## Case Study

# Developing a Dataset with Protein Degrading Enzyme-Substrate Relationships

## The Purpose

A mid-size pharma company sought to create a comprehensive list of human enzyme datasets that have protein degrading enzyme – substrate pairs, along-with their associated effects/degradation levels to support their internal research.

## About the Client

**COMPANY**
Mid-size pharma

**LOCATION**
Europe

**THERAPEUTIC AREA**
Protein degradation

## Client Requirement

The client required a comprehensive list of protein degrading enzymes and their specific substrates to support their internal organizational needs. The objective was to utilize Excelra's proprietary text mining pipeline to identify and create this list.
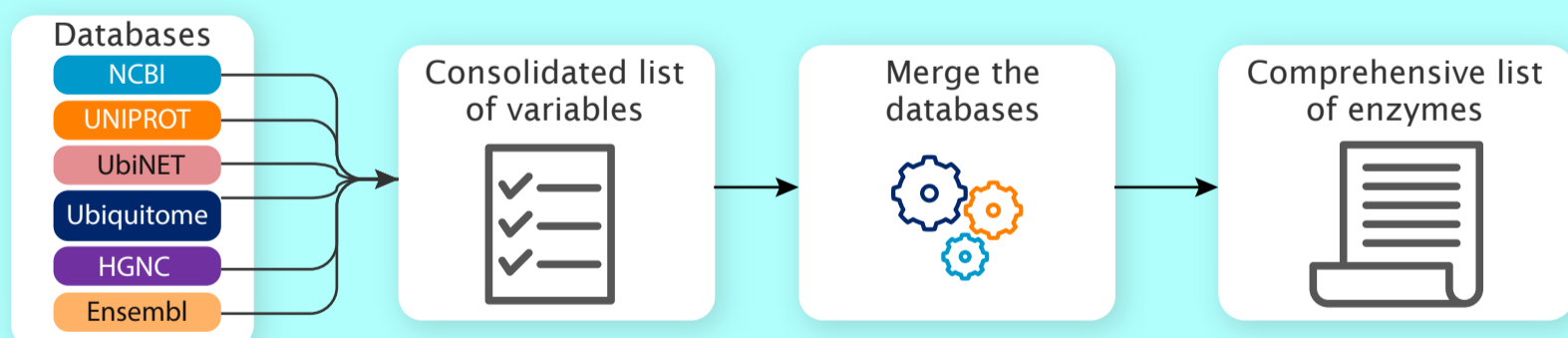
## The Excelra Approach

The objective of the project was to develop a dataset with experimentally proved enzyme -substrate pairs along-with their specific levels of protein degradation. The solution strategy implemented was outlined in 3 different phases.
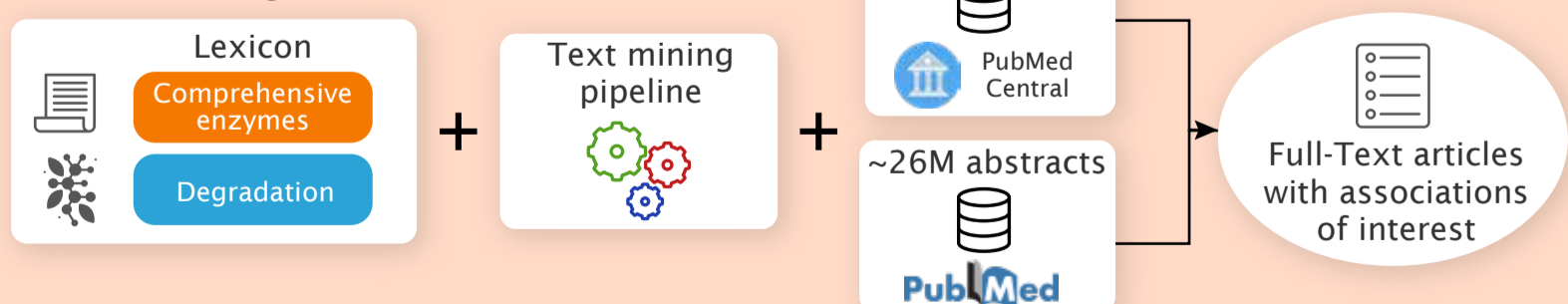
• 'Phase 1' involved developing a comprehensive list of human protein degrading enzymes.

• 'Phase 2' elaborated on building a literature mining tool to extract relevant articles.

• 'Phase 3' focused on manually curating the relevant articles to develop a dataset and a user interface (UI) to visualize the output data.
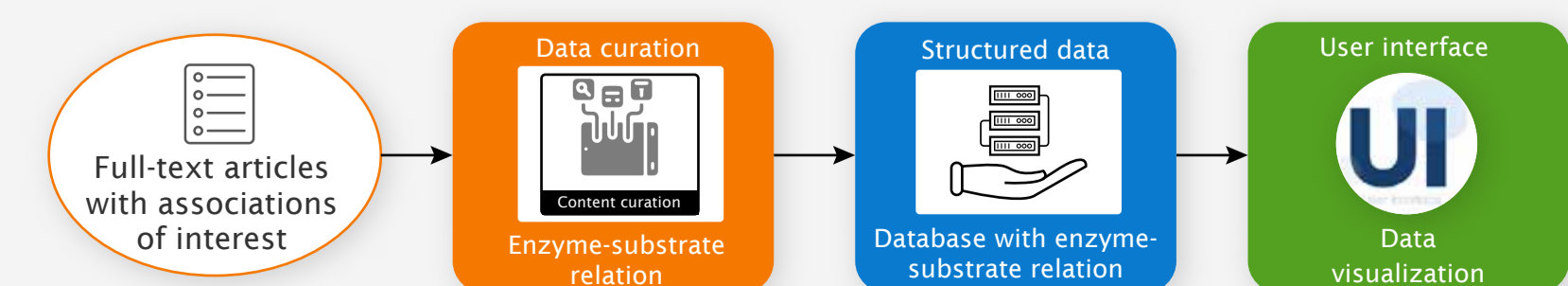
## 3 Phased solution strategy

**Phase 1, Preparing the consolidated list of human proteases**

Databases
- NCBI
- UNIPROT
- UbiNET
- Ubiquitome
- HGNC
- Ensembl

→ Consolidated list of variables → Merge the databases → Comprehensive list of enzymes

**Phase 2, Developing a text mining pipeline and extracting the relevant data**

Lexicon
- Comprehensive enzymes
- Degradation

+ Text mining pipeline +

Dataset
- PubMed Central
- ~26M abstracts
- PubMed

→ Full-Text articles with associations of interest

**Phase 3, Manual curation of extracted relevant articles**

Full-text articles with associations of interest → Data curation
Content curation
Enzyme-substrate relation → Structured data
Database with enzyme-substrate relation → User interface
UI
Data visualization

# Excelra's Contribution

Using our proprietary text mining pipeline, we provided the client with a comprehensive list of relevant enzymes and experimentally approved enzyme-substrate pairs, along-with their associated levels of degradation effects, which helped the client to start a new program at their end.

# Excelra's Service Portfolio

| | **Insights** | **Data** |
|---|---|---|
| **Discovery** | • Data Science Driven Drug Discovery<br>• Target Identification<br>• Target Dossier Services | • Chemistry Curation Services<br>• GOSTAR Structure Activity Relationship database |
| **Translational** | • Biomarker Discovery<br>• Drug Repositioning<br>• Life Cycle Management<br>• Systems Biology Informatics | • Biology Curation Services<br>• GOBIOM Biomarker intelligence database |
| **Clinical** | • Precision Oncology Informatics<br>• Clinical Pharmacology | • Clinical Trial Outcomes Database |
| **Value Evidence** | • Outcomes Research<br>• Epidemiology Modelling<br>• Economic Modelling<br>• Value Evidence Communication | • RWE & Big Data Realization<br>• SLR & Meta-analysis |
| **Technology Solutions** | • Enterprise Data Strategy<br>• Enterprise Cloud Ops<br>• Enterprise Digital Transformation | |

For more information, visit https://www.excelra.com/translational/#biology_curation

**excelra**