

Oghma compiles a ML-ready database of 30-year R&D internal data

Problem statement

Most R&D data is proprietary and compiled over many decades of experimentation. Such a long timeframe makes data standardization complex: different data-recording practices overlap, conventions change, old measurement types are replaced by newer ones, and the data quality may reflect the practice of multiple operators. Furthermore, it is not uncommon that access to operators is no longer possible (researchers may have left the company), so that ambiguities are difficult to resolve. Yet, there is enormous value in historical proprietary data.

A rational archive of data allows a company to navigate their product design history. Data can be incorporated into new R&D development, avoiding duplications and reducing the volume of testing and new experimentation to carry on. Data, when abundant, can be at the foundation of predicting machine-learning models, valuable for internal R&D and for customers. These models can be further improved by fusing proprietary data with public ones.

The challenge

Over the last 30 years of R&D, the personal care division of BASF has produced a large volume of data, relating product recipes with physical/chemical testing. These are produced through extensive experimental work, requiring long testing times over a complex design space. Historically, the experimental work was documented manually by operators, whose standards and ontology have evolved over the years, as a result of technology development and changing regulatory constraints. BASF has manually created a small structured dataset of such large data volume, *control dataset*, but this proved to be an extremely time-consuming exercise. The grand challenge was *to produce full digitalization of personal-care recipes and their corresponding physical/chemical properties*. This poses two main challenges:

- 1) Perform automated data extraction at an accuracy comparable with manual work
- 2) Scale up the process to a large corpus of documents, while keeping accuracy

The solution

Our adaptable in-house agentic framework built around scientific information extraction, *Oghma Extract*, was used to perform the extraction task. After calibrating the model on a small sample of representative BASF documents, we proceeded by performing data extraction over the same corpus of documents used by BASF for constructing a manually curated dataset (control dataset: 7 multi-sheet excel files). Independent verification from BASF revealed an accuracy of 99.7%, with the remaining 0.3% error margin to be attributed incomplete data in the source document. Following this validation, *Oghma Extract* was used to perform the extraction on an entire corpus related to a set of person-care products (465 excel documents).

Breakthrough results

Metric	Without Oghma	With Oghma	Improvement
Dataset size (entries)	154 recipes 14,200 properties	5,667 recipes 168,500 properties	37x data volume 12x data volume
Documents processed	7	465	66x documents processed
Total extraction time (control database)	10 weeks*	5 hours	7 documents processed in 10 weeks against 5 hours.
Extraction time per recipe and associated properties	78 minutes**	1.9 minute***	40x speed
Extraction Accuracy (control database)	90%	99.7%	Accuracy improvement over manual quality
Total extraction time (final database)	unfeasible	184 hours***	Volume of the dataset created beyond manual capability

* this is the total manual extraction time for one person and does not correspond to full-time work

** estimate based of 4h/day extraction time

*** based on serial extraction (no parallelization)

Competitive Advantage

Oghma agentic system enables the creation of a structured dataset from a large corpus of experimental reports describing 30 years of R&D. This contains data difficult and lengthy to reproduce, so that its value is high. A small control dataset was enough to construct predictive machine-learning models, a capability largely increased by the complete dataset. The accuracy of the extracted database exceeds that of the manually curated one, whose construction required frequent interaction with the lab operators. Scaling the manual extraction to the full data corpus will require 368 weeks from a single operator (estimated on a 4h/day extraction commitment), a technologically and operationally unacceptable time. A reduction of such time to 6 months will simultaneously mobilise 14 operators, making the process economically unfeasible.

“Oghma combines high extraction accuracy with the scalability needed to transform large volumes of complex historical R&D data into a reliable, machine-learning-ready resource. It is a practical tool for unlocking the value of historical scientific data for current R&D.”

Sandip Bhattacharya, Manager Data Science, Global R&D Personal Care Solutions, BASF

Note on BASF

Funded Headquartered in Ludwigshafen, Germany, BASF is a multinational company and the largest chemical producer in the world. It comprises subsidiaries and joint ventures in more than 80 countries, operating six integrated production sites and 390 other production sites across Europe, Asia, Australia, the Americas and Africa. BASF has customers in over 190 countries and supplies products to a wide variety of industries.