SciBite

Standardizing taxonomies to unlock *deeper scientific insights*

ELSEVIER

Advancing human progress together

Across the life sciences and broader R&D landscape, taxonomies are vital for capturing, structuring and retrieving scientific data

By connecting related concepts and standardizing scientific language, taxonomies provide the scaffolding needed to extract accurate insights and maximize ROI from advanced analytics, including AI. In the absence of industry-standard taxonomies, data scientists must invest significant time and effort in aligning external datasets with the internal standards their companies have established. This adds complexity to data integration efforts and slows discovery. To fully realize the value of data, scientific teams need an interoperable, scalable approach to taxonomy management.

Spanning more than a million concepts and tens of millions of synonyms, Elsevier’s taxonomies enable deep, high-quality indexing and discovery for its customers

Many research-intensive and pharmaceutical organizations are spending heavily on internal taxonomy initiatives. Often, public standards are used as a starting point, such as GMDN for medical devices, UniProtKB for protein targets and the Gene Ontology. These taxonomies make data FAIR (findable, accessible, interoperable and reusable), enriching how data can be searched, linked and analyzed.

Elsevier is an example of a company that has long recognized the strategic value of taxonomies within its product portfolio. The company has developed a wide array of rich, domain-specific vocabularies to support platforms including ScienceDirect, Reaxys, Embase and PharmaPendium. Spanning more than a million concepts and tens of millions of synonyms, Elsevier’s taxonomies enable deep, high-quality indexing and discovery for its customers. However, as proprietary structures, these taxonomies were typically not aligned with public standards. For customers, this misalignment meant added time and effort to manually harmonize Elsevier’s datasets and product data with internal taxonomies.

Additionally, Elsevier faced internal challenges managing its taxonomies. Taxonomy development was siloed across domains within Elsevier, including life sciences, health and engineering. Each team operated independently with its own processes for maintaining and updating taxonomies. This resulted in a heavy reliance on manual processes and limited reuse of curated taxonomy data across products, for example, duplicated efforts for disease taxonomies in PharmaPendium and EmBiology.

To support taxonomy growth, enable automation and support its mission to provide customers with high-quality, trusted data, Elsevier recognized an opportunity to build a more unified and scalable approach to taxonomy management. By more closely aligning its data with public taxonomies, Elsevier’s data would be more interoperable with its customers’ systems and accelerate insight generation from AI and other analytics technologies.

Introducing *The VOICE Project*

To address the growing complexity of taxonomy management and the need for greater interoperability, Elsevier's SciBite launched **The VOICE Project: Vision for Ontological Interoperability and Content Enhancement.**

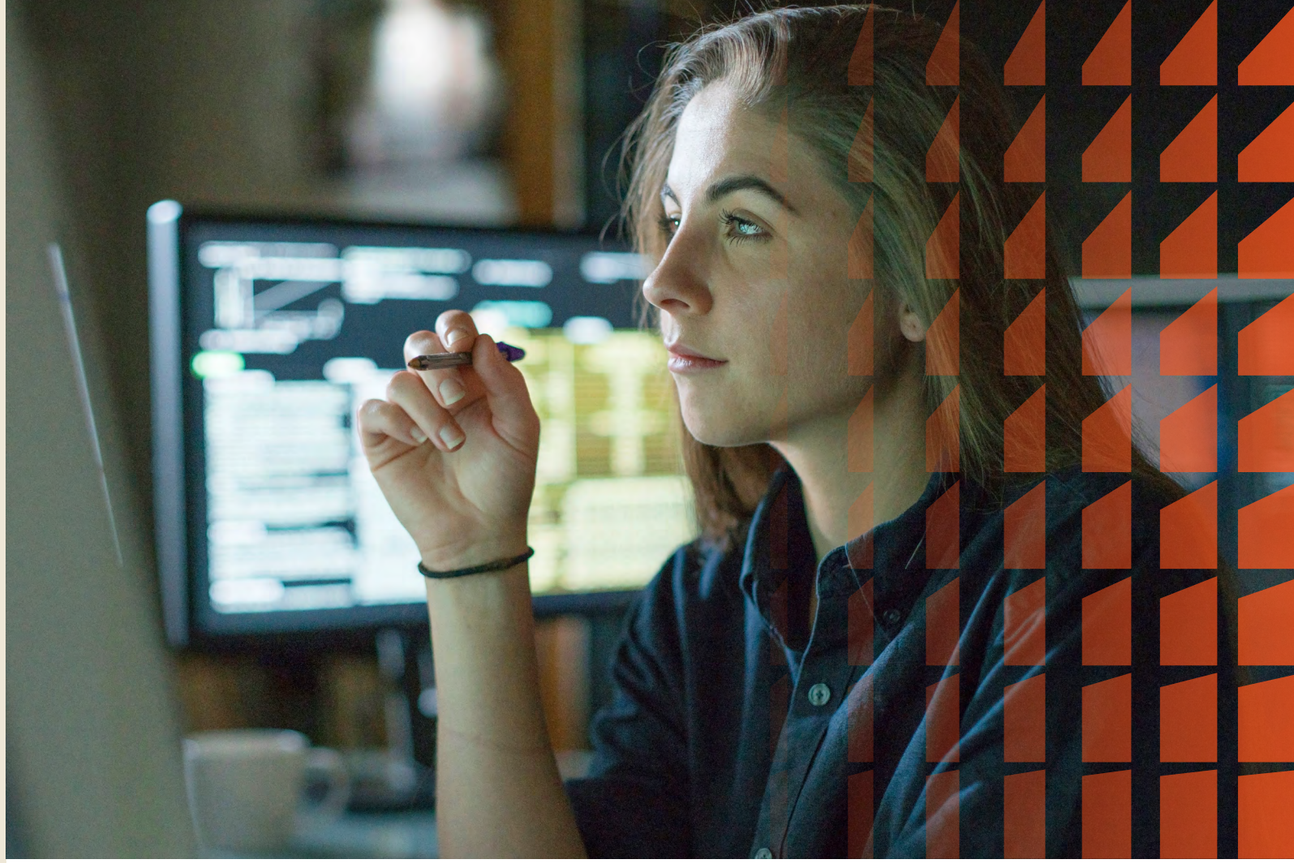
VOICE represents a strategic transformation in how taxonomies are created, maintained and delivered across Elsevier's extensive product ecosystem.

The objective was to unify Elsevier's siloed taxonomy landscape and lay the foundations for scalable, interoperable and FAIR-compliant taxonomies that could support both internal workflows and customer-facing integrations. VOICE aims to:

Consolidate taxonomy management across domains and products onto a single, centralized platform.

Automate previously manual processes to improve scalability and reduce redundant effort, such as candidate term identification, mapping and enrichment.

Deliver taxonomies that are FAIR and easily mapped to public standards, ensuring smooth integration into customer data pipelines.



Defining a FAIR taxonomy

One of the first steps in the project was to define what a FAIR taxonomy means in practice. This involved a rigorous review of taxonomy governance, usability and data stewardship across Elsevier's estate, resulting in several guiding principles:

Access rights and compliance:
Taxonomies needed to be structured with clear privacy controls and GDPR compliance, especially important when dealing with sensitive health and biomedical data.

Vocabulary and data reuse:
Taxonomies must support both internal and external vocabularies and ontologies (e.g., MeSH, UniProtKB, Gene Ontology) to promote data reuse.

Streamlined mapping and ingestion:
VOICE must standardize processes for sourcing, validating and integrating new terms to ensure consistent terminology. It was also important to include customers in this loop to inform taxonomy evolution.

Resolve ambiguities:
Where needed, qualifiers must be introduced to resolve conflicting meanings and ensure semantic precision. For example, 'cold' as temperature vs illness.

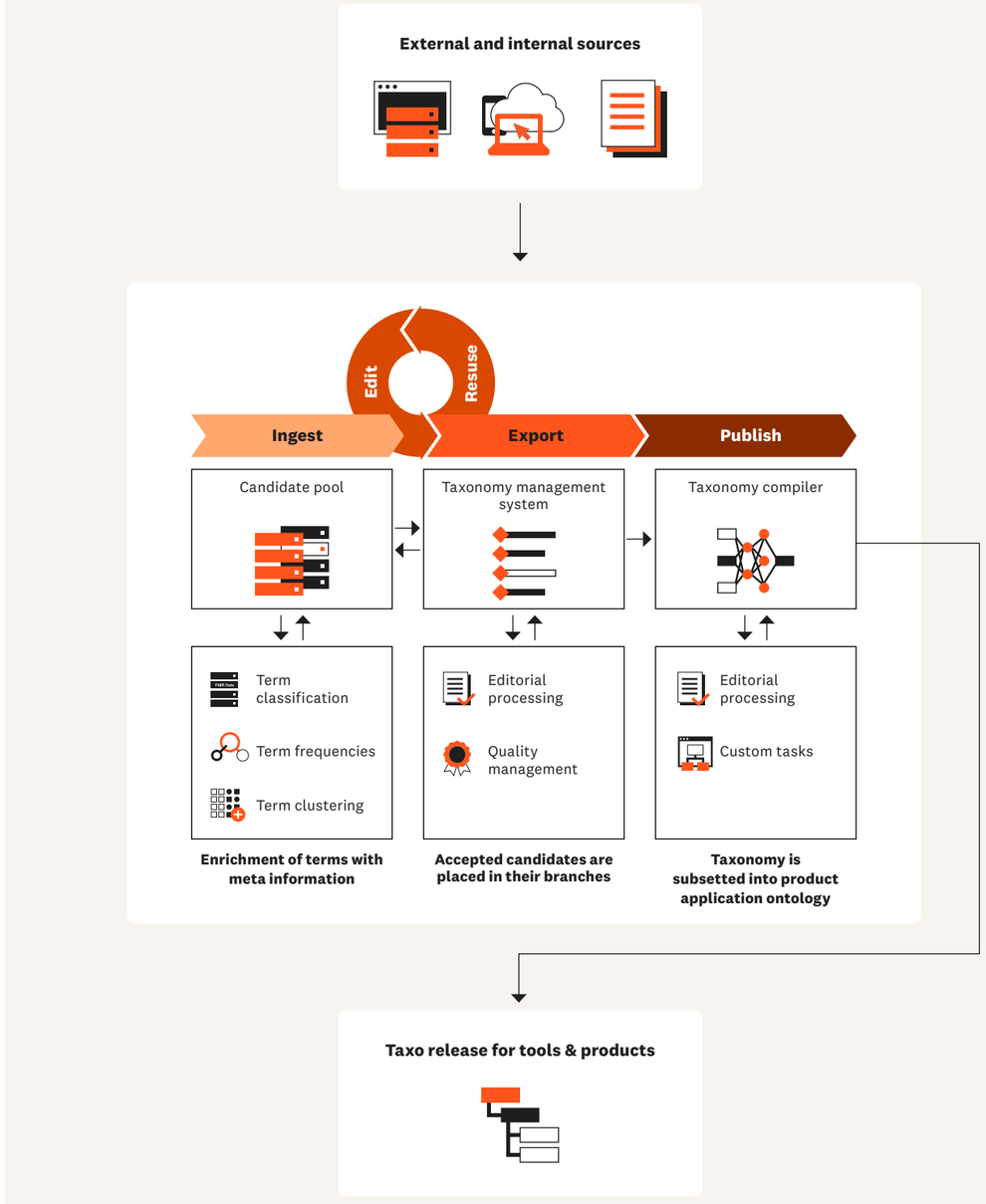
Prioritize usability and change management:
To support long-term sustainability and adoption, VOICE placed a strong emphasis on making taxonomy content more accessible to both internal teams and external partners. This included addressing change management, documentation and governance challenges common in large organizations.

Taxonomy production process in three stages

The taxonomy production workflow is shown in *figure 1*. VOICE introduced a modular architecture for taxonomy development, comprising three core stages: the candidate pool, the taxonomy management system and the taxonomy compiler. Each stage plays a specific role in transforming raw terminology into high-quality, FAIR-compliant taxonomies ready for customer and Elsevier product use.

Figure 1:

The taxonomy production workflow in the VOICE project



Candidate pool

The candidate pool is the initial intake point for new and potentially relevant terms, capturing concepts before editorial review. VOICE standardized the candidate selection process by introducing a robust process designed to improve consistency, traceability and efficiency:

Variety of data:
Source terms are drawn from a wide range of internal and external data, including scientific literature, product databases and public ontologies, to determine their relevancy.

Historical tracking:
All candidate terms are logged with full provenance, enabling editors to trace sources and avoid duplicate effort.

Term normalization:
Lexical variants are harmonized (e.g., dashes removed, Greek characters converted, spellings standardized) to support consistent comparison and matching.

Categorization:
Terms are routed to appropriate subject matter experts (e.g., drug, disease, medical device) for validation.

Clustering:
Related terms are grouped to identify whether a candidate is a synonym or a distinct new concept.

Frequency analysis:
Each candidate is scored based on its occurrence across trusted literature corpora to assess significance and relevance.

Modular workflow:
Editors can select and combine steps flexibly, enabling efficient review and enrichment.

Figure 2:

Candidate pool process in action

Input	Normalization	Categorization	Clustering	Frequency counting	
Ibuprofen	Ibuprofen	**Ibuprofen**	Rheumatoid arthritis Arthritis deformans	Rheumatoid arthritis Arthritis deformans	👍
Arthritis, rheumatoid	*Rheumatoid arthritis*	*Rheumatoid arthritis*			
Advil	Advil	**Advil**	Ibuprofen Advil	Ibuprofen Advil	👍
Arthritis deformans	Arthritis deformans	*Arthritis deformans*			
Tamoxifen	Tamoxifen	**Tamoxifen**	Tamoxifen	Tamoxifen	👎

Figure 2 shows the candidate pool process in action. Terms are visually categorized: bold for drugs and italics for diseases. Advil and Ibuprofen are clustered since they are synonyms and appear multiple times, so they would be accepted as a new concept. Tamoxifen only appears once in the input literature and has no synonyms, so it might not be prioritized for inclusion as a new concept.

Prior to VOICE, taxonomy teams operated in silos, each using different tools and models. Now, there is a unified taxonomy management environment that supports the FAIR principles

Taxonomy management system and compiler

Once candidate terms have been enriched and validated, they are passed to the taxonomy management system for formal inclusion. Elsevier editors review accepted terms, assign them to the appropriate branches and ensure each term is correctly positioned within the broader taxonomic hierarchy. This stage is critical for maintaining consistency, traceability and editorial quality across products.

Prior to VOICE, taxonomy teams operated in silos, each using different tools and models. Now, there is a unified taxonomy management environment that supports the FAIR principles:

Findable
All taxonomies are now hosted on a single platform, making them easier to locate across the organization.

Accessible
Editors and product teams can now work from the same system, reducing duplication and improving collaboration.

Interoperable
The platform supports mapping between taxonomies. Users can now identify differences in how terms are represented, e.g., “How does Emtree refer to pancreatic cancer compared to Omniscience?”

Reusable
Editors can now borrow and repurpose concepts across taxonomies, fostering consistency and reducing maintenance effort.

CENtree: A purpose-built foundation for FAIR taxonomy

Scientific language is highly nuanced and constantly evolving; off-the-shelf content management platforms aren't equipped to handle this complexity. To power VOICE's taxonomy management system and compiler, Elsevier needed a solution that could support both the scale of its taxonomy estate and accommodate the precision and editorial control required for scientific content.

SciBite's CENtree provided an ideal foundation; it is purpose-built for life sciences, but adaptable across domains. CENtree offered the flexibility to support Elsevier's wide-ranging product portfolio, including its engineering products like Knovel.

CENtree also met Elsevier's requirements around usability for editors of all technical backgrounds. CENtree has an intuitive interface and search capabilities that enable users to easily filter and locate terms by label, concept properties, tags and synonyms. Editors can track who has made changes and when, creating an auditable taxonomy management environment that ensures quality and consistency.

From an infrastructure standpoint, CENtree's API-first architecture means it can be integrated seamlessly into any organization's broader data ecosystem, which significantly accelerated Elsevier's project. The platform supports W3C web standards, ensuring the taxonomies produced are both technically robust and future-ready.

Finally, bringing together multidisciplinary experts from SciBite and Elsevier ensured internal and customer needs around taxonomies were met. The joint effort between the two teams ensured a smooth transition and equipped Elsevier's editors with the tools needed to manage taxonomies with greater speed, clarity and control.

A scalable taxonomy ecosystem for cross-domain discovery

The VOICE Project delivers tangible benefits for both Elsevier's customers and internal editorial teams. The project supports easier data integration today, which will enable more advanced, cross-domain insight in the future.

For R&D organizations using Elsevier data, VOICE enables:

- A more intelligent, context-aware search experience. For example, see figure 3: searching for "multi-drug resistance" in Embase automatically surfaces conceptually linked entities such as targets, proteins and disease subtypes.
- Faster integration and enhanced taxonomy interoperability, by improving alignment between vendor and internal taxonomies.
- Future-facing benefits, such as facilitating cross-domain insights. For example, VOICE could connect data from life sciences and engineering sources to answer advanced research use cases, such as: "Are there insights from materials science that could enhance my drug delivery mechanisms?"

"Early indications suggest VOICE could save data scientists a fifth of the time they spend integrating external data – which we expect to increase as more taxonomies are standardized."

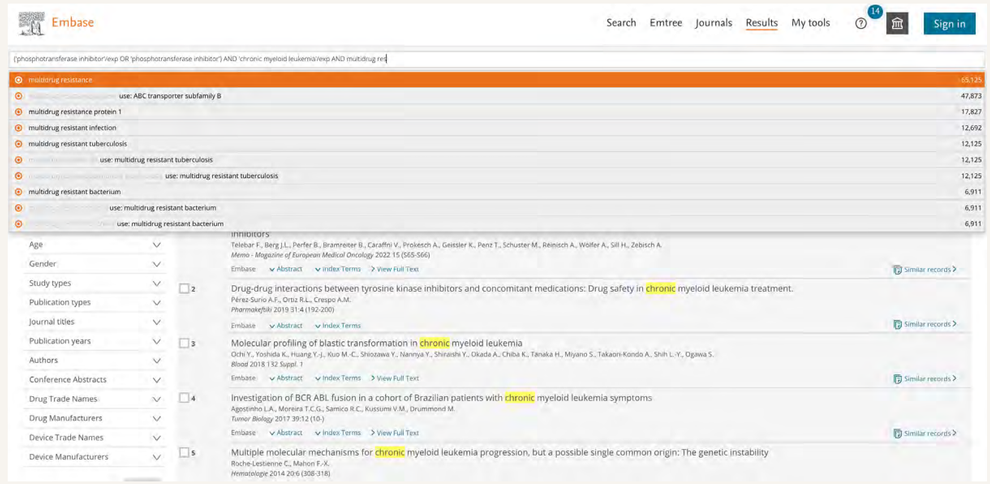
George Georghiou
VOICE Project lead

For Elsevier editors, VOICE has:

- Reduced technical demands as teams no longer need to maintain multiple taxonomies across disparate systems.
- Improved transparency and collaboration through shared governance and streamlined tooling.

Figure 3:

Searching for "multi drug resistance" in Embase automatically surfaces conceptually linked entities such as targets, proteins and disease subtypes



Summary

Strong data foundations like taxonomies are an essential scaffold for downstream technologies like AI. Without that foundation, insight is limited and integration is costly.

Change management is challenging, but worth it. Optimizing Elsevier's taxonomy processes was a necessary step toward building scalable, sustainable taxonomies.

Bridging data and domain expertise unlocks value from data projects faster, as demonstrated by the CENtree partnership with SciBite.

Technology doesn't replace expertise. Automation enhances efficiency, but subject matter experts remain vital to validating meaning, resolving ambiguities and evolving taxonomies over time.

About SciBite

SciBite is an award-winning semantic software company offering an ontology-led approach to transforming unstructured content into machine-readable, clean data. Supporting the top 20 pharma with use cases across life sciences, SciBite empowers customers with a suite of fast, flexible, deployable API technologies, making it a critical component in scientific data-led strategies. Contact us to find out how we can help you get more from your data.

To learn how SciBite can unlock the value of your data, speak to one of our experts today or email us at contact@scibite.com.

ELSEVIER