Contents lists available at ScienceDirect

Journal of Pharmaceutical Sciences

journal homepage: www.jpharmsci.org

Review

The Future of CMC Regulatory Submissions: Streamlining Activities Using Structured Content and Data Management

Kabir Ahluwalia[a,b], Michael J. Abernathy[a], Jill Beierle[a], Nina S. Cauchon[a,*], David Cronin[c], Sheetal Gaiki[d], Andrew Lennard[e], Pradeep Mady[f], Mike McGorry[g], Kathleen Sugrue-Richards[h], Gang Xue[i]

[a] *Department of Global Regulatory Affairs − CMC, Amgen, Inc., One Amgen Center Drive, Thousand Oaks, CA 91320, USA*
[b] *University of Southern California, School of Pharmacy, 1985 Zonal Ave, Los Angeles, CA 90089, USA*
[c] *Cognition Corporation, 24 Hartwell Ave, Lexington, MA 02421, USA*
[d] *Biotherapeutic Development & Supply, Janssen Pharmaceuticals, 1000 Route 202 South, Raritan, NJ 08807, USA*
[e] *Department of Global Regulatory Affairs − CMC; Amgen Ltd, 1 Uxbridge Business Park, Sanderson Road, Uxbridge UB8 1DH, United Kingdom*
[f] *Product Quality Management, Janssen Pharmaceuticals, 1000 Route 202 South, Raritan, NJ 08807, USA*
[g] *Biotherapeutic Development & Supply, Janssen Pharmaceuticals, Barnahely, Ringaskiddy, Co.Cork, Ireland*
[h] *Department of Global Regulatory Affairs − CMC, Amgen Inc., 40 Technology Way West Greenwich, RI 02817, USA*
[i] *Biotherapeutic Development & Supply, Janssen Pharmaceuticals, 200 Great Valley Pkwy, Malvern, PA 10355, USA*

ARTICLE INFO

Article history:
Received 26 August 2021
Revised 29 September 2021
Accepted 29 September 2021
Available online xxx

Keywords:
Regulatory science
Automation
Standards
Artificial intelligence
Database(s)
Machine learning

ABSTRACT

Recent advancements in data engineering, data science, and secure cloud storage can transform the current state of global Chemistry, Manufacturing, and Controls (CMC) regulatory activities to automated online digital processes. Modernizing regulatory activities will facilitate simultaneous global submissions and concurrent collaborative reviews, significantly reducing global licensing timelines and variability in globally registered product details. This article describes advancements made within the pharmaceutical industry from theoretical concepts to utilization of structured content and data in CMC submissions. The term Structured Content and Data Management (SCDM) outlines the end-to-end scientific data lifecycle from capture in source systems, aggregation into a consolidated repository, and transformation into semantically structured blocks with metadata defining relationships between scientific data and business contexts. Automation of regulatory authoring (termed Structured Content Authoring) is feasible because SCDM makes data both human and machine readable. It will offer health authorities access to the digital data beyond the current standard of PDF documents and, for a review process, SCDM would "enrich the effectiveness, efficiency, and consistency of regulatory quality oversight" (Yu et al., 2019). SCDM is a novel solution for content and data management in regulatory submissions and can enable faster access to critical therapies worldwide.

Abbreviations: AI, Artificial Intelligence; ANDA, Abbreviated New Drug Application; CBER, Center for Biologics Evaluation and Research; CDER, Center for Drug Evaluation and Research; CDL, Consolidation Layer; CDS, Chromatography Data System; CMC, Chemistry, Manufacturing, and Controls; CMC-UDM, CMC specific Unified Data Model; CMO, Contract Manufacturing Organization; CPP, Critical Process Parameters; CQA, Critical Quality Attributes; CRO, Contract Research Organization; CTD, Common Technical Document (eCTD, electronic CTD); DMAP, Data Modernization Action Plan; DP, Drug Product; EC, Established Condition; EDL, Enterprise Data Lake; eLN, electronic Laboratory Notebooks; EMA, European Medicines Agency; ERP, Enterprise Resource Planning; ETL, Extract, Transform, Load; FAIR, findable, accessible, interoperable, and reusable; FDA, Food & Drug Administration; FHIR, Fast Healthcare Interoperability Resources; HL7, Health Level 7; HTS, High Throughput Screening; ICH, International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use; IDMP, Identification of Medicinal Products; ISO, International Organisation for Standardisation; IT, Information Technology; JoS, Justification of Specification; JSON, JavaScript Object Notation; KASA, Knowledge-Aided Assessment and Structured Application; LIMS, Laboratory Information Management System; MES, Manufacturing Execution System; OPQ, Office of Pharmaceutical Quality; PDF, Portable Document Format; PLCM, Product Lifecycle Management; PQ/CMC, Pharmaceutical Quality/CMC; PQRI, Product Quality Research Institute; PQS, Pharmaceutical Quality System; QOS, Quality Overall Summary SCA, Structured Content Authoring; SCDM, Structured Content and Data Management; SPQS, Structured Product Quality Submissions; TMAP, Technology Modernization Action Plan; XML, eXtensible Markup Language.

Declaration of Competing Interest: Kabir Ahluwalia (Summer Intern), Michael J. Abernathy, Jill Beierle, Nina S. Cauchon, Andrew Lennard, and Kathleen Sugrue-Richards reports financial support was provided by Amgen Inc. Sheetal Gaiki, Pradeep Mady, Mike McGorry, and Gang Xue reports financial support was provided by Janssen Pharmaceuticals Inc. David M. Cronin reports financial support was provided by Cognition Corporation.

[*] Corresponding author.
E-mail address: ncauchon@amgen.com (N.S. Cauchon).

https://doi.org/10.1016/j.xphs.2021.09.046

Introduction

Albert Einstein said that "the basis of all scientific work is the conviction that the world is an ordered and comprehensible entity."[1] The current regulatory submission and review of drug applications is currently not considered to be an ordered process as it requires significant manual and repetitive labor by both sponsors and health authorities which delays the speed at which novel therapeutics become available to patients. Pharmaceutical companies generate abundant volumes of data, content, and, ultimately, electronic or paper documentation for regulatory submissions involving clinical trial applications, new drug approvals, and post-approval lifecycle management activities. In turn, each health authority must receive, review, and respond to these submissions, initiating further document generation between a health authority and sponsor throughout the lifecycle of the product.[2] In the process of developing Chemistry, Manufacturing, and Controls (CMC) content for regulatory submissions, industry sponsors must incorporate CMC data expectations that are aligned to globally harmonized guidance and regional guidance into the regulatory strategy for each individual product. The CMC data are typically accessed and manually transferred into the dossier from various systems, e.g., LIMS (Laboratory Information Management System), electronic Laboratory Notebooks (eLN), batch records, certificates of analysis, or from a data lake or other repository. From there, data and content flow into technical reports prepared by the company, and then finally into the CMC sections of the regulatory filing dossier. Throughout this process, significant redundancy is incurred when developing the original submissions, which is then further adapted for different health authorities across the globe. Generally, the product-specific CMC information remains the same as a single global product progresses through a commercialization process. However, specific regional regulatory expectations add to the burden of information management.

Digitization (the process of converting physical data into a digital format), digitalization (the conversion of human-based processes to computer-operated processes), and automation are all expected to help with management of region-specific documents by having a greater focus on the data and less on the accompanying narrative.[3] By integrating electronic narrative, data, and visuals, into a "single pane of glass", and by providing this information in a readily usable format outside of a PDF (Portable Document Format), health authority reviewers and sponsors can draw conclusions in a more timely and efficient manner. In addition, providing this information in a virtual cloud-based solution will enable multiple health authorities to perform reviews collaboratively and in parallel across regions.[4] Developing strategies to automate and reduce redundant labor will allow sponsors and health authorities to focus on true risk-benefit analyses, ease the burden of regional adaptation of regulatory expectations, limit the "flavors" associated with variations in globally registered details, and accelerate approvals thus bringing new medicines to patients faster.

One strategy to increase efficiency and speed of drug development is to restructure a sponsor's current data management ecosystem and content authoring process to facilitate improved regulatory submission and review. Technological maturation has now turned what was once theoretical into practical data management with the advancements made to structured content management tools. This technology can enable individually authored and approved content in a format that is both human and machine readable.[5] Structured content is connected outside of a specific application or submission such that it can be reapplied to any interface. This structured format retains the multiple layers of the data, which permits the content to be readily used in multiple documents without the need to reauthor and reverify the information, as opposed to the current rigid and unstructured formats, e.g., PDF. Simply put, structured content is the individual building block of a document. The information block is fully defined and structured which enables the block to be readily available in an approved state for use or reuse to build another document. Structured Content and Data Management (SCDM) is an emerging field which is closely related to structured content management. Here, SCDM is defined as the integration of structured content with structured data and the management of those integrated components, currently specific to CMC activities that involve high volumes of data used to author CMC submissions. At the heart of SCDM is a core design principle which aims to shift a company's focus to managing data instead of managing documents.

The current unscalable method of manual tracking and tracing data must be addressed to allow the sponsor to pull reproducible and verified data for rapid regulatory strategic decision-making.[6] This transformative change is needed due to the growing size and complexity of company portfolios and the corresponding increase in applications managed by health authorities. The pharmaceutical industry is highly regulated and must abide by numerous, heterogeneous, and complex regulations across multiple regions. While new regulatory guidances are continually in development across health authority interest areas, the rate of change is often delayed as substantial evidence and data are needed to support deviation from standard processes already in place. As a result, pharmaceutical companies and health authorities are lagging in the implementation of technologies, such as SCDM, despite the advancement of novel technologies and analytical capabilities that support crucial innovation. While there has been significant discussion on the implementation of automation technologies over the last few years, there is now an acute need for global industry participation.[7−10]

This article provides considerations on transforming the pharmaceutical industry's submission process that could be realistically achieved within the next few years by utilizing intelligent automation and SCDM to support the standardization of information pertaining to the submission process in terms of integrated electronic narrative, data content and its analysis. This paper will address several topics of CMC data management and the development of SCDM. First, it will define the challenges currently facing the pharmaceutical industry regarding regulatory content authoring, responses to health authority information requests, and CMC data management. It then describes the regulatory developments related to health authority initiatives to standardize CMC data and automate the review process. This review of challenges and initiatives is not meant to be exhaustive, but to illustrate key aspects being addressed by industry and regulatory collaborators. This article will detail the current state of SCDM technology with use cases which are being collaboratively developed to provide solutions addressing the deficiencies in the current data collection and submission process. In addition, the benefits of SCDM and Structured Content Authoring (SCA), as well as near and long-term objectives under development, are highlighted, thereby shedding some light on what the future might hold relative to the ongoing efforts to transform the regulatory filing and review landscape.

Regulatory Challenges From an Industry Perspective

The pharmaceutical industry lags behind other sectors and even other highly regulated industries in their digital maturity.[11−13] The need to address challenges associated with the current regulatory environment is being discussed across the industry and by health authorities. Challenges in regulatory submissions, health authority initiatives for data standardization, select cases of how SCDM is being implemented internally in industry and challenges of SCDM implementation in CMC applications were previously discussed in a 2020 publication.[9] Recently in May 2021, another article was published as a review on digital innovation in regulatory submissions in the pharmaceutical industry.[14] The authors discussed several challenges and opportunities in the regulatory submission process centering on the theme that static, PDF-locked documents produce unique challenges for automation, dynamic access and review of data, and intelligent analysis of data. Taking advantage of data and content formats, which are not as restrictive as PDF formats, would digitally transform the pharmaceutical industry. This article expands on the previous discussions and promotes the use of technology to modernize current methods of coordinating and retrieving the required information from the immense volume of data produced and reviewed in regulatory decision-making. Specifically, there is a focus on the possibilities and developments in automating data for CMC sections of pharmaceutical regulatory submissions to health authorities. CMC data are amenable for automation and this article will propose rules and processes which can later be extended to other datasets such as pre-clinical and clinical data. Additionally, digitization, digitalization, and automation provide an opportunity and platform by which true global standardization of regulatory filing content and data can be realized. In the following sections, the structure and challenges of these regulatory submissions will be described from an industry perspective.

Brief Overview of Regulatory Submissions and Applicability of SCDM

The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) developed the Common Technical Document (CTD) describing the content required and where the information should be located in a regulatory application using a standardized format.[15] The CTD is organized in five modules and CMC information is placed into Module 3 with a summary in Module 2. In some regions this information is also transcribed and registered within Module 1 which contains regional-specific administrative information. While the harmonization of the CTD across the ICH-participating jurisdictions did improve the submission and review processes, the subsequent rapid acceleration of therapeutic development, introduction of novel modalities, and accumulation and maturation of data necessitates a further transformational change in the management and reporting of CMC data. The dynamics of the pharmaceutical industry cannot sustain the outdated, isolated, and static data retrieval methodologies utilized by both sponsor and health authorities in strategic decision-making and risk-based analysis.

The information contained within the CMC modules is highly repetitive and heavily data driven, and, therefore, represents a prime candidate for SCDM and automation. As shown in Figure 1, the submitted data are defined by the CTD guidelines and the CMC data in Module 3 are highly interconnected and dispersed into separate packages as needed. These separate submission packages account for region-specific requirements, revisions based on health authority information requests, new product presentations, and amendments throughout the product lifecycle. This requires significant tracking, knowledge, and effort to ensure accurate and updated information is submitted to each health authority in each separate package. The applicability of SCDM to CMC data is based on the feasibility to convert data into a structured format, which would further enable automated content authoring including reuse and traceability of submitted data.[16] Targeting Module 3 data and narratives will provide a standardized infrastructure enabling a sponsor to provide an accelerated, accurate data source to a health authority review division and automate the authoring process in general. To meet the current standards of drug applications outlined in the electronic CTD (eCTD), pharmaceutical companies have undergone data digitization while several health authorities have developed electronic submission portals. However, in the context of CMC data most companies and regulators are only in the beginning steps of digitalization which includes the development and implementation of SCDM and SCA.

To show the importance of digitalization in pharmaceutical submissions, the eCTD initially includes approximately 45 independent granules (sections of the eCTD) that can be used for ICH and regionally specific sections in Module 3. Dissecting these granules, there are typically an average of 5 to10 documents per granule, each of which are typically created via 3 authoring events, 3 review events, and 3 data verification events, totaling approximately 5000 internal sponsor events in building the core Module 3. This immensely complex and time-consuming process only constitutes the first CMC module sent to a health authority, following which regional customization results in other required variants for a product's global approval which can include over 80 individual health authorities. All the previously described manual processing, manipulation, and verification of data, in addition to global variation in registered details, is for a product that is essentially the same for all global markets. These authoring and verification efforts continue throughout the product lifecycle, which for many products can extend to 20 years on the market, which further highlights the need for digitalization and global standardization. Despite the need for digitization as a solution for streamlining data collection and regulatory authoring workflows, several prominent regulatory challenges preclude broad implementation in the pharmaceutical industry and are discussed further below.

Challenges in Regulatory Content Authoring

Figure 1 displays the multiple CMC submission packages that industry sponsors need to generate and update throughout the lifecycle of a product. This content requires multiple rounds of review and data verification for each authoring event to confirm accuracy, consistency or interpretation, and sufficient content for regulators prior to its internal approval and submission. In addition, significant resources are required to reauthor or repurpose the same section to meet the requirements or preferences for individual countries or regions. This results in multiple variations of the same section containing slightly different content because of differences in regulatory expectation, negotiations including specific requests, and post-approval commitments for each region. Maintenance of these multiple variations is a challenge in lifecycle management and in identifying and tracking the precise documentation submitted to each health authority.

To add to the complexity, filings across regions are often staggered, sometimes by years, for a multitude of reasons. During the time between submissions to different health authorities, lifecycle changes are frequently implemented in the regions where the marketing authorization has been approved while an original application is being prepared for a follow-on market. To prepare for the follow-on market, it is critical to have access to the most current information including statistical assessments, which requires knowing what is approved by region and having traceability back to what was revised, why, and for which jurisdiction or health authority. It is imperative to be able to trace which information was included in the approved submission to manage the lifecycle content properly for each region. The complexity of tracking this for every CTD section in every country is laborious, manual, repetitive, and can lead to misalignment of information if the correct approved document for the specific country is not used as the basis of a variation. Another layer of intricacy is added when partnering with contract organizations as they may utilize different internal document or data systems that feed into each CTD section. Even when leveraging a health authority's initiative such as the Food & Drug Administration's (FDA) Project Orbis, which aims to leverage a collaborative review format, variation in registered details still occur because of regional health authority preferences.

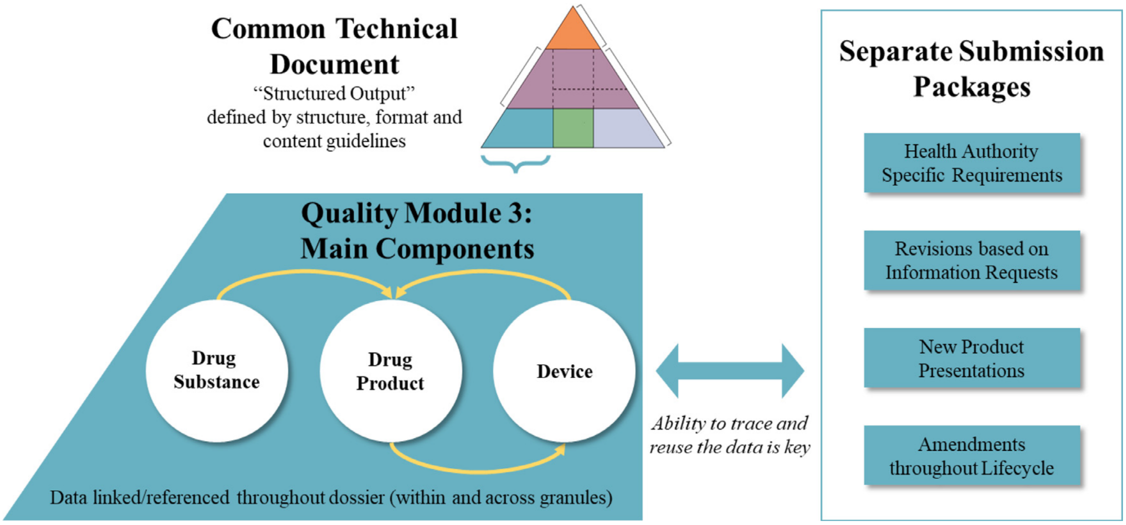

Common Technical Document
“Structured Output”
defined by structure, format and content guidelines
Quality Module 3: Main Components
Drug Substance
Drug Product
Device
Data linked/referenced throughout dossier (within and across granules)
Ability to trace and reuse the data is key
Separate Submission Packages
Health Authority Specific Requirements
Revisions based on Information Requests
New Product Presentations
Amendments throughout Lifecycle

Figure 1. Current CMC Data Management and Content Authoring Process.

Challenges in Responses to Questions from Health Authorities

Additional complexity occurs during the process of gaining health authority approvals with information requests presenting additional challenges for both sponsors and health authorities. Following an initial submission, responses to a health authority's information requests pose a challenge because the turnaround time for responses varies by region and can, in many cases, be as little as a few days. In addition, regulatory dossier approval timelines often depend on the response times being met. Agency response times require that sponsors are diligent in efficiently utilizing and prioritizing resources to compile responses, verify data, review, format, publish, and submit responses in a timely manner. Module 3 holds large datasets and the PDF format of the eCTD granules lack any automated traceability back to the original data sources which makes new analysis, data retrieval, or updates to Module 3 that require manual data input a time-consuming process. As discussed, there are redundant efforts involved in developing documents for submissions to different health authorities. In turn, regulators each have their own review methodologies and may ask the same or similar questions, creating repetitive efforts on the side of both sponsors and reviewers. Ultimately, this extends the timeline for global product approvals. Lastly, the static PDF format of the CTD is not efficient for data mining and information exchange between the sponsor and reviewers because the raw data embedded within is not readily accessible or deconstructable for further analyses. Regulatory reviewers have to manually deconstruct CMC narratives and data from these rigid PDF files in order to populate information into their own data systems and assessment reports, which involves additional cycles of copying and pasting data into other software platforms.

Overall, the ability of sponsors to maintain efficient and user-friendly knowledge management systems is degraded by the presence of multiple archives of documents throughout many storage systems and repositories. This creates a complex array of submissions per region and further complicates subsequent change controls involving agency information requests and updates to Module 3, as well as tracking individual CMC commitments with each health authority. Throughout the product lifecycle, the discussions and evaluations between sponsor and health authority demand fast access to data that is findable, accessible, interoperable, and reusable (FAIR). However, industry document management standards have not been universally adopted across companies because substantial resources and a cultural shift in the industry are required for their development and implementation.[17]

Challenges in CMC Data Management

The abundance of data contained within the thousands of pages being authored for regulatory submissions is heavily reliant on narratives written by a large team of subject matter experts. There can be many different authors, reviewers, and data verifiers per product submission which can lead to more subjectivity with narrative-based submissions, resulting in inconsistency even from product to product within a company. A further complication arises from the interrelationships between many of the granules within a module of the CTD and even across modules. These interrelationships can result in different authors using the same data with the potential for misalignment in interpretation between sections.

In addition to narratives, sponsors must generate data tables, process schematics, statistical plots and visuals. These are often manually created and populated from a source document into Module 3 or ‘copy-pasted’ from an electronic source document. Given the vast amount of data provided in Module 3, this is a resource-dense activity for both the author and the data verifier. If data change during lifecycle management, it is difficult to trace the correct data to the submitted and approved version of the CTD. Furthermore, the time lapse between regional filings often results in additional independent statistical analyses based on changing data availability, which can potentially impact sponsor conclusions and health authorities' assessments leading to further filing detail variations. Additionally, the heterogeneity of data systems in use across departments or functions within a company has made it challenging to implement more efficient document creation capabilities. Some important drivers of data variability include source data origin (internal to the sponsor or from external partners), data gathered in an unstructured format, regional inconsistencies (spelling, significant figures, formats), product modalities, and number of manufacturing sites and batches. The lack of consistent structure in the data results in significant resources placed in manual data transformation, aggregation, verification, and error correction.

While there have been improvements in the management of data, it is currently insufficient to incrementally advance the authoring process. Some companies have begun compiling raw data into “data lakes”, which allows different systems to store their data in a single repository. However, many of these datasets are not easily accessible as it remains unstructured with minimal standardization. Unstructured data create additional challenges for creation and timing of submissions as well as tracking the data throughout the product lifecycle. Achieving the ultimate goal of submitting a single, harmonized global filing with parallel reviews would be challenging because it relies on the implementation of data standardization and automated processes by sponsors and regulators. The lack of digitalization towards Structured Content Authoring (SCA) puts unnecessary burdens on both industry and regulators, delaying regulatory approvals and the rate at which novel therapies can be delivered to patients. In the following sections, the efforts being made by health authorities towards improving efficiencies in their review processes are described.

Current Regulatory Landscape

Recently, industry and health authorities have begun to discuss and develop an intelligent automation-based approach to modernize the regulatory submission and review process. This modernization includes content authoring automation in parallel with data automation which will merge submission narratives with data and images from internal disparate sources to drive submissions in a more efficient and accurate manner. Additionally, health authorities are involved in digital transformation approaches to increase information technology infrastructure and security for data management to meet the needs of their research modernization efforts.[18] Below, major developments from the FDA, European Medicines Agency (EMA), and ICH initiatives are described.

FDA Initiatives

The FDA initiated several workshops and action plans to contribute to a SCDM-based approach for data collection and regulatory review of CMC content and data. In 2018, the FDA established the Knowledge-Aided Assessment and Structured Application (KASA) initiative, headed by the Office of Pharmaceutical Quality (OPQ).[8] The KASA initiative was created in response to increasing numbers of Abbreviated New Drug Application (ANDA) submissions and included a public FDA meeting in September 2018 and a discussion at a 2019 Product Quality Research Institute (PQRI)/FDA Conference.[8,19–21] Structured assessment templates and risk-ranking algorithms have been developed to support the large volume of applications as part of the KASA initiative. In 2020, the OPQ developed, tested and began using structured data interfaces for drug substance information and liquid dosage forms.[22] In 2021, the KASA initiative progressed further as the OPQ proceeded to evolve from text-based to increasingly data-based quality reviews in developing the tool.[22,23] Although the initial focus was on generics because of the large number of ANDAs, the use is now being expanded to new drugs and biologics applications.[24]

The FDA released the Pharmaceutical Quality/CMC (PQ/CMC) initiative on how to structure data with eXtensible Markup Language (XML) and Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) data formats.[25] PQ/CMC and KASA work synergistically; while KASA intends to change how submissions are evaluated by the FDA, PQ/CMC aims to transform presentation of CMC data in submissions by harnessing a structured data approach. To support KASA, PQ/CMC provides structure by organizing data and standardization of terms to feed into structured applications. A PQ/CMC draft guidance is proposed for 2022 and will likely be limited to data-driven sections such as Specification, Batch Analysis and Stability.[25] While these initiatives are specific to the FDA currently, there is increasing discussion of global harmonization.

While older systems were once sufficient for enabling regulatory review, the current era of technological innovation has demonstrated that significant improvements can be made to drive efficiency. The FDA has acknowledged the need for newer technologies to meet increasing demand for a modern infrastructure by piloting a variety of supportive initiatives. It recently announced two modernization plans for data technology transformation. In 2019, the FDA provided the Technology Modernization Action Plan (TMAP) for technology modernization in the agency's strategy for data and management of data. Additionally, the Center for Biologics Evaluation and Research (CBER) is taking the first step to integrating the CBER's Information Technology (IT) and data platforms by aligning with the FDA's TMAP.[26,27] Following the TMAP, in 2020, the Data Modernization Action Plan (DMAP) provided a strategy concentrating on newer approaches, IT, and the process of using data to accelerate pathways to better therapeutics.[28] The DMAP focuses on identifying solutions and allowing the reviewer and sponsor to discuss the critical scientific rationale and development of capabilities rather than focusing on collecting data first and then looking for questions the data can answer.

In April 2021, the FDA Center for Drug Evaluation and Research (CDER) published an article on Industry 4.0 for pharmaceutical manufacturing. Industry 4.0 refers to the fourth industrial revolution which combines advanced technologies including the Internet of Things, Artificial Intelligence (AI), robotics, and advanced computing with the aim of revolutionizing the manufacturing landscape to create autonomous and self-organizing systems which require little human involvement.[7] However, the FDA acknowledges that the current stage of pharmaceutical manufacturing is primarily, aligned with Industry 2.0, which involves manufacturing with pre-determined and static settings.[29] Reaching the level of digital maturity necessitated by Industry 4.0 will require the integration of multiple data sources to allow for process controls to be connected to process performance. As well, Industry 4.0 will identify control points to ensure quality in drug development for both industry and health authorities. SCDM is among the key enabling technologies which puts the industry on the path to Industry 4.0.

EMA Initiatives

The EMA and Heads of Medicines Agencies provide strategic direction in five-year strategy documents. The EMA Network Strategy to 2025 identified six strategic focus areas including data analytics, digital tools and digital transformation.[30] The goals of this strategy include accessing and analysis of healthcare and clinical trial data, building EU network capabilities to analyze big data, promoting dynamic regulation and policy learning, leveraging bot technology, and ensuring data are managed and analyzed securely. Importantly, the EMA intends to engage with the FDA and other regulatory agencies for global alignment.

The EMA has also developed an electronic format for product information in adopting the ISO IDMP (International Organisation for Standardisation, Identification of Medicinal Products).[31] Like the FDA's FHIR standards, this EMA initiative is also based on HL7 for reliable exchange of product information by using a common language for product name, ingredients, and pharmaceutical form for example. This aligns with the FDA's PQ/CMC initiative as previously described. The purpose of the ISO IDMP and the digital technology for implementation have many similarities to SCDM and, therefore, the IDMP is suitable for future expansion of scope to CMC. Indeed, the scope is envisioned to soon be expanded for use in GxP inspections. Both the IDMP and SCDM are designed for improved, accessible data exchange with assured data integrity that can be reused for varied purposes; they allow for streamlined, simplified, efficient regulatory filing and faster regulatory decision-making.[32] The data may also be exchanged between regulators in addition to between the applicant and agency. The data elements of the IDMP can bring in source information from across the CTD and thereby has the potential to encompass all relevant CMC information.

In 2017, the EMA held a workshop on the use of prior knowledge.[14,33] In 2018, at a joint EMA/FDA workshop, the application of prior knowledge to accelerate development of products in expedited review procedures, such as Breakthrough Designation and PRIME, was discussed and led to an EMA draft guideline.[34,35] In 2020, the EMA also implemented a new regulatory submission portal (IRIS) that is being piloted with selected types of submission such as Scientific Advice and Orphan Designation.[36] IRIS includes an electronic, cloud-based data lake repository of information intended to facilitate information exchange between the EMA and applicant. Combining the IDMP and IRIS has the potential for a platform to support SCDM, sponsor submissions, and health authority reviews. SCDM holds exciting potential for the use of prior knowledge by pulling together cross-product information from a single data lake repository. Furthermore, the use of prior knowledge requires justification in using data that could also be appropriately structured within the data lake and by applying AI, it should be possible to filter prior knowledge according to predetermined criteria that can learn differences and similarities between products relative to a desired application.

ICH Efforts

In 2014, the ICH Steering Committee endorsed the development of a guideline to address post-approval CMC challenges. The 2019 finalized guideline, ICH Q12: *"Technical and Regulatory Considerations for Pharmaceutical Product Lifecycle Management"* provides a framework for an improved use of quality data and a critical step towards harmonization and standardization of post approval changes.[37] The core of the endorsed guideline establishes a variety of related documents to be included with the original application that describes planned changes to the manufacturing process with anticipation that these lifecycle management plans simplify the review process for predicted post-approval changes. This provides an excellent opportunity for the utilization of SCDM, which can efficiently facilitate and synergize with the regulatory tools described in ICH Q12. For example, ICH Q12 defines Established Conditions (ECs) as the elements in an application which would require a regulatory submission if changed post-approval. With digitalization, binary decisions can be made relative to what is and what is not an EC, and through automation details related to the specific determined ECs can be autopopulated into a standardized and agreed upon format. Updates to ECs are then captured within the Product Lifecycle Management (PLCM) document which is updated as needed throughout the product lifecycle. SCDM can use the ECs as rules to flag and automate content authoring and submission of updates to the PLCM document which would typically be a hands-on and time-consuming process. Thus, the proposed harmonized approach to managing manufacturing and analytical procedure changes in ICH Q12 will encourage unification of post-approval submission requirements across regions and enable more effective solutions, such as SCDM, to provide streamlined access to higher-quality products by avoiding redundancies in manufacturing and testing.

The ICH Assembly recently endorsed a proposal on revision of *"The Common Technical Document For The Registration Of Pharmaceuticals For Human Use: Quality – M4Q(R1),"* and a proposal for a new guideline on Structured Product Quality Submissions (SPQS).[38] The ICH 2020 Annual Report conveyed the revision of ICH M4Q(R1) as aiming to reorganize the application in a manner that would be compatible with current quality assessment platforms in use in various regulatory agencies and would facilitate a more highly structured and standardized assessment as compared to the current narrative approach to filing submissions and regulatory review.[39] The SPQS document identifies sections containing CMC information and data that can be internationally standardized into a structured data format in order to create a common set of data elements, vocabularies, and an electronic exchange format for those sections of Module 3. Overall, the initiatives would minimize variability and duplication in global data. This harmonization will also work well in parallel with SCDM to improve the efficiency of collection, analysis, reporting, and review of CMC data.

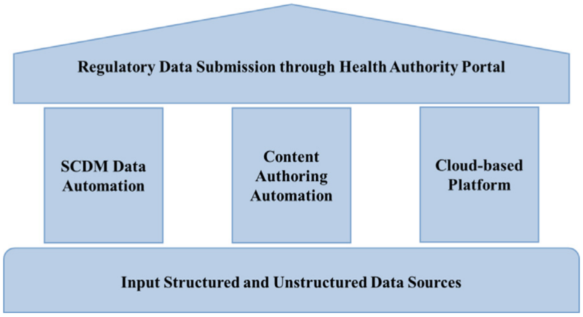

Regulatory Data Submission through Health Authority Portal
SCDM Data Automation
Content Authoring Automation
Cloud-based Platform
Input Structured and Unstructured Data Sources

Figure 2. Three pillars for a Structured Content and Data Management based Knowledge Management Process.

Structured Content and Data Management and Authoring Based on a CMC Unified Data Model (CMC-UDM)

Though the simplest of solutions would be to build a single CMC submission that is reviewed and approved by a single global health authority, this scenario is highly unlikely because of geopolitical and economic limitations. However, with the recent advancements in SCDM technology and solutions, a single virtual submission housed within a cloud-based ecosystem is a real possibility. By leveraging SCDM and SCA in an integrated fashion, the industry will likely see a shift from rigid filing formats to a usable and much more efficient data exchange platform.

Three key pillars in using data in a knowledge-based management system for a submission process and agency decision-making are: 1) data automation using SCDM, 2) electronic regulatory component or content authoring automation and 3) leveraging standardized data from a structured cloud-based platform. The relationship between these three pillars is illustrated in Figure 2. Implementing SCDM and SCA will reduce manual and redundant labor throughout the product lifecycle. It avoids potential replication errors when manually copying data to new submissions and allows automation of regulatory content authoring. Additionally, SCDM has the potential to enable improved use of AI technologies, such as machine learning, to analyze the collection of data as part of the company's Pharmaceutical Quality System (PQS) and to provide further insights for development strategies, thereby improving decision-making.[40] Lastly, SCDM in combination with cloud technologies will enable the ability to construct regulatory filings concurrently, submit filings simultaneously to multiple health authorities, allow collaboration between agencies, and simplify the information requests from each health authority. The following sections will describe technical aspects of a CMC specific Unified Data Model (CMC-UDM), SCDM implementation, and examples of SCA.

The CMC-UDM and Semantics

A key element of SCDM and SCA is to establish the structure that would be consistently used to capture, store, and consume data. In computer science, this structure is referred to as a UDM.[41] An example for a CMC-UDM that would support product specification and the analytical data of a pharmaceutical drug product is shown in Fig. 3. The core of the analytical data resides in the "Measurement".

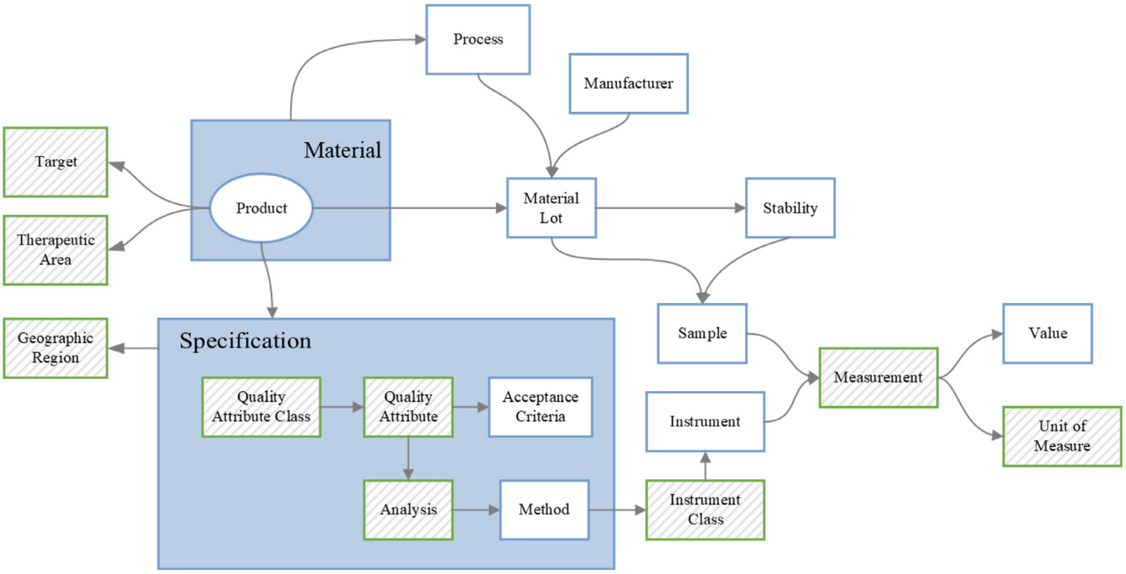


Figure 3. CMC-Unified Data Model for Product Specification and Analytical Data. Each green box with diagonal lines represents a controlled vocabulary specific to a taxonomy domain; each blue outline or blue shaded box represents a structured data container such as registries for Product, Material, Equipment, Protocol, Results, etc.

However, to enable the accurate and automated content authoring from the data, complete and consistent contextual information (metadata) needs to be captured and logically linked to the “Measurement”. For example, and in reference to the Module 3 Batch Analysis sections (S.4.4 and P.5.4 for drug substance and drug product respectively), the Product (e.g., molecule, manufacturing stage, dosage form, etc.) and its specific Material Lots must be selected to filter down to the corresponding batch release testing measurement results. In addition, to support the conclusion of pass or fail for a batch release, the batch should be compared against the Country/Geographic Region and its individual Specification for the Product. Defining the data and metadata, as well as their semantic relationship (the basic meaning and interrelationship between structured data) in a way that is unambiguously understandable by humans and reasoned by machines is crucial to automate the data query and content authoring. In addition, controlled vocabularies need to be established for each of the data elements such as Geographic Region, Quality Attribute Class, Quality Attribute, Analysis, Measurement (parameter) as indicated with green boxes with diagonal lines in Figure 3. In the computer-science domain, the controlled vocabulary and its hierarchical structure is named taxonomy, while the relationships between the various taxonomies are defined as ontology.[42] The taxonomy and ontology collectively define the UDM accurately such that computers are able to reason and support various data queries based on the data consumption and authoring needs. To achieve effective SCDM and SCA, the CMC data domains, or collection of values that a data element may contain within the CMC-UDM, must be defined.

In comparison to Batch Analyses, the Stability Data sections (S.7.3 and P.8.3) require similar data elements but also need additional metadata related to stability studies such as stability time points, storage conditions, etc. Instead of requiring a separate data model specifically for stability, the CMC-UDM that supports the Batch Analysis data (typically also including ‘time zero’ values in the stability protocol), needs to be extended, as shown in Figure 3. The same data elements as well as semantic relationships between Product, Process, Material Lot, Manufacturing, Specification, Instrument Class, Instrument, Sample and Measurement can all be reused by the Stability section data query to automatically build the stability lot summary as well as the stability data tables.

To enable SCDM and SCA for the entire Module 3, the CMC-UDM needs to be extended to cover all the various data elements that feed into these CMC sections. As the internationally harmonized standard for regulatory filing, the CTD provides the complete content guideline to be extracted and semantically organized to the CMC data model. To properly illustrate the entire scope of Module 3, the data model is elevated to the data domain level as shown in the Figure 4 ‘honeycombs’ scheme. Each hexagon represents a data domain, which is semantically connected to the neighboring hexagon. For example, the Material Lot in the center is a physical instance of a Product and produced by a Process. It then gets aliquoted for release testing Measurement, Stability Study, or Product Characterization. Through these connections, the Material Lot is further related to a Target and Specification via Product, and Raw Material and Manufacturer via Process, as shown in the detailed data model in Figure 4. Compared to the traditional rigid schema-based relational database, the knowledge graph, which uses graph-structured data models, perfectly supports the extensibility of the semantic data model with great flexibility.

A closer examination of the CTD sections would allow the construction and extension of the data domains. For example, the product specification sections (S.4.1, P.5.1) clearly span across Product, Country/Region and Specification, while the Stability sections (S.7.1, S.7.3, P.8.1, P.8.3) contain data from Material Lot, Stability Study, Specification, and Measurement. On the other hand, the contents for most data domains are used by many different sections such as Measurement which feeds into Characterization (S.3.1, S.3.2), Justification of Specification (S.4.5, P.5.6), Batch Analysis (S.4.4, P.5.4), and Stability (S.7.1, S.7.3, S.8.1, P.8.3), while Product definition and attribute data are required by Nomenclature (S.1.1), Structure (S.1.2), General Properties (S.1.3), Formulation (P.2.1, P.2.2), Specification (S.4.1, P.5.1), Analytical Procedure (S.4.2, P.5.2), and many more. For this reason, even though Module 3 is complicated and data rich, the semantic data domains required to provide complete coverage of Module 3 are manageable in quantity with careful semantic engineering. These same principles can be applied to any product entities including combination products with devices.

Establishing an authoritative single data source for each logical data domain not only avoids the repeated manual transcription of the same data, but also allows for a one-time only data integrity verification of the input source information. While individual companies can build their own CMC-UDM, it would be more efficient and impactful if the model could be harmonized across the pharmaceutical industry in partnership with global harmonization efforts such as

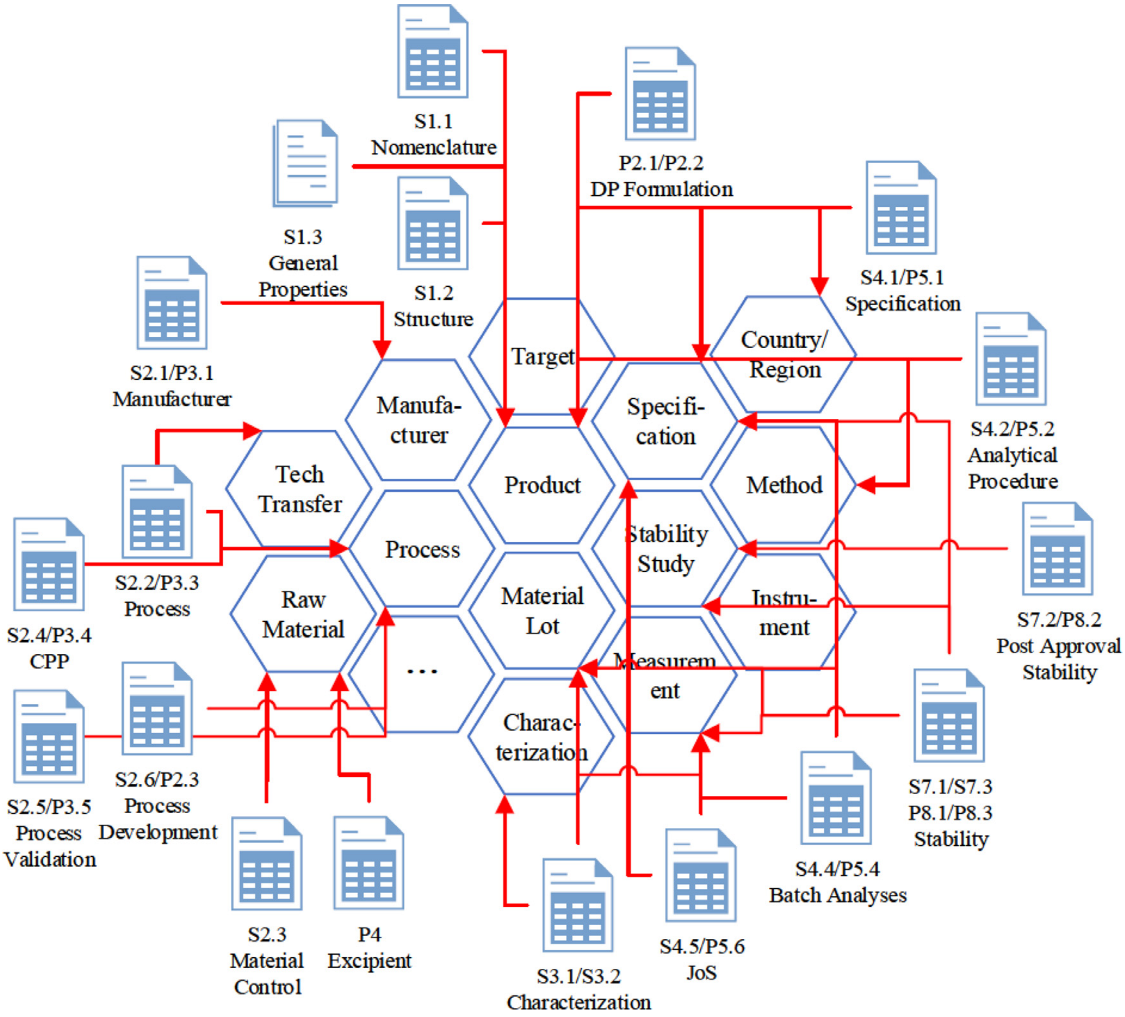


Figure 4. CMC Data Domains from CTD Module 3. Abbreviations: CPP, Critical Process Parameters; DP, Drug Product; JoS, Justification of Specification.

those within the ICH. Not only would a harmonized CMC-UDM allow the collective expertise to build a more robust data model for SCA, but it would also allow the submission of a digital data package for the statistical assessment, data trending, and analytical comparison by a health authority.

Structured Content and Data Management (SCDM)

Once the CMC-UDM is established, SCDM becomes not only possible but optimal. When feasible, the semantic data model and associated controlled vocabularies should be built into the laboratory informatics ecosystems such as Compound Registry, eLN, and LIMS. The scientific data could be captured in a FAIR manner with consistent and complete metadata from the beginning. Structuring data in this manner enables the ability to pull verified data and visuals directly into authoring templates. Although, it is recognized that in the highly regulated pharmaceutical industry it would typically be expected to take years before such a sweeping change could be fully implemented, global agencies are phasing in the standardization of key drug product terms through the ISO IDMP, which is a step in the right direction. In the near future, especially considering the decades of legacy data, the Enterprise Data Lake (EDL) approach, including multi-source aggregation and data pedigree, would be pragmatic in implementation of SCDM and SCA tools.

Multi-Source Data Aggregation

As shown in Figure 5, the first step of SCDM is data aggregation via the established Extract, Transform, Load (ETL) process into the EDL. Most pharmaceutical companies have established such consolidated data warehouses either on the premises or in a private cloud environment. Typically, the data stored in the EDL are consolidated from many disparate data sources and therefore follow the same data model as the data sources. For this reason, it is often referred to as the Consolidation Layer (CDL). Due to the heterogeneity of the data sources (typically over a dozen), the data in the CDL vary significantly in structure and even nomenclature. Duplication and sometimes conflicting data are also common, which makes it difficult to assemble and parse the data to support data analytics and automated report authoring.

With the CMC-UDM, the aggregated data from the CDL can be re-engineered and the data elements segregated into the various semantically connected data domains. In a way, the CDL resembles the agricultural harvest process where produce is collected from various farms into warehouses after which a semantic engineering tool would sort the produce into the different supermarket departments such as vegetables, fruits, seafood, and meat. For example, the Enterprise Resource Planning (ERP) systems are commonly used to manage the input and output materials for clinical batches. After the ERP data go through the ETL process into the consolidation layer, they are parsed into the Raw Material, Material, Process, and Manufacturer

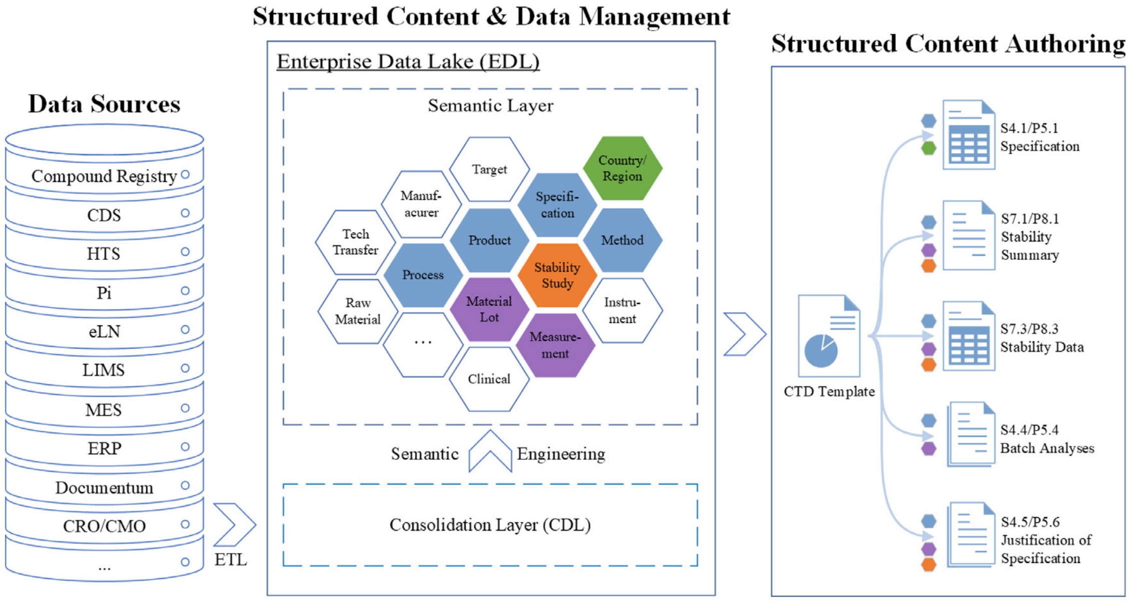


Figure 5. Structured Content Authoring of Module 3 via a CMC-Unified Data Model (CMC-UDM) Enabled by Structured Content and Data Management (SCDM). Abbreviations: CDS, Chromatography Data System; CMO, Contract Manufacturing Organization; CRO, Contract Research Organization; eLN, electronic Laboratory Notebooks; ERP, Enterprise Resource Planning; ETL, Extract, Transform, Load process; HTS, High Throughput Screening; LIMS, Laboratory Information Management System; MES, Manufacturing Execution System.

data domains. Here the Process domain would keep all the data domains connected semantically, providing accurate lot genealogy and linkage to the process master batch record. The semantic engineering process would also detect any duplication or conflicting data elements and prompt for immediate remediation, some in the form of automated reconciliation by validated rules while others may require human intervention. As a result, only one verified true copy of the same data element would remain in the semantic layer, serving as the authoritative source for data analytics and SCA.

Data Pedigree

CMC data not only increase over time but also evolve and mature throughout the product lifecycle thus increasing the complexity of data management. For this reason, one of the key elements of SCDM is to manage the traceability (pedigree) throughout the data lifecycle from the initial data capture to transformation, storage, and consumption, for both compliance and change management reasons. The traceability must support bi-directional genealogy queries. Within the authored filing sections, each data element should clearly point to the data element in the Semantic Layer as well as the raw data captured in the original source system. In case there is any change in the source system, the SCDM and SCA need to provide a summary of all impacted downstream data blocks, reports, and filing sections to allow updates and version control. While the bi-directional query is feasible for SCDM because it has a complete chain of custody of the data, proper breadcrumbs or data trails need to be built into the system to allow the efficient pedigree search in both directions.

Structured Content Authoring (SCA)

Since the CMC-UDM is built by reverse engineering the eCTD contents, the SCA can be performed from the SCDM by connecting the various data elements in the standardized filing templates to the corresponding data domain in the Semantic Layer. As shown in Figure 5, the color-coded data elements in the EDL are automatically authored into CTD sections based on the data pedigree defined by the section templates. Product, Country/Region, and Specification get automatically fed into the Specification sections (S.4.1, P.5.1). Additional metadata from data domains, such as Process which is semantically related to Product and Method which in turn are related to Specification, are available to be populated in the Specification sections (S.4.1, P.5.1) when needed. Similarly, Product, Process, Method, Material Lot, Measurement and Specification may autopopulate the Batch Analysis section data tables (S.4.4, P.5.4), while all these data domains plus the Stability Study data domain populate the Stability sections (S.7.1, S.7.3, P.8.1, S.8.3). One distinct illustration is that the SCDM maintains a single true copy of a data block for overlapped contents such as Product, Process, Specification, etc. The verified data blocks become the common sources for many Module 3 sections, enabled by SCA templates with clear data pedigree for straightforward compliance and change control management.

Figure 6 further illustrates the end-to-end SCDM and SCA lifecycle, which starts from the multi-source aggregation of scientific data (eLN, LIMS, ERP, etc.) and semantic transformation, then autopopulates the eCTD and technical report templates with optional data loaded into the SCA. The autogenerated reports get published for compliance review and approval in the cloud before being released for submission. Finally, any added content or interpretation results can be consumed back into the SCDM layer. The SCA (step 3 in Fig. 6) requires the digitization of the eCTD or technical report template which semantically links each data column and data selection filter to the SCDM data elements defined by the CMC-UDM as shown in the Specification and Stability examples below.

Specifications

Specifications are set and monitored based on analysis of Critical Quality Attributes (CQAs) across the life of the product, starting with the Target Product Profile prior to the initiation of First in Human clinical trials. Specifications need to be clinically relevant, adequately justified (which may include leveraging prior or platform knowledge), and supported by continued analysis. Specifications can vary from region to region based on a multitude of factors such as regulatory requirements (e.g., peptide mapping is required for specifications filed in Japan but not the US or EU), and negotiations during review (e.g., commitments and responses to information requests). In the meanwhile, the specifications can change over the life of a

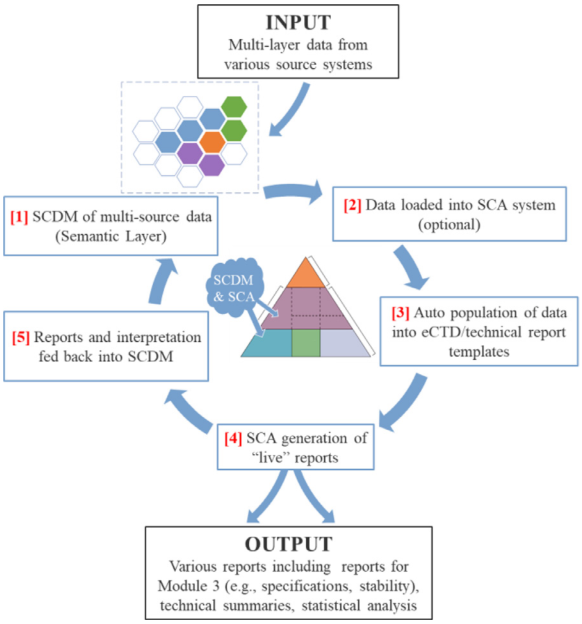

INPUT
Multi-layer data from various source systems
[1] SCDM of multi-source data (Semantic Layer)
[2] Data loaded into SCA system (optional)
SCDM & SCA
[3] Auto population of data into eCTD/technical report templates
[5] Reports and interpretation fed back into SCDM
[4] SCA generation of "live" reports
OUTPUT
Various reports including reports for Module 3 (e.g., specifications, stability), technical summaries, statistical analysis

Figure 6. The Structured Content Authoring (SCA) Process Cycle. Abbreviations: eCTD, electronic Common Technical Document; SCDM, Structured Content and Data Management.

product for various reasons including new testing technology or 'validating a test out' (e.g., sunsetting a test because a CQA no longer needs to be monitored). Specification acceptance criteria are referenced by multiple CTD sections within Module 3 including Specifications, Stability, Batch Analyses, and Justification of Specifications, as well as other CTD modules (e.g., Module 2 Quality Overall Summary [QOS]). This makes it challenging to track the individual specifications approved by each health authority and difficult to lifecycle all relevant sections when there is a change that affects specifications.

For these reasons, the SCA Specification template is designed in three digital layers: the top layer links to the Product and Process data domain, which defines the metadata for the product specification including process and specification versions; the middle layer relates to the Country/Geographic Region and may contain multiple entities, each defining the regional attribute for one specification section; and the bottom layer is the core of the specification and may have one or multiple sections, each of which is a data table containing Quality Attribute Class, Analysis, Method, Quality Attribute, and Acceptance Criteria as discrete data columns based on the CMC-UDM (Fig. 3). During authoring, the user would select the Product and Process via the available dropdowns populated from the SCDM data query results. The SCA would use the filtered Country/Regions to autogenerate one specification section for each Country/Region. The Specification data contents then get populated into each specification section with the product/process/country/region specific Quality Attributes and Acceptance Criteria. With accurate metadata in the SCDM, the country/region variations could all be automatically reported with minimal user input. When there is an update to all or part of the product specification, the SCA could be automatically alerted of the impacted sections and prompted for version updates of either the entire specification section or individual sections, leveraging the SCDM and SCA data pedigree. This creates traceability of the lifecycle changes to the product specification and other parameters, including full audit trails and justification for the changes.

Batch Analyses, Stability, and Justification of Specifications

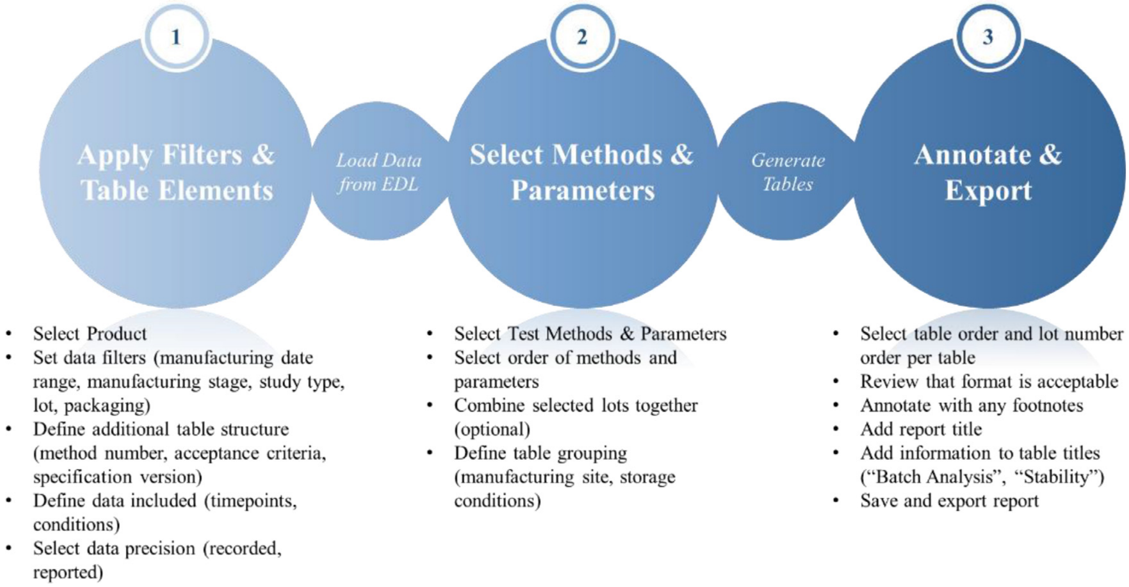
Batch Analysis (S.4.5, P.5.4) and Stability (S.7.1, S.7.3, P.8.1, P.8.3) are all data-rich sections with reasonably consistent tabular structures from ICH guidelines and minimal narrative. They also need to be continuously updated as new data are obtained throughout the product lifecycle. Additionally, the Justification of Specifications sections (S.4.5, P.5.6) largely reuse Batch Analyses and Stability data. Therefore, these sections were selected as the next target for SCDM and SCA automation. Similar to Specification, the Batch Analysis and Stability authoring follows the workflow in Figure 6, where the Process and Material Lot metadata from the ERP systems and the Stability Study and Measurement data from the eLN and LIMS systems are aggregated and semantically transformed using SCDM based on the CMC-UDM as shown in Figure 3. Typically, these data are sourced across multiple data systems which makes data compilation and verification time consuming, especially with the trend to outsource release and stability testing to contract organizations. However, with the data consistently stored in a structured format provided by SCDM, including data provided by contract organizations, SCA tools can directly populate Batch Release and Stability templates.

The process flow for SCA to generate a stability report is shown in Figure 7. In Step 1, the user selects filters for the Product, Material Lot, Specification, and Stability Study including the available stability time points and analysis results. This will load the structured content and data from the EDL into the SCA system. In Step 2, the methods and parameters of the loaded data are selected. The order can be manually adjusted to establish a combination of lots and definition of how to group tables before generating the tables. Finally, in Step 3, additional table formatting controls can be applied before finalizing and exporting the tables. Narrative can be added to the footnote of the table to annotate some observations or clarifications, but no reported data can be modified to assure the data integrity and pedigree link back to the SCDM data elements. As the specifications may vary by country/region, different versions of the Batch Release and Stability reports could be easily generated by selecting the different Specification sections or more specifically the country/region metadata. When specifications require an update by the company PQS, the data pedigree will trigger similar alerts to the impacted Batch Release and Stability reports.

When new lot release data or additional stability time points are accrued, these reports may be easily updated with the expanded data selection. In addition, trending analyses are typically required for stability data over time and compared against the stability specification. A separate SCA template could use the same selected dataset to autogenerate the Stability Summary and Conclusions sections (S.7.1, P.8.1). Similarly, the Batch Analysis data is continuously monitored in the company PQS for any trends as new lots are manufactured. Batch Analyses (S.4.4, P.5.4), Comparability (S.2.6, P.2.3) and Justification of Specifications (S.4.5, P.5.6) all feed from the same SCDM data blocks to autogenerate the reports, some involving additional data processing steps such as statistical analysis. As the SCA system links all these sections back to the common source in the SCDM layer, it natively assures the consistency across the variety of reports as well as complete traceability. Furthermore, since the data contents in the SCDM system are established as a true copy of the raw data from various sources, the validated SCA system eliminates the requirement for manual data verification for any of the automatically authored reports and sections.

Benefits of SCDM and SCA

So far, the current challenges faced in data management, regulatory authoring, and lifecycle management of CMC content have been described. SCDM has the ability to transition and transform the current siloed data management systems and unstructured data lakes to an interactive and structured data fabric. As well, SCDM allows for the creation, capture, and reuse of information as product and process development progresses, addressing the manual and repetitive



- Select Product
- Set data filters (manufacturing date range, manufacturing stage, study type, lot, packaging)
- Define additional table structure (method number, acceptance criteria, specification version)
- Define data included (timepoints, conditions)
- Select data precision (recorded, reported)

- Select Test Methods & Parameters
- Select order of methods and parameters
- Combine selected lots together (optional)
- Define table grouping (manufacturing site, storage conditions)

- Select table order and lot number order per table
- Review that format is acceptable
- Annotate with any footnotes
- Add report title
- Add information to table titles (“Batch Analysis”, “Stability”)
- Save and export report

Figure 7. Structured Content Authoring (SCA) Workflow for a Stability Report. Abbreviations: EDL, Enterprise Data Lake.

challenges associated with constructing Module 3. It offers consistency and traceability of information across integrated data sets and reduces the manual efforts of accurately reusing data across documents and data sources. Importantly, increasing business efficiencies can be achieved via design and implementation of SCDM by enabling process automation and content authoring tools. Finally, transitioning the CMC content and data from the PDF file format to a functional, structured and exchangeable data format, such as JSON (JavaScript Object Notation) or XML, allows sponsors and health authorities immediate access to usable content and data, streamlining the submission and review processes and, most importantly, accelerating access to therapies for patients.

To meet the challenges in the existing CMC regulatory landscape it is necessary to develop a more templated approach to CMC content with the use of libraries and SCDM templates which will introduce a standardized language and data presentation across all regions. Standardized language can be integrated with PQ/CMC and similar initiatives. Within these libraries, sponsors can store the narratives as components and multiple components can be compiled together to assemble each Module 3 CTD section in a consistent, reproducible, and easily reusable fashion. If necessary, further “transcription” of Module 3 details into Module 2 QOS and Module 1 Regional Information becomes automated with SCA and SCDM, avoiding the pitfalls associated with repetitive and tedious ‘copy and paste’ activities. A SCDM-based dossier preparation process could help to understand the dossier strategy, visualize the final output earlier in the process, and identify gaps and conflicts among several variations of a narrative. SCDM allows reviewers immediate access to supporting data which increases the efficiency of the internal review processes and reduces inconsistencies as comments and edits are incorporated.

There is also the need to develop a system or tool capable of repetitive SCDM to allow real-time CMC data automation, mapping, and authoring. Deployment of this tool would automate the flow of existing, table-ready data from the EDL to finalized documents to health authorities’ review tools. As such, the tool enables simultaneous submission of data, electronic narrative, and visuals across multiple regions allowing a sponsor to submit data once to all health authorities for data access and review. In this case, health authorities could confer with each other and see the questions and decisions of other regulators. This would reduce the amount of reauthoring and redundant activities for sponsors and health authorities. SCA improves quality and efficiency of report deliverables and aids teams to establish a predictable product development process, mitigating risks, improving quality, and decreasing time to market.

Extensive time is spent on manual sourcing and production of reports, leaving little time for analysis and insights. If institutions synthesize their data to tell a story using analytics, they can cut through the noise and provide clarity into trends and risk-based assessments. Data analytics platforms provide a time-saving solution and effectively incorporate and assess massive data volumes, reducing effort required by resources to create reports and increasing time for much-needed analysis. Removing the redundant aspects of regulatory filings will create more time for scientific rationale discussions and true risk-benefit analyses. Also, reducing redundant activities will increase job satisfaction for industry and health authority employees. Overall, this improves the efficiency of global product submission, review and approval, accelerating the access of new medicines for patients. This technology to reach digitalization in Regulatory CMC is currently being developed and can transform these submission and review processes in the coming years.

Trends and Perspectives

The pharmaceutical industry and health authorities are undergoing a transition towards digitization of data and are at different stages of digitalization.[43] The ultimate goal of SCDM implementation would be to achieve a digital transformation of all data and information supporting regulatory filings from a global perspective, supplying interconnectivity between CMC data in Module 3 and other modules of the eCTD. Incorporating the aspects of digitization and digitalization with AI technologies will fundamentally change strategies and operations for both industry and health authorities. In this article, examples have been shown of how SCDM can be utilized in heavily data-driven CMC sections of the CTD to automate the submission process.

SCDM has the potential to be expanded beyond data from the specific product under development by incorporating prior knowledge of data obtained through product development of ‘like-molecules’. Indeed, combining SCDM across products and platforms, with advanced analytics including AI deep learning, could aid the selection of applicable prior knowledge by automated access of all relevant information for the designated criteria in order to achieve an objective assessment of the transferability of data across ‘like-molecules’.

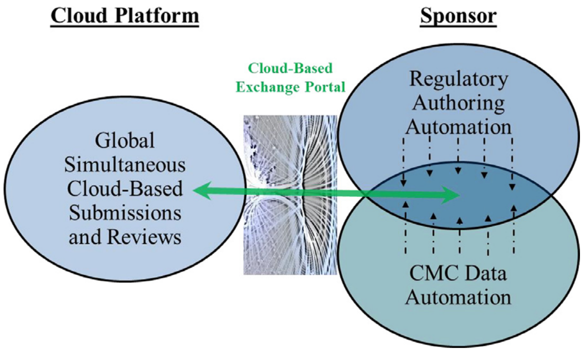

Cloud Platform
Sponsor
Cloud-Based Exchange Portal
Global Simultaneous Cloud-Based Submissions and Reviews
Regulatory Authoring Automation
CMC Data Automation

Figure 8. Integration of Industry Structured Content and Data Management/Structured Content Authoring (SCDM/SCA) with Global Information Exchange Cloud Platforms.

Moreover, deep learning may help identify differences in criteria between products that are relevant to product stability and determine the acceptable limits for those criteria. In parallel to submission and regulatory advancements, SCDM enables machine learning and similar technologies which can transform the development process by providing novel insights from the data accumulated throughout and between drug development platforms.

Over the next few years, additional SCDM-generated filings will be developed and tested in other regions for developing harmonization efforts after successful internal industry incorporation efforts are established. Within 10 years, SCDM can be extended to all modules of the CTD providing a standardized data and submission format which is globally harmonized and accelerates the development of therapies. However, the ultimate goal for submissions is to eventually create the opportunity for a single global submission of a drug application that can be accessed and reviewed by various health authorities simultaneously and reduce the time for a novel medicine to be accessible to patients in need. As stated earlier, a single submission is challenging because of geo-political, socio-economical regional differences. However, SCDM, SCA, and a CMC-UDM with proper rules can manage these differences and enable such a concept. This is especially applicable for CMC data and content since typically a single global product is characterized, manufactured, and tested in the same manner. SCDM provides the capability to manage the regional differences of data reporting for the same product that is supplied globally. This creates the basis to ground these technologies in CMC applications and branch out to other sections of the CTD.

Information exchange between biopharmaceutical companies and health authorities and between multiple agencies in a cloud environment containing the most recent data as they emerge will be a step towards achieving the goal of a single global submission. To this end, Accumulus Synergy, a non-profit company sponsored by leaders in the biopharmaceutical industry, is developing a cloud-based platform to facilitate the information exchange to achieve real-time review in a global setting.[44] Within the Accumulus platform there will be separate and shared spaces for sponsors and health authorities to both work independently and collaboratively. Sponsors can utilize SCDM and SCA technology to automate document generation and push filings or have filings pulled to the Accumulus cloud to facilitate the submission and review process (Fig. 8). Additionally, there will be a strong focus on data privacy and cyber-security to protect patients, sponsors, and agencies. Currently, Accumulus is testing their platform using a Parallel Review Shared-Space and a CMC Data and Analytics use case. As discussed previously, parallel review will reduce the total time to approve a drug in multiple regions while the Accumulus platform will facilitate the collaboration between health authorities and communication between the sponsor and health authorities. The CMC use case is focused on structured data submissions and real-time data exchange, as well as leveraging the aspects associated with collaborative and parallel reviews. CMC data were chosen because of the amenability to a structured approach, and once established the principles developed can be extended to clinical and preclinical data. Accumulus aims to have developed their use cases and additional components within three years and have a functioning cloud marketing application within a decade.

Conclusions

There are a number of drivers for establishing more efficient, reliable and automated SCDM ecosystems, notably the direction health authorities are heading in evolving data standards and compliance requirements.[45] In addition, demands on pharmaceutical and biotechnology companies are increasing, including cost, efficiency, speed, new regulatory requirements, novel modalities, new manufacturing paradigms, and the vast amount of data which accompanies these changes. A further impetus is the growing pressure on the life sciences and health industry to be more transparent, with readily available answers to inquiries on demand and with more details. SCDM provides order and structure that will enable concurrent, collaborative review of submissions among global health authorities, or within agency committees and divisions. Reviewers can more quickly access data earlier as components are completed and allocated to submissions with real-time dynamic and iterative assessments. Also, data provided for multiple health authorities in the cloud will encourage unified submission requirements and collaboration between regulatory agencies facilitating more efficient and timely assessments.

Data drive scientific and regulatory decision-making, and technology advances the interconnectivity of data sources. These connections help identify, curate, and govern the data as well as automate the sourcing of events for tracking on demand requests for specific data sets for any type of quality data query in a product lifecycle. To match the initiatives taken by health authorities, SCDM and SCA will be effective tools to address deficiencies in the current CMC regulatory submission and review processes while improving data integrity, ensuring chain of evidence, and increasing health authority confidence in sponsor submissions. This concept has moved from theoretical discussions to practical application in ongoing testing phases as illustrated by the CMC-UDM, Specification and Stability examples described in this article. Within the next 5 to 10 years, there is optimism that the concepts of SCDM and the supporting technologies will expand beyond Module 3. The long-term aspiration and vision are to attain a single cloud-based global regulatory submission for new drug applications utilizing SCDM, SCA, and a seamless, yet secure information exchange cloud platform where health authorities can review collaboratively and in parallel.

Acknowledgements

The authors would like to thank Simon Hotchin, Rita Algorri, Gerd Kleeman, David Enck, and Thomas DiFeo for their helpful suggestions throughout development of this review article.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Einstein A, Shaw B. *Cosmic Religion: With Other Opinions and Aphorisms*. 1931.
2. EFPIA. *Optimising Post-Approval Change Management for Timely Access to Medicines Worldwide*. 2017. https://efpia.eu/media/25953/efpia-post-approval-change-position-paper_final_feb2017.pdf.
3. Schmelzer R. *AI Runs Into the Document and People Barrier: Digitization and Digitalization*. Forbes; 2020. https://www.forbes.com/sites/cognitiveworld/2020/06/23/ai-runs-into-the-document-and-people-barrier-digitization-and-digitalization/.

4. Robertson AS, Malone H, Bisordi F, et al. Cloud-based data systems in drug regulation: an industry perspective. *Nat Rev Drug Discov*. 2020;19(6):365–366. https://doi.org/10.1038/d41573-019-00193-7.
5. DitaExchange. *How Structured Content Management (SCM) Is Revolutionizing the Life Sciences Industry*. 2015. https://ditaexchange.com/wp-content/uploads/2015/06/SCM-White-Paper.pdf.
6. Braun R. *What does the Future Hold for Regulatory Information Management?* 2019. https://www.pharma-iq.com/regulatorylegal/articles/what-does-the-future-hold-for-regulatory-information-management.
7. Arden NS, Fisher AC, Tyner K, Yu LX, Lee SL, Kopcha M. Industry 4.0 for pharmaceutical manufacturing: preparing for the smart factories of the future. *Int J Pharm*. 2021;602: 120554. https://doi.org/10.1016/j.ijpharm.2021.120554.
8. Yu LX, Raw A, Wu L, Capacci-Daniel C, Zhang Y, Rosencrance S. FDA's new pharmaceutical quality initiative: knowledge-aided assessment & structured applications. *Int J Pharm*. 2019;1: 100010. https://doi.org/10.1016/j.ijpx.2019.100010.
9. Algorri M, Cauchon NS, Abernathy MJ. Transitioning chemistry, manufacturing, and controls content with a structured data management solution: streamlining regulatory submissions. *J Pharm Sci*. 2020;109(4):1427–1438. https://doi.org/10.1016/j.xphs.2020.01.020.
10. Cauchon NS, Oghamian S, Hassanpour S, Abernathy M. Innovation in chemistry, manufacturing, and controls-a regulatory perspective from industry. *J Pharm Sci*. 2019;108(7):2207–2237. https://doi.org/10.1016/j.xphs.2019.02.007.
11. Fox BP, *Amit*, Prevost M, Subramanian N. *Closing the Digital Gap in Pharma*. McKinsey & Company; 2016. https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/closing-the-digital-gap-in-pharma.
12. Braun R. *Why it's Time to Get Smarter about Content Management and Document Production*. @PharmExec; 2018. https://www.pharmexec.com/view/why-it-s-time-get-smarter-about-content-management-and-document-production.
13. Braun R. *What Life Sciences Firms can Learn From Other Industries About Optimizing Routine Document Production*. 2018. https://www.pharmamanufacturing.com/articles/2018/what-life-sciences-firms-can-learn-from-other-industries-about-optimizing-routine-document-production/.
14. Macdonald JC, Isom DC, Evans DD, Page KJ. Digital innovation in medicinal product regulatory submission, review, and approvals to create a dynamic regulatory ecosystem-are we ready for a revolution? *Front Med*. 2021;8: 660808. https://doi.org/10.3389/fmed.2021.660808.
15. ICH. *The Common Technical Document for the Registration of Pharmaceuticals for Human Use: Quality – M4Q(R1)*. 2002. https://database.ich.org/sites/default/files/M4Q_R1_Guideline.pdf.
16. Benedictus J. *Structured Content in Life Science*. 2019. https://janacorp.com/structured-content-in-life-science/.
17. Goran J, LaBerge L, Srinivasan R. *Culture for a Digital Age*. @mckinsey; 2017. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/culture-for-a-digital-age.
18. Cox B. *Woodcock: the US FDA Sets the Stage for Global Quality Dossiers*. 2020. https://pink.pharmaintelligence.informa.com/PS141465/Woodcock-The-US-FDA-Sets-The-Stage-For-Global-Quality-Dossiers.
19. Yu L. *FDA's New Initiative: KASA*. 2019. https://pqri.org/wp-content/uploads/2019/04/PQRI_KASA-Presentation_V4.pdf.
20. Ahmed S. *KASA To Support Generic Drug Review*. 2019. https://www.contractpharma.com/issues/2019-04-01/view_fda-watch/kasa-to-support-generic-drug-review/.
21. Chatterjee B. *Understanding the FDA's Knowledge-Aided Assessment & Structured Application (KASA) Framework*. BioProcess Online; 2019. https://www.bioprocessonline.com/doc/understanding-the-fda-s-knowledge-aided-assessment-structured-application-kasa-framework-0001.
22. FDA. *2020 Annual Report*. 2021. https://www.fdanews.com/ext/resources/files/2021/02-11-21-OPQAnnualReport2020.pdf?1613087522.
23. Eglovitch J. *FDA Gives Generics Updates at DIA Town Hall*. RAPSorg; 2021. https://www.raps.org/news-and-articles/news-articles/2021/6/dia-town-hall-fda-discusses-onsite-inspections-int.
24. Cox B. *Newly Aligned Teams Sped US FDA's Drug Quality Reviews Over Pandemic Hurdles*. 2021. https://pink.pharmaintelligence.informa.com/PS143801/Newly-Aligned-Teams-Sped-US-FDAs-Drug-Quality-Reviews-Over-Pandemic-Hurdles.
25. Fitzmartin R. *Prescription Drug User Fee Act (PDUFA) VI: Electronic Submissions and Data Standards*. 2021. https://www.fda.gov/media/147696/download.
26. FDA. *2021-2025 Strategic Plan*. 2021. https://www.fda.gov/media/81152/download.
27. FDA. *FDA's Technology Modernization Action Plan*. 2019. https://www.fda.gov/about-fda/reports/fdas-technology-modernization-action-plan.
28. FDA. *Data Modernization Action Plan | FDA*. 2021. https://www.fda.gov/about-fda/reports/data-modernization-action-plan.
29. Binggeli L, Heesakkers H, Wölbeling C, Zimmer T. 2018. Pharma 4.0$^{\text{TM}}$: Hype or Reality?, ed.: ISPE. https://ispe.org/pharmaceutical-engineering/july-august-2018/pharma-40tm-hype-or-reality.
30. EMA. *Network Strategy to 2025*. 2020. https://www.ema.europa.eu/en/documents/report/european-union-medicines-agencies-network-strategy-2025-protecting-public-health-time-rapid-change_en.pdf.
31. EMA. *Products Management Services - Implementation of International Organization for Standardization (ISO) Standards for the Identification of Medicinal Products (IDMP) in Europe*. 2021. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/products-management-services-implementation-international-organization-standardization-iso-standards_en.pdf.
32. EMA. *Introduction to ISO Identification of Medicinal Products, SPOR programme*. 2016. https://www.ema.europa.eu/en/documents/other/introduction-iso-identification-medicinal-products-spor-programme_en.pdf.
33. EMA. *Joint Biologics Working Party /Quality Working Party Workshop With Stakeholders in Relation to Prior Knowledge and its use in Regulatory Applications*. 2017. https://www.ema.europa.eu/en/events/joint-biologics-working-party-quality-working-party-workshop-stakeholders-relation-prior-knowledge.
34. EMA. *Workshop with Stakeholders on Support to Quality Development in Early Access Approaches (i.e. PRIME, Breakthrough Therapies)*. 2018. https://www.ema.europa.eu/en/documents/report/report-workshop-stakeholders-support-quality-development-early-access-approaches-ie-prime_en.pdf.
35. Oakes K. *New Toolbox Available for EMA's PRIME Designees*. RAPSorg; 2021. https://www.raps.org/news-and-articles/news-articles/2021/2/new-toolbox-available-for-emas-prime-designees.
36. EMA. 2020. New online platform for scientific advice. ed. https://www.ema.europa.eu/en/news/new-online-platform-scientific-advice.
37. ICH. *Technical and Regulatory Considerations for Pharmaceutical Product Lifecycle Management Q12*. 2019. https://database.ich.org/sites/default/files/Q12_Guideline_Step4_2019_1119.pdf.
38. ICH. *ICH Press Release*. 2020. https://admin.ich.org/sites/default/files/2020-06/ICH40MayTC_PressRelease_2020_0603_FINAL_0.pdf.
39. ICH. *2020 Annual Report*. 2021. https://admin.ich.org/sites/default/files/inline-files/ICH_AnnualReport_2020_2021_0602.pdf.
40. Kamiński R. *AI in Pharma. What does Artificial Intelligence Bring to the Pharmaceutical Industry? - nexocode*. @nexocode_com; 2021. https://nexocode.com/blog/posts/ai-in-pharma/.
41. IBM. *IBM Unified Data Model for Healthcare*. 2021. https://www.ibm.com/downloads/cas/KOV7YZPE.
42. Breitman KK, Casanova MA, Truszkowski W. Ontology in computer science. *Semantic Web: Concepts, Technologies and Applications*. London: Springer London; 2007:17–34.
43. McKinsey & Company. *The state of AI in 2020*. @mckinsey; 2020. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020.
44. Accumulus Synergy. *Accumulus Synergy White Paper*. 2020. https://www.accumulus.org/wp-content/uploads/2021/06/Accumulus_Synergy_White_Paper.pdf.
45. Cwienczek A. *Capitalising on Structured Content Management in R&D/Regulatory Affairs*. 2020. https://www.pharmiweb.com/article/capitalising-on-structured-content-management-in-rdregulatory-affairs.