapshots of metadata over time to gain insight into the BI life cycle;
- Highlight usage and non-usage to inform decision making;
- Measure and control data quality with regression testing that compares historical metadata, including information about the source;
- Clean up metadata in order to keep the size of the data mart under control;
- Display data lineage; and
- Perform impact analysis.

Finally, 360Suite creates data that can be used as a data source for further automation and analysis so that AI (e.g., IBM Watson) can be used on top of it. This provides contextual information around data and makes it possible for AI to leverage metadata.

## Conclusion

Whether or not “data is the new oil”, there’s no doubt that data is extremely valuable. As the volume of data (whether structured, semi-structured, or unstructured) grows exponentially, it becomes increasingly difficult to discover and understand the *potential* in potential information. The only way to effectively manage big data is by managing metadata as a semantic layer that can be optimized for AI. Leveraging metadata and non-usage as a semantic layer supports “descriptive analytics,” “predictive analytics,” and “prescriptive analytics.” Metadata brings agility and insight. In fact, data without metadata is virtually unusable.

Interoperability is key to reach the goal of machine learning. As the number of BI sources and connections continues to grow, metadata interoperability will pose a greater challenge, and metadata services will become increasingly important. Metadata services will represent a new source of information to which other services can connect. This will generate new metadata and create a greater need for metadata on metadata -- BI on BI -- to make the information available for machine learning, artificial intelligence, and business analytics.

The importance of metadata for BI and business analytics cannot be overstated. Metadata allows BI managers and information scientists to make data-driven decisions: to understand user adoption, compare usage to non-usage, manage risk, optimize platforms, and better allocate resources. Ultimately, metadata supports the shared goal of IT and business -- an optimal end-user experience at the lowest possible cost.

Auteurs: Bruno Masek  

Kristen Champagne Gray 

# Metadata for BI and Analytics Solutions

GET STARTED!



 360 Suite



Visit <https://360suite.io>