A BASE life science whitepaper
Published May 2020

Transforming regulated documents to structrured content: Exploring the capabilities of Artificial Intelligence

A life science challenge

Keeping regulatory documents updated and fully aligned across multiple health agencies throughout a product life cycle can be a headache posing challenges to many pharmaceutical companies. Often, documents are kept in a narrative format, which means that each document is uniquely composed and therefore not based on any reusable content. When health agencies review content, feedback is similarly provided in a narrative format. Consequently, pharmaceutical companies handle large bodies of text compiled by multiple authors, languages, and agencies in scope, requiring manual processing, thus adding a risk of human errors caused by sheer complexity.

These challenges are pushing companies to rethink their way of authoring regulatory documentation to improve quality, enhance writing efficiencies and increase accessibility to the content. An effective measure that can help achieve this, is by transforming narrative formats to structured content. Simplistically put, such transitioning allows for harmonization and standardization of contents submitted to Health Authorities (HAs), through content reusability across documents, regions, and medicinal products. The transitioning of narrative texts, which are typically crafted in word-based documents, to structured content in an XML-based format, is a process driven by HAs. As an example, the FDA implemented the structured product information called Structured Product Labelling (SPL). The main objectives for the implementation of the SPL specification was to ensure a uniform approach to the development of labelling content, thus providing various benefits for the industry, such as:

- Promoting the use of standard terminology and coding across labelling content
- Reducing manual efforts for reviewing and approving product information
- Reducing the amount of content redundancy
- Decreasing the risk of non-compliance by reducing incidences of product information inconsistency

However, implementation of structured content is not an easy task, and issues regarding impracticalities inevitably arise:

- In order to maximize the value of structured content, what is the right balance between reusable text components and what will still require manual updates?
- Where is the standardization of text a necessity before structured content is providing value?
- How to create and maintain documents from structured content?

The answer to the questions is simple. By enabling a new set of tools, Artificial Intelligence (AI) and Machine Learning (ML) can support you in the process of resolving the issues. Let us dive into how.

Turning sausages back into pigs

Artificial Intelligence and Natural Language Processing as tools for generating structured content from thousands of narrative documents.

The life science companies that are aiming to structure their labelling content often stumble upon a variety of change management challenges, including a few not-so-typical problems that require special tooling to resolve. Consider a standard labelling document hierarchy:

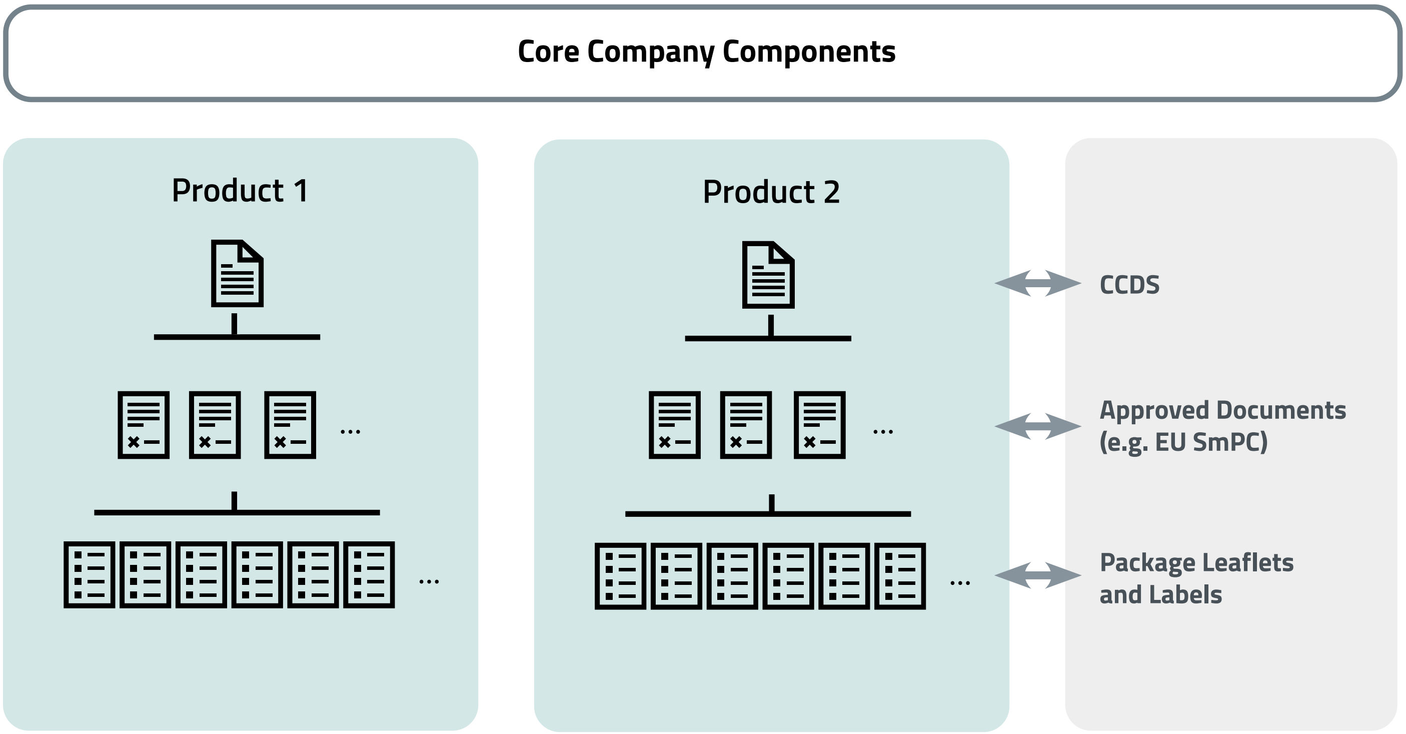

Core Company Components
Product 1
Product 2
CCDS
Approved Documents (e.g. EU SmPC)
Package Leaflets and Labels

Figure 1

The Companies Core Data Sheets (CCDS) and EU summary of product characteristics (SmPCs) Package Leaflets are documents that contain redundant information with a predetermined purpose for reuse. The CCDS dictate the content that should be included into the documents approved by regulatory authorities, which then determines the content on labels and packages. The central task in transforming a body of narrative documents to structured content is to determine the right chunks of reusable content both for 'horizontal' reuse between products, within or across product families, and 'vertical' reuse, tracing the information on labels and packages back through the approved documents to the CCDS. Such reusable information is gathered as core company components which can then be referenced when needed, however potentially with slight modifications. It is desirable to anchor as much information as high in the hierarchy as possible to achieve the best result.

About the Authors

Badr Fathi, Senior Principal Regulatory Lead, Switzerland

Badr is a customer-centric consultant focused on solving regulatory processes and IT systems challenges for pharmaceutical companies through advisory, assessment or implementation services. As Regulatory SME and Project Manager, Badr has a strong expertise in business processes and extensive experience within various of Regulatory and Clinical projects. Badr has 15+ years of experience within Regulatory and clinical areas.

☎ (+41) 76 439 26 29 ✉ bafa@baselifescience.com

The journey towards structured content

The BASE approach to transform regulated documents to structured content consists of five simple steps:

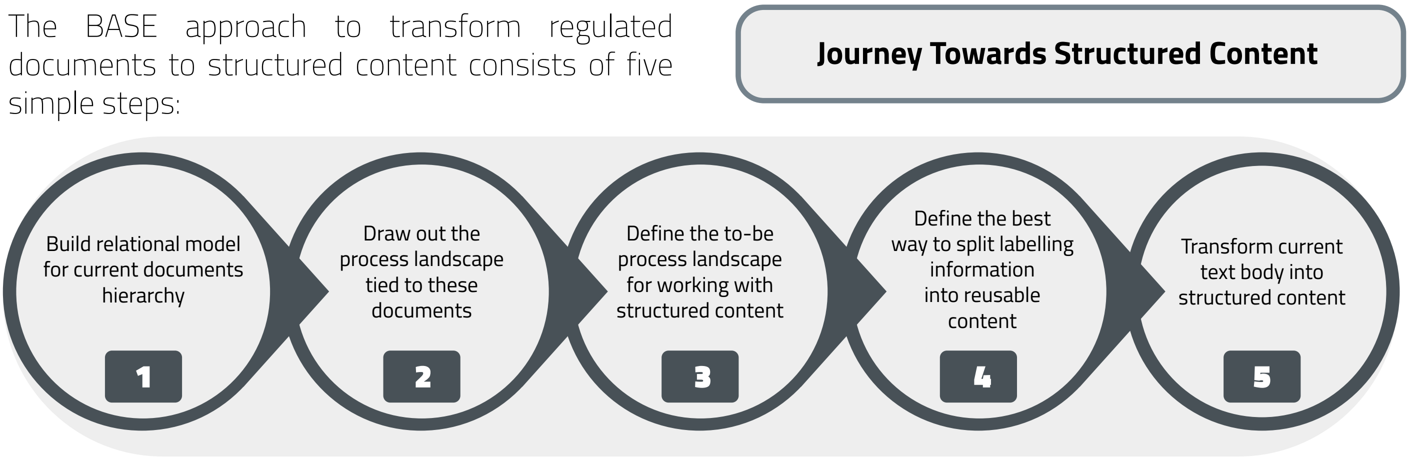

Journey Towards Structured Content
Build relational model for current documents hierarchy
1
Draw out the process landscape tied to these documents
2
Define the to-be process landscape for working with structured content
3
Define the best way to split labelling information into reusable content
4
Transform current text body into structured content
5

Figure 2

Step 1 to step 3 cover classical change management tasks, which can typically be lifted with the input of the stakeholders directly involved in the current processes.

Step 4 is where the journey towards structured content becomes difficult. While input and guidance can be generated from information management and textual analyses, it is typically difficult to get to the point of having specific and concrete break-down suggestions for the documents within a text body, or even a product for that matter. In practice, a common example being manual solutions revolving around hanging documents on the walls and using highlight markers. Defining a path to make actual operable suggestions for document break-down is crucial to optimize business processes, and we highly suggest a dual AI and natural language processing (NLP) approach, based on the following elements:

1. The entire body of labelling documents should be parsed, thus machine-read
2. Content should be broken into tunable information chunks, thus splitting different sections of different classes of documents based on, for instance, sections, sub-sections, newlines, full-stops, and bullets
3. Similarity between all chunks should be explicitly quantified, both horizontally and vertically, thus illustrating where identical and similar sections have been applied to the text body
4. Based on this, the chunk-sectioning strategy should be updated such that:
   - As many chunks as possible can be traced all the way from labels and packages and back to the CCDS
   - As many chunks as possible can be anchored in the core company components, with slight adjustments to the wording
   - The chunk size remains large enough to be operable

Having all text and information parsed and operable in a data science framework, we are provided with a unique opportunity to make significant headway with the fifth and final step. Step five can be conducted because the resulting chunks of parsed text can be directly used as output for structured content, generating everything from rough drafts to final versions of the XML text and map used in the structured labelling management system. The degree of completion of the output chunks typically varies between different areas of the total labelling text body, depending on the initial degree of text alignment.

Step 5: One technical standard for modelling structured content is DITA (Darwin Information Typing Architecture), which proposes a way to structure content as topics and maps. A topic is considered a self-contained chunk of content that can be reused, and a map is the composition of topics, or in other words, the architecture imposed on a set of topics. Topics are therefore context-free, and maps provide the methods for defining the order of topics and the relationship between topics. Even though it is not a DITA map, a table of content can similarly be considered a map, as it informs the reader of the expected sequence and structure of sections. Applied to the context of DITA, this would be equivalent to a map of topics. If sections contain subsections, this would be considered a collection of nested topics. The table of content, therefore, imposes an architecture of topics for the reader.

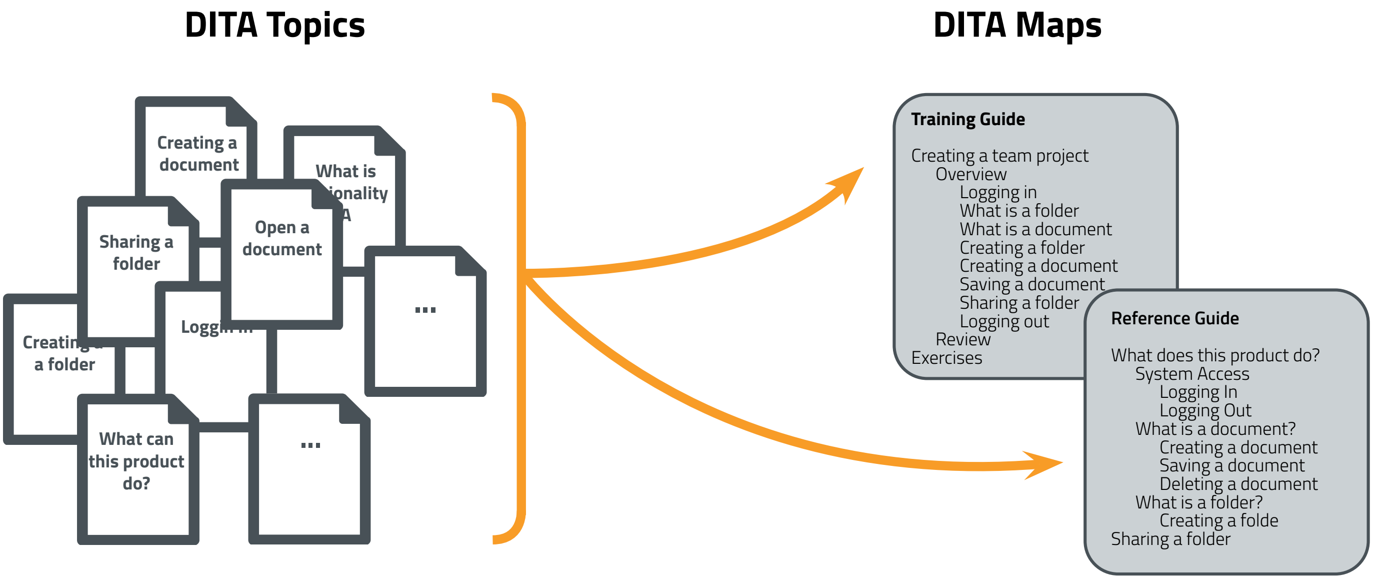


Figure 3

A powerful feature of maps and topics is the ability to refer to maps inside other maps. Additionally, this makes the maps flexible with topic keys and attributes that determine what content is related and not, so when we compile a document for a specific domain, only topics that have been marked for that domain will be compiled in the map. This feature is what enables us to define maps for documents at different business levels and for making maps applicable for different domains without having to keep unique documents for each, which is common practice today.

The process of figuring out the correct maps, the composition of maps, keys and attributes can be puzzling. The correct way to structure these and their granularity largely depends on business requirements. So where do we even begin?

Returning to step 4, we found our topics by identifying the desired chunk size that makes content meaningful, self-contained, and reusable. With the help of step 4, turning an entire body into reusable content becomes programmatically trivial. Chunks are compiled into topics and maps are created from the original documents table of content. By doing so we turn each document into a map of reusable topics. This, of course, is probably not the desired level to keep our maps, but by following respectively step 4 and step 5, we have managed to drastically reduce the amount of content to consider and now have a collection of reusable topics as well as the first stab at a collection of maps, which requires far less time to reduce into fewer more meaningful maps.

About the Authors

Thomas Røhme, Partner, Co-Head of BASE Analytics, Denmark

From his experience with large and complex corporate projects within Life Science, Thomas has accumalated extensive knowledge and competence. He is an expert at project initiation, change management, project management and strategy implementation. He has experience with implementation of a wide range of IT systems, including Workflow Management, Regulatory Information Management and Performance Management.

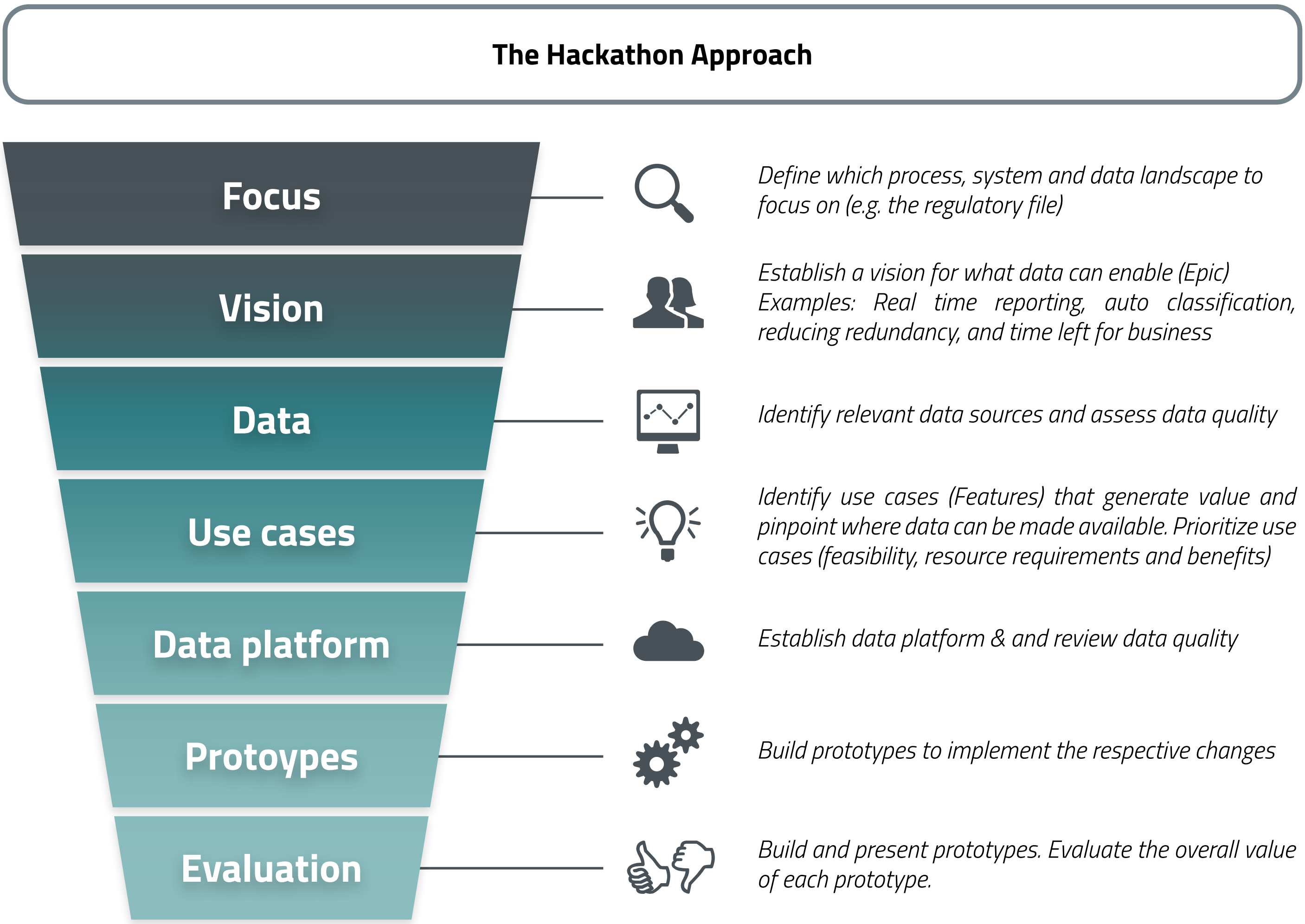
☎ (+45) 20 75 67 17 ✉ trhm@baselifescience.com

Hackathon approach at Novo Nordisk

This above-described approach was initiated at Novo Nordisk, who is in the process of acquiring a new tool for managing the labelling process. An important part of this is to make the whole authoring and review of documents more effective by transitioning from narrative texts to structured content.

BASE conducted a hackathon with Novo Nordisk Regulatory Affairs involving a technical team and Business Subject Matter Experts with the purpose of:

- Identifying use cases for using AI / NLP to understand, classify and predict labelling data
- Verifying business opportunities
- Achieving hands-on prototyping and development experience



The Hackathon Approach

Focus — *Define which process, system and data landscape to focus on (e.g. the regulatory file)*

Vision — *Establish a vision for what data can enable (Epic) Examples: Real time reporting, auto classification, reducing redundancy, and time left for business*

Data — *Identify relevant data sources and assess data quality*

Use cases — *Identify use cases (Features) that generate value and pinpoint where data can be made available. Prioritize use cases (feasibility, resource requirements and benefits)*

Data platform — *Establish data platform & and review data quality*

Protoypes — *Build prototypes to implement the respective changes*

Evaluation — *Build and present prototypes. Evaluate the overall value of each prototype.*

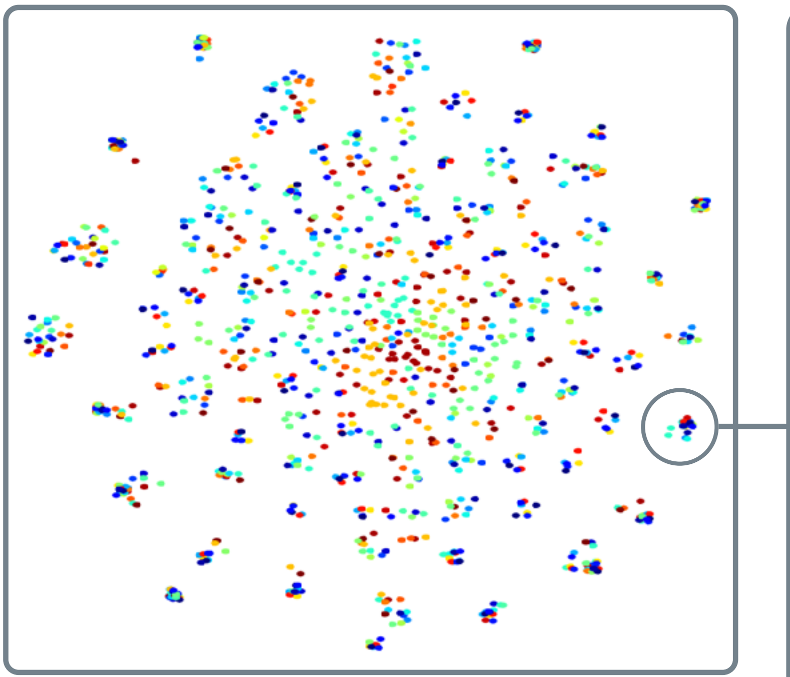
Figure 4

To conduct this Hackathon, we addressed two use cases: a reusability analysis and transformation to structured content.

For the first use case, the reusability analysis, the team developed a model for how CCDs data should be structured considering the right portion sizes to support processes and minimize workload, thus defining the potential for reuse by deploying the approach described above.

In conclusion, more than 50% of all sections proved to be identical or almost identical to at least one other section. Additionally, results showed that more than 80% of the analyzed sections were comprised by significant content of at least one other section.

Use Case 1 - Similarity Analysis



Insulatard - 1.2.1 - Origin of active substance(s) - optional, dist = 0.00
Origin of active substances(s) - optional
*Insulin human, rDNA (produced by recombinant DNA technology in Saccharomyces cerevisiae).

Fiasp - 1.2.1 - Origin of active substance - optional, dist = 0.36
Origin of active substance - optional
Insulin aspart is produced by recombinant DNA technology in Saccharomyces cerevisiae.

Mixtard - 1.2.1 - Origin of active substance(s) - optional, dist = 0.00
Origin of active substance(s) - optional
*Insulin human, rDNA (produced by recombinant DNA technology in Saccharomyces cerevisiae).

NovoRapid - 1.2.1 - Origin of active substance(s) - optional, dist = 0.36
Origin of active substance(s) - optional
Insulin aspart is produced by recombinant DNA technology in Saccharomyces cerevisiae.

Levemir - 1.2.1 - Origin of active substances(s) - optional, dist = 0.31
Origin of active substance(s) - optional
Insulin detemir is produced by recombinant DNA technology in Saccharomyces cerevisiae.

Actrapid - 1.2.1 - Origin of active substance(s) - optional, dist = 0.00
Origin of active substance(s) - optional
*Insulin human, rDNA (produced by recombinant DNA technology in Saccharomyces cerevisiae).

NovoMix - 1.2.1 - Origin of active substance(s) - optional, dist = 0.49
Origin of active substance(s) - optional
Insulin aspart produced by recombinant DNA technology in Saccharomyces cerevisiae
P1 Description and composition of the drug product
AaHv009.202

Observations

- Each dot represents a section in the NN CCD's
- The color specifies what product the section come from
- Distance between dot represents their similarity, such that a cluster of dots shows a collection of similar sections

Figure 5

This result indicates that the business case for implementing a solution to support structured content at Novo Nordisk is solid. Furthermore, it guided Novo Nordisk to where the focus should be in the initial text cleaning and transformation when preparing for the implementation of structured content.

For the second use case, the transformation to structured content, we were able to split the word document into sections using BASE parsing and compile each section into a dita topic. Based on the table of content, a ditamap was generated, using each content entry as a reference to the dita topic. By doing so, we were able to transform the selected CCDS into dita structured XML. From this, we found that turning regulatory documents into structured content and mapping becomes trivial once the desired chunk size has been found. This also enables fast prototyping and experimentation with real content, as we can start compiling ditatopics and ditamaps as soon as we have identified chunks and documents that share these chunks.

Use Case 2: From Word to structured contentusing BASE Artificial Intelligence

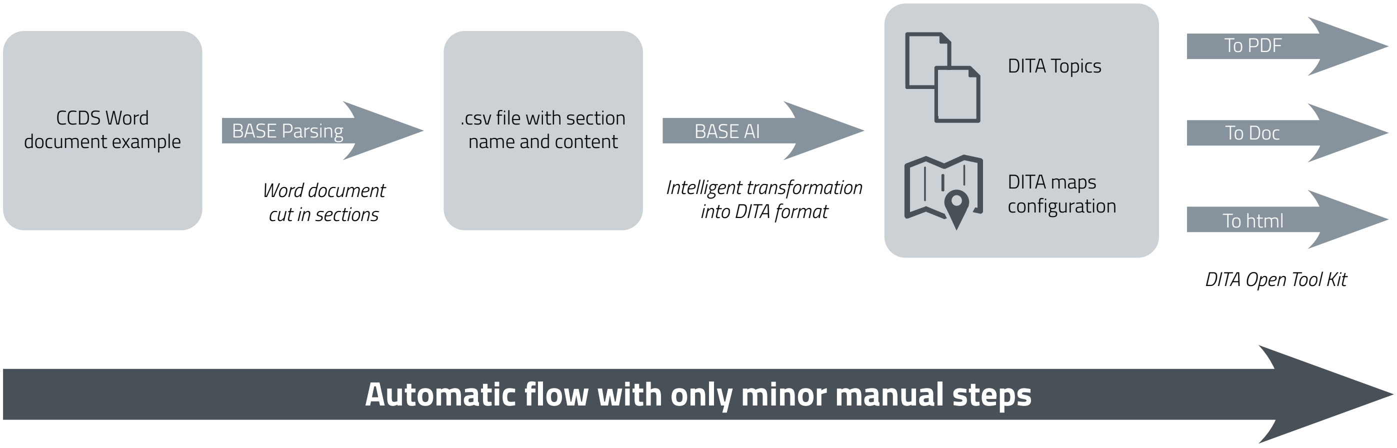

CCDS Word document example
BASE Parsing
Word document cut in sections
.csv file with section name and content
BASE AI
Intelligent transformation into DITA format
DITA Topics
DITA maps configuration
To PDF
To Doc
To html
DITA Open Tool Kit
Automatic flow with only minor manual steps

Figure 6

As take away

The key challenges of the implementation of structured content have been identified and can be summarized as follows:

- In order to maximize the value of structured content, what is the right balance between reusable text components and what will still require manual processing?
- Where is the standardization of text a necessity before structured content is providing value?
- How to create and maintain documents from structured content?

Natural Language Processing and Machine Learning have successfully been used to mitigate the challenges. Identifying areas where text can be reused, and where text can partially be reused, is an important step in order to create the right balance between reusable and not reusable text components. NLP and ML technologies enable fast lookup for "similar but not identical" pieces of text held against other documents, thus making them convenient tools for determination of where the standardization of text is a necessity before structured content can provide full value.

Finally, we successfully transformed regulatory writing documents to XML format, effectively making it possible to create and maintain documents based on reusable text components using ditamaps as templates. Through this method, documents can effectively be optimized simply by updating the relevant text component.
BASE would like to thank Novo Nordisk for good collaboration in an effective Hackathon that ended with promising results for Novo Nordisk's transition to structured content.

About BASE life science

BASE life science is a fast-growing, fast-paced consultancy focused on the life science industry. Established in 2007 and based in Copenhagen, Denmark, BASE targets a local as well as a global customer base.

At BASE, we are experts at helping life science companies create real business value from digital platforms and data within areas of Commercial Excellence, Clinical, Regulatory Affairs and Quality & Compliance. Since 2007, the company has been active globally from Denmark and Switzerland with more than 60 employees.

Contact Us

Jakob Winkler, Partner, Co-Head of Customer Engagement, Denmark

Jacob has 20+ years of experience within global project and program implementations in leading biotech and pharma industries. He has extensive business knowledge across the areas of R&D, Manufacturing, Safety, HR, Quality and Regulatory. Jacob is an expert within complex stakeholder management and communication, IT governance, project/program controls and vendor selection, and management.

☎ (+45) 53 73 70 34 ✉ jwin@baselifescience.com

Luca Morreale, Head of Operations, Switzerland

Luca has experience solving commercial and pricing challenges for life science companies through advisory, assessment or implementation services. Luca's approach is based on a pragmatic and result-driven approach, allowing him to lead teams in complex environments to achieve project goals. Luca is capable of providing in-depth insights, both at the strategy level as well as the operational level.

☎ (+41) 76 503 87 14 ✉ lumo@baselifescience.com