Text analytics reduces regulatory affairs costs, speeds compliance

Overview

Pharmaceutical companies need tools to enable rapid and effective responses to regulatory challenges. Text analytics can bring value in a wide range of regulatory use cases:

- **Text analytics for compliance and master data management**: Extracting data attributes from regulatory documents (SMPC, eCTDs, CMC documents) for compliance with standards (e.g. IDMP, xEVMPD) or master data management.

- **Identifying discrepancies in regulatory documents**: Finding errors in documents prior to submission, to reduce time for manual checking, including automated cross-checks for MedDRA adverse event coding consistency.

- **Response to regulatory questions**: Text analytics to capture and analyze Response to Questions for more effective data re-use.

Introduction

The pharmaceutical industry is among the most heavily regulated in the world. Text analytics speeds up and reduces the cost of regulatory compliance, as the huge volume of text-heavy documents means that manual efforts are often slow and expensive.

Linguamatics I2E provides a text-analytics solution that can be deployed to extract key data from unstructured text, find and highlight key information within regulatory documents, check for MedDRA coding, detect inconsistencies across documents, and more.

Recent and upcoming changes in regulation mean that companies require new tools and solutions to assist with regulatory review and compliance. In some cases, meeting the regulators’ requirements is straightforward, while in other cases, accessing the necessary data can take a significant amount of time, money, and effort, all of which increases costs, but does not necessarily increase revenue. The case studies below demonstrate the value that text analytics can bring for regulatory affairs.

Text analytics for compliance and master data management

IDMP (Identification of Medicinal Products) is a set of international standards developed by ISO that will become mandatory in Europe in a phased approach, fully effective from 2021; it is expected to be adopted by the FDA and globally over the next few years.

Capturing the hundreds of data attributes required per product, 70% of which lie in a variety of unstructured text sources, demands time, resource, and investment. But structuring this valuable data can benefit the
broader organization, enabling better data governance and master data management across discovery, development, clinical, and manufacturing.

Linguamatics’ I2E text-mining solution can save organizations time and money by rapidly finding, extracting, standardizing, and structuring the required data elements from IDMP-relevant unstructured text documents, including Summary of Product Characteristics (SmPC) documents (see Figure 1); manufacturing licenses; Chemistry, Manufacturing and Control (CMC) documents; and regulatory and compliance documents, e.g. electronic Common Technical Documents (eCTDs).

Figure 1: Example of text analytics from SmPC document for IDMP data elements. Top: I2E can extract structured results for key data elements, e.g. pharmaceutical dose form, name, strength, from a set of SmPCs. Bottom: Cached copy of the SmPC document, showing the highlighted mark-up for the extracted text around Levitra. Clicking on the “hit” mark-up in the tabular results takes the user directly to the correct place in the document, enabling rapid and efficient review.
The challenges of IDMP data capture from unstructured data containers are many, including:

- data extraction from internal and external documents;
- differing document types and formats;
- different styles from document authors;
- content that can be verbose or tabular;
- MedDRA coding needed for harmonization of adverse events or indications;
- different languages; and
- flexibility needed, as IDMP framework, timelines, and scope are still evolving.

**Benefits of text analytics for extraction of IDMP data elements**

A text-analytics approach brings multiple benefits over manual data extraction from unstructured text. Copying and pasting relevant data from documents into spreadsheets is intensive, repetitive, and tedious work, and is also prone to errors. Text mining uses business rules and standard vocabularies to systematically create a consistent, normalized set of product data, and can be used across tens, hundreds, or thousands of documents. Business rules can be rapidly translated into search queries, and this flexibility is key when the IDMP framework is still evolving. This approach can also be used for other reporting frameworks such as xEVMPD, or to provide data suitable for enterprise master data management.

**Identifying discrepancies in regulatory documents**

**Regulatory QA summary**

Quality control of regulatory documents before submission is an important step in the drug-regulation process. Consolidation of the various reports and documents into the overview document set required by the regulator necessitates significant volumes of manual checking and cross-checking, from the subsidiary documents to the master. The process is generally manual and, therefore, both slow and error-prone. Errors can result in applications being delayed.

Linguamatics has worked with our top 20 pharma customers to develop workflows to improve the quality control of regulatory documents. This can include cross-checking MedDRA coding, checking references to tables, highlighting format and calculation errors, and finding discrepancies between the summary document and source documents. These checks require the use of advanced processing to extract information from tables in PDF documents, as well as natural language processing to analyze the free text.

Using I2E to identify inconsistencies within submissions can save weeks of tedious manual checking and prevent a re-submission request, potentially saving months of time and millions of dollars.

**Workflows for error detection**

Typical workflows implemented include the following steps:

- Process PDF documents: This includes OCR to convert the PDFs to standard HTML or Word DOCX format; table processing to add metadata to tables within the documents; and indexing to create a corpus of documents ready for querying.

- I2E querying: Errors of various types are identified in the documents via business rules encoded in I2E queries (see Table 1).
Cross-checking MedDRA coding, for MedDRA labels, hierarchies, and identifiers (see Figure 2).

Post-processing, to generate a report dashboard describing the errors. Rendered versions of the documents can be highlighted to show the errors, with different error types colored distinctively (see Figure 3).

Email notification: Once complete, workflows can be created to send an alerting email to the registered user with a link to the results dashboard.

Table 1: Example error categories include searching for specific single terms, or comparing terms within a sentence, a table row, or a document, or across the entire document set.

<table>
<thead>
<tr>
<th>Error category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing source tables</td>
<td>Identify references in the summary document to source tables not appearing in the same document bundle</td>
</tr>
<tr>
<td>MedDRA label errors</td>
<td>Identify source tables that should contain MedDRA terms and then check to ensure that accurate MedDRA terms appear</td>
</tr>
<tr>
<td>Incorrect formatting</td>
<td>Identify cells in tables that contain values with inconsistent formatting, such as doubled period, incorrect number of decimal places, addition of percentage sign</td>
</tr>
<tr>
<td>Incorrect calculation or threshold</td>
<td>Identify cells in tables where the numeric value for the particular cell is incorrect, or where the table title threshold is not met</td>
</tr>
<tr>
<td>Inconsistent units</td>
<td>Identify cells in tables where the units are not appropriate for the measurement reported, e.g. haematocrit, haemoglobin level, platelet count, etc.</td>
</tr>
</tbody>
</table>

Figure 2: MedDRA coding consistency. I2E text analytics is used to check MedDRA coding consistency for regulatory documents, e.g. for clinical trial data. Checks can ensure that the MedDRA label is correct (e.g. “Nausea” rather than “feeling queasy”), that “Nausea” is correctly associated with “Gastrointestinal disorders,” and that the MedDRA ID is correct. This schematic shows the five MedDRA levels relevant to the adverse event “Nausea,” showing the lowest level term (synonym) of “feeling queasy” and the associated MedDRA identifiers for each MedDRA concept.
Figure 3: Sample table and text highlighting, to show inconsistencies between data. The example table (left), text (middle) and highlighting key (right) show the types of error found. The highlight color makes it easy for the reviewer to rapidly assess where there are errors, what type of errors they are, and then correct them appropriately.

### Table: Most Frequently Reported Medical Conditions (>5% in Any Treatment Group)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number Subjects</td>
<td>Rx N=897</td>
<td>Pbo N=927</td>
<td>Rx N=1021</td>
<td>Pbo N=956</td>
</tr>
<tr>
<td>Number (%) of Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiac disorders</td>
<td>70 (7.0)</td>
<td>32 (3.5)</td>
<td>108 (16.6)</td>
<td>101 (10.6)</td>
</tr>
<tr>
<td>Angina pectoris</td>
<td>4 (0.4)</td>
<td>5 (0.5)</td>
<td>74 (7.2)</td>
<td>71 (7.4)</td>
</tr>
<tr>
<td>Dyspepsia</td>
<td>174 (17.5)</td>
<td>120 (12.9)</td>
<td>3 (0.3)</td>
<td>2 (0.2)</td>
</tr>
<tr>
<td>GERD</td>
<td>83 (8.3)</td>
<td>52 (5.6)</td>
<td>30 (2.9)</td>
<td>27 (2.8)</td>
</tr>
<tr>
<td>Metabolic / nutritional disorders</td>
<td>253 (25.4)</td>
<td>165 (17.8)</td>
<td>194 (21.2)</td>
<td>19 (2.0)</td>
</tr>
<tr>
<td>Dyspnoea</td>
<td>1 (0.1)</td>
<td>0 (0)</td>
<td>15 (1.5)</td>
<td>19 (2.0)</td>
</tr>
<tr>
<td>Hypercholesterolaemia</td>
<td>65 (6.5)</td>
<td>50 (5.4)</td>
<td>88 (8.6)</td>
<td>103 (10.8)</td>
</tr>
<tr>
<td>Hyperlipidaemia</td>
<td>147 (14.7)</td>
<td>79 (8.5)</td>
<td>56 (5.5)</td>
<td>66 (6.9)</td>
</tr>
<tr>
<td>Osteoarthritis</td>
<td>102 (10.2)</td>
<td>67 (7)</td>
<td>12 (1.2)</td>
<td>11 (1.2)</td>
</tr>
<tr>
<td>Nervous system disorders</td>
<td>628 (63.0)</td>
<td>409 (44.1)</td>
<td>28 (2.7)</td>
<td>19 (2.0)</td>
</tr>
<tr>
<td>Headache</td>
<td>413 (41.4)</td>
<td>280 (29.7)</td>
<td>0 (0)</td>
<td>7 (0.7)</td>
</tr>
<tr>
<td>Psychiatric disorders</td>
<td>137 (13.7)</td>
<td>81 (8.7)</td>
<td>14 (1.4)</td>
<td>15 (1.6)</td>
</tr>
<tr>
<td>Insomnia</td>
<td>84 (8.4)</td>
<td>47 (5)</td>
<td>9 (0.9)</td>
<td>8 (0.8)</td>
</tr>
</tbody>
</table>

Commonly reported conditions included Seasonal allergies, Back pain, and Hypercholesterolaemia. The majority of AEs were considered treatment related in all cohorts and the relationship between treatment groups and between cohorts was similar to that observed for all-causality AEs. Permanent discontinuations were reported at higher rates in the Rx group than in the placebo groups in the 3 pooled cohorts. The majority of AEs leading to permanent discontinuation were considered treatment related in all treatment groups in all cohorts. The single most frequently reported event was headache, which was reported in approximately 40% of Rx subjects and 29% of placebo subjects in the 2000 Pooled cohort. Other AEs reported across all cohorts at rates greater in Rx subjects than placebo subjects included: Seasonal allergies and Insomnia (2000 8.4% vs 6.9%, 2003 0.9% vs 0.8%, 2006 14.0% vs 10.1%; Rx vs placebo respectively).

### Key
- **Incorrect formatting:** doubled period, incorrect number of decimal places, addition of percent sign
- **Incorrect calculation:** number of patients divided by total number does not agree with percent term
- **Incorrect threshold:** presence of row does not agree with table title

---

Response to regulatory questions

Responding to regulatory questions can be a challenge. Companies often have a goal to respond to questions within a certain time frame. Short turnaround cycles can lead to a messy submission with lots of appended information and orphaned responses. Capturing and analyzing the Response to Questions (RTQs) sent to regulatory agencies around the world enables pharmaceutical companies to gain insights on:

- frequently asked questions based on current new drug submissions;
- current regulatory questions and concerns;
- trends in regulatory concerns by product type (e.g. antibody-drug conjugates) and therapeutic area; and
- geographical pattern of regulatory concerns.

By mining past questions, product-development teams can anticipate future regulatory questions and concerns, and proactively address them in the initial submission, thereby shortening approval times. Linguamatics I2E is used to mine RTQs in order to answer complex questions such as: "Was a request
made for more information on the mechanisms of action of a product?” “Was product quality a concern?” “Were there questions around stability, clearance, model validation?”

Effective use of these RTQ responses by the product-development teams reduces the number of errors prior to submission, and allows the teams to anticipate what the different regulatory requirements are, and how that can influence current and future development.

Summary

Linguamatics I2E can find, highlight, and extract structured data elements from regulatory documents, which can provide rapid, systematic, repeatable analysis within regulatory applications. I2E can handle a broad variety of document formats and types, and provide agile querying and integration into enterprise workflows, enabling the flexibility needed to rapidly address critical business issues across regulatory affairs.

If you are interested in learning more about Linguamatics I2E and text analytics, please contact us at enquiries@linguamatics.com

© 2018 Linguamatics Ltd. The Linguamatics logo is a trademark of Linguamatics Ltd. All rights reserved. All other trademarks mentioned in this document are the property of their respective owners.