Unlocking the Business Benefits of Text Mining in Regulatory Operations

As IDMP implementation advances in Europe, data is becoming a key asset for the life sciences industry. Currently, pharma companies are focusing on the heavy burden of initial data collection, but ultimately it is data maintenance that will become the prevalent challenge. This can only be tackled by paying continuous attention to the quality and integrity of data. Advanced text mining technologies, applied in the context of end-to-end regulatory information management, can help considerably both with accurate data ingestion and ongoing data maintenance. Amplexor's Renato Rjavec explains.

The protracted business of collating and cleaning up data, in preparation for the new target operating model for EMA regulatory submissions, in which original product data must be submitted alongside eCTD dossiers, has triggered a whole raft of activity and resource use for European and global life sciences organisations.

The road to full IDMP compliance has not been an easy one; nor does it end with initial registrations. On the one hand, many marketing authorisation holders (MAHs) are still trying to locate source data, vet its quality, and plug any gaps. The information they need may straddle regulatory information management (RIM) systems, Excel spreadsheets and any number of static documents (labelling, CMC documents, and so on). This may be strewn across functions as diverse as Regulatory, Supply Chain, Pharmacovigilance and Commercials – each department frequently employing its own preferred formatting and terminology. Extracting and cleaning up all of these fragments of data to form something meaningful and usable is a massive undertaking.

Yet the work to this point – building a complete and viable data set - is just the tip of the iceberg. The job of maintaining and updating all of this information, and keeping it tightly aligned with anything appearing in document form, will be never-ending. Under the emerging target operating model (TOM) for regulatory submissions, once IDMP is live and mandatory in the EU, any discrepancies between the product data and the dossiers filed in parallel will immediately spark Agency queries and set back registration timelines.

Ensuring that data and content remain in sync and up to date, and that FHIR messages (conforming to the Fast Healthcare Interoperability Resources standard data formats/API requirements for exchanging electronic health records) are fully aligned with the content of submitted eCTD sequences, will be essential to efficient process management and registration progress.

Assessing the Technology/Process Options

To keep on top of Agency expectations, teams responsible will need to harness technology strategically. Processes must be established to ensure that the contents of the dossier match the contents of the IDMP/SPOR dataset for each submission. (Under IDMP, Substances Products Organizations and Referentials – SPOR – data services provide the vehicle for implementation of ISO IDMP standards in the regulatory and e-health worlds.)

Continuous Data Extraction

One option is to pull data from documents as an ongoing operational process, but this approach is likely to be very labour intensive and offers companies very little additional benefit beyond compliance.

Structured Authoring

The opposite option is to leverage well-structured data to generate content. Structured content authoring technology (in which documents are assembled automatically from pre-approved content fragments/data sets) would appear to be the optimal long-term option, however the technology is not yet mature enough to offer a failsafe and simple-to-use solution, allowing dossiers to be created intelligently using approved source data.

Advanced Text Mining

A better approach, at least for the time being, is to establish parallel processes to prepare documents and data, keeping both in tight alignment and ensuring this is the case as a quality control requirement until the final submission.

In this context, companies would do well to harness an already proven technology – advanced text mining. This has strong potential for application both at a data extraction/quality checking level, and for ongoing data and content maintenance.

The accuracy of such tools has reached around 95 per cent in the context of automated data extraction, meaning that teams can place a lot of trust in it – saving human resources for an oversight role or to home in on more complex use cases.

So How Does it Work?

Text mining technology uses machine learning and natural language processing to help teams detect patterns or data points in existing documents, such as content around the composition of a drug, any counter-indications, or manufacturing detail. Once identified, it can extract this information and encode it properly using the correct controlled vocabularies and flow it into the company's RIM system for onward processing. On top of the technology's strong track record at doing this accurately, text mining tools are also very good at detecting whether the original data used in the documents was wrong, flagging this as a potential quality or consistency issue.

Double the Potential: Aiding Data Extraction & Ongoing Data Maintenance

In initial data collection use cases, where text mining is already gaining traction, the technology is helping greatly to improve

the efficiency of IDMP data extraction from a range of different documents, automatically populating RIM data records directly from those static files. This provides teams with a good foundation for data enrichment, allowing skilled professionals to focus their time on populating additional fields that are now needed.

At an ongoing data maintenance level, advanced text mining tools support proper data validation and user guidance to ensure that data is and remains complete, consistent, and properly encoded, ready for a final review and approval by a human supervisor. This vital validation step ensures that discrepancies are picking up and gaps identified and flagged to the experts overseeing the data quality.

Tangible ROI

The return-on-investment potential of advanced data mining tools in both data extraction and data maintenance use cases is impressive – as long as the technology is harnessed appropriately within the context of end-to-end regulatory information management processes.

In a data extraction context, where a set of documents must be read to populate product records, potentially taking someone four hours per record, a text mining solution can (very conservatively) halve that time. With a potential saving of 100s of euros/dollars per record, companies processing tens of thousands of authorised records per year could see cost savings run into the millions.

For data/content validation – checking the consistency of data and eCTD dossiers, which becomes critical in the IDMP era – the potential to use smart text mining to compare product records with submission document content is enormous too and can generate considerable business value. By at least halving the current error/discrepancy rate via automated content validation, companies could on average yield a cost saving of 100s of euros per submission. Multiply that by 10,000 submissions annually, and the math stacks up robustly.

Across the two use cases then, text mining can play a vital and direct role in improving efficiency, reducing costs, improving quality and minimising errors. More than that, advanced text mining has a meaningful role as part of a broader, end-to-end RIM capability – aiding planning, editing and formatting throughout, through its ability to validate data across the entire lifecycle.

Latent Potential

If more companies were aware of text mining and how mature the technology is today, its take-up would be considerably higher than it is currently. Once responsible teams are made aware of such solutions, it usually takes only a small proof-of-concept study to showcase the potential and lay to rest any concerns about the technology's accuracy and efficacy.

Ideally, text mining technology should be deployed seamlessly as part of a broader RIM project – as part of an IDMP data migration initiative, as companies press on with data cleaning, structuring, and importing, ready for the IDMP go-live date. Similarly, for ongoing data validation (maintenance, updates), text mining needs to be integral to RIM too, so that the benefits can be leveraged effectively in everyday regulatory operations.

Whether the teams responsible are exposed to the technology directly or not, it is something that should be on their radar when assessing how they will accomplish their projects and keep within their allotted timeframes and budgets.

Renato Rjavec

Renato Rjavec is Director of Products, Life Sciences at Amplexor. Amplexor Life Sciences helps organisations that are developing pharmaceutical drugs, medical devices, and biotechnology to launch products and break into new markets quickly using end-to-end regulatory and quality management solutions. Its solutions and services expedite the management of highly structured data and the creation and delivery of consistent, compliant global content. Amplexor's services include technology consultancy, implementation, and management services. Its partner in the area of text mining is Averbis.

Email: renato.rjavec@amplexor.com