ACCURIDS

HOW A LEADING PHARMA COMPANY IMPROVED THE DATA CULTURE IN R&D

Supporting interoperability of R&D Data Assets with linked reference data

Heiner Oberkampf
ACCURIDS

Armin Meinel
ACCURIDS

Within the context of digitalization, data has become the center of both traditional and research-intensive industries. For over 120 years, a world leading pharmaceutical company has been researching and developing innovative medications and new therapeutic approaches that help make a difference in people's lives. Combining data from different sources inside and outside the organization plays a critical role in this innovation process. This means that the pharma company must be able to produce high-quality data assets at scale to be successful. The R&D Data Assets team is part of the Digital Transformation. Their main objective is to ensure data fluency across the R&D value chain by enforcing reusability and interoperability. Thus, it is of utmost importance to provide always up-to-date code lists, and to make this multitude of large datasets searchable. This simplifies the identification of internal and external terms, codes, and their relationships.

THE IMPORTANCE OF CREATING A GOOD DATA CULTURE

Master and reference data are essential for research and development in the pharmaceutical industry, because they provide structured context to information about experiments, materials, devices, and regulatory submissions. This context is crucial to meaningfully analyze data and quickly understand what factors make or break a future product. In this sense, the establishment of a data culture based on good reference and master data is critical for a successful R&D organization.

At the world leading pharma company, effective data management is particularly important, because it lets researchers tap into the wealth of information already available in the company. Only by leveraging these data assets is the pharma company able to develop truly innovative medications efficiently. As with other organizations, this pharma company embraces a FAIR approach to managing data assets. As a first step, the data assets team ensures that different user groups can easily find existing data, based on semantic annotations with standardized terminologies. This is increasingly important, since scientific innovation today requires combining data insights from many different sources.

At a glance

About R&D:
Teams of experts focused on the research and development of innovative medications, with research centers in Germany, US and other countries. Data takes a central role in their work.

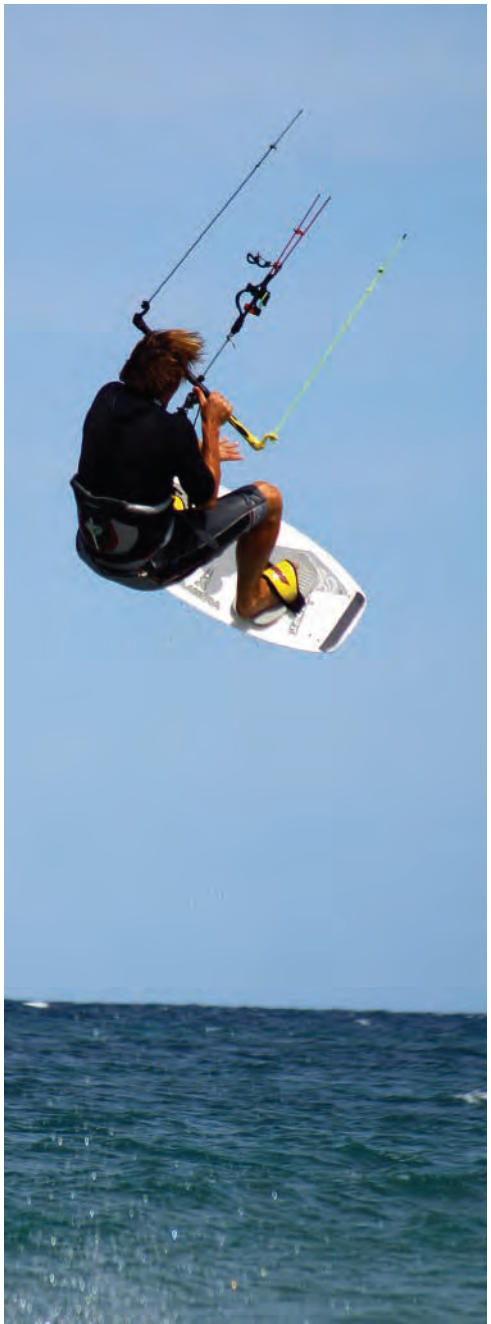
CHALLENGE:
Reference and master data are distributed across applications, with high redundancies and low adoption rate of internal or external standards.

SOLUTION:
Semantic reference and master data registry, with simple and scalable lookup services to increase adoption.

RESULTS:
Interlinked information of more than 250 ontologies, with a total of more than 20 million registered concepts, provides context for automated integration and allows scientists to generate more insights faster.

BARRIERS TO MAKING THE MOST OF DATA

R&D data comes in all shapes and sizes, so it is not always possible to fully standardize it. This is especially true when we include external data into our research. Working with this diverse information is challenging for a number of reasons.



FINDING THE RIGHT CODES FOR ANNOTATING DATA IS DIFFICULT AND TIME CONSUMING

Finding the right terminology is crucial for creating high-quality semantic annotations. Unfortunately, searching for the appropriate codes is challenging in large organizations because many different terminologies are used. Additionally, large taxonomies are time-consuming to navigate, or sometimes even impossible to manage, without good searching and browsing functionality. As a result, businesses may not widely adopt standards that would improve data interoperability.

THE NEED FOR INTEROPERABILITY BETWEEN INTERNAL AND EXTERNAL DATA ASSETS

Scientists at the pharma company need to integrate internal data and external information for conducting analyses that lead to new insights. However, meaningful and efficient comparison and integration is only possible if data annotations from various provenances are interoperable. For instance, using consistent data standards (e.g., for compounds and assays) across internal and external data is possible with the appropriate information mapping. Creating these links between information, however, can be a difficult process.

COORDINATING THE DATA STEWARDS WHO MANAGE TERMINOLOGIES IS CHALLENGING

In any large organization, there are many data stewards working within the different business functions. This means that, without a central governance registry where terminologies can be easily published, it is impossible to coordinate data stewardship activities. A natural consequence of the lack of coordination between different data stewards is that multiple versions of the same terminology end up being used internally. Also, there are cases where multiple terminologies for the same data domain are created and managed redundantly. To overcome this issue, data stewards need to align in their management efforts, with transparent roles and responsibilities. Therefore, it is absolutely necessary to adopt a consistent platform that lets everyone browse across internal and external terminologies, retrieving relevant information.

THE NEED FOR SEARCHING ACROSS DATA STANDARDS

Out of the many roles that regularly interact with data at a pharma company, the following user groups have particular requirements for high-quality data annotations.

DATA SCIENTISTS AND BIOINFORMATICIANS

Data scientists and bioinformaticians use data to analyze and understand biological processes, helping scientists navigate a wealth of scientific information. To deliver the best possible results, they need:

- The ability to quickly look up codes for a term, so that they can deliver accurate clinical reports.
- Data that is appropriately linked to the terminology, so that they can further analyze a given concept as required.
- A well-organized repository for the data, so that the relevant information can be efficiently found.

LAB SCIENTISTS

Lab scientists examine samples and carry out experiments, comparing results with previous data to contextualize their findings. To progress their research as productively as possible, they need:

- A very intuitive search function, so that identifying the right terms is effortless and does not disrupt their train of thought.
- The ability to find specific codes in very large ontologies, such as the NCBI Taxonomy – which has around 2 million concepts –, the NCI Thesaurus, or MeSH.
- An easy process for annotating experimental data with these standard terms, so that information can be thoroughly described.
- A system that integrates terminology searches, so that they do not have to switch applications and can instead run queries directly from within their electronic lab notebooks.

DATA STEWARDS

Data stewards share biomedical ontologies across the organization, promoting the standardization of terminology. To efficiently handle the very large ontologies they work with, they need:

- An easy way of updating the content, so that this is done regularly and the information stays up to date.
- The ability to create and, crucially, to share terminologies, so that they can work collaboratively with other stewards.
- A method for providing their work to other users in an easily consumable way.

CHANGING THE GAME WITH ACCURIDS

For broad user adoption, a robust and highly scalable tool with fast and intuitive tree visualizations is needed.

To be successful at scale, the R&D team needed to:

- ✓ Provide up-to-date terminology through enterprise-scalable tools, while also maintaining low operational costs.
- ✓ Bridge internal and external information by annotating data with public terminology, which increases reusability and interoperability.
- ✓ Increase the adoption of public standards by broadly sharing them, e.g. by making CDISC related terminologies easily available to access internally.
- ✓ Extend public standards with internal information, tailoring the terminology to support specific data assets.
- ✓ Connect applications to an integrated system, providing users with the right terms as part of their usual business workflow, e.g., to an Electronic Lab Notebook.

As no solution in the market could fulfill all of these needs, the team decided to co-develop a product with ACCURIDS, with the goal of closing important gaps in their tool stack. The most important feature was a highly scalable lookup service, allowing users to browse large biomedical ontologies and other large sets of reference and master data. Today, ACCURIDS is a key component in the success of the pharma companys new data asset management strategy. It provides all necessary functionality for addressing their specific needs, yet it is far cheaper than building and maintaining a custom solution.

Why ACCURIDS?

SCALABILITY:
ACCURIDS makes it possible to register and load all UMLS and OBO ontologies, supporting multiple versions within a simple and robust platform.

TREE VIEW:
The user interface displays data in a tree-like view, providing a user-friendly and responsive experience even for huge biomedical ontologies.

CUSTOMIZATION:
ACCURIDS can be customized according to users' needs, with settings for matching the look of the UI to that of other systems.

MAPPINGS:
We connect similar master and reference entities from different sources, providing a unified management strategy.

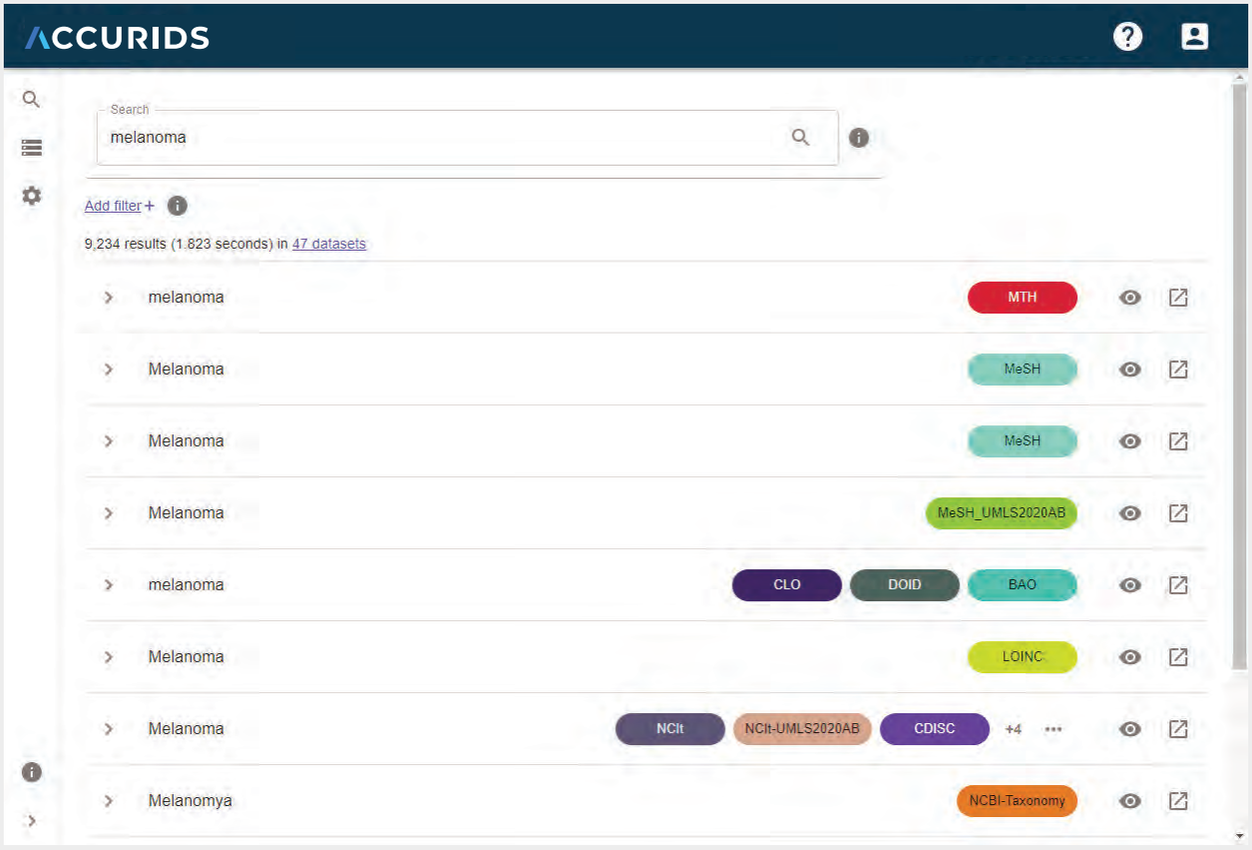


Figure 1: *ACCURIDS search is intuitive and allows to find terms across terminologies.*

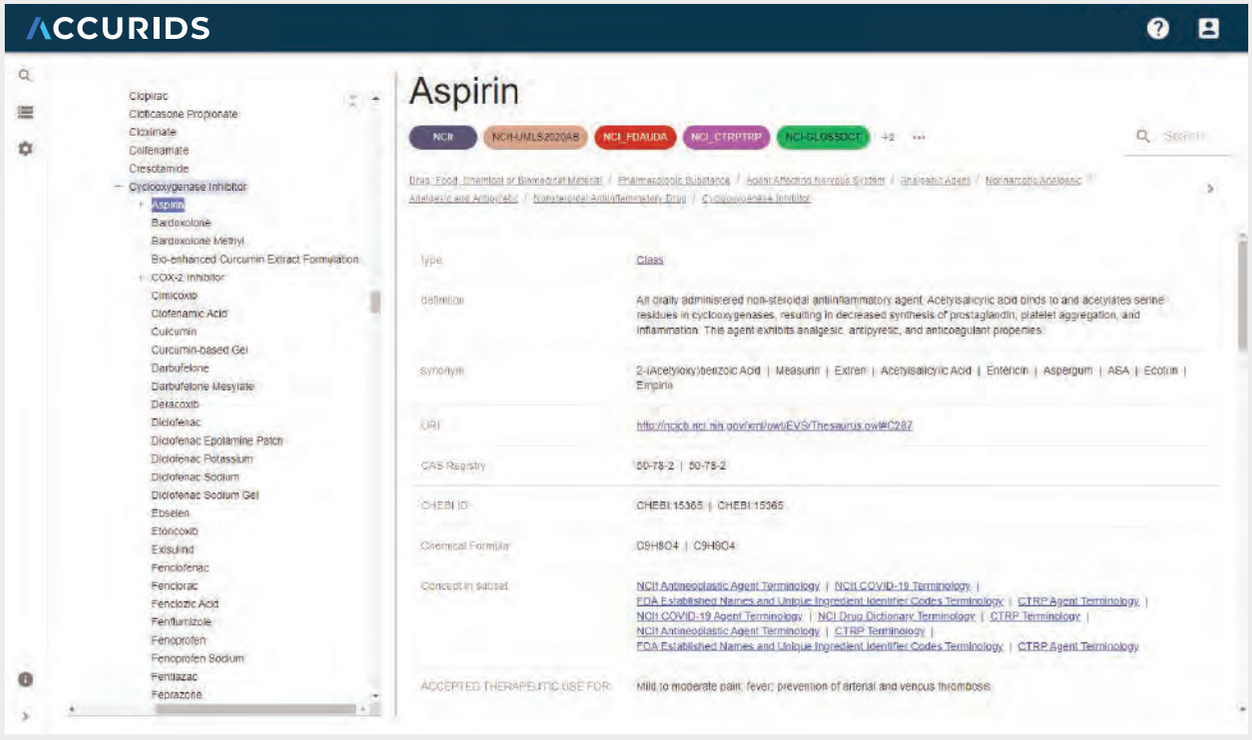


Figure 2: *ACCURIDS global entity view that shows the hierarchical information and data from 6 different terminologies*

HOW THE WORLD LEADING PHARMA COMPANY BENEFITS FROM USING ACCURIDS

With ACCURIDS, the Pharma R&D team can easily look up all kinds of terms, making it much easier to tap into the wealth of existing knowledge within the company.

With ACCURIDS, the R&D team can easily look up all kinds of terms, making it much easier to tap into the wealth of existing knowledge within the company. Thorough data annotations help users quickly find additional relevant information, from both internal and external sources. Additionally, applications such as the Electronic Lab Notebook Signals from PerkinElmer are now connected to ACCURIDS. With that, around 600 lab scientists can easily annotate experiments directly in their workflow, using standardized terms.

The new system provides the following benefits:

1. Data stewards have a simple way of publishing up-to-date terminologies and code sets avoiding costly redundant management.
2. Lab scientists can easily identify the right terms to describe information.
3. Data scientists can quickly find linked records for analysis and reports.
4. Business users have a scalable solution that is fully compliant with regulations.
5. Data governance managers can ensure high-quality information is produced and maintained.

In light of these results, the pharma company will extend the use of ACCURIDS to other divisions. While the terminology is different in those domains, following the data asset management approach employed for the pharma R&D side will bring similar rewards.

"With ACCURIDS Data Registry, we set a foundation for broad adoption of internal and external standards, because it is really easy to use by different user group, e.g. data stewards, data scientists and researchers."

Head of Pharma R&D Data Assets

Learn more about how ACCURIDS can help you govern and publish reference and master data from distributed sources.
Visit accurids.com