

SEMANTIC SEARCH 2.0

From big data to smart knowledge

The new era of life sciences data research



ONTOFORCE

 **MAIN OFFICE**

Moutstraat 108
9000 Ghent
Belgium

 ontoforce.com
 +32 9 292 80 37
 info@ontoforce.com

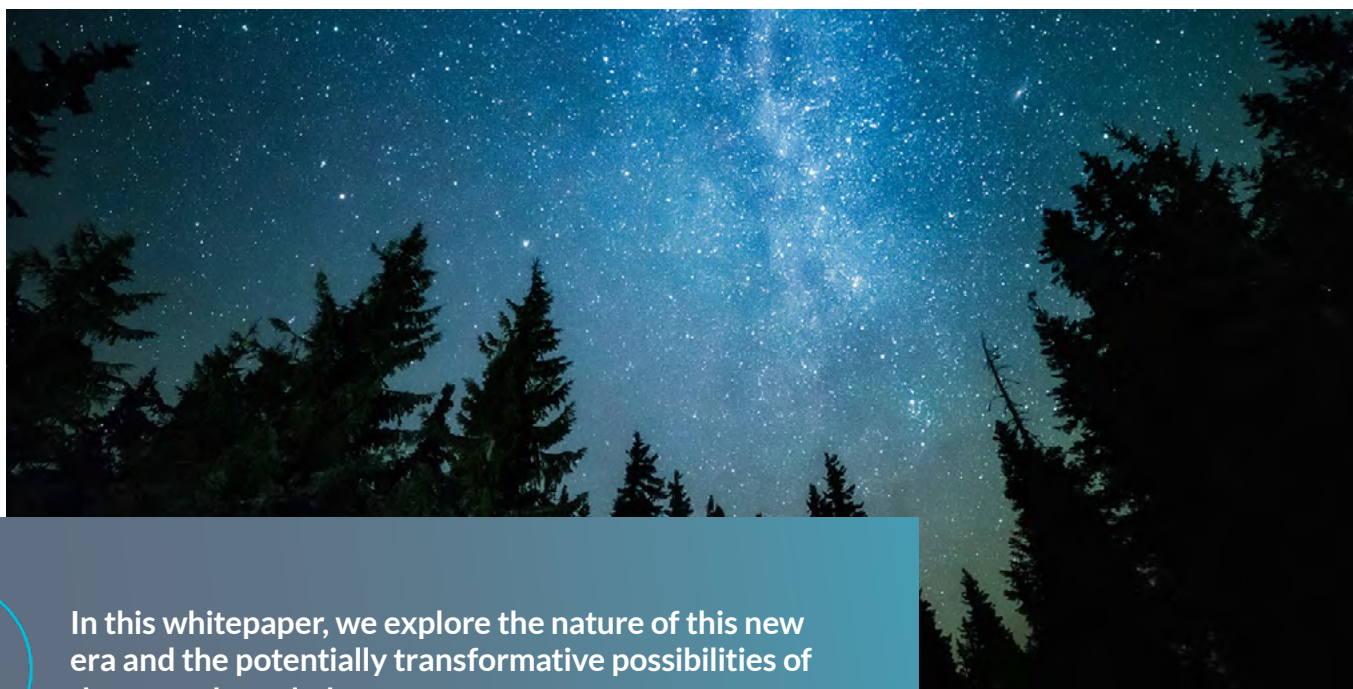
Summary

The term ‘big data’ seems old school now that ‘machine learning’, ‘deep learning’ and emerging concepts such as ‘edge AI’ are the hypes of the day. However, despite our general familiarity with the concept of big data, **challenges related to data-driven decision making** still remain.

Several key learnings have emerged over the last decade. The tension between the culture of generating massive volumes of data and the culture of applying that data to achieve meaningful outcomes is stronger than ever. How do we use our expensive data processing and analytics tools to **generate actionable insights**?

Giving more attention to the first principles of data management is essential to ensuring sustainable solutions for data processing and analytics. The first step is to lower the barrier to making data ‘findable’ before introducing artificial intelligence capabilities to data consumers. To integrate and link concepts, even within different data sources, it is critical to first properly distinguish and label them.

Semantic search is changing the game by making complex searches accessible to many kinds of data users. The combination of and interaction between technologies such as flexible context-sensitive search, ontologies, linked data and navigable visualization are fueling research and innovation – and the life sciences sector is the vanguard of this movement.



In this whitepaper, we explore the nature of this new era and the potentially transformative possibilities of the smart knowledge movement.

The big data hype

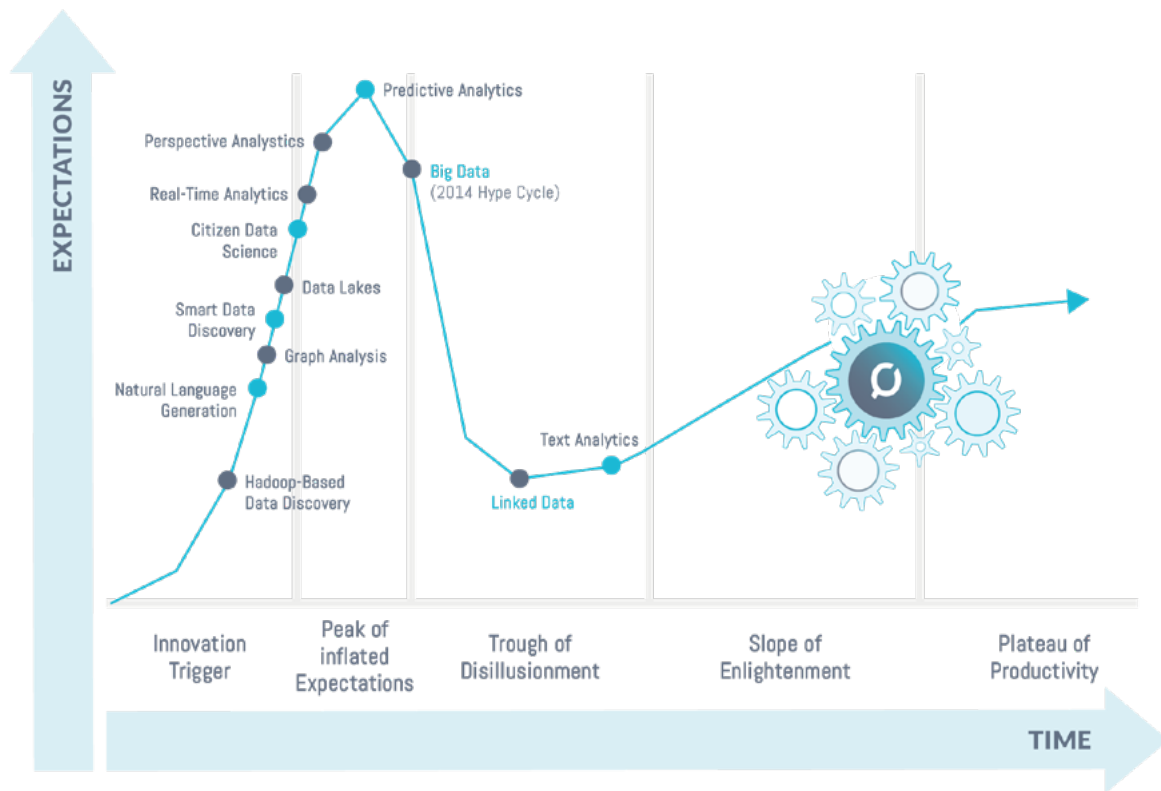


Figure 1: Reworked Gartner's 2015 Hype Cycle for Advanced Analytics and Data Science

A decade ago, 'big data' was the buzzword on everybody's lips. Expectations were soaring, and Gartner predicted that the hype would soon plunge into what it labels the 'trough of disillusionment' – when formerly high-spirited hopefuls feel the pang of disappointment when those expectations fail to be met.

But around five years ago, 'linked data', a component of big data, had already begun rising from the trough of disillusionment and moving toward the 'slope of enlightenment'.

'slope of enlightenment' – when hypes transform into real solutions that begin to generate value.

Source: <https://www.gartner.com/en/research/methodologies/gartner-hype-cycle>

Does this mean that the majority of digital-age organizations are reaping the benefits today?

Most people are unaware that the term 'big data' was coined nearly 30 years ago. The big data hype can be considered the externalization of technological advancement that really started to accelerate in the 90s. Today, available technology allows organizations to easily generate and store data on a scale orders of magnitude beyond the possibilities of the early 00s. Why, then, is going a step further – generating insights from these data oceans – proving to be so hard?

The answer is relatively simple, but the solution itself is not. However, by stepping back and strategically **applying common sense and basic data management principles**, we can bring solutions within reach, with some effort.

The confused psychology of data management

The biggest challenge isn't the technology but the psychology of data management, and the ease with which data is generated, stored and logged has exacerbated this issue. We exist in an era characterized by the consumption of **massive amounts of easily digested, Time-dependent bits of information**, which seems to clash with our underlying quest for deep knowledge.

Life sciences leans on the shoulders of – and combines – various fields of basic science. An enormous body of knowledge has already been generated, and to uncover new insights, one must obtain or gain access to a critical mass of background information first. This implies looking at a research question or problem from multiple angles, and fundamentally requires the **holistic, integrated use of available data**.

However, as users, we are often unable to find the data we seek. The stream of information is just too vast to

plow through manually. In the past, it was acceptable for a user to prepare their experiment to test a hypothesis, generate the relevant data and process it, isolating the data and communicating these rather limited results.

Today, we must integrate our data to reveal deeper, broader, more impactful links, insights and nuggets of truth. Our interactions are digitalized – rapid and easy – but this fact seems to backfire when it comes to managing information. **Making data findable** is the first step in the right direction towards holistic insights.

To make data optimally available to all stakeholders, users need to adhere to the **FAIR-principles** for data management.

FAIR-principles

IN OTHER WORDS THE DATA IS:

FINDABLE,
ACCESSIBLE,
INTEROPERABLE, AND
REUSABLE.

This means, a.o. that each data set is registered, has unique and persistent identifier, is enriched with knowledge from ontologies and can be retrieved easily.

The future of data consumption

For decades, new technologies have touted the promise of tackling challenges related to the volume, variety, velocity and veracity of big data. However, **big data has remained more of an experimental topic** – rather than a practical one – for most organizations. Despite our oceans of data, linking data silos together in meaningful ways while making holistic sense out of it generally remains out of reach.

Next-generation semantic search intuitively simplifies the process of linking and searching through massive volumes of data spread across multiple datasources. It's not exactly a standard concept yet, but it's rapidly gaining traction.

Shining a light on semantic search

Semantic search is a core technology in resolving the stubborn challenges of big data. Current search possibilities are extended by considering the **intent of the searcher** as well as the **context of the search**. Combined with identifying, distinguishing and labeling concepts, including context is a game changer here. When properly integrated, more refined and better-attuned search results are generated.

If applied broadly and generally (such as to the internet), semantic search fails to generate meaningful results because of the extremely wide and complex range of possible contexts. But when used in, for example, a life sciences context, the capture of intent and context using semantic approaches **has demonstrated impactful, ground-breaking results**.



Search engines in a perfect world

The ideal search engine is capable of matching search queries to the precise context, and returns results relevant only to that context. Although traditional search engines (such as Google) are still the most frequently used, a semantic-based approach that considers the meaning of the query generates far better results.

With these approaches, the searcher no longer relies on preset keyword groupings or inbound link measurement algorithms. Instead, by adding context and meaning, a semantic query leads to far more pertinent, refined, detailed results.

The components of a semantic search

A useful semantic search platform combines the **effective use of ontologies, linked data and graph-style visualizations** to illuminate the meaning found in data from many disparate sources in a highly effective and accurate way.

Ontologies

MAKING DATA READABLE AND UNAMBIGUOUS

Essential components of semantic search are ontologies, **formal naming conventions and meta descriptions** that are used to identify different entities and properties. An ontology is used to categorize and organize information while simultaneously reducing complexity. In the process of creating and applying an ontology, data is standardized and different descriptions for the same entity receive a uniform resource identifier (URI), making data more readable for machines and unambiguous for humans.

Creating ontologies requires multiple iterations, for example, to set up basic naming conventions or to identify ambiguities across different descriptions. In short, an ontology is **a means of standardizing and cleaning up data**. Today, an increasing number of automated or semiautomated software programs can run this curation process.

Compared to manual ontology creation, automated ontology creation significantly extends the possible impact of standardizing data and disambiguating terms.



BRINGING ORDER TO THE “CHAOS” OF NATURAL LANGUAGE

Natural language – or sentences, paragraphs and concepts expressed using words – tends to be ambiguous and unstandardized.

For example, one source of clinical study data may use the terms ‘non-recruiting’ and ‘recruiting’ to categorize the status of the clinical study, while another source may use more nuances, such as ‘active, non-recruiting’,

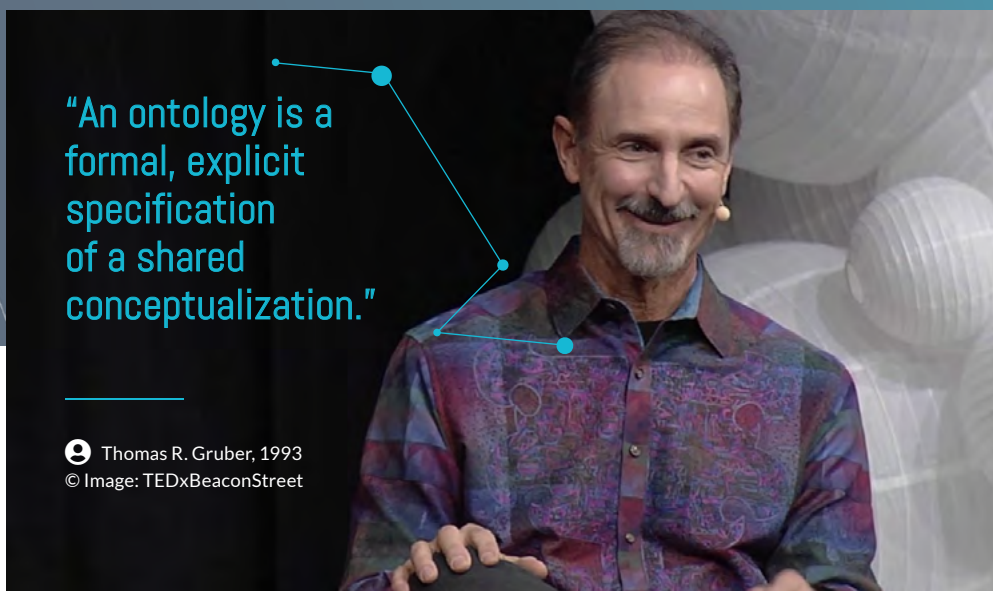
‘terminated’, ‘not yet recruiting’, etc. By applying ontologies, categorizations with the exact same meaning receive an unambiguous uniform resource identifier (URI).

DEFINING ONTOLOGY IN INFORMATION SCIENCES

In computer and information science, an ontology is a specification of a representational vocabulary of and relations between concepts for a shared domain of discourse. In plain terms: a standard set of terms used within the same context to refer to and link between concrete bits of information in the same way.

It includes the formal naming and definition of the types, properties and interrelationships of fundamental entities that exist within a specific domain of discourse.

As such, an ontology provides a shared vocabulary that can be used to model a domain – including the types of objects and concepts that exist as well as their properties and relationships with each other.



Sources: Thomas R. Gruber, “A Translation Approach to Portable Ontology Specifications”, Stanford University, 1993 | Fredrik Arvidsson (LIBLAB/HCS/IDA) & Annika Flycht-Eriksson (NLPLAB/HCS/IDA), “Ontologies I”

Linked Data

A STEP ABOVE ONTOLOGIES.

The advantage of using ontologies is that **different ontologies can be mapped onto one another**, enabling us to integrate multiple, disparate fields of data or data sources within the same context.

Data entities residing in different and hard-to-connect data sources can be linked quite easily. As a result, different angles of the same real-world concepts can be mapped onto others – **unlocking unexpected insights**.

With linked data, searching among more diverse datasets offers more relevant query results while also generating more original outcomes, and a continuously **growing number of datasets is rendered interoperable**. Data silos are glued together and made available for everyone to access and consume.



BUILDING BRIDGES BETWEEN RICH RESOURCES

Imagine that the ontology relating to clinical studies mentioned above concerns disease research.

These diseases, although they may be named differently across databases, have received the same URI via the ontologies established earlier. Thanks to this, the semantic system knows that the information about Disease A related to a clinical trial in the clinical trial database concerns the same Disease A found in different disease databases. As a result, the disease data can be linked together. In the process, a whole volume of additional information is automatically included about Disease A, enriching the meaning of the results.

Entities are linked to other entities automatically, regardless of the database in which they reside. By gluing data together, we are able to uncover information found in other databases that we might not have **known existed**. The challenge of making all of this data searchable is solved by navigable visualizations, which make complex queries very simple to understand (see further).

Groundbreaking research requires access to cutting-edge scientific resources. However, those resources are often locked away in the laboratories or university departments where they were developed.

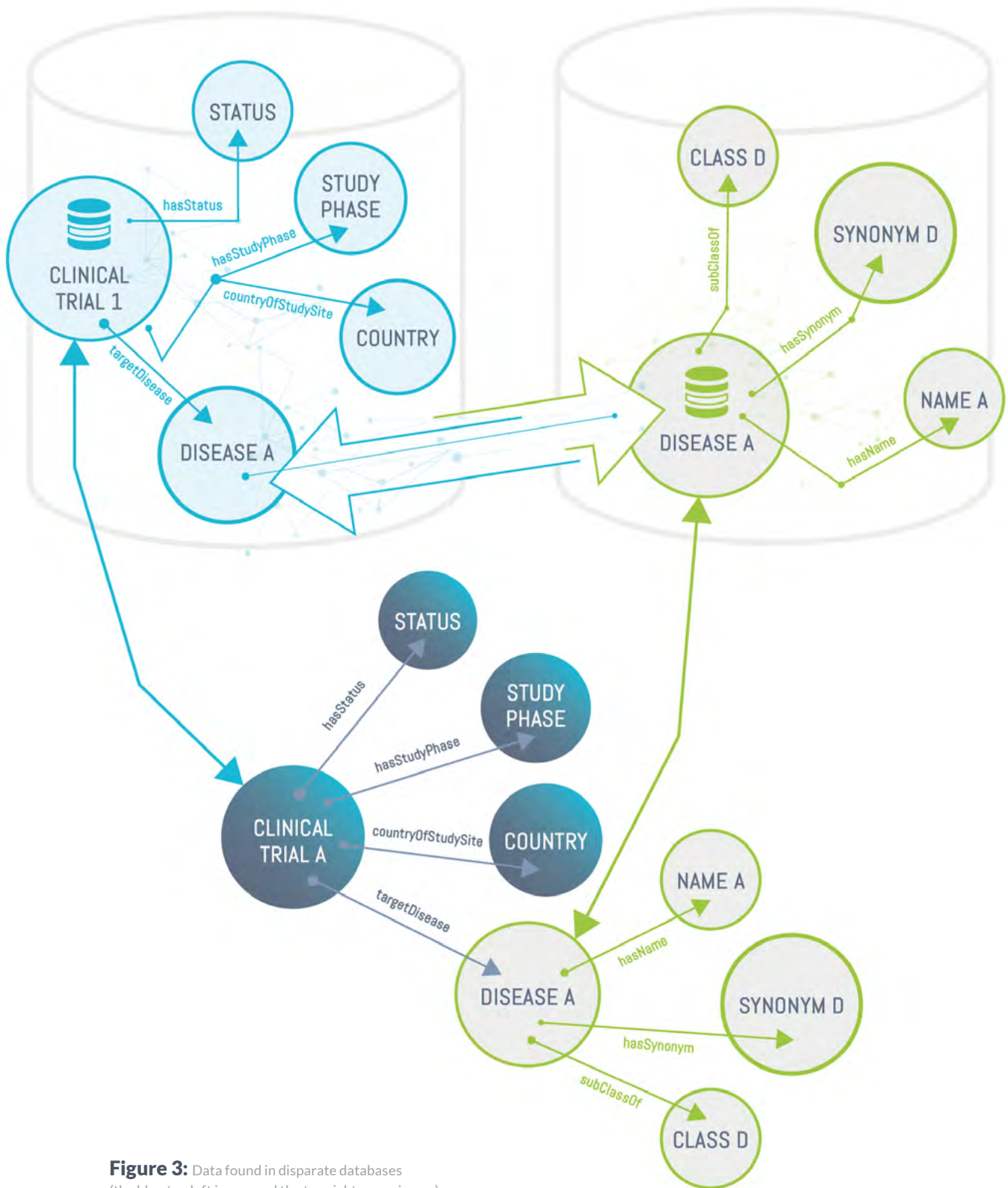
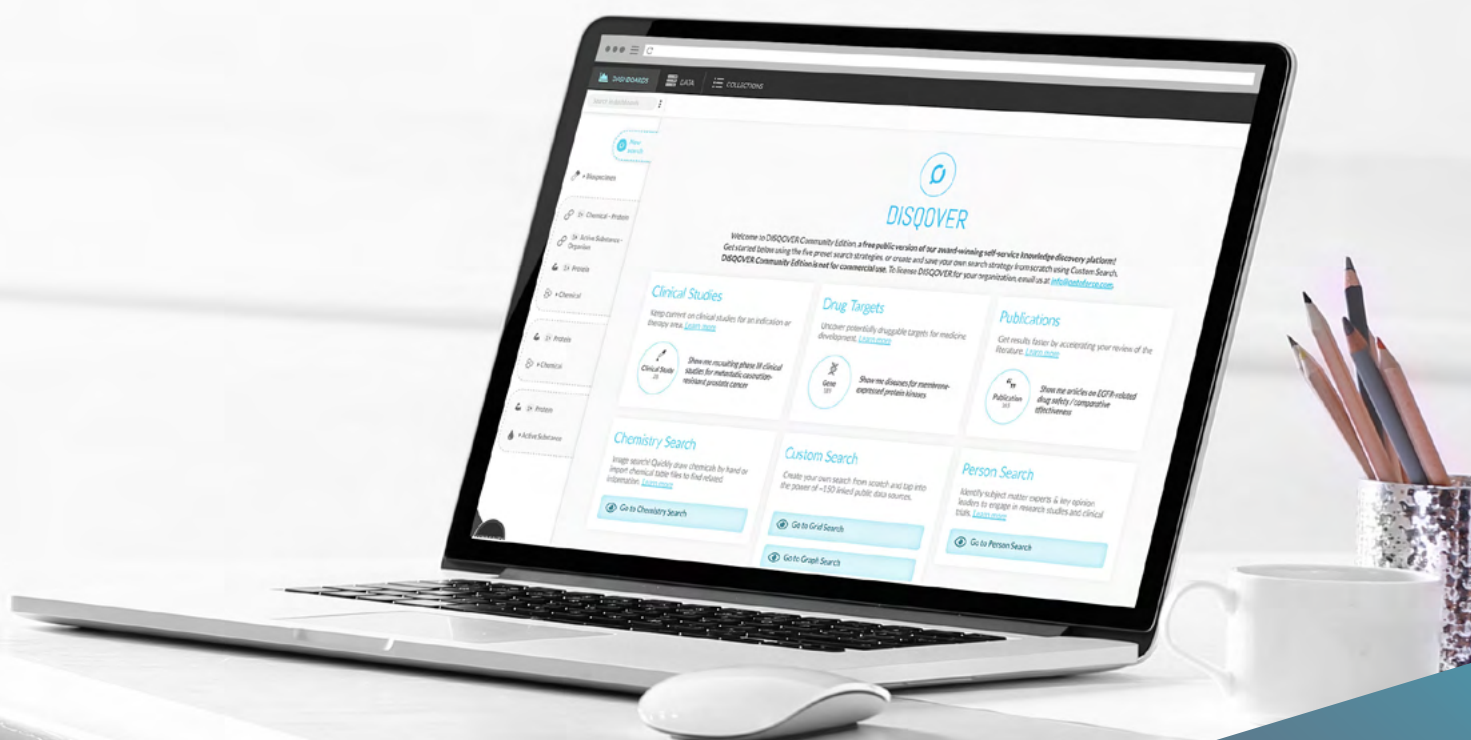


Figure 3: Data found in disparate databases (the blue top left image and the top right green image) is linked together, creating a far richer dataset.

The power of navigable visualization

With so much data to sift through, proper visualization becomes indispensable. Semantic searches stretch across numerous databases, and the returned results can be displayed in myriad ways – including visual representations such as infographics, plots, tables, charts, graphs, maps and more.

Effective visualization makes searching through complex data a lot more efficient. By simply changing the view, new insights emerge which would otherwise stay hidden in inscrutable tables. Through intuitive visualization, data becomes **more comprehensible, its characteristics are highlighted and relationships and connections between data points stand out** more clearly. In this way, exploratory searching and navigation through vast amounts of linked data becomes possible.





INTUITIVE SEARCH AND DATA EXPLORATION

With semantic search engines such as DISCOVER, the act of searching becomes a visual experience. Although a search begins with a (combination of) keyword(s), each results page offers a targeted variety of filters and related keywords that the searcher can use to refine and dig deeper.

Each of these elements can be clicked on. Even within complex graphical representations, every single element becomes a clickable filter. For instance, searching for clinical trials on a specific condition or drug can be visualized as a bar chart that includes the number of existing studies by clinical trial phase.

Clicking on one of the bars refines and reshuffles the results page and offers the user a further set of associated visualized filters to choose from. As a result, navigable visualization transforms searching into a user-friendly experience that can be performed by just about anybody.

Visualized searching makes browsing and navigating extremely large, heterogeneous data sources a simple task. Users with no data science or statistics backgrounds **can easily perform complex queries** without the assistance of a data scientist – a huge breakthrough for research-intensive sectors such as life sciences.



Figure 4: searching for clinical trial phases within semantic search platform DISCOVER.

Making the case for semantic search

Consider a researcher working for a pharma or biotech firm who needs to find out whether a certain gene, biological target or active compound has already been studied in previous projects carried out by the company.

The goal of the researcher's quest is to discover if the company **has already generated knowledge** about the topic and has **invested in certain resources** such as reagents, cell lines or antibodies that could be immediately useful in initiating new experimental work.



FINDING INSIDE INFO LIKE THIS COULD GIVE A HEAD START TO THE RESEARCHER AND THEIR COMPANY:

1. Identifying the results of previous experiments can help the researcher prioritize what is most interesting to test. The researcher can avoid repeating experiments that were already proven unsuccessful.
2. If certain resources are already available inside the company, the researcher can assess how long it would take to obtain them depending on where they are stored, and which application will be used.

WHEN SILOS ABOUND IN LIFE SCIENCES

In many companies, inside information like this still resides in different databases, with **each system having its own interface and search logic**. Different systems may also apply diverse nomenclature for the same concepts. These are data silos.

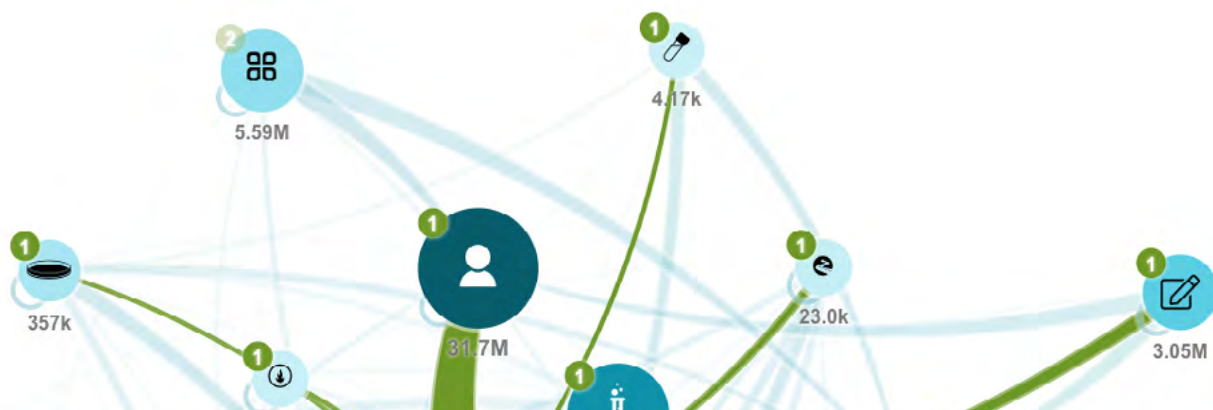
When these data silos exist, researchers must access multiple databases, be familiar with the specificities of each system to generate useful results, and then combine these results into a single document – often an Excel file. Depending on the complexity of the query and the fragmentation of the data, such a query can be a huge time sink, taking hours or even days.

THE ADDED VALUE OF SEMANTIC ORGANIZATION

Using a semantically organized and well-curated data source combined with a semantic search platform such as DISCOVER, the researcher only needs to access a single platform with **an intuitive and consistent user experience**. Navigable visualization transforms the researcher's query into an intuitive experience that is simple to follow, understand and modify on the fly.

Disparate internal data sources have already been connected to the semantic search platform and all data has been categorized through the application of ontologies. Using linked data technology, **hidden connections between various data points are brought to light**, revealing previously invisible insights.

With semantically linked data and effective semantic search, a complex query spanning different databases the process of compiling findings into a cohesive document **takes a matter of minutes**.



Enter the world of DISCOVER

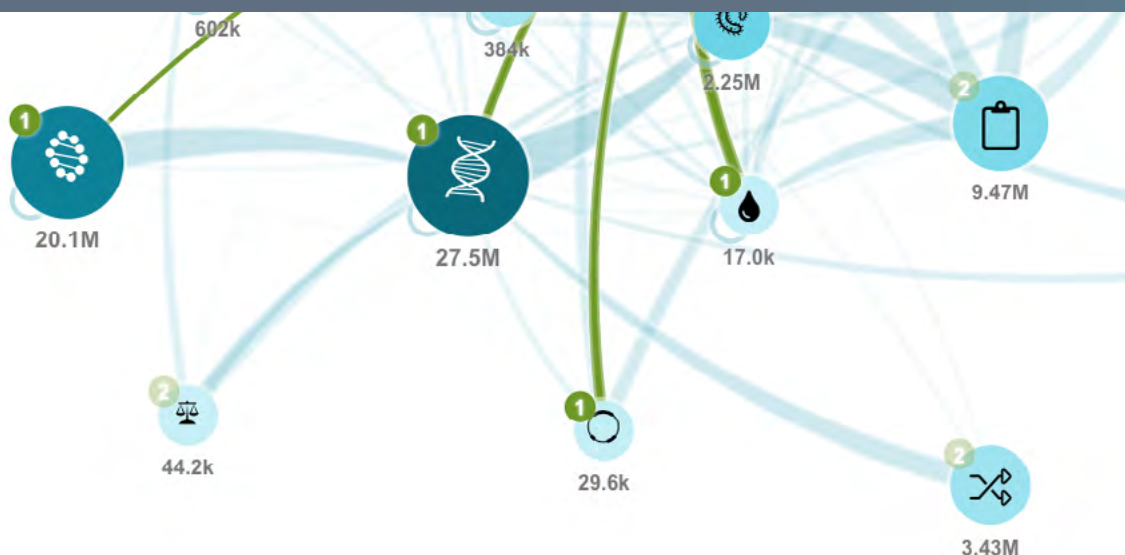
The DISCOVER self-service knowledge discovery platform gathers, transforms and orchestrates information from internal, third-party and public sources, transforming it into actionable insights and unlocking self-service knowledge discovery. DISCOVER delivers actionable insights by creating data literacy and search accuracy, making the platform an integral part of your enterprise ecosystem.

Massive volumes of information spread across multiple sources. New analytics tools popping up every day. The daunting tasks of data harmonization and integration. From drug discovery and clinical research to literature analysis and chemistry, every

aspect of life sciences is fraught with data-related challenges.

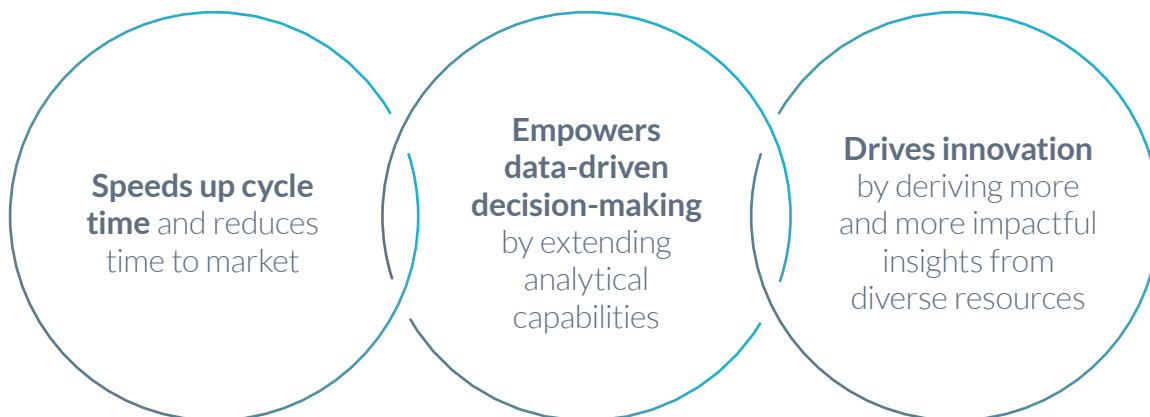
Powered by semantic search and an intuitive user interface, DISCOVER lets you explore and connect data from disparate sources to uncover new insights in a matter of minutes. This enables you to:

- master the flood of information in healthcare and life sciences
- speed up the research and go-to-market of new treatments and products
- homogenize, link and reveal hidden correlations.



Entering the new era of semantics

Semantic search is extremely valuable for research-intensive organizations, as it:




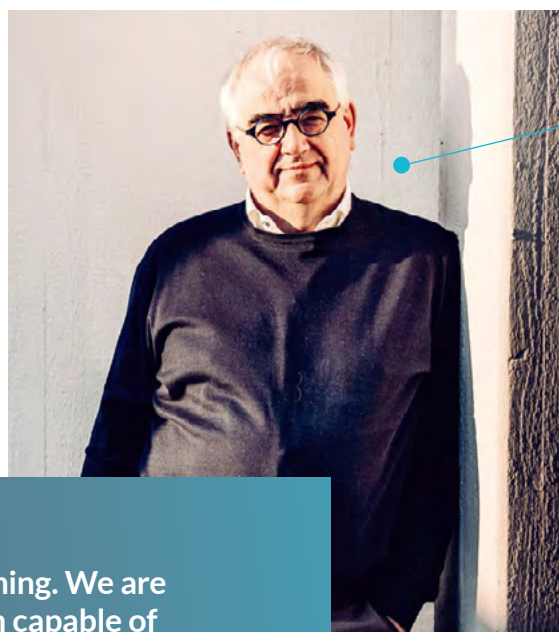
DATA IS NOT THE NEW OIL – IT’S EVEN BETTER

Unlike natural oil reserves, data is never depleted. It’s more like reusable wind or solar energy: there’s plenty of it, we’ll never exhaust it, and through technological optimization, we can create more of it and deploy it ever more effectively. **Making data sources interoperable is of critical importance.**

As more life sciences companies, institutes and organizations adopt linked data technologies, the research process is optimized, **ultimately leading to more rapid advances in healthcare.** This is clearly where a semantic network like the [EBI’s Samples Phenotypes and Ontologies Team](#) can operate as a catalyst together with industry-driven initiatives such as DISCOVER. This applies not only to life sciences, but to every data-intensive industry.

“Patients out there are waiting, let’s do it well and fast.”

 Paul Stoffels
Vice Chairman of the Executive Committee and Chief Scientific Officer at Johnson & Johnson
© Image: Newsweek België



This is the start of a new era of searching. We are on the brink of a new query paradigm capable of fundamentally transforming the way we do research.

About ONTOFORCE

ONTOFORCE helps people transform data into knowledge, gather insights across all data sources and make an impact on your business.

Our ultimate goal is to enable everyone to create value out of data in an empowered and flexible manner.

This will result in direct user benefits, including less repetitive work, less time spent on searching, and faster insights. Achieving this will result in significantly accelerated time to value, increased productivity and higher probability of success for the organization!

 www.ontoforce.com

 [@ontoforce](https://twitter.com/ontoforce)

 <https://www.linkedin.com/company/ontoforce/>

Sources used

On Gartner's 2014 Hype Cycle:

<http://www.gartner.com/newsroom/id/2819918>

On Gartner's 2015 Hype Cycle for Advanced Analytics and Data Science:

<https://www.gartner.com/doc/3087721/hype-cycle-advanced-analytics-data>

<http://www.kdnuggets.com/2015/08/gartner-2015-hype-cycle-big-data-is-out-machine-learning-is-in.html>

On Big Data:

<http://www.ibmbigdatahub.com/infographic/four-vs-big-data>

<http://www.gartner.com/newsroom/id/2819918>

Quote on Big Data:

http://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Big_Data.pdf

On ontologies:

<https://en.wikipedia.org/wiki/Ontology> (information_science)

<http://tomgruber.org/writing/ontolingua-kaj-1993.pdf>

<http://www.ida.liu.se/~janma56/SemWeb/Slides/ontologies1.pdf>

MAIN OFFICE

Moutstraat 108
9000 Ghent
Belgium

 ontoforce.com
 +32 9 292 80 37
 info@ontoforce.com