CAS CUSTOM SERVICES℠
CASE STUDY

UNRAVEL THE POTENTIAL OF INTERNAL R&D DATA: CURATE AND CONNECT FOR SEARCHABLE INSIGHTS

CAS
A division of the
American Chemical Society

Idea in brief

The knowledge you own has value, but what if you can't access it?

A large health-tech organization had decades of key research trapped in unstructured documents. As a result, relevant individuals struggled to locate, access, and apply this valuable information to their work. CAS helped this organization overcome data management challenges by leveraging the same expertise and technologies that make the world's scientific data available in CAS SciFinder^{n}. As a result, internal experimental data is accessible, connected to the world's science, and enables better insights for data-driven decisions.

The challenge

When valuable data is inaccessible for research, discovery, and decision-making

Over decades, R&D teams amass large amounts of experimental results and findings in a variety of formats. This momentous amount of data that remains unknown, unused, and untapped across an organization can be known as 'dark data'.[1] It is estimated that 55% of an organization's data is 'dark'.[2] Generated by systems, devices, and interactions, the knowledge cannot be easily accessed due to its unstructured nature.

90%

of industry leaders agree that every organization will need to extract value from dark data to be successful in the future.[2]

Splunk, The State of Data White Paper, 2019

80%

of the knowledge inside dark data is missed due to inefficient analysis.[3,4]

IBM, 2018

R&D data may sit in silos in several disconnected systems, making it difficult for individuals to access and utilize the information they need. This can lead to:

- Loss of legacy data gathered by individuals who have since left the organization.
- Limited ability to collaborate with individuals and teams across the organization.
- Inability to make connections between disparate datasets due to a lack of harmonization in scientific terminology.

When a knowledge management system is not fit for its purpose, it hides your own data from you and prevents you from making connections to external sources, leaving the potential for key insights to be missed.

While out-of-the box solutions promise fast results, they are rarely able to accommodate the unique way organizations capture and store scientific data. The capabilities of any data platform are limited in their ability to deliver meaningful results without properly cleaning, normalizing, connecting, and enhancing the data first. Without specialized scientific and technical expertise, automated solutions will only go so far.

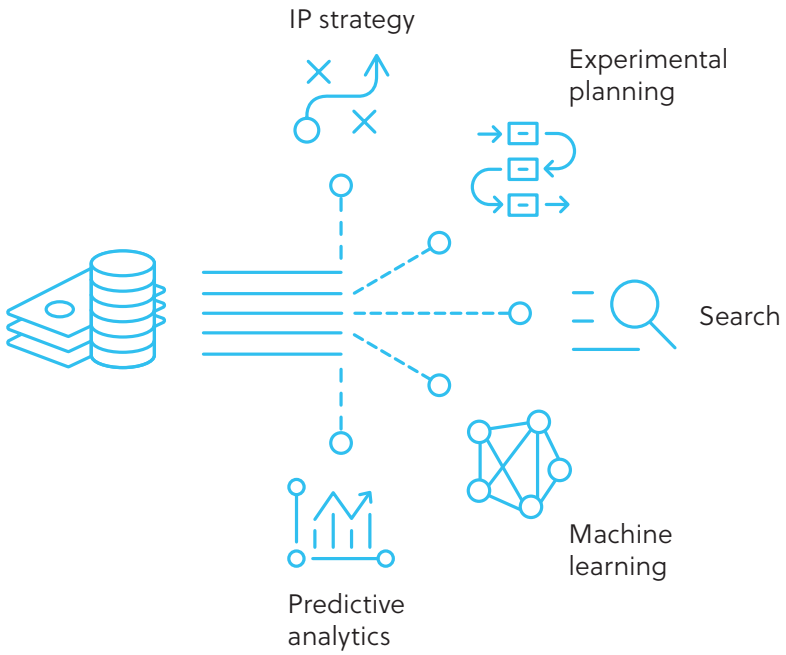


Figure 1. Unstructured, inaccessible data cannot be leveraged to drive future discoveries

Scientific complexity and connections: An ongoing challenge for in-house IT

The IT knowledge within R&D organizations is typically insufficient to deal with the complexities of scientific information; this is due to nuances with critical representations, nomenclature, and relationships. For instance, if the name and taxonomy of a drug has changed over time, a traditional database may not be able to make connections between varying information sources that have used inconsistent naming conventions. This may result in key information being omitted from searches. To access the information central to informing future discovery, technologies must be scientifically aware.

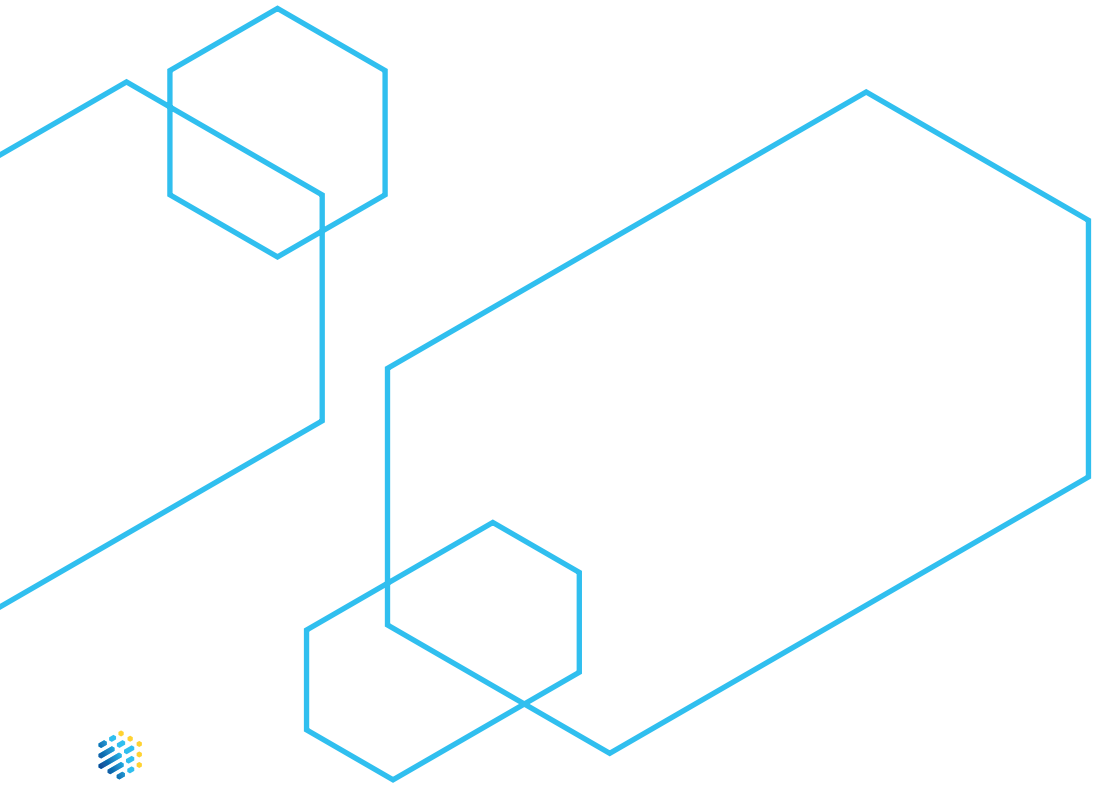
One organization's struggle

A large health-tech organization had decades of key research findings housed in several disconnected, legacy systems. There was no systematic way to search the documents; scientists had to know what they were looking for and manually sift through the files until they found it. The organization sought to better utilize decades of stored research through improved searching and retrieval of data, helping to drive more efficient innovation.

Stored proprietary knowledge was not being utilized often as veteran scientists retired and newly hired researchers were unaware of the available insights. Without access to these findings, many experiments were being unnecessarily repeated across R&D groups which cost considerable time and resources. Without the ability to access and utilize the knowledge gained by other researchers across the organization, innovation was hindered, costs increased, and schedules were pushed out.

The lack of organization, limited search capabilities, and an insufficient data model meant that valuable information remained hidden in their systems. This rendered it useless for future research, discovery, and decision-making.

> "Having clear visibility to what data is already available is critically important. You need to be able to harvest the available information to reduce workload and speed up processes."
>
> Director, R&D



The solution

Enabling access to your data through a custom solution

To enable the proprietary data to be readily leveraged, CAS built the client a custom knowledge management system to curate and connect their internal R&D data. This enabled the organization to search the full text of their documents, connect similar concepts and substances, and search by concepts tailored to their R&D efforts, such as indicating if a substance is a natural product for the health-tech organization.

This solution was made possible by utilizing the same process CAS uses to curate the world's published science that powers leading scientific information solution, CAS SciFinder^{n}.

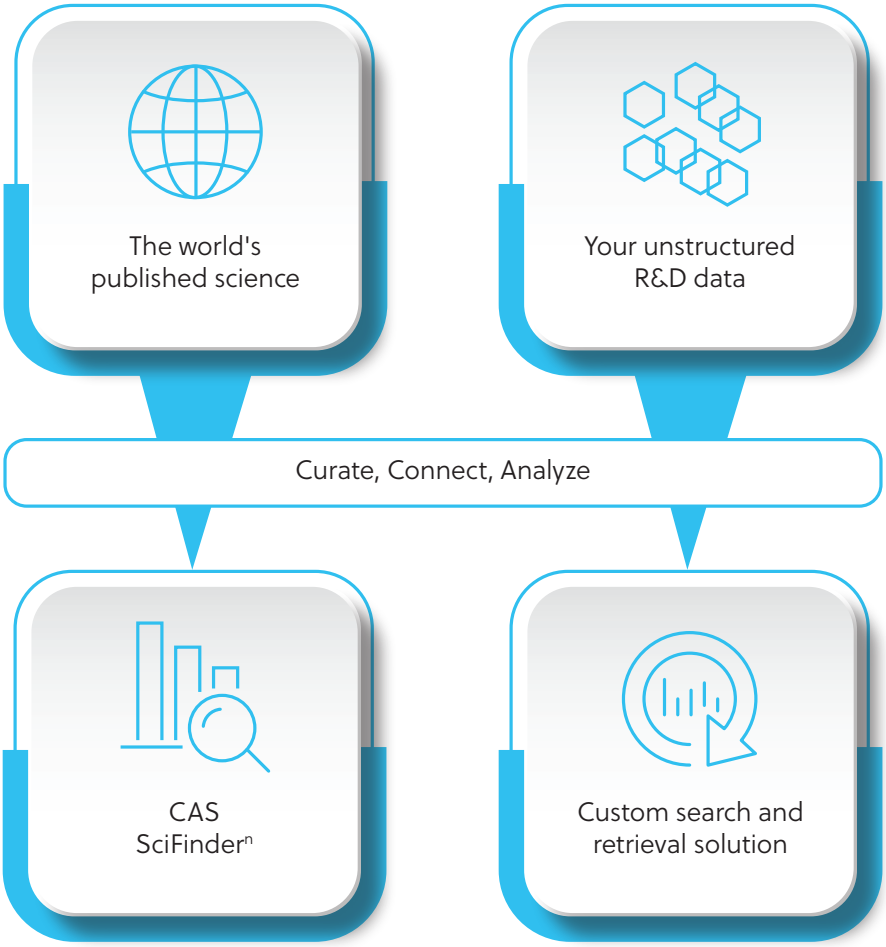


Figure 2. CAS reveals the full potential of your digital assets with similar processes that it uses to curate the world's published science

The CAS curation process utilizes the expertise of scientists alongside specialized technologies and approaches, such as:

- Tailored data governance and data management frameworks to ensure data is consistent and trustworthy
- Specialized lexicon, ontologies, and taxonomies to alleviate gaps from information nuances
- Machine learning algorithms, trained with CAS comprehensive data, that provide an accuracy edge to extract and connect data
- Manual curation for delivery of precise, high-quality data
- Optical character recognition (OCR) to digitize content
- Named-entity recognition (NER) & natural language processing (NLP) to identify substances and concepts

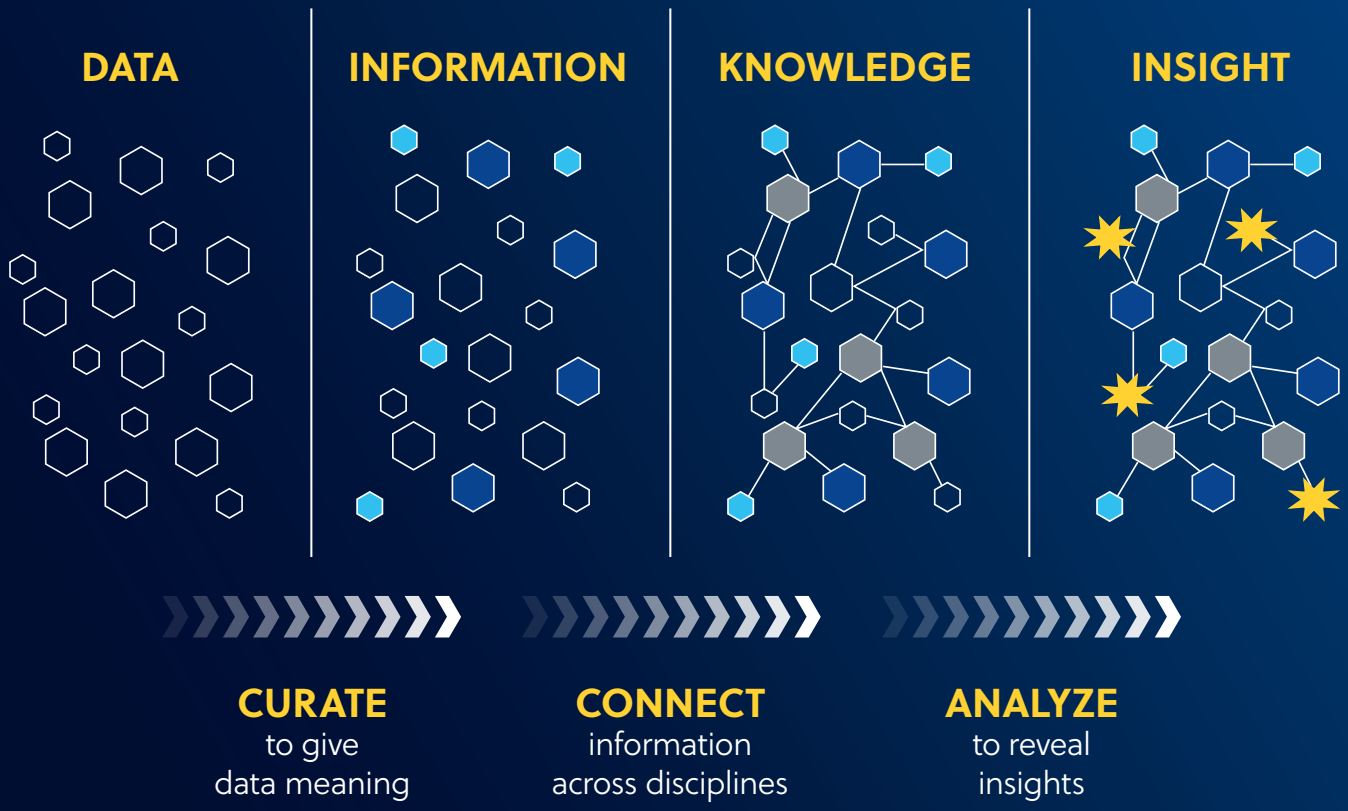


Figure 3. Data is valuable only when it is transformed into insight. Unlock the potential of your R&D data through capturing insights to help make data-driven decisions

A scientifically aware search for meaningful results

To access the R&D information needed to inform decision-making, knowledge management solutions from CAS move beyond standard keyword search and enable scientific context. Information extracted from documents is associated with key concepts in chemistry, material sciences, life sciences, and custom criteria based on a client's innovation needs. A researcher can immediately find where a concept, such as bioactivity, polymers, or specific diseases, is referenced in the document.

To overcome the challenge of variations in synonyms amongst substances, scientifically aware searching enables users to take a trivial name and associate it with a structure not mentioned in the document. Refined lexicon, ontologies, and taxonomies allow for matching on synonyms (trade names, systematic names, internal names) or even matching on classes of substances. This sophisticated feature is far beyond the capabilities of any out-of-the-box solution.

To yield high-quality, impactful results, CAS utilizes sci-tech expertise paired with specialized tools and algorithms. Where human intervention is required, CAS can systematically route documents to trained domain-specific curation experts for conceptualization accuracy that only a scientist can provide.

Our proprietary substance comparison technologies use CAS Registry Numbers®, the worldwide standard for identification of substances, to disambiguate substance data. CAS Registry Numbers also connect internal data to the world's published science; this includes alerts of publications related to a user's internal research pushed to them in near real time.

The CAS Content Collection

Your internal R&D data

Associated data

- CAS Registry Number
- Molecular properties
- Chemical structure
- Chemical names
- Regulatory information
- Reaction schemes
- Experimental procedures
- Additional data fields...

Figure 4. Scientifically aware search is enabled by connecting your organization's internal data with the world's published science.

Linking insights for internal collaboration and connections to the world's science

Knowledge graphs were constructed based on the available data enabling the organization to link information on substances, scientific information, and concepts, highlighting connections between the data across the organization. By utilizing knowledge graphs, researchers can find others in the organization working on similar areas of science, creating collaboration opportunities.

Data in internal documents is further enhanced when connected with the world's published science. Our unique data model securely links your propriety documents to the CAS Content Collection™ or custom-curated data relevant to your industry, enabling you to identify and fill gaps in your data. Substance information becomes more robust when connected to data such as the CAS Registry Number, property data, and related research.

Referenced areas of science can be analyzed to determine trends in patents or publications, enabling you to understand who is publishing in your field and identify collaborators or key competitors.

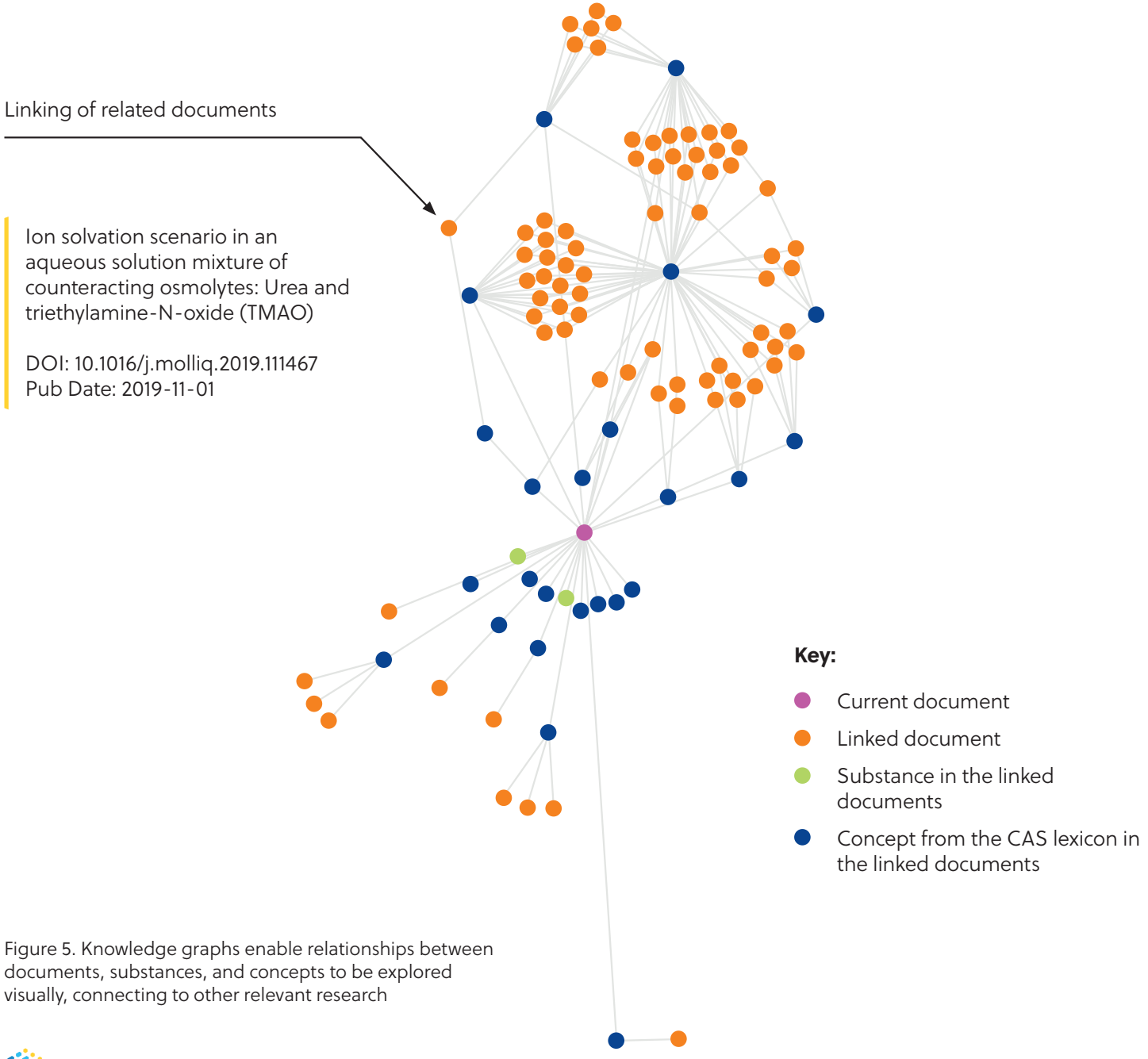


Figure 5. Knowledge graphs enable relationships between documents, substances, and concepts to be explored visually, connecting to other relevant research

Examples of a customized solution for searchable insights

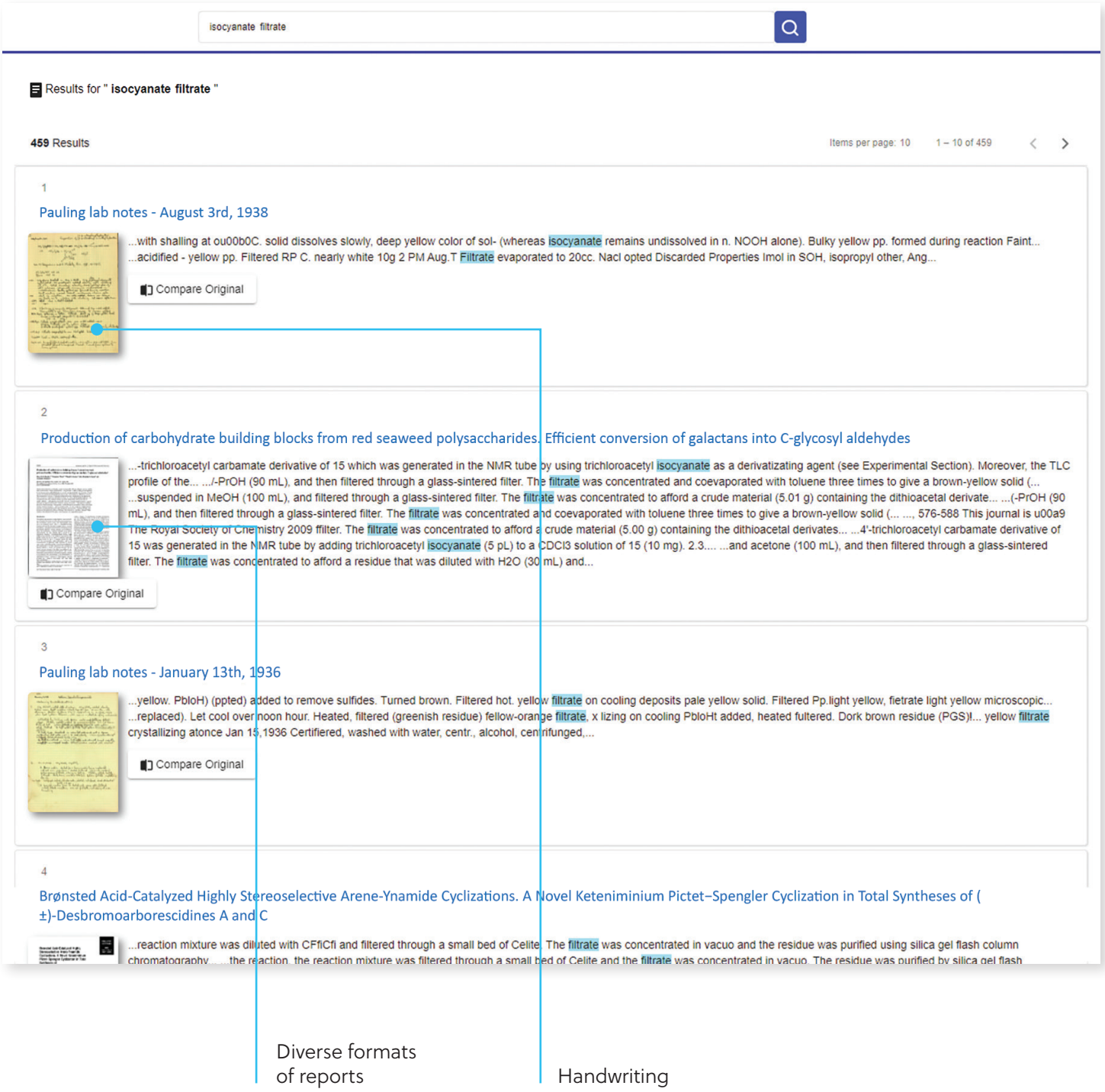


Figure 6. Search the full text of internal research documents that have been digitized and curated, including hand-written notes, tables, and a variety of data formats

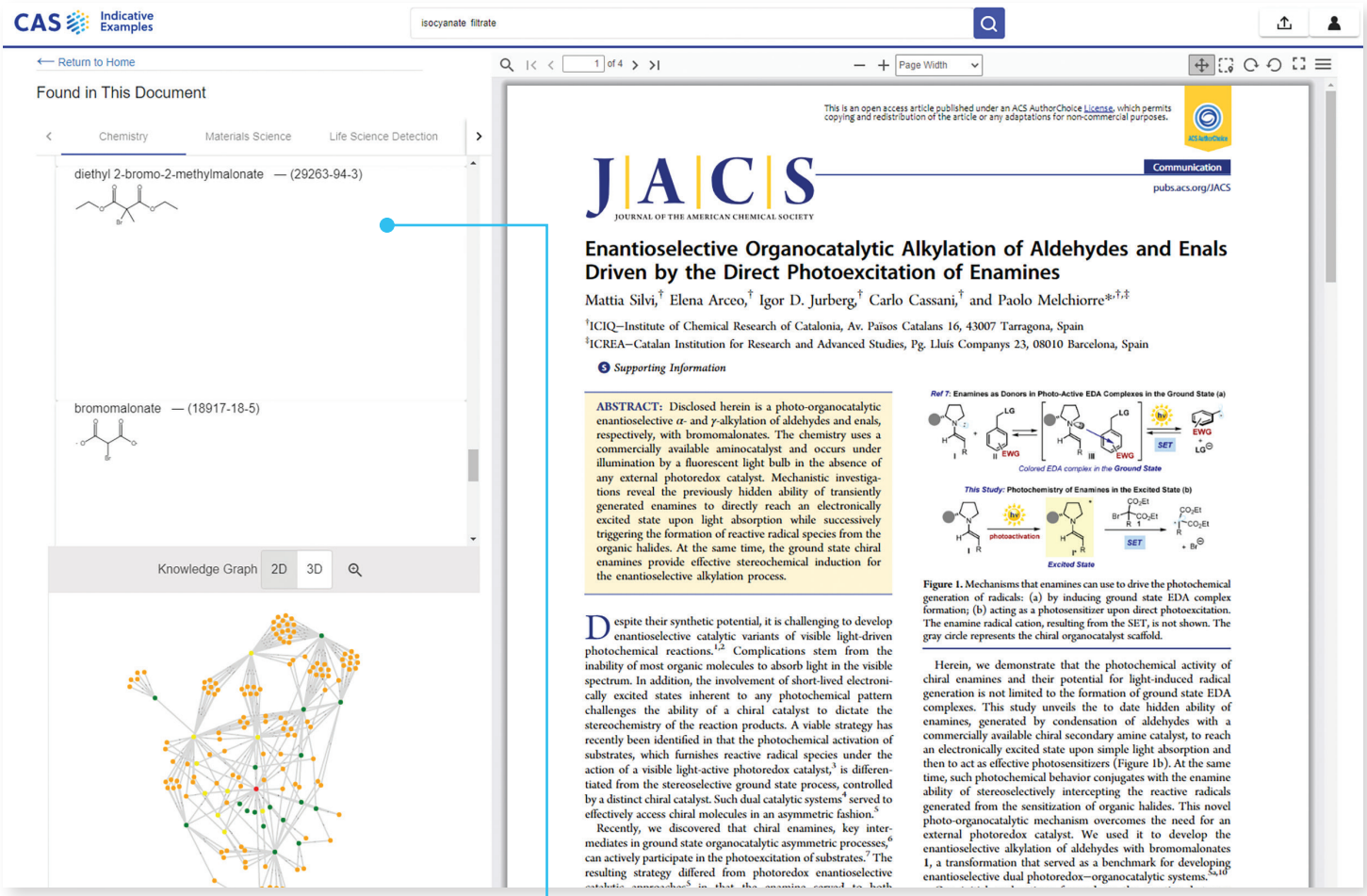


Figure 7. Substances, terms, and concepts are algorithmically extracted and embellished, linking key scientific data

References

1. *What is Dark Data?* https://www.splunk.com/en_us/data-insider/what-is-dark-data.html (accessed 2022-09-30). **2.** *Splunk>. The State of Dark Data.* https://www.splunk.com/pdfs/dark-data/the-state-of-dark-data-report.pdf (accessed 2022-09-30). **3.** *Dark Data Analytics.* https://articles.xebia.com/dark-data-analytics (accessed 2022-10-05). **4.** *Illuminate Dark Data – Part 1.* https://medium.com/ibm-data-ai/illuminate-dark-data-part-1-c3c2028f91af (accessed 2022-10-05).

The impact

Time and cost efficiencies

With proper strategy and structure, research findings previously hidden in legacy technologies are now readily accessible to the R&D team. Before implementing the customized search and retrieval tool, one search for a specific data point within the unstructured files took a senior scientist approximately eight hours; now it takes minutes. This not only frees up resources for more specialized tasks, but it also creates cost savings by preventing unnecessary experimentation and repetition of work.

A competitive advantage

Data drives decision-making. With access to the full breadth of the organization's knowledge, teams can now leverage their valuable R&D data to accelerate the innovation of new health-tech solutions. Our customized solutions enable organizations to quickly respond to challenges that arise, from supply chain problems to the duration of reformulation. As a result, time to market can be reduced, ensuring your organization remains at the forefront of cutting-edge R&D.

Looking to unlock the potential of your data?

CAS Custom Services designs solutions so you can take advantage of the information you own to drive discovery. Reach out today at **cas.org** for an initial consultation.

> One search took a senior scientist ~8 hours; now it takes minutes with scientifically aware search from CAS.

CAS is a leader in scientific information solutions, partnering with innovators around the world to accelerate scientific breakthroughs. CAS employs over 1,400 experts who curate, connect, and analyze scientific knowledge to reveal unseen connections. For over 100 years, scientists, patent professionals, and business leaders have relied on CAS solutions and expertise to provide the hindsight, insight, and foresight they need so they can build upon the learnings of the past to discover a better future. CAS is a division of the American Chemical Society.

Connect with us at cas.org

CAS

A division of the
American Chemical Society

SVCGENENGCSS100847221010