



Vision

A unified AI agent that finds, processes and uses decades of scientific discoveries, both privately owned and in the public domain, to actively drive R&D in materials science.

Key Technology

Overview

Oghma unlocks valuable scientific data trapped in technical documents. We convert reports, papers, spreadsheets, scanned records and other fragmented sources into trusted, structured data that R&D teams can search, analyse and use for modelling. Every extracted value remains traceable to its original source, quality-controlled against agreed requirements, and ready to integrate with the scientific and enterprise systems that the organisation already uses.

Building machine-learning-ready databases at scale

Scientific organisations hold valuable data in journal articles, internal reports, spreadsheets, tables, charts and scanned records. Manually finding, extracting and standardising this information is slow, expensive and difficult to scale.

Oghma builds bespoke multimodal agentic workflows that convert these fragmented sources into structured scientific databases aligned with each customer's schema or ontology. These workflows extract and connect information from text, tables and figures across formats including PDF, HTML and spreadsheets.

Every extracted value remains linked to its original source and passes through configurable quality-control and validation workflows. The resulting datasets can be searched, analysed, integrated into existing scientific systems and used as a foundation for statistical analysis and machine-learning models.

This allows R&D teams to:

- Recover valuable information from historical and external documents
- Replace repetitive manual extraction with scalable workflows
- Standardise data across inconsistent formats and terminology
- Create traceable datasets suitable for analysis and modelling
- Expand databases continuously as new documents become available

Automated model construction

Creating a structured dataset is only the first step towards data-driven R&D. Researchers must then determine which variables, representations and modelling approaches reveal meaningful relationships, a process that often requires extensive manual experimentation.

Oghma develops agentic workflows that propose, implement and evaluate scientific descriptors and modelling strategies within reproducible pipelines. These workflows can reduce manual feature-engineering effort, broaden the range of hypotheses tested and help reveal relationships that might otherwise remain hidden.

This allows R&D teams to:

- Find relations between data
- Understand the data structure and its value
- De-risk models construction in data-limited regimes
- Reduce the time for model construction



Scientific Sources

Reports, PDFs, Spreadsheets, Tables, Figures, Scanned records



Oghma Workflow

Find relevant information; Extract from text, tables & figures; Map to your schema / ontology; Validate against agreed requirements



Structured Output

Quality-controlled dataset; Source-traceable records; Ready for integration



R&D Value

Search & reuse knowledge; Analyse relationships; Support modelling; Continuously expand databases



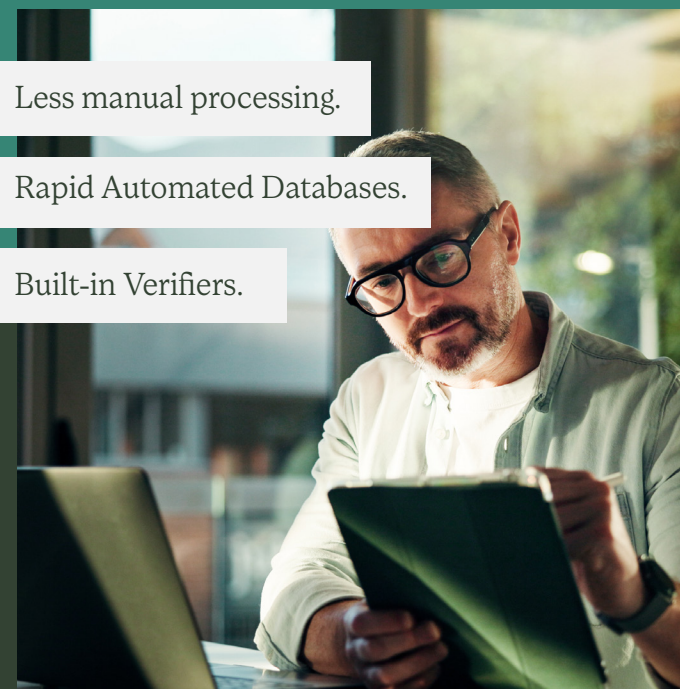
Less manual processing.



Rapid Automated Databases.



Built-in Verifiers.



User cases

Build ML-ready datasets from public data

About 1M peer-reviewed materials articles are published every year, but their information remains embedded in prose, tables and figures rather than available in directly usable form.

Working with APC Ltd., Oghma screened more than 20,000 publications and assembled a source-traceable solubility dataset containing approximately 25,000 data points across 250 small molecules. The resulting database provides a foundation for solubility-prediction models that can support pharmaceutical process-development decisions.

Unlock legacy R&D data

Industry R&D builds over legacy data accumulated over the years. This body of information is stored across reports, spreadsheets, laboratory records and databases created under changing standards.

Oghma worked with a chemical company to convert 30 years of R&D legacy data into structured source-traceable datasets, mapped to the organisation's preferred schema or materials ontology. This supports future experiments, predictive modelling and customer-facing information services.

Onboard third-party scientific data

R&D organisations routinely receive data from customers, contractors, research partners and analytical laboratories. These records are often compiled in unstructured formats (lab-books, reports, etc.) and follow a terminology and standards that differ from the organization's internal system.

Oghma extracts and maps this information to the customer's ontology while preserving traceability to the original evidence. In a toxicology project for a chemical-industry client, reports produced by multiple analytical laboratories were converted into a unified structured dataset despite differences in format and reporting conventions.



Contact us

Prof. Stefano Sanvito,

Oghma at CRANN Institute, Trinity College Dublin,
sanvitos@tcd.ie



A powerful engine for materials R&D

sanvitos@tcd.ie