EBOOK

Your Guide to Quality Drug Data

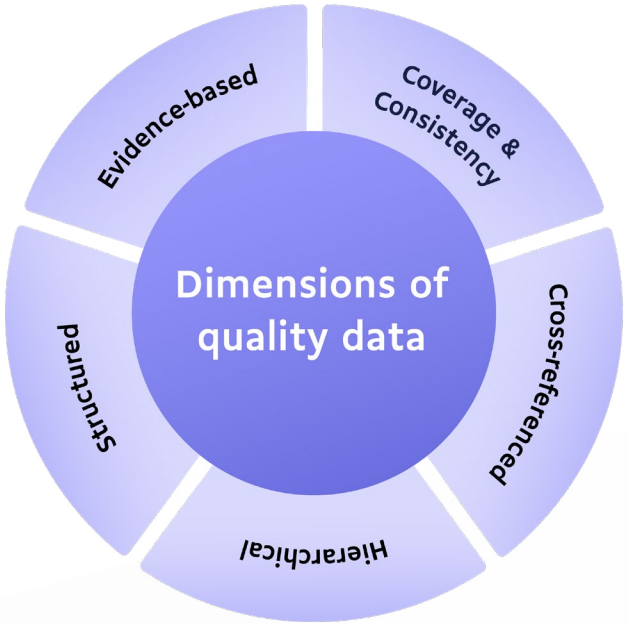
The whats, whys, and hows of quality drug data, according to our experts.

DRUGBANK

Contents

Dimensions of Quality Data

Quality data will never be the result of a single metric or of executing on one dimension of quality perfectly.



Rather, it is the sum of many crucial elements all working together, which is kind of a convoluted way of saying that none of our metrics for quality can be seen as the be-all-end-all of high-quality data.

It's also worth remembering that quality data for one user could be defined quite differently than for another, so we can't rely too heavily on one metric more than another. It is for these reasons that we take into careful consideration so many elements of quality.

In the following chapters, we'll be taking a deep dive into numerous dimensions of quality, how we define quality at DrugBank, and why we think you should care.

CHAPTER 1

DrugBank's Philosophy

When we say garbage in, garbage out we all know what we're talking about.

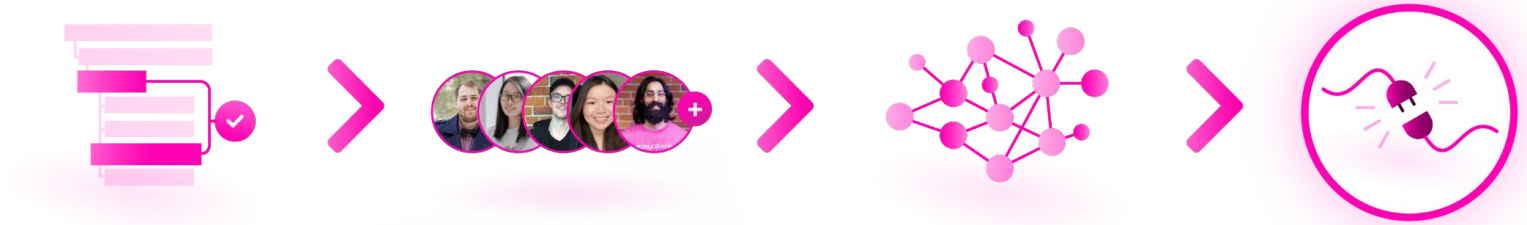
In order to make good decisions in our work and in our research we need high-quality data. Without quality data we're left making assumptions, bad decisions, and risking people's health.

Unfortunately, quality data can be hard to come by and sometimes even harder to recognize. With the rate at which biomedical findings are being published—reports have this somewhere in the range of two million journal articles being released every year—it is increasingly more challenging to manage and maintain reliable, up-to-date data.

Should you decide to take on this herculean task yourself, there are tough questions you need to navigate. What do you include or exclude? Where do you stop? How much data do you really need? How will you keep your data up to date? And, how do you know if the data you're collecting is accurate and evidence-based?

The challenges mount further when it comes to ambiguous and conflicting data. Now you aren't just responsible for collecting and organizing information, you need to be knowledgeable enough to discern and interpret findings so that you can be decisive. If this is done inconsistently your work, once again, suffers.

Knowing all this, DrugBank resolved to be relentless in our pursuit of high-quality data. This commitment meant laying out and adhering to the strictest criteria to ensure our data can always be trusted and relied on.



We use AI and machine learning to seek out the latest biomedical info

Our in-house experts verify everything and author new content

All approved data is organized, connected, and structured

Now it's ready for you to discover more

We work every day to uphold these standards by asking ourselves a series of questions:

Does it have quality coverage and is it consistent?

Coverage means knowing that our data sufficiently captures all relevant medical information while consistency focuses on how we input and maintain that data. At DrugBank we have strict curation specifications that all data must meet before it's incorporated into our datasets. Additionally, we've standardized a multi-step peer review process that's aided by automation and helps us deliver consistency and accuracy with speed.

Can you cross-reference related data?

With every piece of new data we add to our datasets, we're able to create and strengthen new connections between data points. These additional connections make our data more powerful and usable.

Is it hierarchical?

Quality data enables you to zoom in or out at the appropriate level to adapt the data to the problem you're trying to solve. In order for our data to qualify as high quality it needs to encompass the full variety and complexity of information available. To achieve this we incorporate as many levels of detail as possible.

How structured is it?

Structured data is easier to search, use, and reason with. We work to create highly structured, detailed data so that our users have total control over how they manipulate and explore it.

Is it evidence-based, can you follow the data-lineage, and does it have the appropriate meta-data?

Quality data is evidence-based. By ensuring that all our data is based on evidence, and its lineage can be traced, we are able to review and adjust as the underlying evidence changes over time. We also make sure that we can trace all curation actions so that we can internally audit and question our data. At any time, we can review who added or updated data and when they did it.

Now that we've got good data, does it work?

Similar to any other tools or products on the market, if our data is complicated to use it'll never be adopted and its potential can easily be lost. Once we had established strong metrics and processes for ensuring that our data is of the highest quality, we turned our focus to usability.

For us, that means offering data with a point of view. To accomplish this, our team started thinking deeply, exploring our data, and asking hard questions about our users and what they need to do their best work.

We pushed ourselves to provide an exceptionally usable structure that elegantly captures the greatest possible breadth of information without adding anything distracting or nonessential.

Then, because we know that DrugBank will never be a finished product, we created processes that cycle our team through these deep thinking phases so that we will always be growing and improving what we offer.

Data with a point of view isn't just about how meticulously we structure it, it's that we make intentional decisions about how we structure it. This has allowed us to build explicit relationships within our data in a consistent fashion that removes ambiguity. We obsess about the best ways to organize our data so that our users can quickly and easily start exploring and discovering more.

We also ensure that data never gets added to our knowledgebase without our in-house experts verifying its accuracy and integrity.

We pride ourselves on our holistic and iterative approach to building and evolving. Instead of recreating or deleting existing versions, we seek ways to add value to what we already have.

This pursuit for value frames how we do our work and has instilled in our team the highest sense of responsibility to continue learning and incorporating each lesson back into our systems. As a result, our team has become an invaluable resource that continues to sharpen their skills and build stronger datasets that are even more useful in the hands of our customers.

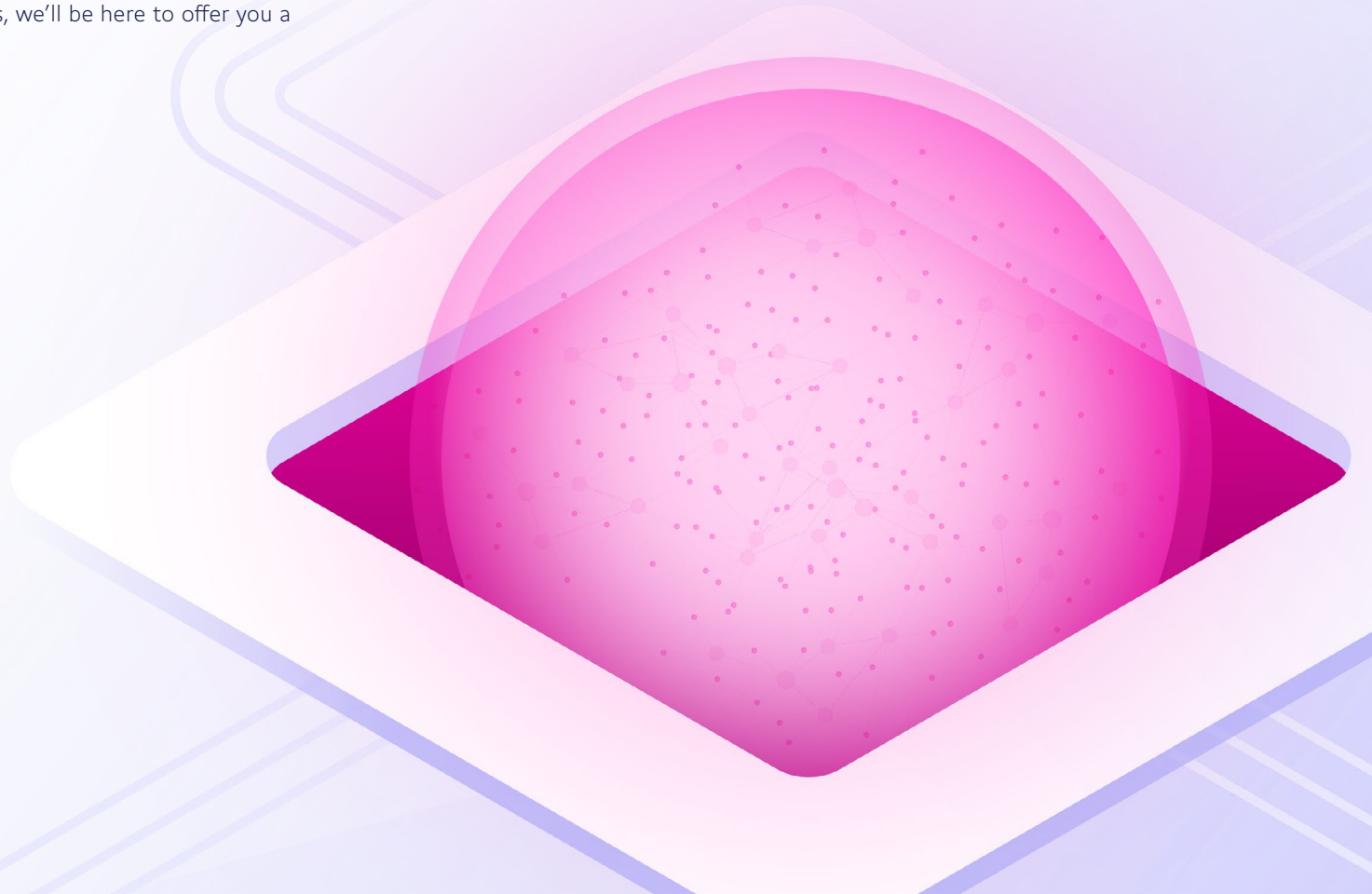
The manual curation process that our experts practice takes our data beyond merely being a compilation of the world's biomedical knowledge.

It transforms it into a highly trustworthy source that has been shaped with our users' needs in mind.

Of course there will be times when it makes the most sense to curate your own datasets. From our experience, this may be the case when your research or work is focused on a finite problem that you have a deep level of understanding about. In those instances, the amount and complexity of the data you need will be much smaller. If the scope of your work becomes much broader so too will the challenges you face. And in those instances, we'll be here to offer you a strong foundation that you can build from.

Our knowledgebase is now nearly 20 years in the making and it, and our team, has been growing and learning every step of the way.

We know the importance of high-quality, highly-usable data, and we know good research and confident decisions can't be done without it.



CHAPTER 2

Coverage & Consistency

The only limitations our data has is in our strict quality standards.

Quality through Coverage

WHAT IS IT?

In its simplest form, quality coverage at DrugBank means that we ensure our data covers the biomedical knowledge necessary to solve the problems at hand.

HOW DO WE DO IT?

To achieve appropriate coverage, DrugBank doesn't simply attempt to collect the most information. We aim to collect the most information while also checking every piece's accuracy and relevance. Theoretically, you could collect every single piece of data the world has to offer and you'd have perfect coverage, but it could be unusable or misleading if it isn't carefully vetted. When we seek data sources we work hard to collect and maintain coverage of data that is both true and valuable.

Next, we continuously ensure all of our data is well-maintained and up-to-date. Attaining quality coverage isn't an end in itself, but an ongoing process that we are always working on.

We have specifically designed proprietary AI that is constantly analyzing and seeking new information to add and improve our coverage. Then we integrate human expertise into the loop to provide feedback, vital checks and balances, and to ensure the overall accuracy of every piece of information that our AI brings in.

Our team also spends their days authoring novel content and supplementing our knowledgebase with information that they find through their own reading and researching. This multi-dimensional approach ensures that we have the greatest coverage possible.

WHY SHOULD YOU CARE?

First, we understand that each of our users have unique needs, and the data that one researcher or clinician might need can differ greatly from the next. For this reason, we sweat the details and obsess about overall coverage and consistency to ensure that no matter what you're looking for, we have it and you can trust it.

We also see how quickly biomedical information is growing and know how unmanageable a task it is for our users to source, analyze, and compile high-quality data on their own.

This growing body of evidence is becoming increasingly difficult to use, and when faced with such an overwhelming amount of data it can feel impossible to navigate what is evidence-based and useful information, versus what is contradictory and distracting.

We focus on maintaining non-redundant information that removes the burden of trying to extract meaning from the depths of the cumulative data available in the world. DrugBank's coverage goes beyond merely maintaining a vast scope of data, and ensures our coverage can aid in bridging gaps in knowledge.

With each additional connection made there will be the possibility for stronger evidence-based decisions and more confidence in research outcomes.

Quality through Consistency

WHAT IS IT?

At DrugBank, data consistency means that our customers can trust and expect that equivalent data will be presented the same way regardless of how, when, and where it is consumed. We ensure this by normalizing external sources and connecting the same data together rather than storing it redundantly. These connections are maintained through rigorous processes, regardless of who creates or integrates information on our team.

HOW DO WE DO IT?

For us, consistency is about rigor.

It's about having relentless standards for our data's completeness and structure, and then doing everything it takes to deliver on it every day.

In a lot of ways, consistency is a direct result of the fundamental activity of structuring data. At DrugBank we've established a strict process for standardizing our data so that it reliably translates third-party information into a similar and predictable format. We will cover structure more in-depth in a later chapter, but it's worth mentioning here because without it our data cannot be consistent.

Basically, in order for our data to be shaped the same way across datasets and sources, it all must be structured the same way. This ensures that no matter what data you access, you will experience it similarly to any other piece. And, it allows you to manipulate and put our data to work reliably across datasets.

Another vital element of consistency is completeness. Because some of the data and evidence we collect is from third parties, it is important that once we've verified and structured it, we then assess its completeness. With the help of automation, each new piece of data will pass through a minimum of two in-house experts that utilize multiple sources to identify inconsistencies, mistakes, and gaps in the information. Then, our team of in-house experts, aka our Curation Team and Data Review Specialists, look for ways to fill those gaps and right the inconsistencies.

This enables us to anticipate future problems, needs, or changes that are necessary to guarantee quality and consistency in our datasets.

WHY SHOULD YOU CARE?

Consistent data saves you from frustration and lets you focus your time and resources on what you do best.

When data is consistent it is easier to use and can rapidly integrate into research or applications.

When it's easier to use, it is faster to extract meaning from and emerge with reliable results or land on evidence-based decisions.

By maintaining well-organized datasets we are working to enable better decision-making by reducing errors and lowering the risks associated with unreliable information. But again, consistent data is only as useful as it is accurate and up-to-date with the most relevant and valuable evidence.

Conclusion

Data quality is about more than perfecting one singular metric. It's about a greater, more flexible set of interwoven dimensions of quality working together to solve problems. And, depending on a specific user's needs or the problems they are trying to solve, the metrics they prioritize can skew in a number of different directions.

At DrugBank we're not aiming to unlock some absolute form of quality.

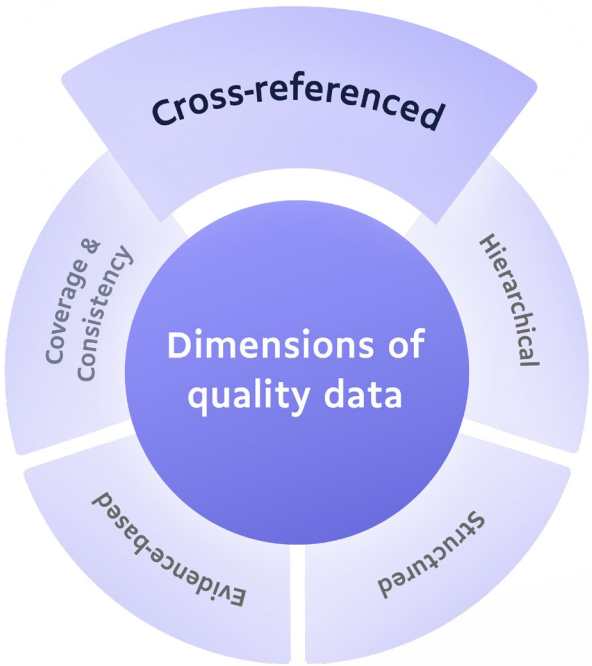
Instead, we are intent on obsessing over our customers' needs so that we can offer a multi-dimensional approach to quality that equips them with the best tools to solve their problems.

As we've discussed earlier, quality comes from having the right level of coverage within your data and strong consistency built into it. These two elements ensure you have the range of information you need as well as reliable, usable data you can trust.

CHAPTER 3

Cross-referenced

Or, ‘How I Learned To Stop Worrying And Love Common Data Entities’.



At DrugBank, we take data quality extremely seriously. When we aren’t honing our philosophies on quality, we’re inspecting our data with a fine-toothed comb to ensure each and every piece lives up to our strict standards.

Unfortunately, there’s no single metric that will deliver perfect, high-quality data. Instead, data quality is hugely dependent on the user’s needs and what they’re hoping to achieve.

As a result, we work tirelessly to remain flexible and responsive to changing sources and to our users’ various needs as a way of curating better, stronger data.

Common Data Entities and Cross-references

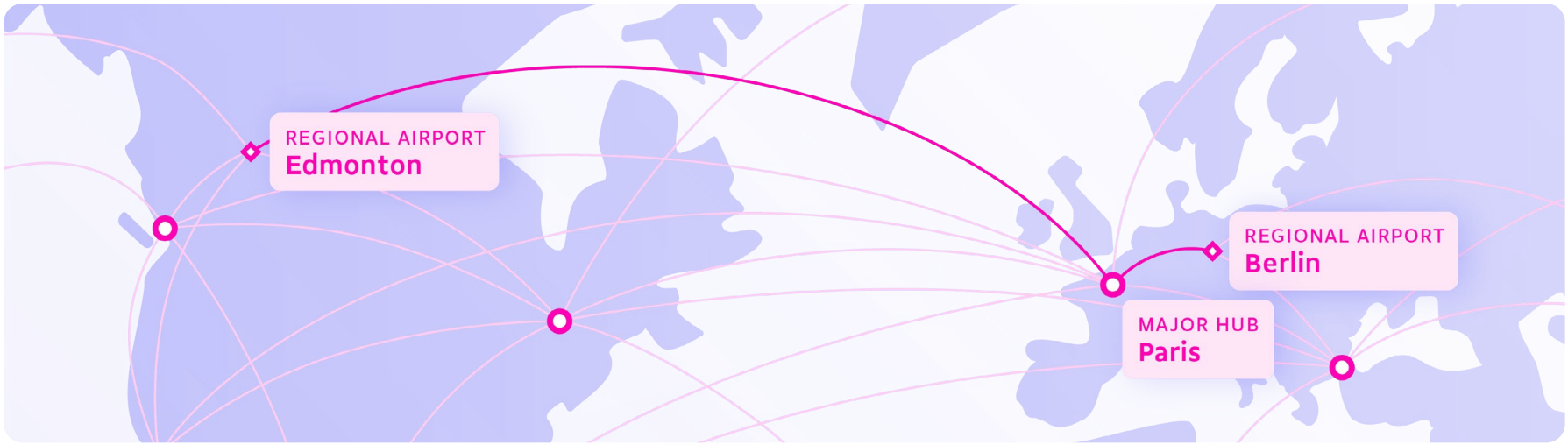
WHAT IS IT?

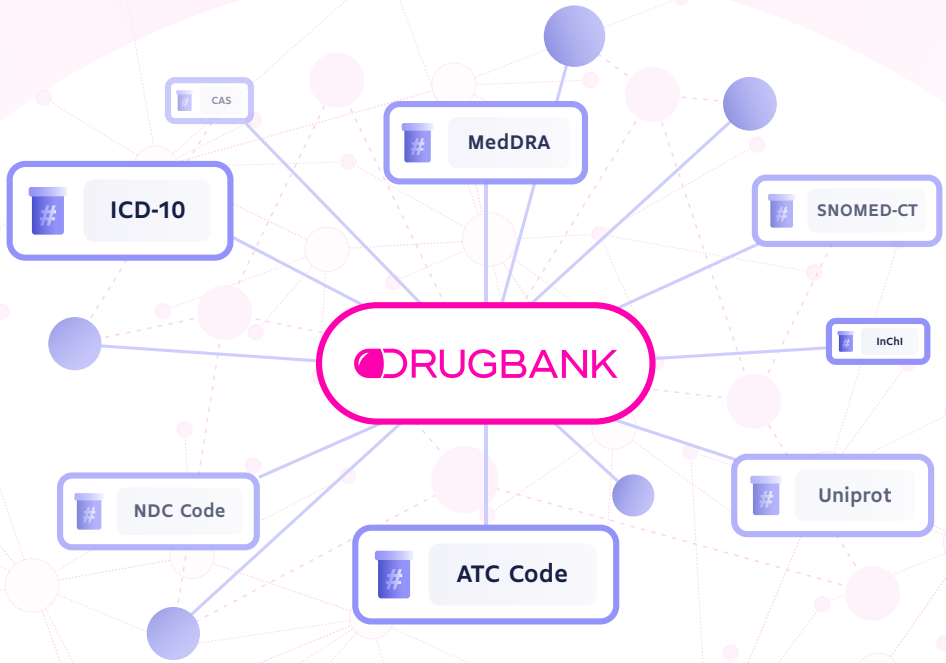
Our data is organized into datasets that our users can download or access through an API. Common data entities are simply pieces of data that are shared between multiple datasets. Each common data entity increases the number of potential connections between data, and the more common data entities we have, the more highly connected and cross-referenced our data can be.

To think of this another way, common data entities are kind of like airports:

> *There are smaller **regional ones** and **major international hubs**. The regional airports connect you to smaller cities as well as to the larger hubs, which can connect you to nearly any corner of the world.*
>
> *If our home base team in Edmonton wanted to get to Berlin, we might have to fly through a connecting city, such as Paris. In this analogy, the airports (Edmonton, Paris, and Berlin) would be our **common data entities** and each flight a vital connection between them.*
>
> *The **major hubs** would be the **common data entities** that have the most cross-references, whereas **regional airports** would be **less connected data** with fewer cross-references.*
>
> ***Well-connected hubs with ample route options makes it easier to explore more of the world faster.***

Similarly, well-connected data with ample common data entities makes it faster and easier to explore and learn from large amounts of data.





HOW DO WE DO IT?

At DrugBank, we excel at normalizing and enhancing drug data. Our team of experts work diligently to build connections between established and trustworthy external data, such as **RxNorm** and **NDC** so that our users can explore it with confidence.

To look at a specific instance in our knowledgebase, we can explore our conditions data. This structured hierarchical collection of medical terms and concepts is connected to a number of other datasets that can be used to 'jump' between different data points.

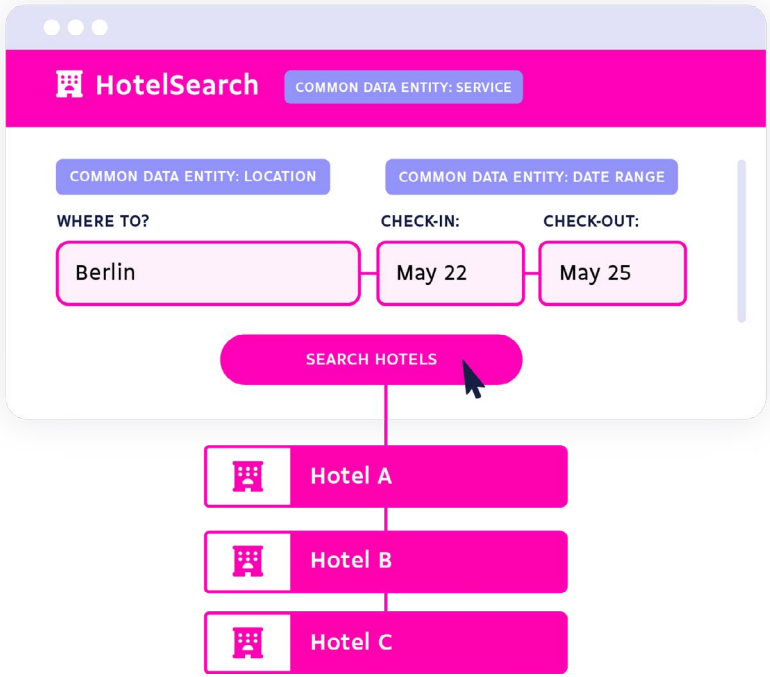
Whenever possible, we map our conditions to external ontologies, such as **SNOMED**, **ICD**, and **MedDRA**, to ensure both internal and external connectivity. This way, regardless of how you handle data internally, you will always have a common means of building external connections.

We've also developed an in-house tool that has made cross-referencing our data easier. With it, we can help our users map our data to their existing drug data (as long as the data they're cross referencing meets a few of our requirements). This tool was developed with the help of proprietary AI, which enables it to tolerate some level of imperfections in drug data and will find matches where traditional mapping methods typically fail. It also uses specially trained language models that do Named Entity Recognition (NER) in a scientific and pharmaceutical domain, thus automating the process of mapping raw text to common entities.

An example of this technology is search query analysis. Imagine searching:

> *I'm looking for* ***hotels in Berlin from May 22-25****. A NER model for a service that connects users and travel resources might identify* ***"hotels" (service), "Berlin" (location), and "May 22-25" (date range) as common entities.***
>
> *Once it does this, it can more easily connect that search to existing data it has on hotels in Berlin that are available on those dates.*

The machine learning model acts as the glue between the natural language a person would use to describe what they're looking for and the more traditional structured information that is better handled in a database.



WHY SHOULD YOU CARE?

All databases will have internal connections and cross-references to some extent, but it is important to understand the degree of connectivity within the data you're using.

If we take our airport analogy one step further and imagine that we're traveling around the world to solve a puzzle (discover a new drug or prescribe the correct medication) and each city (common data entity) holds a clue, it would track that the quicker we can move between cities, the faster we'll have a solution. Connectedness is what allows us to uncover novel information and insights.

To better illustrate the power of highly connected data, imagine a patient presenting with **condition X_3.** Utilizing our **conditions data** and its numerous connections to other datasets, we can explore potential causes of their condition.

For example, we might want to look into the patient's medications to search for drugs or drug products which are known to cause condition X_3 as an adverse effect:

Perhaps one of their medications is causing an allergic reaction, or two of their medications are interacting with each other to result in ***condition*** X_3*. All of these question can be explored by leveraging connections between* ***condition*** X_3 *and other datasets.*

Then we can go even further. By searching for drugs indicated to treat ***condition*** X_3 *(and then looking at the drug-protein targets for those drugs), we can perhaps glean the mechanism through which the condition develops.*

On a broader scale, this knowledge can aid pharmaceutical companies in developing new therapies (or repurposing old ones) for the treatment of ***condition*** X_3*.*

Suppose ***condition*** X_3 *is rare and still poorly understood, we can still explore potential experimental treatments by examining clinical trials in which the condition is being investigated. Barring that, the hierarchical nature of DrugBank's conditions data allows us to simply navigate up to the parent (or less specific)* ***condition X****. From here, we can explore drugs indicated for the treatment of* ***condition X*** *generally, rather than* ***condition*** X_3 *specifically.*

Essentially what we're saying is that the whole is worth exponentially more than the sum of its parts. With this in mind, DrugBank has created an elaborate network of common data entities within our knowledgebase.

This intricate level of connectivity makes our data much more valuable than if it was merely listed and unconnected.

Conclusion

Assessing a dataset's quality in terms of cross-references or common data entities comes down to more than just the existence of common connections.

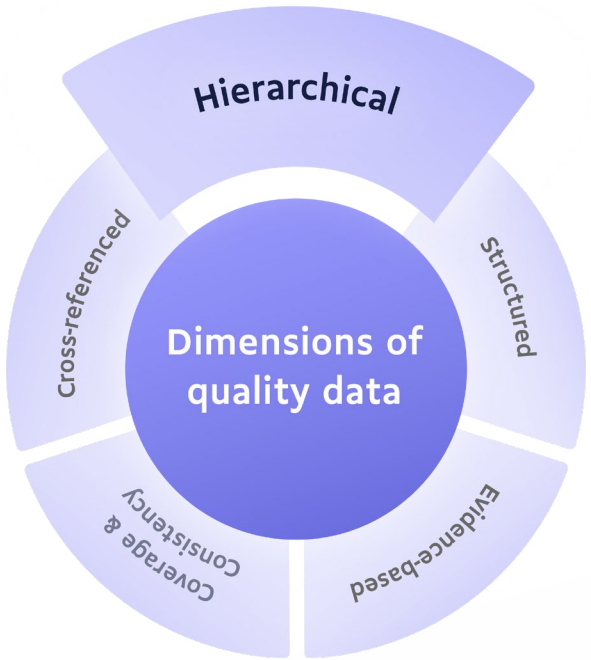
Instead, it is more important to ask how extensive those connections are and what you will lose if you work with data that doesn't prioritize them.

These vital hubs of data enable greater exploration, better flexibility in your work, and faster routes to uncovering insights that would otherwise be challenging, if not impossible, to make.

CHAPTER 4

Hierarchical

In the game of quality data, hierarchies reign supreme.



Data, and how to make ours undeniably great, is one of those things that keep us up at night.

Not because we're worried we can't do it, but because we want to be sure we're not missing opportunities to iterate and improve what we do. One thing that ends up being tricky is that quality is a moving target depending on who you ask.

Every user has unique needs and how they define quality will depend on how they plan to use the data. For this reason we work tirelessly to remain flexible and responsive to changing data sources and to our user's needs. We also work to balance the many dimensions of quality data so that we are always delivering quality you can count on.

The third dimension that we take into consideration is how hierarchical our data is, so in this chapter we'll be exploring what hierarchical data is and why it is such a vital piece of all quality datasets.

WHAT IS IT?

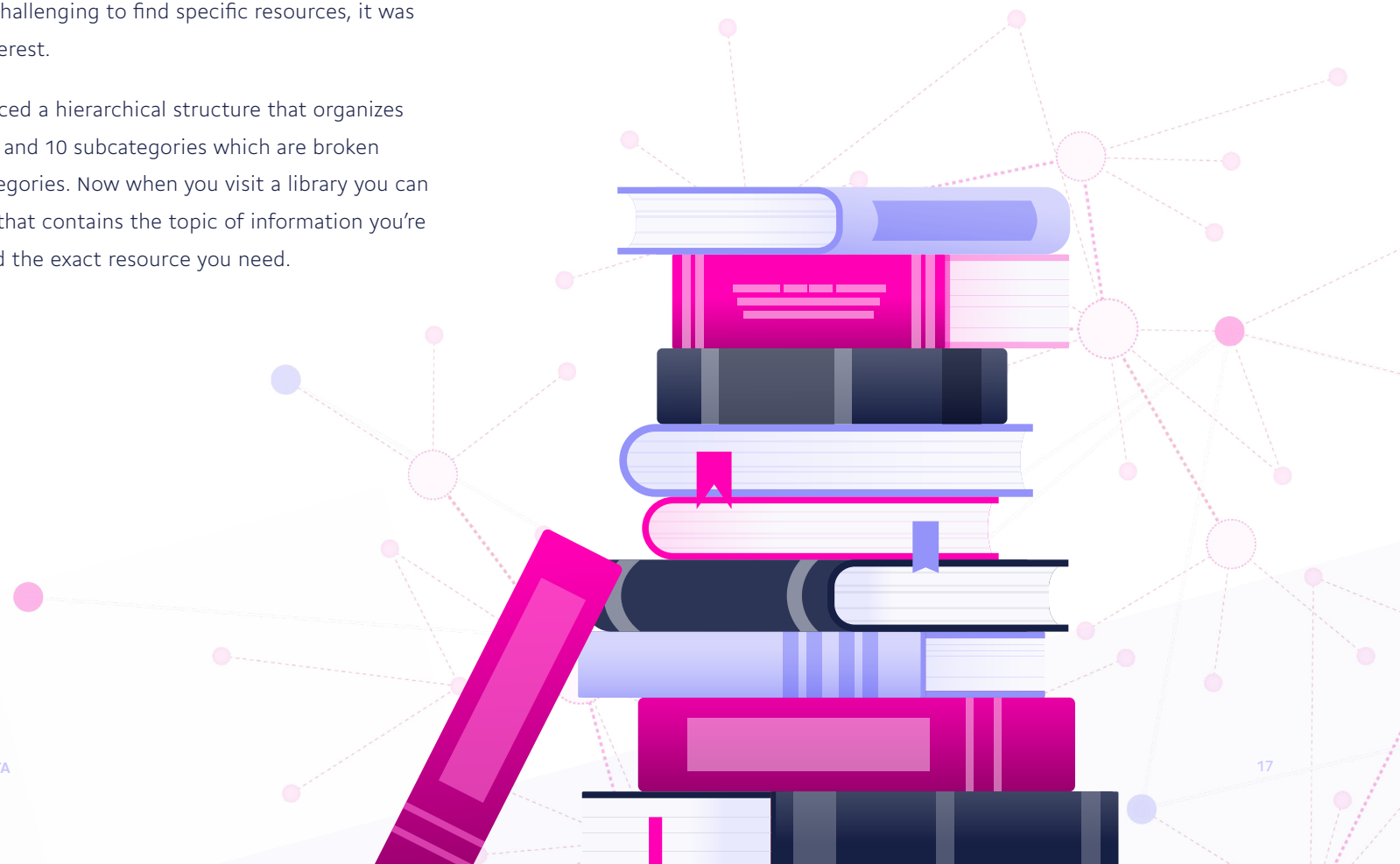
Hierarchies are a means of organizing information or items based on status, importance, or lineage. An obvious and very common example is a family tree. Another great example is a phylogenetic tree. This branching diagram is used to illustrate evolutionary relationships between organism, and with all life on Earth organized in a singular phylogenetic tree it is easy to trace through the evolutionary history of any known species.

Another great example is the Dewey Decimal System. Prior to its creation, books were stored in a permanent location based on when they were added to a library. Not only did this make it challenging to find specific resources, it was impossible to browse by area of interest.

The Dewey Decimal System introduced a hierarchical structure that organizes resources based on 10 main classes and 10 subcategories which are broken down further into 10 additional categories. Now when you visit a library you can navigate to an area of the building that contains the topic of information you're looking for or quickly and easily find the exact resource you need.

Hierarchies make it easy to work with large amounts of information, track connections, and identify relationships.

As a result, they are also extremely useful in the biomedical field and have proven to be helpful in simplifying the complexity of biomedical data and phenomena.



HOW DO WE DO IT?

At DrugBank we spend a lot of time thinking about how we can best organize our data so that it brings the most value and benefit to the researchers and clinicians who rely on us every day. We lean on the organizational effectiveness of hierarchies to structure our data in a way that provides both utility and ease of use.

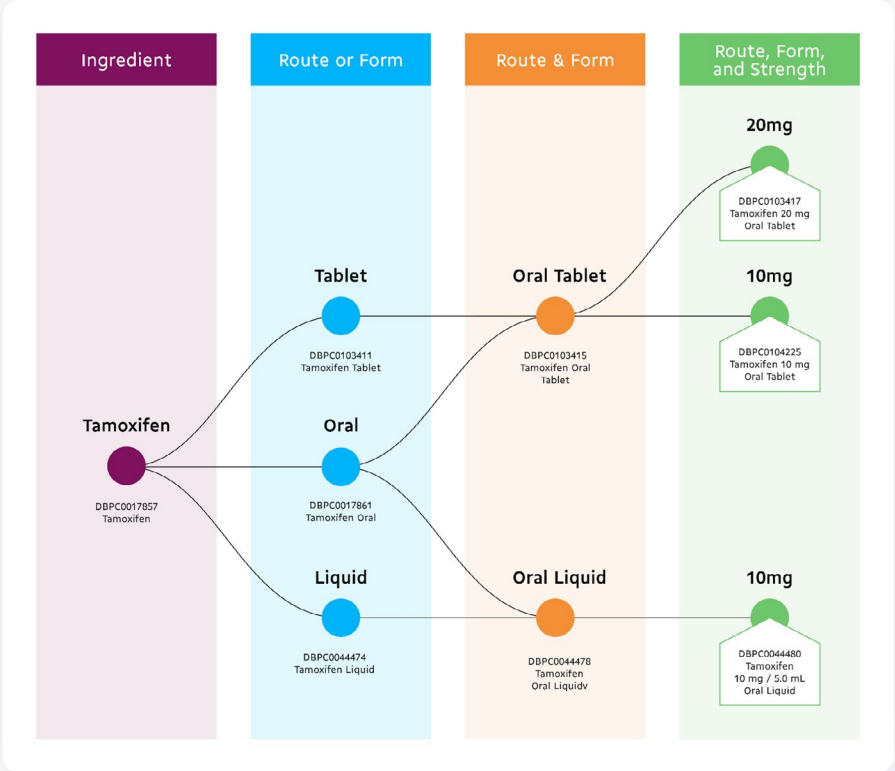
For every dataset we model we look for levels, groupings, and any natural hierarchies that are inherent to the data. Parent-child relationships are common, especially in biomedical data where concepts exist at varying levels of specificity. To further demonstrate this we'll take a look at two specific examples.

First, our **product concepts**. In its simplest form, product concepts are a hierarchical index of the many elements of a drug product.

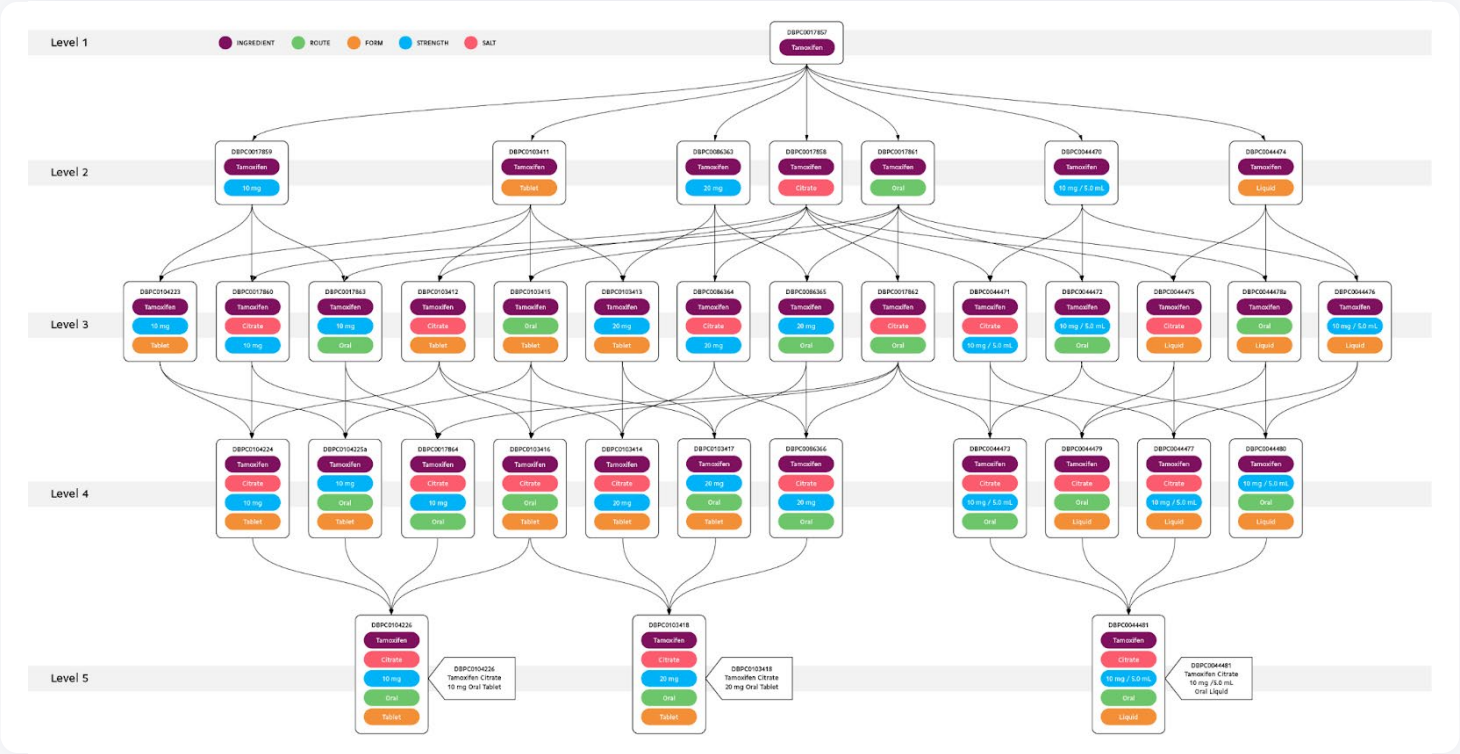
For example, a drug product may have one or more components used to treat a specific condition (**ingredients**), it may be available as something that can be injected or taken by mouth (**route**), and employ different methods of delivery, such as tablets or liquids (**form**). Within product concepts we've established levels, where each level indicates an additional piece of information that in turn makes the product concept more specific.

In this example we can see how a new property is added at every level, going from left to right.

*Starting with ingredient (**tamoxifen**), the next level adds either route (**oral**) or form (**liquid/tablet**). At the next level, both route and form are present. Finally, strengths (**10mg, 20mg, 10mg/5.0mL**) are added to each of the concepts in the previous level. As the level of detail increases, the product concept matches fewer products.*



(Extremely simplified example)



(Zooming out, we can see just how complex these hierarchies can be)

Although we have shown an extremely simplified version of the extensive product concept hierarchy we maintain, it can quickly end up looking quite complex. However, it is this complexity that provides users free rein to explore with very few limitations. Once expanded with all relevant concepts our product concepts end up being a much more elaborate hierarchical web.

The hierarchy of product concepts provides different degrees of granularity and allows users to search through products at whatever level of specificity is most valuable to them.

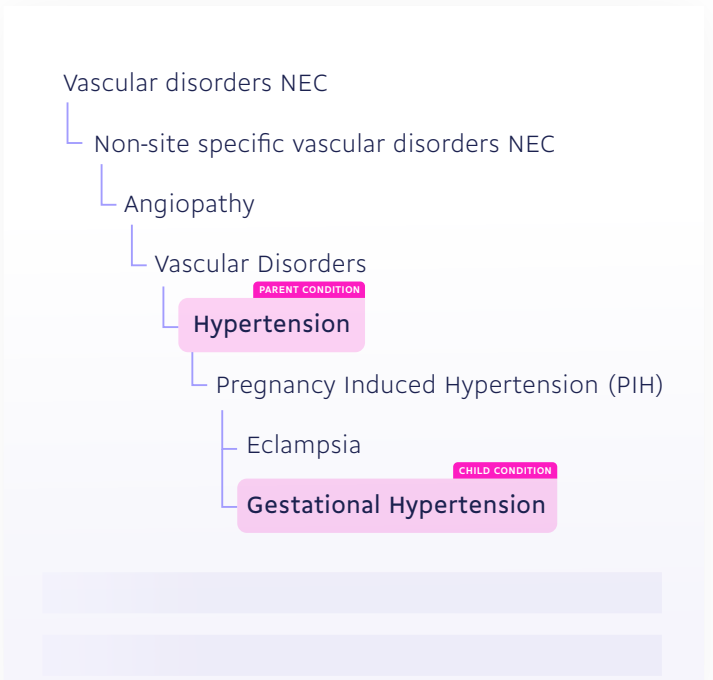
This makes it much easier for our users to manipulate such complex information. It also creates a great deal of flexibility for users to search within the data, or associate our product concepts to external datasets.

Next, we're going to look at our use of **conditions**. **Conditions** in DrugBank represent anything related to patients or people in general. They include diseases such as asthma or hypertension, symptoms and adverse events such as headache or pain, and other characteristics relevant in a medical setting, such as weight or lab test values. Each condition is associated with a specific ID, making queries in DrugBank easier.

We organize conditions in hierarchies where each one is assigned parent-child relationships.

One example of this would be the relationship between **Hypertension** and **Gestational Hypertension**, where Gestational Hypertension is a type of Hypertension, and therefore the "child" of this condition. Parent-child relationships can also be built based on combinations or modifications of conditions.

Thanks to this hierarchy, queries for conditions can be made based on the names of conditions, or more specific or general forms of those conditions.

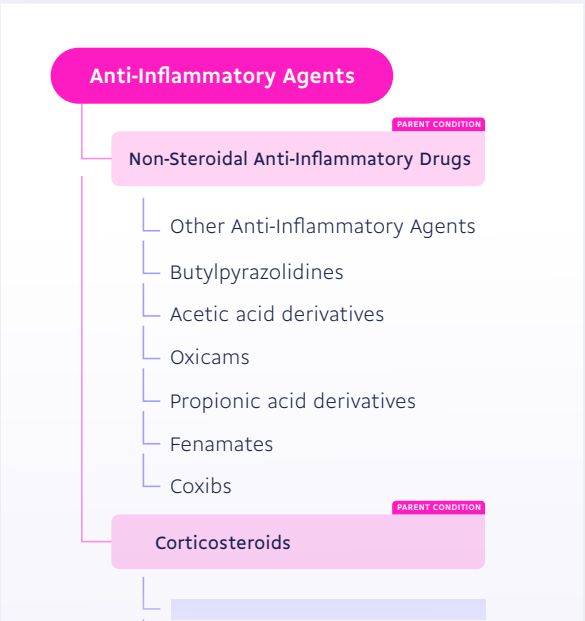


WHY SHOULD YOU CARE?

The organization of data into a hierarchical structure provides a number of benefits. Firstly, data organized in this way is intuitive. Humans (and, indeed, computers) exist in a world full of hierarchical structures, and by adopting these structures as a means of organizing data we can use the ubiquity of hierarchies to our advantage. Let's look at some simple data regarding drug classes to better illustrate this point.

Without any additional context or domain knowledge, we can draw several inferences from this hierarchy.

> *For example, **NSAIDs** as a whole are a type of anti-inflammatory agent, as are **corticosteroids**, although **NSAIDs** and **corticosteroids** are distinct from one another. Similarly, we can immediately intuit that **fenamates**, **coxibs**, and **oxicams** are all examples of **NSAIDs**, but are distinct enough to each warrant their own sub-class. We could also surmise that drugs appearing within '**Other Anti-Inflammatory Agents**' are, as you might imagine, not **fenamates**, **coxibs**, etc.*



With even a quick glance, we can glean a significant amount of information by simply organizing this data into a hierarchical structure.

This holds true even as the amount of data grows—as long as the hierarchy remains intact, it can handle vast amounts of data while ensuring the data contained within remains easily accessible and intuitive.

Hierarchical data is also extremely flexible. The hierarchy provides a structure within which we can view the data at multiple levels of granularity—we can zoom in to view, edit, or otherwise manipulate single pieces of data (the end points) or zoom out for larger swathes of data with ease (such as examining mid-level data points).

Similarly to how the connectivity of our highly cross-referenced data makes it easy to traverse our datasets, the hierarchy of our data provides a means of exploring the different elements of a single dataset.

Hierarchical organization, combined with extensive cross-referencing, provides the ultimate flexibility for our customers to navigate our data in whatever way suits them best.

Conclusion

If a library doesn't employ the Dewey Decimal System there is no straightforward way to explore subject areas and serendipitously make connections that could help you see things differently.

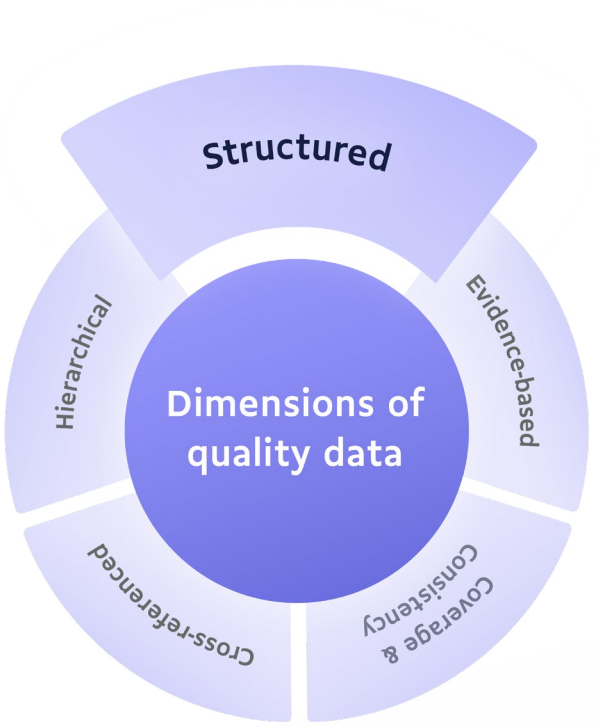
Even if you thought you knew exactly what you were looking for it could be tiresome trying to pinpoint a resource's location. Then, if you did find the resource you were searching for, there would be no easy way to investigate related content or bridge gaps in your knowledge.

The same rules apply to data. Hierarchical structures organize data in a way that puts the power in the hands of the searcher to manipulate, explore, and make discoveries that would otherwise have been exceedingly difficult to achieve.

CHAPTER 5

Structured

Making it easy to find meaning amongst the chaos.



WHAT IS IT?

To understand structured data it can be useful to start with what data looks like when it isn't structured. Most unstructured data is simply raw text that requires interpretation for it to be used, however it can include many other mediums of information such as audio and video files. In this format, it can be efficiently exchanged between people, but computers would find it challenging to work with.

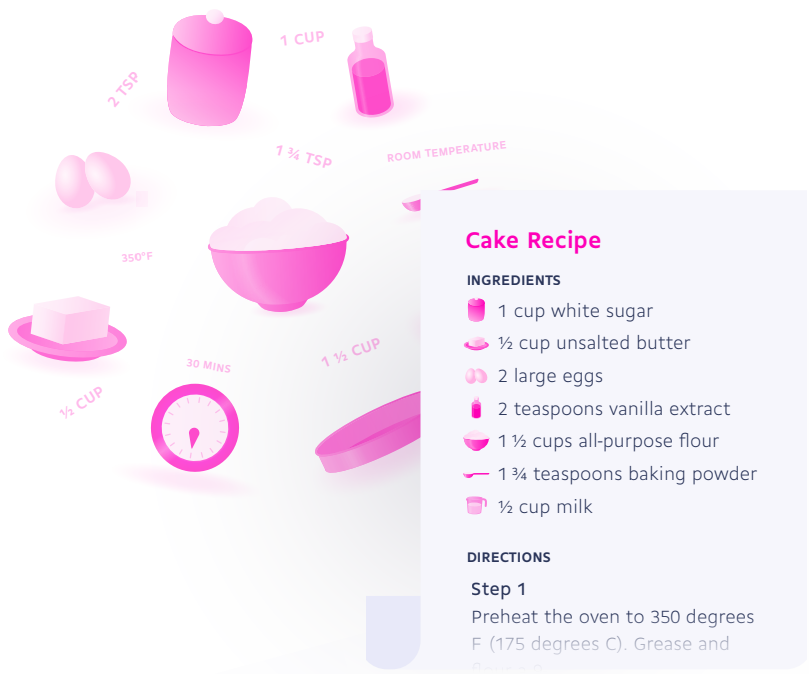
Structured data refers to information that has been organized and classified in a specific way, making it easy to access and process. It can also be thought of as pre-interpreted with key facts and points pulled out. The user is then only responsible for deciding which pieces of data to work with, and what to do with them.

Consider the process of making a cake. An analogy for structured versus unstructured data would be a recipe card versus a how-to video.

You could go online and watch a quick step-by-step video that would help you understand the overall process and what types of ingredients you would need to make a cake. But in order to actually make the cake, you would need to interpret the video to get a specific list of ingredients, then figure out measurements, timing, temperatures, and the order of operations.

Although you can learn to bake a cake in this way, it would require a high level of interpretation. You also wouldn't be able to take a playlist of cake-baking videos and easily pick out a recipe that matches the ingredients and tools you have on hand. To do that, you would have to watch each video and figure out the process one by one.

In a recipe card, all that legwork would be done for you and you could simply follow along the clearly organized format, checking back whenever you needed to. This structured data removes the need for the user to do any interpretation making it an easier process to successfully bake a cake.



The structured drug data provided by DrugBank works similarly. Different types of information about drugs are classified and organized in a way that is easy to find.

This allows clinical and pharmaceutical companies to build their own solutions, academics to do vital research, the general public to easily access information, and individuals in clinical care to make decisions with confidence.

HOW DO WE DO IT?

The structure of data is like the design of an object. The handle on a kettle enables and encourages you to pick it up. By following the design, you're more likely to use it correctly and avoid burning your hand. When we create a data structure, we approach it in much the same way. We consider how our users want to be able to manipulate our data as well as what uses and applications we want to encourage or make easier. Then we structure our data in ways that will ensure these uses are as simple as possible.

To turn unstructured data into structured data we rely heavily on our curation and development teams, which are stacked with subject matter experts. These teams work alongside one another to identify and define common data entities and attributes for different drug datasets (attributes such as **drug name, patient characteristic, route, dosage,** and **form**).

This process involves investigating and analyzing drug data to determine how it can best be connected and cross-referenced. Our team is also always reassessing and establishing strict curation standards to ensure that all our structured data is consistent.

Let's look at a few examples of structured data that we maintain:

Indications, contraindications, and adverse effects

DrugBank's users can view the **indications, contraindications,** and **adverse effects** for every approved drug. Instead of seeing them as text descriptions (unstructured data), **indications, contraindications,** and **adverse effects** are structured in their simplest terms, without losing the level of detail that is provided in a text.

Drug metabolism

This dataset allows people to see the different steps involved in the metabolism of drugs. When a drug is degraded by the body, it produces different compounds or metabolites that may or may not have an effect on a person. DrugBank associates metabolites in an organized manner and maps them to other sets of information such as **pharmacology, drug targets** and **other drug-protein relationships**, and **SNP data**.

Drug-drug interactions

When a patient takes more than one drug, there is a possibility that the two will interact with each other. With **drug-drug interactions**, it is important to know what the severity of the interaction is, the reason why it happens, and how it needs to be managed. DrugBank has structured this information so that it can be easily queried.

WHY SHOULD YOU CARE?

Structured data is important from both a user's perspective and for what it allows us to do on the backend of our knowledgebase.

For us, structured data is foundational and key to maintaining many of our dimensions of quality data. Structured data is built around common data entities, making it easily cross-referenced and interpretable, and it also encourages consistency and lets us easily measure and improve on our coverage.

Further, structure enables us to do validations and ensure quality and usefulness at the time of data creation. If we determine that a property is very important to our customers, we can require it in all instances of that data structure. If we determine that a property can be expressed with a limited vocabulary without losing any value, we will favor this approach as it reduces the burden of interpretation.

The data's structure must align with the function it will be used for, otherwise, it will be difficult to use, easy to misinterpret, and cause frustrations.

Carefully structuring our data ensures we are able to deliver a highly usable product that reduces the number of decisions a user needs to make in order to work with the data.

For scientists and software developers, structured data is simply more usable than unstructured data. It is easy to integrate with existing data and systems, develop with, and requires less time to implement, which saves money.

Structured data is also ideal for software integrations including artificial intelligence, machine learning, and algorithm development. Because it maintains such a strict format, structured data provides a great deal of flexibility for users to manipulate it in ways that meet their unique needs. Whether that is using structured indications data for clinical decision support or as a convenient drug-drug interaction tool.

It also makes connecting to external datasets, ontologies, and resources (such as ICD-10 and UniProt) very easy. The more structured a dataset is, the more straightforward it will be to build relationships between drugs, drug products, and conditions.

Ultimately this makes it that much easier to incorporate all the relevant data one might need for their research or into their decision making processes.

Conclusion

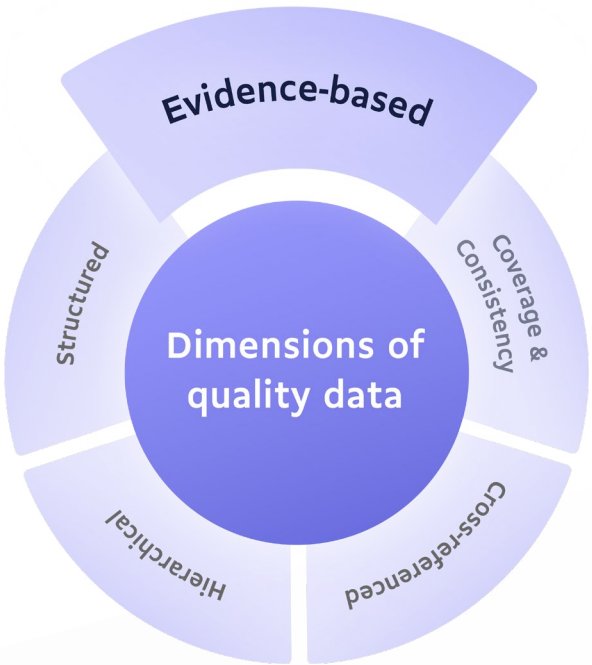
DrugBank understands the value and versatility of structured data. We have organized drug information in a way that is easy to find and retrieve, keeping in mind the different applications and tools our users work with.

By structuring data, we improve consistency and connectivity, and therefore, its quality.

CHAPTER 6

Evidence-based, Data Lineage, & Metadata

Knowing your data's history can be the difference between a big discovery and a big waste of time.



WHAT IS IT?

The scientific process is not linear, it is complex, messy, and requires constant re-evaluation.

At DrugBank we believe that in order for data to be reliable it must represent this reality. Therefore we prioritize data that is structured in a way that models the real scientific perspective, recognizing that sometimes science gets it wrong, and that scientific findings and how we understand them will change over time.

To do this we need to keep track of data changes over time and what has led to those changes, which has led us to intentionally building in data lineage to our knowledgebase.

Data lineage is similar to a timeline, telling us when and where the data was created, when it was connected to other data points, where it is being used, and when big changes were made, among other things.

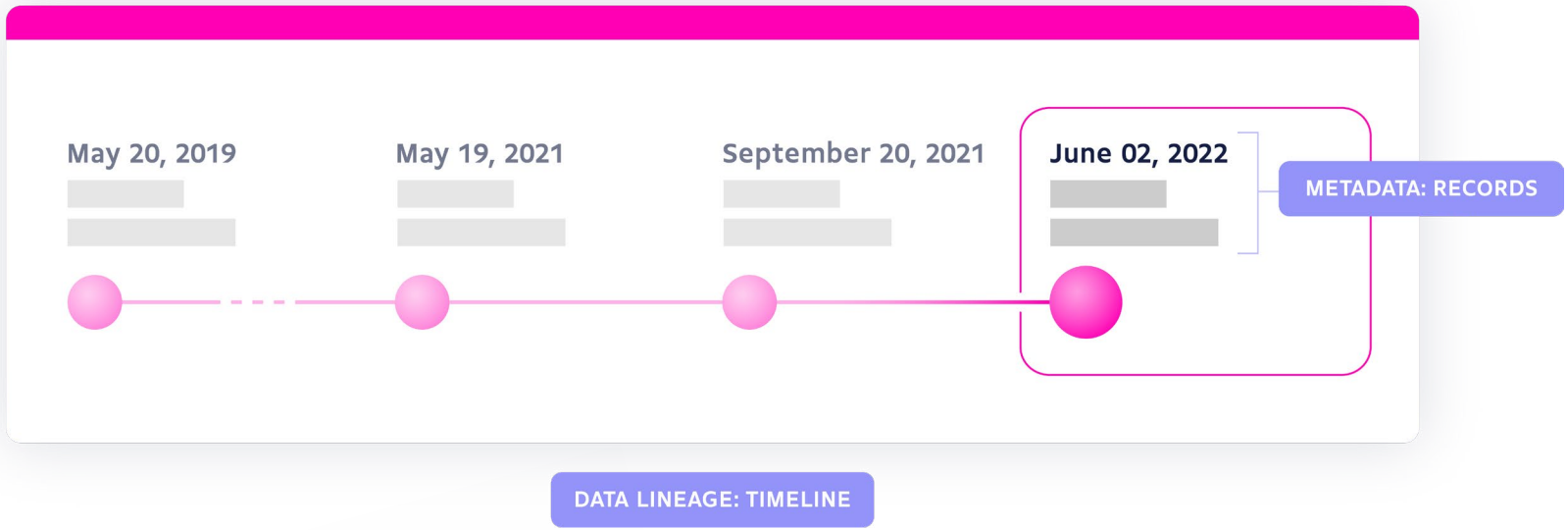
This traceable structure makes it easier for us to understand and respond to new findings and look back to where errors may have occurred. This is where metadata, or data about data, comes in handy.

Metadata can most simply be explained as the record, references, and information about where pieces of data came from.

*If **data lineage** is our **timeline**, then **metadata** are the **historical records** that provide us with a more detailed background. Both of these pieces, combined with our commitment to only using data that is rooted in evidence, create a highly traceable and robust record for us and our users to rely on.*

However, complications can arise when we try to define what evidence-based data is and determine what pieces of evidence to trust. Simply having references of papers that came to certain conclusions is a good starting point, but we have found that we must go beyond this.

Often conclusions are disproven, discredited, or contradicted so instead we make a point to build a base of evidence that meets a certain threshold. We'll explore this approach more in-depth in our next section.



HOW DO WE DO IT?

At DrugBank, we believe that decisions about traceability and data lineage must be made from an organizational perspective, and they must occur before we even begin work on our data.

This ensures that our entire team is delivering the same standard of quality. For us, this can mean negotiating what a sufficient level of metadata looks like, or setting criteria for the extent or depth of information we need to validate our conclusions. We need to be able to agree on what our minimum base of evidence is ahead of our work so that only the pieces that meet or exceed those requirements are included.

Once we've established these metrics we can work on execution. This part of the process is where we rely heavily on our expert team of curators and data review specialists. These talented individuals work together to measure all our data against our strict standards.

When seeking evidence-based data we look for recent, well-established journals, and comprehensive national clinical guidelines with information verified by clinical experts. Then we weigh these findings against our minimum base of evidence requirements to verify validity and accuracy.

If our curation team deems the data to meet our evidence requirements it will be added to our datasets and filtered into its place along the data lineage. This allows us to complete our routine data quality assurance checks, which is a review of older data to see when it was last updated and if it is still up to our curation's standards.

Data lineage also enables us to complete data quality control (which is a process of checking data origin and movement), as well as internal quality assurances (which ensures our curators are capturing data accurately and consistently according to our standards). In addition to building a strong dataset and knowledgebase, these processes ensure accountability for our team and deliver transparency to our users.

These same standards extend to our requirements for metadata. Within our curation practices are baked-in requirements for every single piece of data to have recorded where we derived that information.

This ensures that all data in DrugBank is from the best, most reliable sources.

WHY SHOULD YOU CARE?

Whether working as a researcher or a clinician, it is vitally important to have data that is strongly rooted in evidence. Quite simply, data that cannot be verified is essentially useless. Without rigorous processes and high standards in place, it can be nearly impossible to work with any level of certainty. Each conclusion or finding can quickly be drawn into question.

If, however, there is a robust system in place it creates confidence, increases safety, and improves transparency and accountability. If biases are identified in data it is much easier to look back and determine where they have come from, the impacts they've made, and how to course correct. From a clinical perspective, evidence-based information is essential to confident and quick decision making and also improves the overall quality of patient care.

Further yet, when time is less pressed, a thorough data lineage with ample metadata allows for an in-depth review of underlying evidence and can provide additional nuance that would otherwise be unattainable. And, as decision-making processes evolve with new knowledge, data lineage provides a way to track and learn from those advances.

As a researcher or pharmaceutical data user, reliable evidence-based data means saving time and money by avoiding missteps or incorrect findings. Users will have to suffer through fewer instances of pursuing avenues based on uncertain evidence that later is revealed to be questionable.

Additionally, when we are able to maintain our organizational reputation by publishing findings that are based on solid evidence, there is much lower likelihood that those same findings will later be retracted.

Conclusion

This threefold dimension of quality comes down to answering the question of why. Why does this data exist and why can you trust it and rely on it to produce accurate results?

By having data lineage, metadata, and evidence-based data we are able to document the scientific process which gives us some ability to identify and mitigate bias, as well as an ability to stay flexible to the shifting conclusions emerging from the biomedical world.

These vital steps ensure we are always working toward maintaining the most reliable data possible.

PARTING THOUGHTS

Creating an Ecosystem of Quality

Data users and their needs come in all shapes and sizes, so there's no one right way to measure or manufacture quality.

Throughout this eBook, we tapped into the deep pool of expertise here at DrugBank and have come to a very simple conclusion, defining quality in data is difficult. Fortunately, we were up for the challenge of nailing down this somewhat slippery, amorphous problem. As we pursued a concise definition, we found that our convictions that data quality must be individualized and audience-specific only grew stronger.

It is for this reason that we've continued to reiterate that the individual dimensions are not the be-all-end-all of quality. Quality isn't even the sum of these dimensions. Yes, they should all be present, but it is having the right balance depending on any one user's needs, that results in quality.

The five dimensions we covered (**coverage and consistency**, **common data entities and cross-references**, **hierarchies**, **structure**, and **evidence-based data with data lineage and meta-data**) come together to create extremely dynamic, quality-rich data. Depending on a user's needs they may favour certain dimensions more than others, but ultimately the whole is worth more than each individual part.

Consider that common data entities and cross-references are key for someone looking to make discoveries by combining many sources of information, whereas structured data might be more important for an ML practitioner looking to build a supervised learning model. The five dimensions create an ecosystem of quality even if the user is prioritizing specific metrics in their own work.

Our hope is that this framework makes it easier to assess the quality of your own data, or the data you are considering integrating into your work. We suggest starting your examination not by looking at the data you are interested in, but by closely assessing your own needs and priorities first. Then you can more confidently measure each dataset or data source against your needs.

Ask yourself:

- What dimension(s) of quality are most important to my work or specific project?
- How does this data deliver on those needs?
- How well will this data continue to deliver on those dimensions into the future?

Once you've found data that meets your needs, start to investigate its reliability. When it comes down to it, if you can't rely on your data your work will suffer. We suggest digging into sources to understand where the data comes from, who collected it, how they are dealing with biases and conflicting information, and how regularly they are updating it.

Without these pieces it is easy to slide into unreliable, conflicting, or out-dated data that could discredit any findings or decisions you make.

We know that for many teams it makes sense for them to handle their own data. However, if you're finding that building and maintaining data is occupying more of your time than actually doing the work you set out to do, you might benefit from finding a trusted partner to lighten the load.

Partnerships can take on many different configurations. Some might need a partner that can handle all of their biomedical information, but for others it might make more sense to approach the partnership from a hybrid perspective.

If you have ample in-house expertise maybe there are some gaps in knowledge that external datasets could fill. No one source can be the gold standard, so it will often be a matter of building connections between several datasets. We have built the DrugBank Knowledgebase in a way that can easily act as a backbone for our users to easily integrate multiple sources of in-house data.

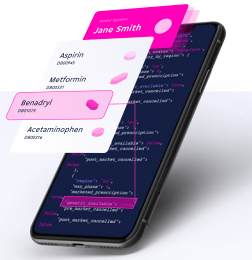
At DrugBank we aim to take care of all the behind-the-scenes parts so that our users can focus on what they do best, without wasting their time and money fussing with data collection and organization. Everyday we work alongside a wide variety of users around the world to help push the limits of biomedical knowledge. Our users range from pharmaceutical and academic researchers, clinical software developers, as well members of the public who are able to access our data through **DrugBank Online,** our free-to-use website.

We’re putting power into the hands of those who can do the most good.

Discover Drugs

We’ve seen pharmaceutical teams speed up their drug discovery process from **12 years to 3 months**

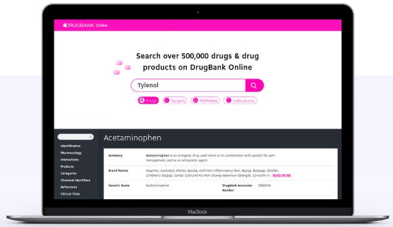
Dataset Downloads



Power Clinical Software

Developers put knowledge in the hands of prescribing physicians to **make clinical decisions with confidence**

Clinical API



DrugBank Online

We’ve even heard from DrugBank team members that their aunts & uncles have landed on DrugBank Online to learn **something new about medications**

go.drugbank.com

Don’t take our word for it, start using our tools today to find out if

we’re the right partner for you

sales@drugbank.com | **drugbank.com**

ABOUT

Hi, We're DrugBank!

DrugBank was founded at the University of Alberta and is the world's first intelligent and comprehensive drug knowledge platform.

With the help of artificial intelligence, our team of medical and scientific experts gather, author, verify, and organize all of the latest, most relevant biomedical information into one machine-learning ready knowledgebase. This platform is constantly updated to include the latest findings and is accessible through **Data Downloads**, our **Clinical API**, and **DrugBank Online**.

We're working to augment human intelligence so that the world's medical information can be used to its fullest potential and ensure that everyone has access to the best possible medical outcomes.

Our datasets and modules draw from a huge range of data including:

- **Adverse Effects**
- **Metabolism**
- **Indications**
- **Targets**
- **Protein Relationships**
- **Chemical Structure**
- **Drug Categories**
- **Pharmacology**

Contact us to learn more about our technology and how you can use it to support better research, healthcare, and patient outcomes.

info@drugbank.com **drugbank.com**

Authors

David Ackerman, *Senior Software Developer*
David has extensive experience with web application development, devops, and data science all of which he draws from to build connections and strengthen DrugBank's data.

Rodolfo Aleixo da Silva, *Data Review Specialist*
Rodolfo is a pharmacist that has more than 10 years of experience in regulatory affairs and good manufacturing practice in the pharmaceutical industry.

Lucy Chin, *Senior Biocurator*
Lucy holds a B.Sc in Pharmacology and is passionate about using high-quality data to improve scientific research and drug discovery.

Jordan Cox, *Senior Biocurator*
Jordan is a Doctor of Pharmacy who holds a B.Sc in Biology. He applies his expertise from working as a pharmacist in both community and clinical settings to his work at DrugBank everyday.

Mark Franklin, *Manager,* Curation
Mark holds an MPharm which he leverages, along with his years of experience, to guide our Curation team and make critical decisions about our data and knowledgbase.

Marysol Garcia-Patino, *Biocurator*
Marysol holds a PhD in Pharmaceutical Sciences, an MSc in Biotechnology, and draws from a wealth of experience in pharmacogenetics to inform her work.

Tim Jewison, *Team Lead*, Data Engineering
Tim holds a B.Sc in Immunology and Infection, and a B.Sc and M.Sc in Computing Science, all of which he applies to leading DrugBank's Data Engineering team.

Therese Karitanyi
Product Development Manager
Therese has a background in chemical engineering and clinical trials. She specializes in technology commercialization with a global focus.

Christen Klinger,
Scientific Support Lead, Bioinformatics
Chris has a B.Sc. double major in Biology and Chemistry, and a Ph.D. in Medicine with a focus on combining evolutionary biology, bioinformatics, and molecular cell biology. Chris focuses on emerging research to understand how informatics can assist the drug discovery process.

Craig Knox, *CTO*
Craig draws on his nearly two decades of experience in computing science, genetics, and metabolomics and bioinformatics research to guide DrugBank's rapid growth.

Authors

Sebastian Paz Vivas
Manager, Customer Success

Sebastian puts all his energy into making DrugBank accessible for our users and ensuring their experience is as seamless as possible.

Matthew Sharp, *Senior Data Review Specialist*

Matthew holds a B.Sc in Ecology and works diligently within our Curation team to collect and structure the most current biological data.

Teira Stauth, *Senior Copywriter*

Teira prides herself on creating accessible content that enables audiences to engage with even the most complex topics.

Rachel Wang, *Biocurator*

Rachel holds a B.Sc in pharmacology and is passionate about the intersection of science and art as a means of breaking down barriers of understanding.

Alex Wilson, *Team Lead,* Knowledge & Insights

Alex holds a B.Sc in Computing Science, with a special interest in machine learning and software architecture. Alex designed and implemented the first version of many key software systems and architectures at the heart of DrugBank's products.

Julie Xu, *Senior Product Manager,* Clinical API

Julie focuses her energy on revolutionizing healthcare using technology as she guides our Clinical API product team.

info@drugbank.com | drugbank.com