Transformation of Regulatory Content Management in Pharma

Rita Shah

Abstract: ***In pharmaceutical industry, managing regulatory content is challenging due to evolving regulations, growing demand of data and content, reusability of data or content across various functions, management of different versions of content and usage of right version across various functions. As a result, there is an impact on business objectives like quality and compliance, target time to submission, right first – time submission, and time to market. Hence, there is a need for the implementation of smarter ways to manage regulatory content in future. This paper describes smarter ways to manage content and enable faster and quality submissions for early to market a drug***

Keywords: Pharmaceutical Industry, Regulatory Content Management, Compliance, eCTD Submission, Drug Approval Process

1. Background

Data on safety and efficacy of medicinal products presented as content in a document and assembled in a dossier as part of electronic common technical document (eCTD) submission for drug approval process. To ensure the quality and compliance of regulatory dossiers, regulatory content needs to be current and compliant. Data that changes from time to time needs to be captured and reflected in various linked document versions for audit and inspection readiness.

2. Challenges

Patient centric needs, increased number of trials, and demand of additional data from regulators results in increased burden of content creation. The terabytes of R&D data **collected is analyzed, summarized,** and put into the right context. It is highly manual, error - prone, laborious, and time consuming. The fact is that **in the NDA** submission process, 20% of the time goes towards content generation and typical addressable spend on authoring content and document amounts between 50 and 110 million dollars. In addition, 60% content is reused across documents in value chain which is redundant, iterative, and non - value - added process.1Complexity is further added in keeping track of changing data and maintaining different versions for quality and compliance.

Typical challenges faced by organizations include:

a) **Inconsistent representation of data**: Non standardized way of data representation impacts quality of submission e. g., Indication and dosage.
b) **Recreation and reuse of content:** Time spent in recreation of content impacts submission timeline e. g., product description and safety caution.
c) **Static content –** Less flexibility to reuse the content of context. Content is copied and pasted from one document to another.
d) **Iterative review approval process** - Effort and time utilized in review and approval process post content reuse overall delays submission timeline.
e) **Content control:** Inconsistent usage of data between parent and child documents lead to non - compliance and audit findings**,** e. g., Protocol amendment versions not coherent with Informed Consent Document (ICDs).

Shift from current state:

Digitization and advancement of emerging technologies are drivers to transform today's highly unconnected, unstructured, static, and non - standardized content into connected, structured, flexible, and standardized content. To enable this transformation, both data and content needs to be standardized, structured, identifiable, and semantically aligned. Mapping and tagging of content and data in defined matrix can make content smart, intelligent, and agile to be used dynamically across functions and improve overall efficiency. Adoption of the following concepts enables content smarter, machine readable, easily searchable, adaptive to be used across.

3. Standard aligned data models

The data in the context of content needs to be smart and flexible to be used across life cycle. For that, it needs to be aligned to data model driven by standards. Standard aligned data models define the organization, relation and dependencies of various data domains and their elements. By understanding the relationships between different data elements in an integrated common data model, faster contextualization and seamless exchange of information becomes efficient. For example, Medicinal product domain is described by seven data elements (using IDMP standards), i. e., medicinal product name; ingredient substances; pharmaceutical product (route of administration, strength); marketing authorization; clinical particulars; packaging. These seven data elements are defined by attributes and interchanged dynamically by various functions. e. g., medicinal product name, dosage, indications can be referred to consistently in marketing, manufacturing, or packaging information in marketing application of a product. This enables flexible use of data and saves time by enabling filling of standard templates simultaneously.

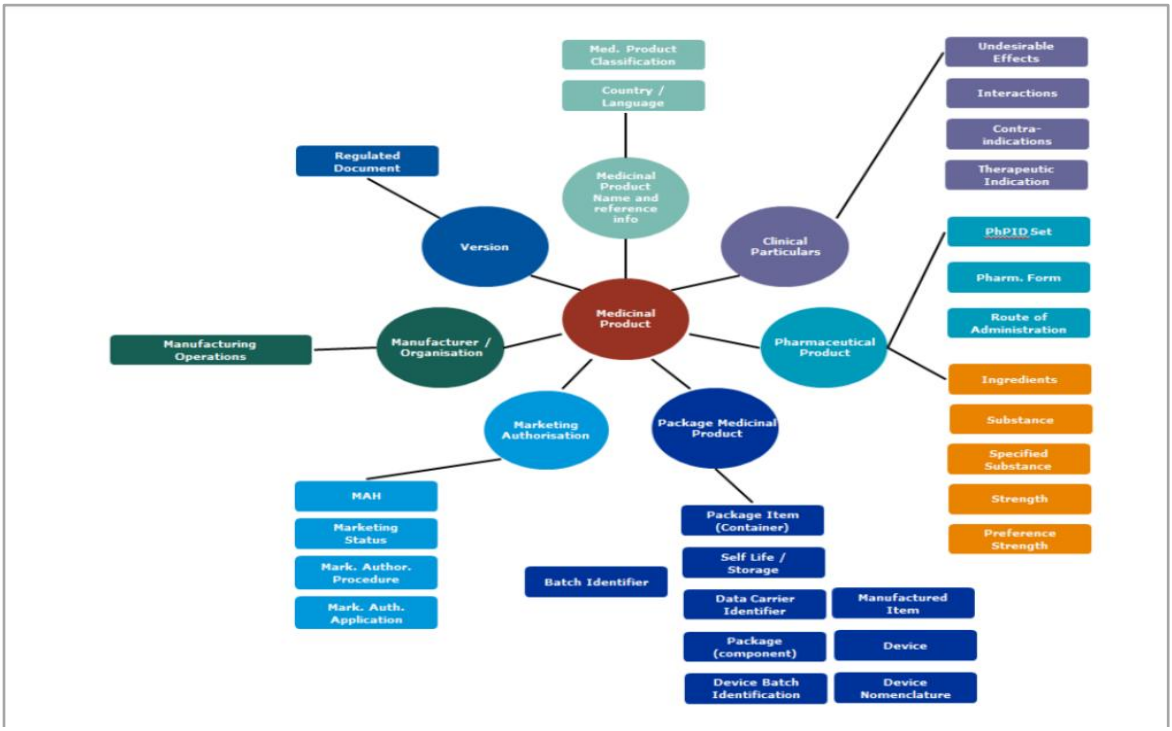

Figure 1: Overview of ISO IDMP data elements - for illustration purposes only)

(Reference - Introduction to ISO Identification of Medicinal Products (IDMP), SPOR program EMA/732656/2015)

1) **Taxonomies and metadata models**: The data attributes within data elements can be uniformly identified using governed taxonomy (classification of information). It will enable consistency, compliance and eliminate complexities of confusion. For example, the underlying taxonomy for the route of administration of pharmaceutical product has a structure specified under administration type which is further classified into local, systematic, and further classified into various components **figure1**. If oral route is selected, it will further map the term for oral ontology among (e. g., chewing, gargling, or dispersion, rinsing or washing, spraying, swallowing) to give more specific and appropriate terms. Tagging of content based on appropriate category of information provides rich metadata of content for its retrieval, access, identification, comparison, and analysis.

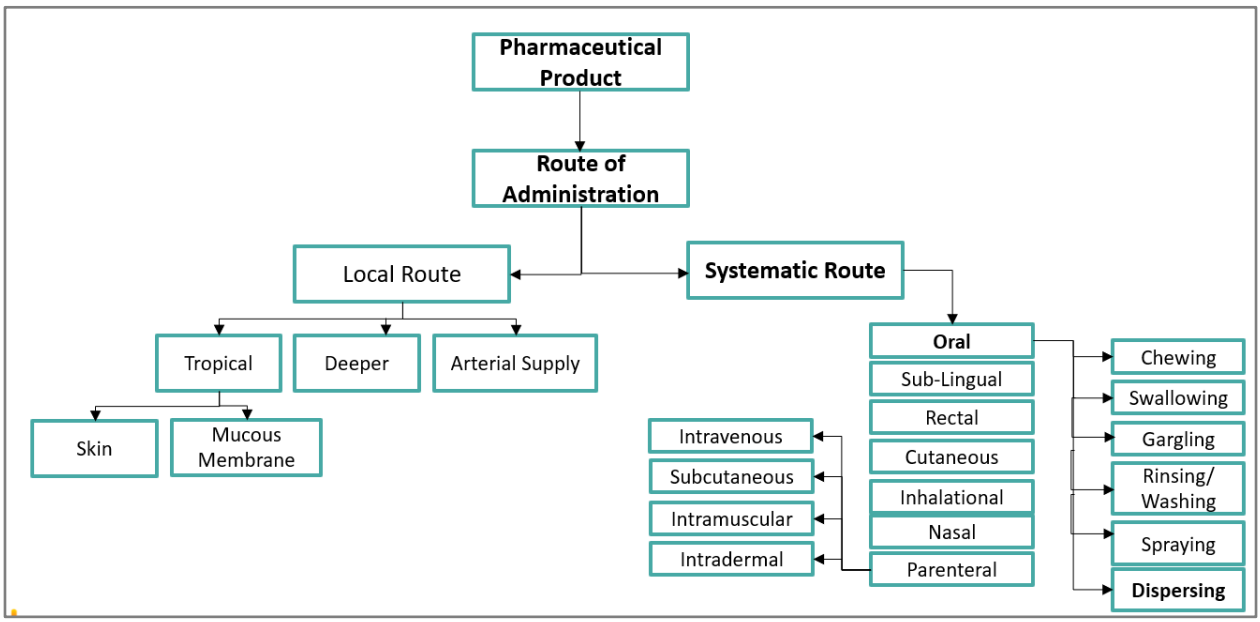

Figure 2: Use of Taxonomy & Ontologies

2) Ontologies and controlled vocabularies:
Taxonomies organize the information through proper classification system and ontologies organize the information based on classes, relation types, concepts, entities, and their specific attributes and help to bring the relationship between the information. Combination of taxonomy or ontology and linking them creates a semantic layer that facilitates machine learning to identify and relate

the terms. **For example**, linking the ontologies of adverse event and drug facilitates the knowledge about their relationship. The use of standard dictionaries and controlled vocabularies enforces standardization and enables consistency. It can be read by machines and intelligently self – corrected through unsupervised Learning. The use of referential and controlled vocabularies brings an elevated level of compliance and quality of data across product lifecycle.

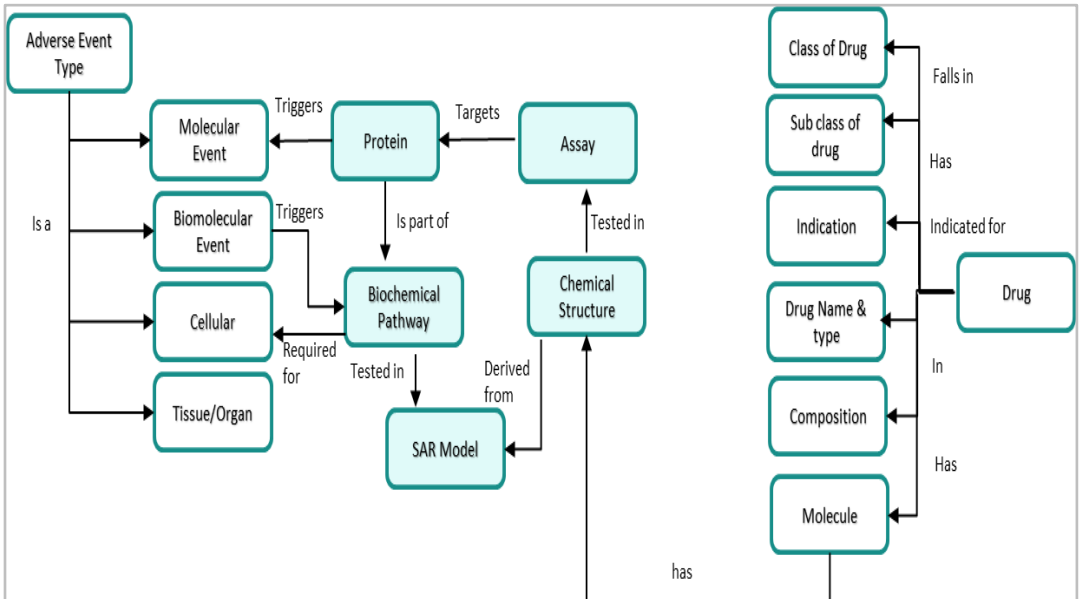

Figure 3: Use of Ontologies in understanding AE and Drug Relationship.

3) Knowledge graphs

A Knowledge graph can extract, identify, and store the entities such as product, indication, substance, country etc. from the text leveraging Named Entity Recognition (NER). It represents semantics by describing entities and their relationships and by leveraging taxonomy and ontologies it virtualizes and enhances data. Semantics in Knowledge graphs facilitate the use of AI/ML technologies to identify, learn and understand relations and enable faster contextualization. The benefit of such knowledge graphs with built - in relations serves connected data matrix that can integrate all the data from varied sources and group information at scale and provide 360 views of data. Defined syntax, structure and semantics are the approaches to present data in the context and connect. Regardless of the underlying sources of data, the solid foundation of data layer with semantics stored in knowledge graphs makes the data flexible and reusable across databases in product life cycle.

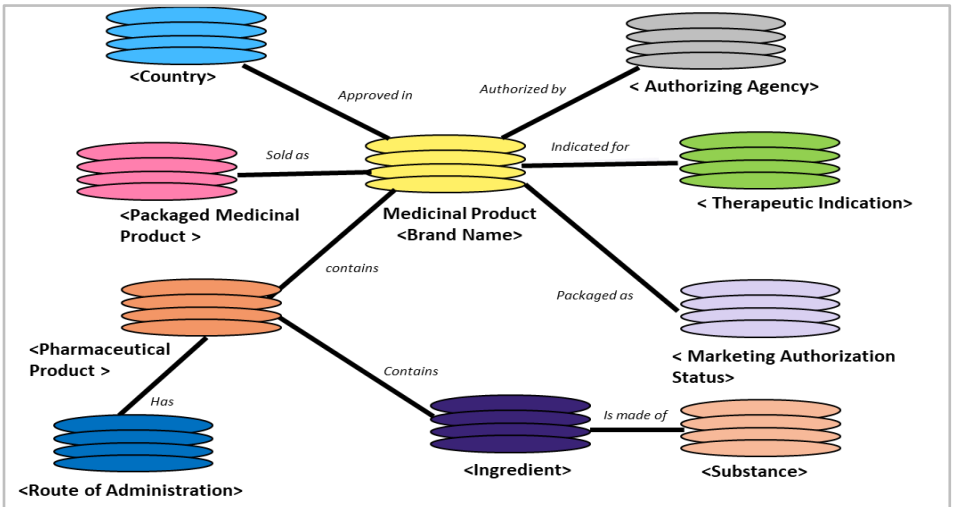

Figure 4: Knowledge Synthesis for faster contextualization

4) Information product concept

Data needs a structural output e. g., component in a document that can be dependent on mandatory content, substitutional data, and Neural Language Generation (NLG) driven. To analyze structural output to frame data, concept of **information product** leveraging ontologies and understanding and analyzing patterns of granular elements is needed.

For example**, in figure 3 -** Information Product has **a type "template" for a protocol**, which has **structure to capture structural details like** ID, title, no. of pages etc. It also captures the details of **organization** from creation and

authorization context and Reference details (no. **Page no.** etc.). Structural details invoke **Structural specifications** to give cumulative number of levels like sections and subsections. Further, it provides **structure type of individual section, its associated levels** and cumulative number of each section, sub sections, paragraphs, tables etc. Looking into granularity of **content** within sections to identify and capture **concepts** and entities and find their relationships using ontologies. For example – Information Product Type "Protocol" has subsection 4.1 "Overall Study Design" under section 4.0 with the heading as **"Study Design. "**The "study design" sub - section has attributes like study type, duration, phase, randomization, patient type structured in a content block that has recognizable style and pattern.

Knowledge graphs can link the information product type, content, and data, define relationships. Information Product Concept can help to define the data that comes into existence, it flows through various work centers, and is assembled to become an Information Product type (like Protocol, CSR etc.) of interest. This way faster content generation can reduce overall cycle time and enhance quality using approved components based on learned patterns.

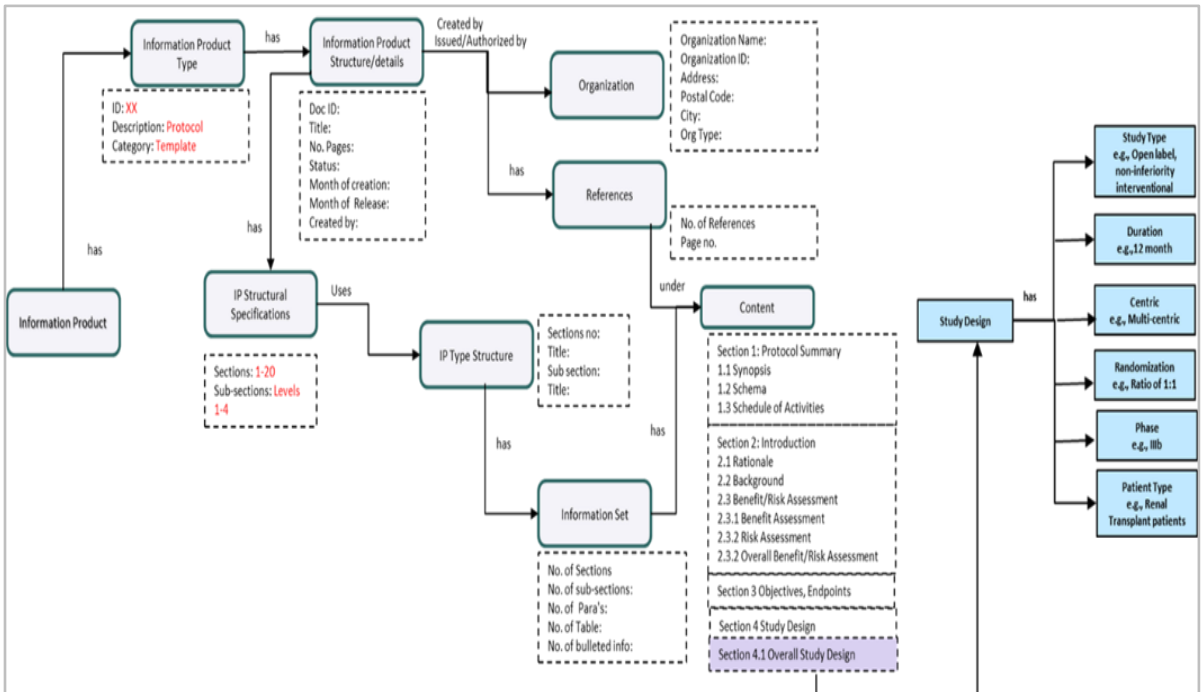

Figure 5: Illustration of Information Product Concept connecting data and Structure

5) Digital assistants and cognitive capabilities

New digital technologies can further facilitate the intelligence of content, for example:

- Patterns and style of content is recognized by technologies AI/ML
- Identification of tagged keywords in the prefilled template and auto fill the templates can leverage RPA (Robotic Process Automation)
- Auto contextualization of data can happen using NLG.
- Understand the structure and context of the sentences by analyzing place of subject, predicate, use of noun, verb, adjectives etc.
- Build xml structure of various sentences having generic syntax by extensible markup Language (xml) which provides the ability to separate content and presentation.
- Store the tagged key data elements (variables) xml in Knowledge graph and identify with metadata tags.
- Position group of sentences in a component and its structure and ontology in a document e. g., section of document
- Create formal structure of content and identification by type, composition including content and metadata element expandable to several document and publication types across the enterprise.
- Identify tables, data in and the relation, based on that can auto generate the content. (e. g., Table to Text) or vice versa conversion of Text to table or Figures. E. g., Patient disposition flow diagram in CSR or figures like response of treatment (in graph).
- Auto correction - standardization, personalization etc.

Automatic detection of specific content like privacy information, automated content creation and suggestions and automated summarizations using large language models **(LLM)** are another way to predict the content for the recognizable datasets.

6) Smart templates

Shift from "static" to 'smart' documents is the need of the day to promote faster and efficient reuse of content for submissions. By replacing the data items and values with intelligent tags, linked to master data/ connected data repositories, information can getup dated in real - time or near real time, hence high productivity, efficiency, data integrity and quality can be achieved. For example – In EMAs DADI project (Digital Application Dataset Integration), the emergence of web - based forms for submission can take advantage of such data layer, the data gets automatically filled like name, type of medicine, organization, application type etc.).

Benefits

- New standards (data and templates) can be accommodated into the scalable knowledge and once the data is updated, it will autonomously update the Information Product (IP) version with linked new data specifications hence enables compliance.
- Once the data gets stored in Knowledge with its properties, it is reused across product life cycle ensuring *consistency and saving time* of creating or copy pasting the data.
- New data changes can enable autonomous changes in respective linked documents when the latest version updates happen, hence *saves time* in managing versions and tracking changes.
- With well governance structure, data change cannot happen anywhere except in the single source of truth. It will ensure the data across value chain is maintaining *consistency*.
- Enhancing quality and compliance by enforcing standardized way of content creation
- Data integrity, cycle time reduction and cost burden by using data at source, less time in content creation, review cycles and reuse.

Implications to consider:

- Shift in the role of content authors to content architects and curators.
- More skill set requirement in Machine Learning & development algorithms to analyze patterns and auto curate.
- Change in Processes and SOPs
- Change in Business Operating Models like Regulatory Content as a Service offering.
- Investment of organizations in implementing Integrated Content Management Systems with automated utilities.

Organizations can achieve business outcomes by utilization of these concepts in a unified ecosystem like Authoring Platform. The organized and standardized, foundational component data layer enabled with ontologies and taxonomies, can generate content through semantic content hub. The content hub is digitally facilitated with process orchestration and presented in the right form and structure can improve overall cycle time reduction and quality submission and has the potential to increase rate of probability of regulatory success.

References

[1] Managing Enterprise Content: A Unified Content Strategy
[2] By Ann Rockley, Pamela Kostur, Steve Manning