expert.ai

Following the Taxonomy Roadmap to Data Discovery Success



Whether it's your employees or your customers, people need to find the information they need when they need it. Data discoverability is a high stakes activity: content that is difficult to find can lead users to have less confidence in the relevance and usefulness of the content itself, defeating the purpose of your content in the first place. The bottom line is this: making your data easy to access is an enterprise imperative. In this white paper, we'll show you why taxonomies are the solution to your data discovery challenges and share important considerations and recommendations to guide you on your journey.

- **Data Discovery:** Why it Matters
- **Obstacles to Data Discovery**
- **What Is a Taxonomy?**
- **How Taxonomies Work with Your Data**
- **The Benefits of Using Taxonomies**
- **Taxonomy Options:** Build, Borrow or Both?
- **Taxonomy Design Considerations**
- **Driving Successful Data Discovery:** Recommendations

Introduction

In your business, insight is what drives innovation. It's the combination of data and analysis that allows you to better understand the markets and customers you serve and to discover where you can deliver a competitive advantage or increase efficiency.

Data discovery is the process of uncovering insights lying dormant in your enterprise – creating value out of unstructured data and getting it to the business users who need it, when they need it. Data discovery is a high stakes activity: content that is difficult to find can lead users to have less confidence in the relevance and usefulness of the content itself, defeating the purpose of your content in the first place. The bottom line is this: making your data easy to access is an enterprise imperative. This helps build what we call actionable intelligence, which is information you can leverage to support decision making.

However, not all of the data in your organization is easy to access. If you're like most organizations, you probably have a process to leverage your structured data (think: sales and financial metrics), but this is just a small portion of ALL the data you manage. So, what about all the other data? Emails, documents, reports, transcripts, PDFs, reviews, social media content—all of the non-numerical data that doesn't fit neatly into rows and columns. This is unstructured data, and according to Forrester, it constitutes as much as 80% of the data that enterprises process and continues to grow two times as fast as structured data.[1]



> "80% of the world's data will be unstructured by 2025."
> **- IDC [2]**

Organizationally, we struggle to drive insights because of the nature of unstructured language data. However, there are ways to leverage this data to build business intelligence to create value. In fact, with the right tools, you can analyze a variety of unstructured data sources to analyze, understand and accelerate business processes that have a significant impact on your business.

For example, we don't have to imagine the critical need for researchers and decision makers to have actionable insights about the spread of infectious diseases or climate change. There is no shortage of examples that apply to any organization.



So, it's time to act. And that's the goal of actionable intelligence: to provide the information that can help people make better decisions and act on them more quickly.

Data Discovery: Why it Matters

Organizations contain a huge amount of knowledge that they depend on to do business, and much of it is contained in those unstructured sources we mentioned above. Let's look at three areas where your data discovery strategy will be invaluable.

1. **Domain knowledge.** Your domain knowledge is the result of all the experience and expertise your organization has acquired over time. It's the R&D information that guides product development; it's the customer data that helps you communicate with them better and provide the products and services they need; it's the identification of risk factors that helps a business navigate market forces and avoid harmful impacts. Domain knowledge is what a company provides to customers and partners, as well as to internal operations that depend on this information to do their jobs. Together, it helps form your competitive edge.
2. **Audience engagement.** Digital transformation has led everyone—your internal and external users/customer and information consumers in general—to expect the same speed of information access and response from the organizations they do business with as they get with the average internet search. Whether it's users who need to find product information or subscribers who rely on your content for news or investigative purposes, you want all of the audiences you serve to be able to easily access the information they need, as quickly as possible. This keeps your customers coming back, again and again.
3. **Efficiency.** Efficiency in accessing information is important, whether it applies to your team or to your data processes. Process efficiency equals faster time to market, which is a key differentiator in a digital environment. Even if you have a large team to help you assemble data, if you want to scale to more sources, more information, more organized intel, you will need to go faster.

Obstacles to Data Discovery

So, why do organizations struggle to drive insights from their data? There are several aspects of information that make data discovery so challenging:

- **Lack of data structure.** Content that is not organized in clearly divided categories makes it difficult to navigate through search or other discovery methods. In addition, data that has not been standardized or normalized, as well as documents that come from multiple sources, in multiple languages, only compound the challenge of being able to find exactly what you need when you need it.
- **Siloed content.** When content is split between different data sources and spread across the enterprise, it's difficult to connect the variety of data sources with an enterprise-wide strategy. Inconsistency across the data structures, as well as the unknown of what is contained in your data archives only adds to the challenge.
- **Inefficient information access.** This happens when your data discovery/search operations return non-relevant content or so much content that you're unable to get exactly the information you need. Simply put, the content is not geared to your needs or preferences.

The solution to overcome such changes is to rely on a taxonomy.

What Is a Taxonomy?

In a business context, we can think of taxonomy as a content management process that allows users to easily find information on your website, your intranet or any other digital repository.

From a more technical perspective, a taxonomy is a system for classifying information into a hierarchical structure that matters to your business. It's written as domain categories that are organized from broad to specific or from specific to generic, and it helps embed terms together so that we understand the logical connections and associations between them. Take COVID-19 for example: it's part of the family of respiratory viruses and it belongs to a broader category of medical conditions. It's also related to SARs-cov2 because they have similar symptoms. In the design of a taxonomy, these relationships are very important.

It's useful to think of a taxonomy not as a fixed structure but as a living organism that evolves with your business. Your taxonomy can grow vertically with new categories, or horizontally, when new terms become associated with an existing term. Think of the words "cloud" and "tablet" and how their meanings have been extended to new domains over the past decade.

The most important thing to keep in mind is that your taxonomy is organization-centric because it structures the knowledge for the domains that matter for your business.



"Companies use taxonomies to more concisely organize their documents which, in turn, enables internal or external users to more easily search for and locate the documents they need. They can be specific to a single company or become de-facto languages shared by companies across specific industries."

- expert.ai Glossary of AI Terms

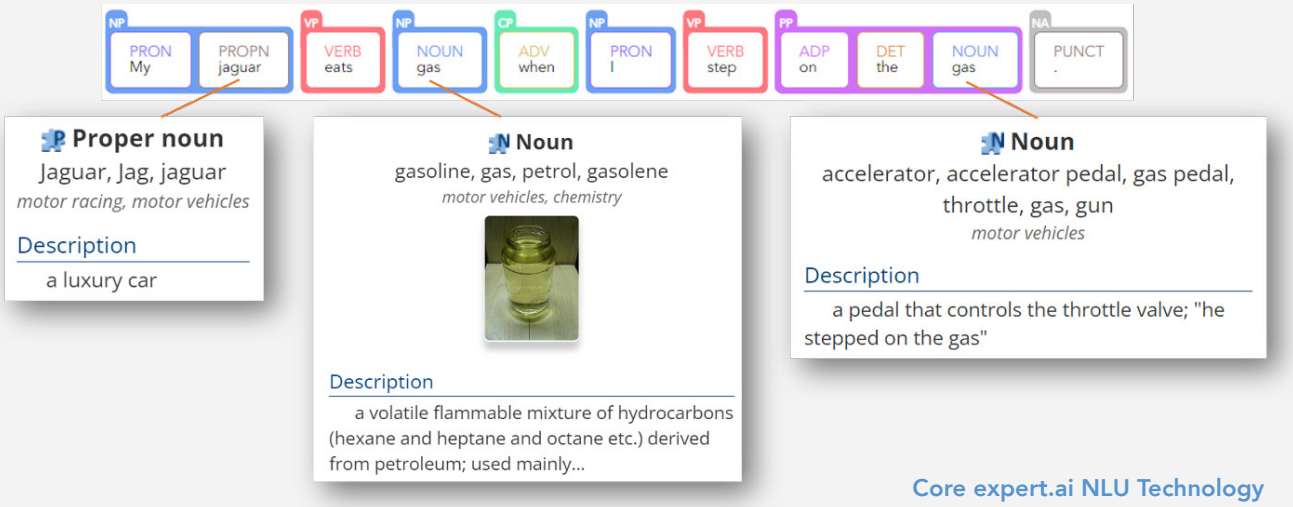
How Taxonomies Work with Your Data

So, let's look at how taxonomies solve the data discovery problems that we previously identified.

First, **they help you deal with ambiguity**. A key goal for any classification system is to minimize ambiguity. To put it another way, it's about minimizing the number of buckets that the given results can be put into, preferably to the point where there are no overlaps.

By its very nature, language is ambiguous. For instance, take the example of "gas" in the automotive domain in American English. The term gas is used for the accelerator as in, "press the gas pedal" and, also, for the fuel that runs the car, as in "fill the gas tank." This ambiguity exists in different ways across all languages.

At expert.ai, our natural language understanding platform is designed to understand the context of the concept to link it to the correct product definition or categorical taxonomy.



Core expert.ai NLU Technology

Next, taxonomies **provide embedded domain knowledge that allows you to designate the meaning of certain terms**.

Thanks to this embedded domain knowledge, we can designate meaning, linking "gas" ("my Jaguar eats gas") to "gasoline," and alternatively, when we say "step on the gas," to designate meaning to the "accelerator." In the world of taxonomy design and implementation, this ability to accurately link concepts together is absolutely critical.

Taxonomies **support data discovery by bringing consistency to how your content is tagged and connected**. In other words, taxonomies ensure that concepts are grouped under the proper category. For example, "salt" is a chemical component and it's also known by a variety of synonyms (chlorate, NaCL, alkali, etc.) that subject matter experts may refer to during experiments.



When applied to all of the content in your organization, a taxonomy physically **helps ensure that the variety of similar concepts are placed under the proper category**.

Lastly, taxonomies can really **transform content into actionable intelligence**. This is the result of being able to connect the dots, to link information and data together. When people are able to more easily find relevant content, this also helps to prevent unnecessary rework.

Providing Data on COVID-19 Risk Factors

During the early stages of the COVID-19 pandemic, there was little information around the risk factors of the virus. A project initiated by the White House [3] delivered information to researchers for free so that they could quickly access real-time risk factor data. Companies were provided with this data to analyze in order to provide information to patients, or to support researchers or others working to develop treatments.

One example is the use of the MeSH taxonomy, a well-known medical taxonomy, to identify different conditions in patient case reports that showed increased health risks when suffering from COVID-19. Knowing that influenza and COPD are respiratory diseases and part of the same category helps link this category of diseases to the risk of factors for COVID-19. This type of data helps drive more insight into COVID-19 to better identify the most vulnerable patients.

Risk factors

▸ **Cardiovascular Diseases (7)**

▾ **Respiratory Tract Diseases (6)**

Name	# of Words	Article	Article Date	Sentence
Influenza, Human	1227	Acute leukemia in the time of COVID-19	May 31, 2020	Bacterial secondary infection can complicate viral infections, a situation well known in **influenza**, and plausible for **COVID-19 infection**.
Respiratory Distress Syndrome, Adult	1841	Effects of host genetic variations on response to, susceptibility and severity of respiratory infections	May 29, 2020	Notably, several genetic loci have been shown to influence the risk of **ARDS**, a condition which is associated with severe **COVID-19 infection**.
Severe Acute Respiratory Syndrome	821	Chronic conventional disease-modifying anti-rheumatic drugs masking severe SARS-CoV-2 manifestations in an elderly rheumatic patient	May 29, 2020	In fact, although the immunological mechanism behind the risk of greater severity of **COVID-19 infection** is unknown, the coronavirus infection (**SARS** and MERS) may induce a cytokines storm especially in patients who developed fatal complications.
Pneumothorax	396	COVID 19: Modified Underwater seal chest drain system	May 29, 2020	There may be a high risk of aerosolization in cases of **pneumothorax** with active air leak and in **COVID-19** patients with intercostal drainage tube in-situ requiring high PEEP mechanical ventilation.
Pulmonary Embolism	6395	Thrombosis risk associated with COVID-19 infection. A scoping review	May 27, 2020	The study outcome was a descriptive assessment of thromboembolism incidence and risk including deep venous thrombosis (DVT)/pulmonary embolism (**PE**), thromboprophylaxis strategy, risk of Disseminated Intravascular Coagulation (DIC) in **COVID-19 infections**, and the role of coagulation parameters in predicting the severity and mortality of the disease.
Pulmonary Disease, Chronic Obstructive	3555	Risk of infection and hospitalization by Covid-19 in Mexico: a case-control study	May 26, 2020	41 In addition, **COPD** did not increase the risk for **Covid-19 infection** in Mexico, however, it increased the risk for hospitalization.

▸ **Nutritional and Metabolic Diseases (4)**

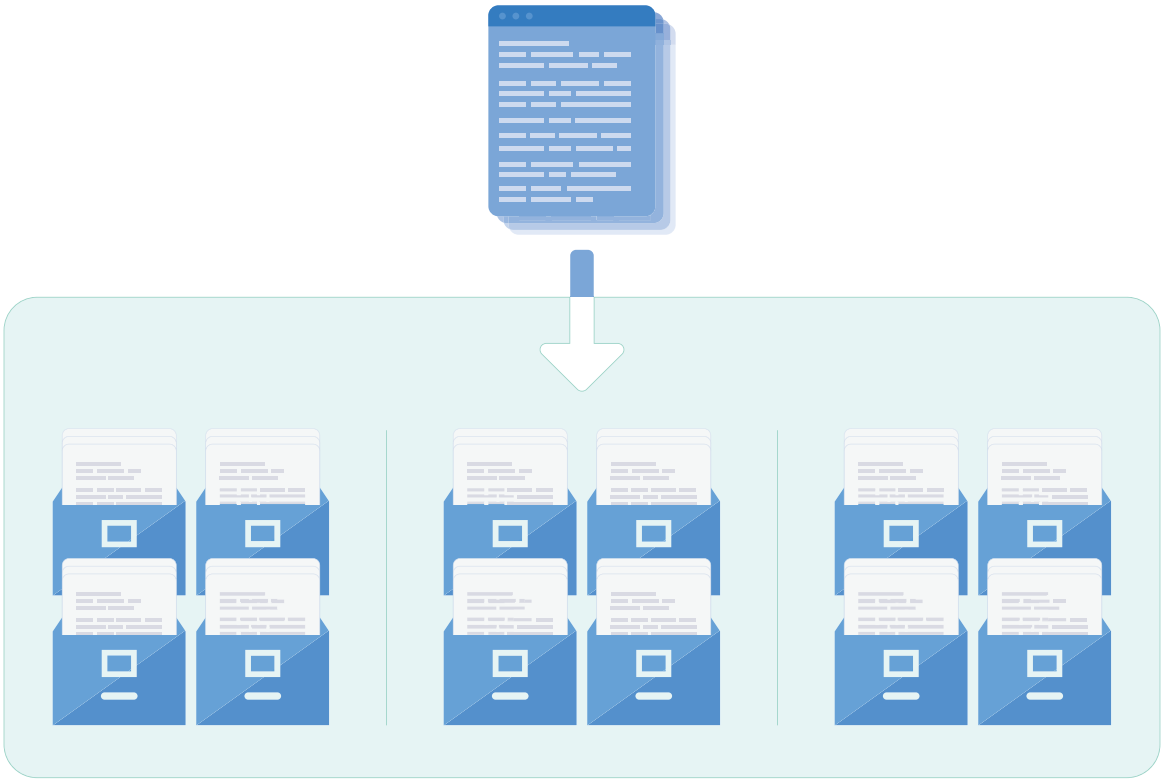
Core expert.ai NLU Technology

By offering a comprehensive taxonomic view of a domain, an organization can provide end users with more tools to discover connections that happen at a higher level. This would not be as easily discoverable if the content was loosely connected and at a very shallow level. By using NLP classification techniques, it's possible to mine thousands of these reports for new insights and connections very quickly.

The Benefits of Using Taxonomies

When you've created an effective taxonomy, there are three main outcomes you can expect:

1. **Users will be able to easily navigate your content.** The tags you associate with your documents will allow users to find the documents they need and to navigate the content they contain. In addition, the taxonomy branches become facets that enable you to add elements like semantic search or related topic recommendations.
2. **You can link content together.** Applying your taxonomy to multiple content silos allows you to link different pockets of content together across document sources. This highlights the importance of consistency in tagging so that you are able to link different sources and databases together. This also allows you to repurpose your archives and give them a second life (and increase their value), thanks to the variable insights they contain.
3. **Your content is made more valuable.** Overall, a taxonomy makes your content more discoverable, supporting easier navigation, increased relevancy of search results and a personalized experience, thanks to content recommendations.



Taxonomy Options: Build, Borrow or Both?

The good news is that you have options depending on your organizational preferences and the technology you can leverage.

A great starting point is to consider the following questions:

- Are there any existing standard taxonomies for your industry or the one you are targeting?
- Do you have any information, or in-house knowledge (aka folksonomies), that you can leverage?

Luckily, you don't have to start from scratch.

There are many industry-specific taxonomies available (MeSH for medical topics, IPTC for media topics are just two examples) that you can use as is, or as a starting point to reduce the time spent developing and customizing your own taxonomy. The TaxoBank database and WAND Inc. are great resources for finding these industry-specific, foundational taxonomies.

However, if you work in a very specialized domain or find yourself somewhere between the standards available, you will need to consider building your own content organization scheme. As you might expect, this can be a labor-intensive task that would require a major investment of time and subject matter expertise to properly tag content. But again, you don't have to start from zero.

Technology exists to help you jumpstart the creation of a taxonomy. Expert.ai offers an out-of-the box feature, called Magic Taxonomy, that will automatically classify any kind of document, such as white papers, news and newsletters, books, frameworks, articles, manuals reviews, etc. This ensures that your content meets the criteria we established above—properly understood, consistently and correctly tagged, connected across all of your repositories—while avoiding the time and expense of manually creating a taxonomy from scratch.

Our knowledge models contain concepts and relationships that are specific to different industries, domains, roles and use cases that can be used out of the box and further customized to meet the specific needs of your project. Our library includes knowledge models for the media, life sciences and healthcare and finance domains, as well as for ESG, sentiment and personally identifiable information (PII).

Taxonomy Design Considerations

Whether you borrow or build your taxonomy, there are several considerations that you'll need to manage to make sure you're successful. Here is a quick checklist:

- **Make sure your taxonomy mirrors your domain knowledge.** Your taxonomy is an opportunity to organize your information—your documents, insights and all your data assets—in a way that supports how your users would navigate your content.
- **Understand your target users. Is your audience the general public or is it a professional audience?** Again, it's important to make sure that the content or insight you deliver is understood by your users and that you classify and deliver your information so that it can be navigated the way that is understood by domain experts.
- **Set yourself up for success in content tagging.** This is a critical part of the process that is best performed by those with the deepest subject matter expertise in your organization. However, content tagging takes time, and these same resources are likely already engaged in other higher value work.
- **Get the right technology.** AI technology specialized in language understanding can provide the high-precision tagging you need, and subject matter experts can be brought in at various stages to help verify accuracy.
- **Enrich your content.** Then we recommend you link your established taxonomy to ALL of your target content, often referred to as **content enrichment**. This will optimize your results. Obviously, your taxonomy needs to be representative of the content it targets. It is designed to cover the variety of topics a dataset is addressing. This requires annotation and testing to improve classification accuracy results. Your taxonomy is a living organism that is fed by your content. Ultimately, you cannot create and test the taxonomy without a representative dataset.
- **Keep your users in the loop.** Users must be involved to improve the relevance of the solution design for information access and discovery. This is where you can really influence user adoption and see the impact on the end solution you create.

Common Taxonomy Q&As

1. **How can I create a balanced taxonomy that's neither too narrow nor too broad?**
   It's important to prevent taxonomies from becoming so narrow that they miss important data but not too broad as to produce unwanted information. If a term is too broad, it will produce so many results that it will not be useful. A term that is too narrow and only returns one piece of content out of a data set of 10,000 isn't useful either. There are two things you can do to avoid this problem:
   - Always test your taxonomy against your content to make sure it's actually useful.
   - Make sure you incorporate regular reviews to make sure that your taxonomy is keeping up with the subject area you're working with. A fast-moving domain where you're regularly adding new content should be reviewed every six months, while slower moving domains can be reviewed annually.

2. **Can I remove idioms and cliches from taxonomy vocabulary to avoid false positives?**
   Yes. You can leverage NLP platforms with semantic analysis to analyze a term to determine whether it is an idiom or a cliche. If you're looking at the term "hand" in a phrase like "handing down," NLP will ensure that "hand" is not interpreted as a body part. While you can also create filters and skill lists to eliminate those idioms, it's easier and more accurate to rely on NLP platforms to differentiate terms in your content.

3. **Can taxonomies be adapted to every business sector?**
   Yes—so long as there is enough information to classify. With enough content, you won't be limited in terms of the domains that can be addressed, nor how you choose to represent the information. If you don't want to build out your own taxonomy, you can use NLP systems to plug into an existing taxonomy (one that you created or an out-of-the-box taxonomy) and use knowledge models specific to your sectors(s), similar to that offered by expert.ai.

4. **Is it easier to create taxonomies for source documents that use a lot of jargon?**
   The short answer is no. Source documents that introduce a lot of jargon make it difficult to create a taxonomy for them, and it's not always easy to disambiguate the terminology—you may have one term used in multiple ways. The best approach is to capture the jargon in your controlled vocabulary as extra information, so if somebody searches using the jargon, they will still find their content. This ensures that your content is discoverable.

5. **How can I simplify taxonomy creation at scale?**
   There are many sources of foundational taxonomies online, both general and industry specific, that you can use as a starting point to create your own. TaxoBank and Wand are great places to start. The catch is that you could end up with a taxonomy that is more complex than what you need, with nodes that don't reflect the actual data being categorized. The advantages of using a platform like expert.ai is that we employ a bottom-up approach that leverages intuitive, semantic tools and driven by your actual content. This provides a solid foundation for a taxonomy that can grow and be easily reused across different knowledge discovery projects, for the life of your business.

Driving Successful Data Discovery: Recommendations

To summarize and review, here are some recommendations to guide you on your journey:

1. **Start with the goal that drives your discovery initiative.** Understand the obstacles you are faced with. Ask yourself: do you have unexplored archives, is your content hard to navigate, or is your search inefficient? Qualify the obstacles.
2. **Remember your options.** You don't have to start with a blank slate. You can rely on existing taxonomies or knowledge models, or you can leverage AI tools to build out a new taxonomy. You can even reuse and recycle taxonomies. This is where technology can really help.
3. **Make sure that your taxonomy properly covers the domain(s) represented by your content.** We can't emphasize this enough. The content that you will use to build your taxonomy must come from multiple representative data sets that reflect your business—don't leave anything out.
4. **Think about your taxonomy like it's a living organism.** As your business, customers and markets evolve, it's important to make sure your taxonomy evolves with it. You will want to occasionally measure how effectively you are solving your business problem with your taxonomy. Are you driving more engagement from your users? Is your team more efficient? Is your search more accurate? Are your archives fully represented in your taxonomy? What about your newest initiatives? These metrics will help ensure that you're on the right track.

References

1. Forrester, An Executive Guide to Text Analytics Strategy Findings from Forrester's 2022 Text Analytics Research September 22, 2022.
2. Solutions Review, 80 Percent of Your Data Will Be Unstructured in Five Years, March 28, 2019. https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/
3. TechCrunch, With a launch of COVID-19 data hub, the White House issues a 'call to action' for AI researchers, March 16, 2020. https://techcrunch.com/2020/03/16/coronavirus-machine-learning-cord-19-chan-zuckerberg-ostp/

Get Started

Contact expert.ai to learn more.



expert.ai

About us

Expert.ai (EXAI:IM) is a leading company in AI-based natural language software. Organizations in insurance, banking and finance, publishing, media and defense all rely on expert.ai to turn language into data, analyze and understand complex documents, accelerate intelligent process automation and improve decision making. Expert.ai's purpose-built natural language platform pairs simple and powerful tools with a proven hybrid AI approach that combines symbolic and machine learning to solve real-world problems and enhance business operations at speed and scale. With offices in Europe and North America, expert.ai serves global businesses such as AXA XL, Zurich Insurance Group, Generali, The Associated Press, Bloomberg INDG, BNP Paribas, Rabobank, Gannett and EBSCO.

www.expert.ai **info@expert.ai**

© 2023 Expert AI. All rights reserved.