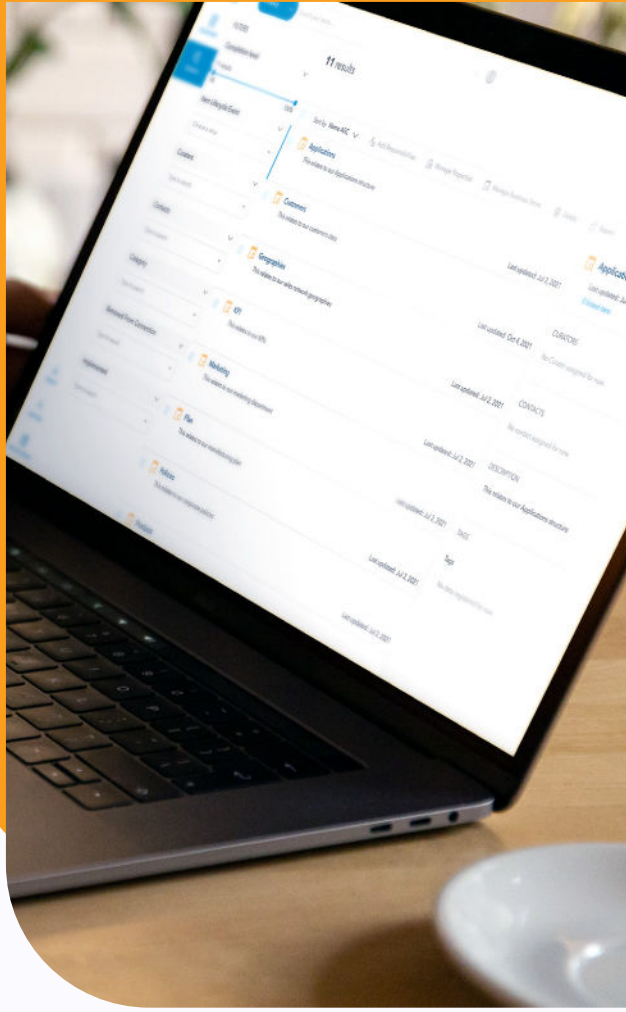
zeenea

Business Glossary:

an essential component of a Data Catalog for data fluent companies



When setting up a data catalog, the first step is to connect the catalog to your data sources and physical systems, on which your assets are stored in order for your users to start inventorying them. This inventory is a necessary step to obtain a first level of information: storage system, location, access modes, formats, types, etc.

Through automation capacities, a certain amount of information is collected which then retrieves detailed technical documentation of what the information system contains. Standard data catalog solutions then enable knowledgeable data users to complete this documentation, by adding classification attributes to better specify the company's technical ecosystem.

However, while this information can answer the questions of some of the more technical users (engineers, architects, etc.), it generally remains unclear to a growing population of data consumers in the enterprise. Indeed, without this knowledge, these consumers are not able to exploit nor govern this data effectively.

In order to provide the necessary context for the consumption of this data, users need different types of information: organizational, statistical, compliance, etc.

> **Precisely, technical documentation must be accompanied by so-called semantic information. This is the objective of implementing a business glossary.**

Building a common language with a business glossary

When business users talk about data, they usually refer to concepts such as *customer address, sales, or 2021 turnover.* They are most likely not referring to a table or a database schema, as they may not know or understand it. A business glossary will help define these concepts and share these definitions amongst all employees.

The addition of semantic metadata thus meets several objectives:

To bridge the gap between business and technical users, by building a common language that allows them to collaborate effectively;

Align business users, especially different entities within the company, with these definitions. In particular, this avoids ambiguities between related terms;

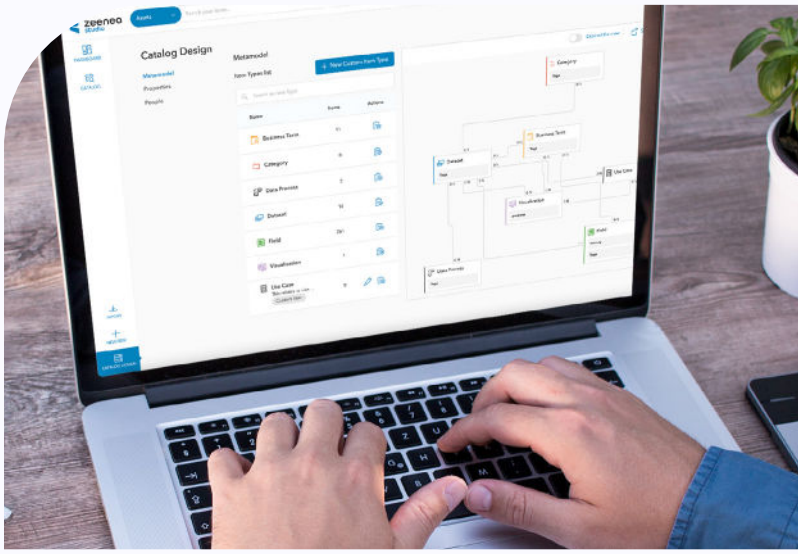
Enable all users to more easily find the data they are looking for, and provide the necessary context to understand and use it.

The business glossary: a productivity lever for a data catalog

Companies handle huge and ever-increasing volumes of technical assets, usually with a lot of duplicate information in various systems. Statista found that in 2020, a total of 64.2 zettabytes of data was captured, copied or consumed worldwide, with a forecast of 180 zettabytes for 2025! Documenting all of these assets one by one, even with the help of automation solutions, is a real challenge that is nearly impossible to overcome for most companies.

In comparison, the number of concepts handled by an organization is generally limited to a few hundred terms. Importing and documenting these business concepts in a data catalog has virtues in terms of the viability and maintainability of the documentation. It is a much more manageable scope that allows the documentation to be centralized, with two advantages: the cost of maintenance is reduced (fewer repetitive tasks on the part of the data stewards) and the consistency and accuracy of the documentation are improved by avoiding human error.

> A good data catalog tool must therefore offer a solution that administers these business concepts, allows them to be linked to the technical assets that implement these concepts, and thus opens up the use of the catalog to the entire enterprise.



The different approaches to building a business glossary

When it comes to describing a domain of knowledge, several more or less advanced approaches can be used depending on the company's objectives.

A Lexicon

The first possible solution is to create a lexicon, i.e. a flat list of terms with their definitions and synonyms, describing a particular field of knowledge. This is a fairly common format that is simply browsed in alphabetical order. A classic lexique by its very nature has few constraints in terms of the organization of its content. It is therefore quick to set up and simple to use.

> **At Zeenea, this was our first approach in response to one of the core values of our product: simplicity.**

The administration and use of a lexique does not require significant effort or expertise in information modeling. Simple to use, it does not require any particular training for end users.

It is an effective tool when users simply want to know that a term exists and its definition. This model is adapted to small or medium-sized structures, or for organizations willing to start quickly on a restricted perimeter.

However, this model does not allow large organizations to represent the full richness of the semantic heritage. Specifically, it does not enable the representation of relations between concepts, especially hierarchical ones, which provide essential context for the manipulation of an extensive document corpus. For this type of structure, the use of a simple lexicon is therefore a real obstacle for scaling up.

A Thesaurus

To overcome this problem, another approach is through a thesaurus. It adds hierarchical and organizational dimensions to a document corpus. Concepts are grouped into coherent sets and subsets, providing a better overview and facilitating the administration of a large number of concepts. This classification of business terms also simplifies searching in the catalog by making the use of similar terms or homonyms less ambiguous. It allows the term to be put into context - for example, the term «closing date» will not have the same definition if we are talking about the accounting year or a bank account. This approach has the advantage of providing a clear structure to the information.

However, the exploitation capacities of the thesaurus remain limited. Indeed, its structure is by definition rather rigid. Like Windows Explorer for instance, where you browse through folders and subfolders, this type of approach does not allow you to represent the diversity of relations that can link different concepts together and therefore exploit their full potential. Moreover, this type of system requires a significant amount of upstream modeling work because of this single hierarchical prism. This slows down its implementation and leaves no room for error - the slightest change in the model can have a major impact on the entire chain, like taking into account new use cases for instance.

> Unfortunately, this is an approach that is often used by data catalog providers on the market. They offer a ready-to-use structure on several levels, into which the company must insert their concepts, with the obvious risk of putting a square peg in a round hole...

A formal ontology

For a richer and more open representation of information, another approach consists in building a formal ontology. This is a representation of a knowledge domain based on definitions of object types or classes, their attributes, and their relationships. There are, for example, applications in the fields of finance (e.g. FIBO – Financial Industry Business Ontology), medicine (e.g. Menelaus) or the Web (e.g. FOAF).

An ontology will allow concepts to be classified along different axes, whereas a taxonomy only offers a single hierarchical viewpoint. It thus offers more extensive possibilities of exploration and exploitation of information. In particular, an ontology introduces the concept of inference, allowing relationships to be interpreted (notably through computer programs) in order to deduce a different representation: my father's brother is my uncle, or a vehicle with four wheels, a steering wheel and an engine has a good chance of being a car.

If this type of information representation opens up possibilities by revealing the meaning of relationships, its implementation remains a particularly difficult and long exercise. Requiring a significant modeling and formalization effort, the construction of an ontology requires expert skills and its interpretation by non-expert users is not easy. Because of its complexity, this type of solution is generally not suitable for deploying a Data Catalog, the objective of which is to democratize the use of data and accelerate the creation of value.

> **In fact, to achieve this objective, a data catalog must provide a rapid and incremental implementation process, as well as limit as much as possible the training costs of its users by encouraging their autonomy in searching and understanding data.**

The graph approach of zeenea 's business glossary

At Zeenea, we decided to design our own solution for constituting this semantic layer in our Data Catalog.

Indeed, we are convinced that the only solution that offers the flexibility, simplicity and scalability necessary to cover the needs of data consumers is by building a graph.

The constitution of the semantic layer in our data catalog

To build this graph, Zeenea allows users to create and customize the types of objects that will constitute this semantic layer. We do not impose a predefined metamodel, and offer the possibility to create only the types of concepts that you need and that correspond to the context of your company. This list can evolve over time as new needs arise. This principle allows you to create a simple lexicon if it is sufficient, or to model a much more complex structure involving different types of concepts (data elements, reports, indicators, fonts, etc.).

Another advantage of this solution is that for each of these types of objects, it is possible to configure their own list of attributes. Indeed, an indicator object can be described by attributes such as its update frequency, its control rules, etc. While these same attributes would not necessarily make sense to describe the concept of B2C Customer. Here again, this list of attributes is not fixed in time and can be progressively enriched according to new uses and the company's ability to maintain more or less rich documentation.

It is then possible to create relationships between these different types of concepts and to configure the way in which they will be transcribed into your physical systems. The first use is generally the representation of hierarchical links between high-level concepts and other more unitary ones that often characterize these «macro» objects. These configuration options provide a great deal of flexibility – in particular, they will help guide Data Stewards in their documentation work by improving the suggestion engine, avoiding input errors and inconsistencies.

A bottom-up approach for efficient deployment

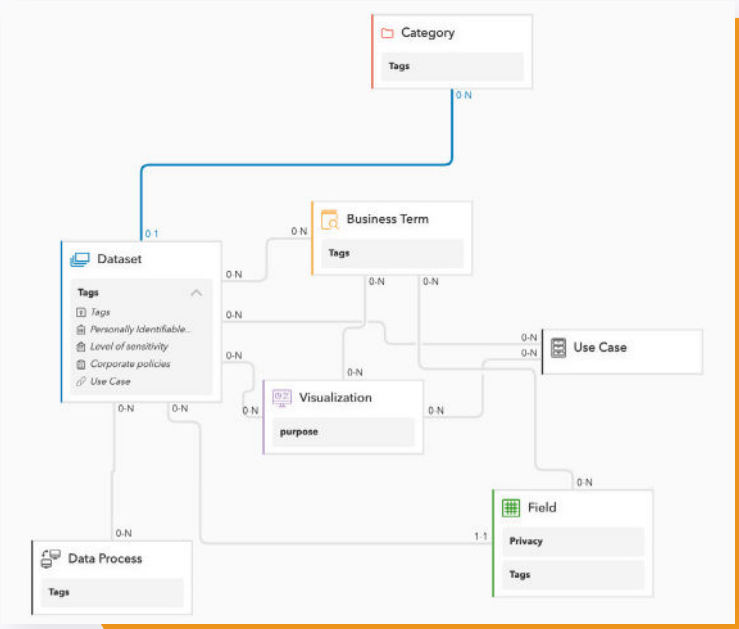
The advantage of this graph-based modeling is that it allows for a quick and iterative start. Indeed, there is no need to think about the entire structure of your glossary in advance, a task that generally requires weeks of design and validation committees. You can progressively enrich your metamodel by prioritizing your use cases, and then correct or adapt it if necessary.

Depending on your priorities, you can, for example, start by documenting the main concepts handled by the company with a Business Object type and associate these concepts with the data sets that implement them. Then, once this global mapping work has been done, go down to a finer level on a case-by-case basis to describe your Key Data Elements. For example, by creating a Business Data object type that will provide precision information at the field level of a table.

Through this flat model, it is also easy to quickly integrate all or part of an existing glossary because by its very nature it adapts easily to your context. This bottom-up approach is at the heart of our product and corresponds to our belief in the successful deployment of a data catalog.

Ergonomics in the service of simplicity

A graph remains a concept that can be quite complex and abstract for end users. Therefore, we had to hide this complexity by paying particular attention to the ergonomics of our applications in order to fully exploit their potential – the functionalities for creating relations are designed according to their semantics to be as natural as possible for users. Graphical tools allow the exploration of the graph without special expertise.



A semantic layer that feeds the search engine

Finally, at Zeenea, the management of this semantic layer is intended to serve the search engine – a fundamental element for exploiting the content of a Data Catalog in an increasingly voluminous ocean of information. Indeed, the performance of the classic indexing and filtering systems used by the majority of Data Catalog providers on the market is limited in terms of search capacity. Especially when the approach is exploratory. The constitution of this graph allows, through a semantic analysis of the relations, to improve the ranking algorithms as well as the suggestion engine (as Google has done for some time).

Take away

To summarize, one of the challenges of a data catalog project is to define a common language around the data at the enterprise level. The construction of this common language must facilitate the search and exploration of the Data Catalog content in order to exploit its full potential.

To build this semantic layer, traditional approaches generally do not allow the objectives of democratizing data to be achieved.

At Zeenea, we are convinced that only a graph-based approach can provide the necessary flexibility, simplicity and scalability.

If you wish to know more about our semantic model, or to obtain more information about our Data Catalog:

Contact us