Semantic Data Exchange Facilities to enable flexible support for data standards and cross-study meta-analysis in the Pharmaceutical Industry

Shree Nath, Ph.D., PointCross Inc.

table of contents

Current and Emerging Needs of the Pharmaceutical Industry

The global Pharmaceutical Industry is facing a number of unprecedented challenges including increased use of global partners and CROs in the R&D process; pressures from the FDA and other regulatory agencies and the public for greater rigor in safety assessments; and a need to accelerate new drugs to market. All of these challenges require a rapid shift towards data centricity, increased transparency, and the ability to look for safety related effects, perhaps overlooked, in past work through meta-analysis spanning multiple studies.

It is also becoming increasingly clear that while study data will need to be captured and communicated in vendor-neutral data standards (such as CDISC and HL7), the emergence of new types of biotechnology drugs and industrial products such as nanotechnology that can affect drug safety, and new uses of existing drugs, imply that such standards cannot remain static for long because information will need to be rapidly constituted or re-constituted on demand. Pharma R&D organizations and Regulatory Agencies alike must be able to semantically identify and convert data from one or more source formats into a current version of a data standard. This need is clear and present. Publishing raw or structured data in paper formats and PDFs are huge impediments to this important change. Likewise, trends towards designing infrastructure based on standards alone are simply not sustainable.

Key Drivers for Change in Current Data Management Practices

The past few years have been marked by a number of significant events that are a call to action for changing current data management practices:

1. Safety-related recalls of commercialized drugs;
2. Increasing awareness of drying drug development pipelines at most Pharma companies;
3. A reduction in the approvals of new drugs as regulatory agencies increasingly err on the side of caution;
4. A continued trend of Pharma company mergers and acquisitions;
5. A strong and continued globalization trend and use of off-shore research partners and CRO facilities;

Contact us at (650) 350-1900
get_started@pointcross.com
pointcross.wordpress.com
www.pointcross.com

6. New pricing realities due to health care reforms in the US and other Western countries.

Items 1 and 3 have caused introspection about the need to increase the speed of approvals, safety reviews and other initiatives – without compromising public safety. Item 2 is causing Pharma companies to look at innovation and to re-structure their research activities to rapidly inject new prospects into their pipelines through internal processes, acquisitions or outsourcing core drug development processes – and an increasing interest to explore legacy information for new insights that could create opportunities for new drug candidates.

Item 4 requires Pharma companies to deal with acquired studies and the clinical and non-clinical data that accompany the acquired companies. Along with the acquisitions, companies also inherit the liabilities of past decisions, and they now must merge this new source of additional information into a single accessible body of knowledge across the new enterprise. Even worse, the whole point of these acquisitions – the potential new opportunities that they represent – are often obscured and missed due to companies' inability to effectively integrate, reconcile, and make the most economically rewarding decisions about the two pipelines.

Item 5 is also a new source of additional disparate data types and formats as these external parties or newly inducted off-shore teams must be able to rapidly flow their study data to their scientist colleagues and sponsor companies.

Item 6 will require companies to collaborate with insurers and health care providers to tap into a range of external patient data both to identify the size of the market for new disease indications, as well as to create reimbursement scenarios sooner in the drug development lifecycle.

All in all, the picture that emerges is a tidal wave of study, trial, patient, and related data in myriad formats that simply cannot be ignored any longer, but which must be considered diligently and be taken up now as a strategic initiative, despite trends in cost-cutting and workforce reduction. Likewise, metadata associated with raw data will increasingly need to be created and mined to understand trends and correlations.

The benefits of managing both non-clinical and clinical data in a holistic and data-centric manner far outweigh costs associated with technology investments and changes to corporate business processes.

Meta-analysis done painstakingly by independent researchers using publicly available data have pointed out major risks in at least one and often more drugs each year. Common people on the street and the media increasingly question why such analysis is not done routinely.

In reality, Pharmaceutical R&D scientists are very diligent, competent, and equipped with the best analytical tools. However, they are stymied by the disparate formats and software applications in which legacy data is stored, with no way to rapidly use them alongside current data that may have been acquired in other studies using these modern analytics. Likewise, extracting metadata is tedious. It is certainly possible to convert each study data manually into a common format, but the time and costs involved make it hugely impractical.

What is required is an airport hub or "Grand Central Station" type of data exchange terminal where data in disparate formats and their versions can be converted into any other format – and versions of applicable data standards. This analogy has been articulated by some of the leading information architecture thought leaders in the Life Sciences industry (e.g., Anglin et al., 2007). Effectively, this approach should create a vendor-neutral, standards-neutral facility where scientists, safety researchers, and statisticians can effectively do what they excel at – with the full complement of institutional knowledge at their fingertips – instead of spending their time manually shuffling data like peons.

The FDA has initiatives, HL7 for instance, which will necessitate the use of data-centric approaches for both clinical and pre-clinical submissions. Initiatives like this exist because regulatory agencies, just like Biopharma scientists, need access to accurate, holistic data. They also need to be able to conduct meta-analysis across submissions from various companies and other sources to assess the potential for safety issues.

The FDA is looking to streamline its processes, allow its stretched resources to rapidly immerse themselves in the submitted data, collaborate with their colleagues, and do all this with workflows that capture the data, annotations, decisions, and internal and external communications as a part and parcel of their normal work on the study data. Agencies such as the FDA are also likely to change the basis for their data standards in line with future regulatory bioinformatics needs. This is particularly true if certain emerging vendor-neutral formats like HL7 demonstrate value in terms of easily and rapidly populating their internal data warehouses like Janus, and more importantly, if they also enable effective cross-study/cross-compound analysis to improve safety assessments and accelerate the review process. The IMI initiative in Europe is looking to create a rich information repository of Pharma R&D data from multiple companies as a way for cross collaboration and to reduce safety risks using predictive analytics and *in silico* modeling tools. Such initiatives will increasingly need companies, academia and regulators to be able to rapidly share data in flexible formats while assuring normalization for cross study meta-analysis and modeling.

Requirements for Study Data Management and Exchange in the Modern Pharmaceutical Enterprise

Some of the key requirements for study data management and exchange include the following:

Adoption of XML Based Data Models

Given the key drivers for change discussed above, pharmaceutical companies ought to take a fresh look at how study data should be managed, and recognize that a fundamentally new, data-centric paradigm for information management and data exchange is not only possible and necessary, but that the technology to do so is already available. To quote a 2004 article in Nature Biotechnology, "Holding the data hostage in proprietary file formats should not be tolerated in the scientific community. The future of data-intensive biology depends on ensuring open data standards and freely-exchangeable file formats."

The ideal *lingua franca* for study data representation is vendor-neutral XML. In layman terms, XML is simply a way to represent data, and describe it so that a machine can understand each section of the data, and how it relates to other parts of the data set. An XML schema can define the way the data will be organized, complete with its nomenclature. It is also extensible (the X represents this aspect) and it is straightforward to write programs to use data represented in this manner. For those who are familiar with RDBMS (Relational Data Base Management Systems) technology, the data tables or ERWIN diagrams used to describe database relationships effectively describe the same thing as an XML schema. But an RDBMS, despite being very fast, is very rigid and inflexible; any changes to an RDBMS require significant time and change control. It is this flexibility in an XML schema, and its vendor neutral nature, that is of high value.

Data Transformation and Organization

Companies must also recognize that for the foreseeable future, they will still have to deal with existing systems and formats, both within their networks and within the extended enterprise (including CROs, co-development partners, and other interested third parties). The simplest approach is one that allows data models from any structured data source (external systems, databases, flat files, Excel outputs and others) to be normalized to a universal set of XML-based data models – essentially a Universal Operational Data Model (UODM) – in (1) any-to-one; (2) any-to-any; (3) any-to many; and (4) many-to-many mappings. This must additionally be done through a vendor-neutral framework for data exchange that can support connectors to commonly used Pharmaceutical R&D applications and data warehouses. The UODM should be flexible enough to handle non-clinical and clinical data, as well as other related R&D or trial information, including unstructured content and legacy paper reports with tabulated data that must be digitized.

Once the transformation rules have been defined (including the mapping of terms between the source and destination formats for meaning equivalency and to create shared taxonomies, it is possible to re-purpose (convert or transform) the source data into vendor-neutral XML with traceability down to the original source. The conversion of "n" disparate data formats into "m" other target formats is very difficult if one attempts to create a one-to-one mapping from each of the source-to-target formats simply because it would require 2(n x m) bi-directional adapters.

On the other hand, a hub and spoke approach requires only (n+m) bi-directional adapters. Versions within each of the "n" and "m" data sets would, of course, further increase the numbers of adapters required. However, by using a common "meta-format" which we refer to as a "Universal Data Model," it makes it possible for any standard to be defined in terms of this meta-format. It is like writing translators for all language to and from a single common language – say Latin. This way, it is possible to convert a passage in any language into any other language. As long as links to the source or raw data are preserved, and as long as no changes are made to the raw source or to the converted datasets (once validated), with proper information security and compliance controls throughout, there is no reason for people to hold on to antiquated metaphors that are a legacy of antiquated paper-based methods of handling data!

Data Model Controls – Versioning, Type Mismatches and Controlled Terminology

Every data model – whether legacy or current – should be able to undergo multiple revisions. Therefore, a strict version control system, and a naming convention both for the data model and terminology within it must be defined and maintained with respect to the meta-format or Universal Data Model. The on-the-fly converters must be capable of recognizing names and version controls, and be able to alert users if there are any unresolved mismatches.

One of the important resources for these scientists will be the capabilities of such a "semantic system" which can relate the terms or nomenclature used in one format to another used in a different data set to facilitate meaning equivalency. If there are controlled vocabularies in use, then this can enable true decision support for scientists, particularly by leveraging industry standards such as NCI's Enterprise Vocabulary Services, among others.

Data Model Extensibility

Study data and their source systems can be quite diverse and change rapidly as more knowledge is acquired; consequently, it is necessary for the data models to be extensible through simple editors that scientists and analysts can use, and that do not require extensive reliance on IT resources.

Data Annotations

Beyond management of data models *per se*, it is also important that certain processes be embedded into systems such that scientists and researchers do not have to remember to organize their data files and annotations. Rather, systems should capture all of their comments in association with the relevant data types as a normal course of their work.

Data Re-Purposing with Governance and Traceability

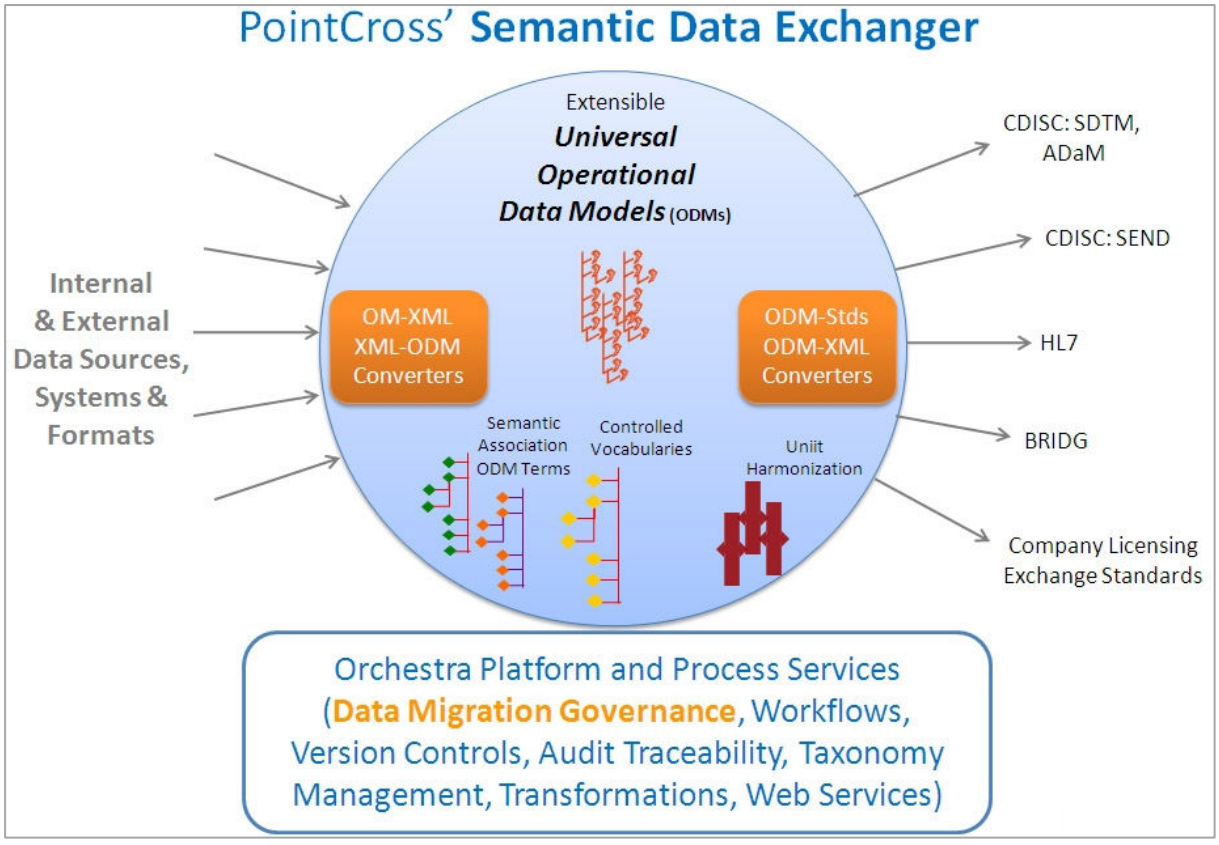
Study data is typically acquired for an original purpose: to experimentally measure parameters that are important to that study. Scientists will agree that this data set could very well be useful again in the future; either by itself or in conjunction with other data sets from other studies to test other hypothesis. Therefore, it should be possible to re-purpose study data for other needs because it reduces costs; it might also point the way to additional trials that must be conducted, and it allows for better regression analysis. However, it is also necessary that the entire infrastructure for data re-purposing and transformation ensure traceability of information back to the source, and that all levels of normalization have complete transparency.

Thematic Data Organization

Study data tends to be multi-dimensional in that a single study will attempt to gather data related to multiple indications. For example, one combination of data may relate to neurological functions while additional data may be gathered to support analysis of endocrinal functions. Each of these can be treated as a separate theme. Furthermore, meta-analysis to look at similar effects or events across multiple studies can be facilitated.

Vendor-neutral XML is an ideal technology for defining such thematic breakouts. It makes it easy to re-constitute the broken-out themes into new data models that can be applied in a future meta-analysis or a future standard. Some of the benefits are:

- ☑ Maximize the present and future value of historical or legacy data
- ☑ Separate themes within the data to aid meta-analysis
- ☑ Ability to meet current and future standards and version of standards
- ☑ Ability to meet multiple standards (US, Europe, Asia and other agencies)



SEMANTIC DATA EXCHANGE SERVICES™ (SDES): A VENDOR NEUTRAL TOOLKIT FOR DATA EXCHANGE IN THE PHARMACEUTICAL INDUSTRY

Semantic Data Exchange Services™ is an enterprise toolkit from PointCross that creates a unified layer for study data exchange services in the Pharmaceutical and other related Life Sciences industries. It has been successfully proven among early adopters interested in exploiting their legacy data, and in regulatory data exchange contexts.

SDES is vendor neutral, in that it supports the creation of standard connectors to Pharmaceutical R&D applications, and also provides the ability for companies and their systems integration partners to build and maintain their own connectors.

In this sense, SDES is a "data exchange terminal" that can securely extract data from multiple systems and formats, convert it into a Universal Data Model, and publish it on demand and with integrity to other formats required for regulatory submissions, or to other applications for analysis and meta-analysis as depicted in the figure below.

SDES provides the following capabilities to pharmaceutical scientists, regulatory personnel and corporate data management groups:

- Data mapping, exchange and consolidation from multiple sources into a common data-centric environment. For instance, a scientist is now able to scan through a library of all potentially available data sources, understand the basis for the data, and the conditions under which data were collected. Even though the old data may be in a format that cannot be read directly into the scientist's current analytics tool, the scientist will be able to read it in and incorporate it into their analysis. This is because SDES will convert the data from the original format to a universal internal format, from which it is converted to the format required by the analytics tool.

- Data exchange with different systems and formats – including data sourced from CROs. This can be done with the conversion tools, and a business process that ensures that the conversion is done with integrity. The workflows will include the ability to view, select, pick, convert, and store the completed analysis.

- Data models managed within the SDES are available to scientists and others through an online browser and editor. All data models are subject to version controls for long-term sustainability, and can be linked to controlled vocabularies to assure meaning equivalency as shown in the above figure.

- All of the data enters a controlled system of record[1] and is managed with complete integrity and traceability in a vendor-neutral XML format

- Audit traceability that links data back to the original source, and traces it throughout the lifecycle of the study through submission and into archival data stores.

- A data governance process to manage semantic normalization, taxonomy remediation, data cleansing, unit harmonization, annotations and related activities, with full decision traceability.

- Data, once consolidated within SDES, can be converted to other formats (e.g., CDISC regulatory standards, HL7 formats, alternate XML schemas, or data warehouse formats); data within an overall data model can also be extracted into thematic groups for reporting requirements. An example would be safety data that needs to be extracted across a common class of molecules.

- Search and orienteering across multiple studies (meta-analysis) is supported, featuring security and authorization, graphing/tabular views and annotation capabilities.

- Scientists have authorized access to the common data pool through relevant dashboards, views and pick lists with filtering, sorting and search capabilities.

- Creation of analysis datasets separate from the original data – but kept in association with them as additional versions – is supported.

- Reviewers can export data for use within other preferred applications.

- Capabilities for annotation (comments, insights or other notes) are provided for study-specific or cross-study analysis at multiple levels: at the data type (attribute) level, and for any results where scientists wish to document assumptions, analysis methodology applied and computations.

CONCLUSIONS

The globally dispersed Pharmaceutical industry is acutely aware of the need to change the way they capture, validate, analyze,

[1] System of Record refers to a system such as SDES that automatically assumes that all data passing through, all access to the data, and all transformations to that data are potentially material to the company and its governance and reporting. Therefore, a System of Record automatically traps all activities and makes these available to internal and external auditors. The System of Record ensures that no one - including IT administrators - can manipulate data without leaving a trace about their activities.

use and report data across both pre-clinical and clinical studies, particularly given their challenges as discussed in this paper.

The industry has, at the same time, not adequately recognized that the problems created by antiquated data management processes cannot be solved with the same level of thinking that created them in the first place (to paraphrase Albert Einstein!).

It is time for a fresh perspective, and for technology to accelerate change across the entire value chain in the industry, such as with the proven vendor-neutral SDES system described in this paper, which is available today.

In conclusion, some of the benefits that can accrue to companies by adopting Semantic Data Exchange technology include:

- Compliance with current and anticipated regulatory requirements, internal policies and records management;
- Creation of a knowledge-based, data-centric framework for study data capture and management – increasing time to market without compromising safety and quality;
- Enablement of search, orienteering and meta-analysis across multiple studies, molecule types or other analysis dimensions for early insights, safety screening and identification of promising leads. This is particularly important given the shift in the industry towards early stage toxicity and efficacy evaluation as a strategy to reduce late stage attrition;
- Facilitation of streamlined interactions and data exchange with partners and CROs alike;
- Streamlining of analysis and review processes by providing scientists with access to consolidated datasets in the required formats; and
- Deployment of such technology with minimal impact to existing corporate IT infrastructure, while extending current technology investments.

References

Wiley, H.S., and Michaels, G.S., 2004. Should software hold data hostage? Nature Biotechnology, 22: 1037 – 1038.

Anglin, G., Burke, J., Ferrante, R., 2007. The Benefits of a Life Sciences Industry Architecture, CDISC Report. 23 pgs.

About PointCross

PointCross is a global provider of advanced strategic business solutions and services to knowledge-rich markets, including the pharmaceutical industry.

Our Integrated Drug Development Suite (IDDS™) specifically addresses the pharmaceutical industry's key concerns. At the heart of IDDS is the Orchestra+Solo™ platform, an adaptive, contextual knowledge environment and personalized client that orchestrates core business processes. This robust solution set delivers the following capabilities:

- ☑ Single point of access to contextualized tacit and structured knowledge across the enterprise, with search and guided navigation within and across contexts;
- ☑ Robust search and orienteering capabilities across studies, emails, documents, meta-data and more across the entire organization, CROs and partners
- ☑ Flexible, fool-proof IP security based on contexts and roles, determined by business rules;
- ☑ Predictive analytics for clinical and non-clinical data;
- ☑ Secure multi-party workflows for knowledge sharing and corporate social networks within and across companies;
- ☑ Semantic Data Exchange Services (SDES) for vendor-neutral data mapping, exchange, and organization;
- ☑ Solutions for business development, acquisitions and licensing, e-discovery, audit, regulatory submissions, compliance, and more;
- ☑ Scalable architecture and development toolkits for additional capabilities.

PointCross represents a new way of doing business. We deliver business ready solutions in 1/10th the time and a fraction of the costs compared to standard technologies, while offering strategic advice from people who know the pharmaceutical industry.

We are headquartered in the California Bay Area. We are a CDISC solution provider to the Pharmaceutical Industry, and a Microsoft Gold Certified partner. We also have a global network of service, consulting and infrastructure partners.

CDISC

Microsoft GOLD CERTIFIED Partner

For more information, visit us at www.pointcross.com and call us at (650) 350-1900. Also, check out our blog at http://pointcross.wordpress.com.