USING DATA DICTIONARIES FOR BETTER, FASTER QUERYING
INTRODUCTION

WHAT ARE DATA DICTIONARIES?

Ever get the feeling that people from different departments in your organization aren’t speaking the same language? Or spent days trawling a dataset that sounded relevant, only to find that its content meant something else entirely?

If so, you’re far from alone. As Big Data rapidly expands and is used by more and more people in more and more ways, you will need a way to ensure that everyone in your organization is on the same page. That’s where data dictionaries come in.

Data dictionaries help users understand any dataset before diving into it. That makes them an essential communications tool for data modeling, curation, governance, and analytics, especially when dealing with datasets that have been collected, compiled, categorized, used, and reused by different people across the organization.

They make metadata information more effective when searching and organizing data, and make it easier for multiple data admins working on the same data model to understand each other. They also help prevent duplications of effort—recreating datasets that already exist.

So how does that work in practice? Well, the ultimate goal of a data dictionary is to help users understand how data will be used and expressed within a certain context—and that means creating a well-organized collection of data element names, definitions, and attributes, arranged in a table. This describes all data held by an organization, in a single database.

More sophisticated versions go a step further, comprising database schema with reference keys and entity-relationship diagrams. These form part of a comprehensive metadata system, encompassing a standardized lexicon for term use, naming conventions, and so on. Not only does this ensure that everyone sticks to the same definitions, but these details also support querying tasks and the creation of business reports—but more on that later.
WHAT KIND OF DATA DESCRIPTIONS ARE THERE?

Within a data dictionary, descriptions can be broadly categorized into:

- **Business Concepts**: definitions of data that are accessed by business users
- **Data Types**: descriptions of which data can be considered valid
- **Message Concepts**: descriptions that create shared understanding between organizations and departments, to make sure business communications are used in the same context

Another way of looking at this is to distinguish between these two categories:

- **Technical data resource data**, which is used to build databases and data models
- **Semantic data resource data**, which is used to support business activities and analysis

Information provided about each attribute or field within a data model is arranged in a spreadsheet format, organized in rows and column labels.

Common elements include attribute name, an optional/required field, and attribute type, while additional fields include information source, table where attribute is contained, and so on.

IS THERE A STANDARD WAY TO CREATE A DATA DICTIONARY?

Yes... and no.

Traditional data dictionaries are indeed standardized: they're always presented as a relational database structure. Typically, they use a database schema that's a graphical presentation of the entire database, with tables connected by external keys and key columns. In terms of conventions and standards, ISO 11179—the international standard used for metadata representation in an organization's metadata registry—provides clear guidance for standardizing, registering and storing data elements, too.

That said, you don't necessarily need a traditional data dictionary at all! If your BI platform has a comprehensive system for tagging and searching metadata, this can provide an excellent, streamlined alternative to the standard solution. We'll come back to this later in this whitepaper.
HOW ARE DATA DICTIONARIES USED IN ANALYTICS AND BI?

Data dictionaries have long played an important role in BI, but recently this role has broadened dramatically.

In the past, data dictionaries were used primarily in a tech context—by database developers and data modelers to help build the infrastructure that supports analysis. Now, however, usage has shifted: data collectors, analysts, and business users have also recognized the value of data dictionaries.

That came about because these non-technical users increasingly work directly with data for analysis purposes. Data dictionaries help establish what’s in a dataset and where it initially came from, without users having to download and search through the whole thing first. In other words, users can tell if the dataset is relevant to a BI query fast, and can find their way to the original data, without wasting time on irrelevant data.

Data dictionaries come with another important benefit: they improve database transparency. That’s because users can trace the structure of the data and tables, viewing where data is located, what it means, and when it was last updated. All of this helps improve the quality and accuracy of data models or reports, as well as the relevance of a particular query or analysis.

WHAT HAPPENS WHEN THERE IS NO DATA DICTIONARY?

Put simply, working without a data dictionary means you’re at risk of losing more in translation or winding up in a situation where you struggle to decipher what you’re looking at.

Without a data dictionary, you’ll likely waste time sifting your data to dig out what you need; which is impractical when dealing with very large or complex datasets. You’ll also struggle to identify and describe data elements or recognize problems like duplicated content and synonymous data, leading to more inefficiencies.

What’s more, you have no record of who has used a particular dataset or subset in the past, how and why they used it, and how useful it turned out to be. This makes it more difficult to make judgments about the value or relevance of a particular dataset without combing through it again for yourself.

As a general rule, mapping data exposes issues with definitions, consistency, and alignment early on, so you can nip issues in the bud. Without this failsafe, you’ll keep making the same mistakes—which will only be exacerbated as your data pool grows.
DOES MY COMPANY NEED DATA DICTIONARIES?

Does your company use large amounts of data? Is this used by people from different teams, departments, and/or disciplines? Are you building data models? Investing in BI? Planning to grow?

If you’ve answered yes to any of these, you probably need data dictionaries.

Why? Because it improves the quality and accuracy of your reports and analysis, which in turn helps you understand your business better and make smarter, more beneficial decisions.

How? Because implementing policies for careful tagging and clear descriptions helps to establish a consistent and coherent use of terminology across your organization, preventing data from going missing, getting overlooked, or being improperly used.

This also makes data document management easier, provides more meaningful metadata for search purposes, and is incredibly valuable when you need to upgrade your database in the future.

Most importantly, getting a coherent, consistent labeling and data storage system in place will prove essential as your data scales. Without data dictionaries to help keep on top of this, data can get lost, corrupted, mislabeled, or misused. In the long run, you’ll struggle to find what you need for detailed, comprehensive analysis, without delays.

If you’re an ambitious, growing organization and you’re serious about rolling out the kind of responsive, data-driven, efficiency-minded BI strategy your company needs to achieve its goals, you should absolutely implement data dictionaries to support this.
WHY ARE DATA DICTIONARIES A CRITICAL BEST PRACTICE?

To recap: data dictionaries are a core part of any best-practice approach to standardization, accuracy, security, compliance, and continual improvement.

From a standardization point of view, using these practices helps you to establish a mutually agreed upon set of clear, consistently defined elements, definitions, and attributes, providing a comprehensive catalog of data definitions, relationships, collection groupings, validation rules, aggregations, and generated reports. This is the cornerstone of any reliable and effective information system.

From an accuracy perspective, by standardizing the way you store and label data in your databases, you improve the availability of and access to, the most relevant, up-to-date information. This is crucial for reporting and analysis.

Security-wise, careful tagging and cataloging underpin a robust authentication and authorization approach, distinguishing between types of users and applying security restrictions to specific data items as appropriate.

Meanwhile, data dictionaries can even ease the data governance and compliance burden in your organization. Pay close attention to tools and labeling within the system that relates to PII, data privacy, and reporting. These may help you to retrieve and store information in ways that keep you in line with GDPR, HIPAA, Dodd-Frank, and other key regulations.

Finally, by incorporating machine learning into your data dictionary tagging system, you can use features like "recommendations" to suggest likely search terms and categorization types, based on past activity. In this way, you keep learning from the data hive-mind in your company, honing the ways you categorize, use, and reference data going forward.
WHAT DOES THE FUTURE HOLD FOR DATA DICTIONARIES?

As more and more types of users adopt data dictionaries, cleaner, faster, more transparent analysis will become the norm.

Today, many data dictionaries are not standalone entities but are integrated with data preparation or analysis, or even fully integrated with all your data-related activities, from preparation and analysis to visualization, governance, and security. As users increasingly demand a seamless experience, more and more data dictionaries and catalogs are moving toward integrated or highly interoperable options.

Expect to see ever-improving data integration, discovery, and sharing, as users gain access to technical and semantic data and improve their understanding of the underlying information system definitions and assumptions. In fact, metadata is already becoming so smart that users can merge files of metadata and get a clear picture of the data inside without opening a dataset at all!

Meanwhile, BI platforms will introduce better ways to interact with your data dictionaries, or will even offer tools that replace data dictionaries altogether. This will allow you to add and manage your own metadata directly into your BI platform, and apply this to your connections modeling, too.
FINAL THOUGHTS:

GETTING THE MOST OUT OF DATA DICTIONARIES WITH SISENSE

You don’t need a dedicated data dictionary per se—at least, not in the traditional sense—but you do need a way to apply descriptions and labels of your data model so that others can understand and refer to it. You need a way to apply rules that all the data within your system will abide by.

With this in mind, Sisense has developed an innovative alternative: tag data and descriptions, and group tables and columns by defining metadata. In short, you can create the basis of a data dictionary without affecting your actual data.

For example, you can tag several tables with a unique word or description, and then locate that group of columns through the Search field to see all the tagged tables across your schema. Both are searchable, but tags are associated with tables, while descriptions are free text fields associated with columns or tables.

From there, you can easily locate tagged columns and tables in the Search field by entering #, which displays a list of all tags that you can select to filter your data.

This function supports data modeling within Sisense’s Elastic Data Hub, where you can mashup live and cached ElastiCube Data Models. The Sisense data dictionaries feature means you can figure out which datasets to work with, without having to open them up and load them into the system first.

As always, this isn’t about having specific tools, it’s about results. To get the most out of your data, you need to be able to do all the things that a data dictionary does—but with Sisense, you can streamline the process from right within your BI platform, bringing the benefits to more and more business users than ever before.

WANT TO LEARN MORE ABOUT HOW SISENSE CREATES DATA DICTIONARIES?

Watch a Demo Here