PhUSE 2017

Paper SI02

Clinical Metadata – Metadata management with a CDISC mindset

Andrew Ndikom, Clinical Metadata, London, United Kingdom
Liang Wang, Clinical Metadata, London, United Kingdom

ABSTRACT

Metadata is now an integral part of the clinical trial process. CDISC provides the framework for this metadata and Define.xml provides a way to share the final product. There is, however, no standard way to efficiently collect, manage and collaborate in the process of creating metadata.

ClinicalMetadata.com provides an end-to-end metadata management solution, including CDASH, SDTM, ADaM and output metadata. It generates SAS code fragments and submission deliverables, automating repetitive programming tasks, accelerating and simplifying the development process.

Clinical Metadata is built as a web-based single page application, using the latest technologies and practices, and is securely hosted in the Amazon Web Services cloud. Its visualisation tools present the user with a high-level metadata overview and allows them to drill down into individual metadata fragments. In-built standards management, traceability and impact analysis tools help users understand the relationships between each metadata element as well as encourage Agile and collaborative ways of working.

KEYWORDS

CDISC, SDTM, ADaM, Clinical Metadata, metadata repository, data standards, define.xml

INTRODUCTION

This paper begins with a short introduction into what metadata is, with a particular focus on metadata within clinical trials. It then provides a short explanation of why metadata and data standards have become an important part of the clinical trial process in recent years. It then moves on to look at the most widely adopted metadata solution and outlines the five key problems with this approach.

Subsequently, the paper provides an overview of Clinical Metadata, an end-to-end metadata management solution. It briefly discusses the technology and software development principles which underpin its design, explaining how much of the inspiration for the tool came from observing how similar problems have been solved in other industries. Finally, the paper explores some of the tool’s features, while demonstrating how it overcomes key problems with existing metadata management tools.

WHAT IS METADATA

Metadata is usually defined as “data about data”. This definition fails to convey its importance. More accurately, metadata could be defined as the information required to contextualize and understand a given data element.

Within a clinical trial, metadata may exist at many levels. For example, dataset level metadata describes the properties of a dataset, for example:

- Dataset name
- Dataset label
- Dataset Class.

Variable level metadata describes variable level properties and may consist of the following aspects:

- Variable name
- Variable label
- Variable format
- Variable type
- Controlled terminology
- Source variables
- Derivation

In the following example we can see how data without metadata has little purpose or meaning.

DATA WITHOUT METADATA

Consider a possible piece of clinical data:

- 192

Viewed in isolation it does not have utility or meaning, perhaps it is the result of some laboratory test, perhaps it is a patient's height measured in centimetres.

If we attach the following metadata to it:

- Type: Numeric
- Length 8
- Format: DATE9.
- NAME: BRTHDT
- Source: DM.BRTHDTC
- Derivation: Equal to DM.BRTHDTC converted to a numeric date.
- Label: Date of Birth (N)

It becomes apparent that this number represents the date of birth of a patient, which in this case is the 11th July 1960.

WHY DO WE NEED ROBUST METADATA AND DATA STANDARDS IN CLINICAL TRIALS?

Technological changes over the past 70 years have enabled the size, scope and complexity of clinical trials to increase. The pharmaceutical industry has moved from relatively simple paper-based trials collecting little information about a small patient population, to large complex trials where data is captured, transferred, transformed, stored and processed electronically.

The number of sponsors, trials and analyses has also greatly increased, leading to the availability of ever-increasing quantities of data. Managing and using this data without industry wide standards was increasingly inefficient and time consuming for sponsors, those working in the drug development industry and regulatory authorities.

In 2004 Lester Crawford, former Commissioner of the Food and Drug Administration (FDA) stated:

"FDA reviewers spend too much valuable time simply reorganising large amounts of data submitted in varying formats. Having the data presented in a standard structure will improve FDA's ability to evaluate the data and help speed new discoveries to the public."[1]

Since then there has been a move within the industry to adopt data standards and the collection, sharing, storage and presentation has become an increasingly important aspect of the clinical trial process.

The Clinical Data Interchange Standards Consortium (CDISC) was formed in 1997. Its aim is: "To develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare." Its Foundational Standards provide the basis for "supporting clinical and non-clinical research processes from end to end." CDISC standards have been accepted across the industry. What's more, the FDA[2] and other regulatory authorities now insist sponsors submit their data in a CDISC compliant format.

CDISC details some of the benefits of adopting its standards as:

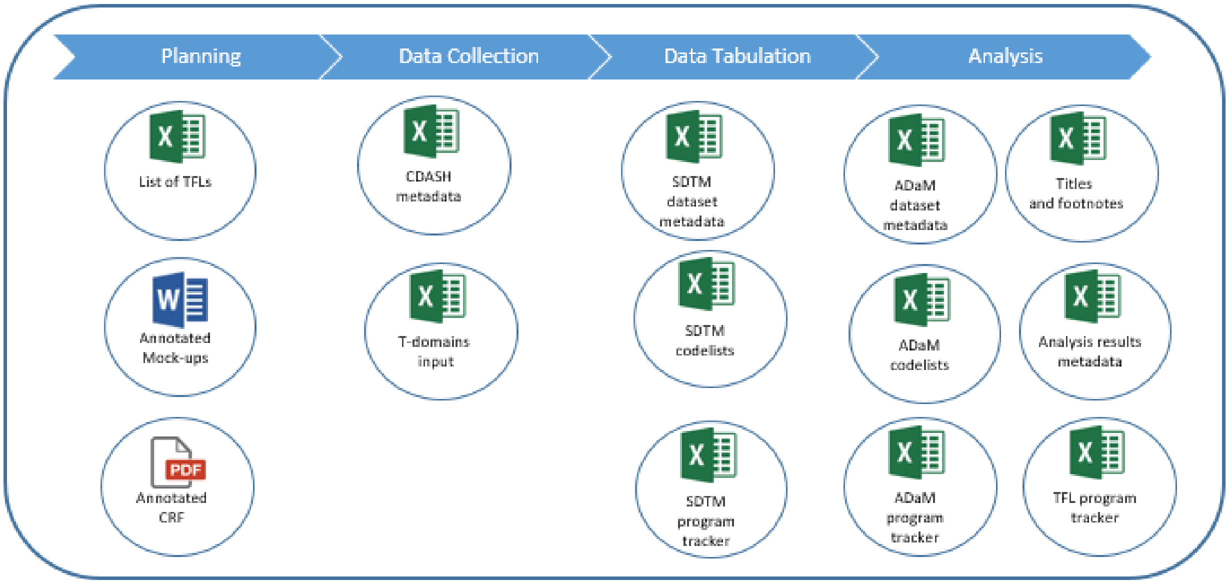
- Facilitated data sharing
- Complete traceability
- Improved data quality
- Streamlined processes.

Over time the number of standards produced by CDISC has greatly increased, see for example, Winnpenny, G (2014) *What's new in ADaM?* [3]. This increase offers opportunities for sponsors to standardize and improve their processes, however, understanding and conforming to these standards has presented challenges to sponsors.

CURRENT APPROACHES TO DATA STANDARDISATION AND METADATA MANAGEMENT SOLUTIONS

The adoption of data standardisation within the industry has been varied. Many sponsors have adopted a responsive rather than pro-active approach. Rather than using data standardisation to transform and improve the way they work, many sponsors have tried to retain their existing workflows and viewed each new metadata standard as a cumbersome add-on to their existing process. Sponsors have often built separate systems to manage each additional class of metadata as it has been introduced by CDISC. In trying to catch up, sponsors have sought the fastest, most readily available tool to deliver the desired solution. In many instances, the most readily accessible solution has been to use supplementary Excel spreadsheets and Word documents. As new classes of metadata have been introduced, this approach has led to a disjointed and complex web of spreadsheets to manage metadata at each stage of the process.

Consider the following potential metadata and programming tracking system. It comprises 12 spreadsheets, a Word document and a PDF and is designed to manage an entire study's metadata, from raw data to reporting, including CDASH metadata, SDTM dataset metadata and annotated mock-ups.



This fragmented solution presents many issues, including these five key problems:

1. Multiple spreadsheets are needed per study, leading to disconnected metadata silos.
2. It is difficult to understand the relationships between metadata elements.
3. Reading and writing information directly from the metadata is complicated.
4. It is complicated for multiple users to collaborate on same file, and for users working on one file to understand how their work impacts other users.
5. Assessment of CDISC compliance and standards management relies on additional tools outside of the system.

In summary, CDISC lists some of the advantages of adopting its standards as:

- Facilitated data sharing
- Complete traceability
- Improved data quality
- Streamlined processes.

Many implementations of CDISC however, have led to opaque metadata managements systems with in-built data redundancy and a lack of traceability. An alternative approach is required. Clinical Metadata provides this solution.

CLINICAL METADATA

AN END-TO-END METADATA AND PROJECT MANAGEMENT SOLUTION

Clinical Metadata is a complete metadata, data standards and project management solution. It uses the latest technologies and practices from the software industry and is built upon the CDISC core principles of traceability and transparency.

Some of its key features include:

1. End-to-end metadata management within a trial.
2. Standards governance.
3. Metadata driven SAS code generation.
4. Evaluation of CDISC compliance.
5. Creation of submission deliverables.
6. Real-time assessment of study status.
7. Communication tools.
8. Metadata visualization.
9. Impact analysis
10. Metadata analysis.

THE TECHNOLOGY

As mentioned previously, technological developments in the financial services have been an inspiration for this new tool. During the past few years, as internet technologies have matured, the financial services industry has gradually moved from using desktop-based applications to web-based applications. This change is being reflected in other industries, including the pharmaceutical industry, where similar concerns have held back this transition in the past, such as:

- Information security and privacy
- System availability and network connectivity
- User authentication and authorization
- Data storage capacity.

There are now effective solutions to these concerns. Companies making the transition from desktop applications to web applications benefit in terms of cost and efficiency, and individual users benefit from simpler workflows:

- Software does not require installation and can be accessed with a web browser, so there is less system maintenance overhead.
- The latest features are always immediately available and software upgrades are not necessary.
- Data can be accessed anywhere with an internet connection and does not rely on an individual PC.
- Data is stored securely in the cloud and back-ups are created automatically.
- Storage capacity is not limited by available hardware.
- Collaboration is improved as multiple users can access the same set of data concurrently.

Clinical Metadata has been developed using an Agile methodology, with a focus on good User Experience (UX). It learns from software projects across many other industries. Features are continuously added and improved, those that add the most value to end users are prioritised, and regular deployments mean that they are quickly available. Automated testing means that problems are identified before going live and development time is reduced to a minimum. Clinical Metadata’s focus is to deliver an application that is clean, simple and easy to use.

Data is secured through a variety of measures, including:

- Traffic to the web application is secured using HTTPS.
- Users accessing the non-secure HTTP website are automatically redirected to HTTPS.
- Industry standard security practices are followed for login credential storage and session management.
- The database is secured, it is not visible to the outside world and access is restricted to within our network.
- Access to server configuration is secured using 2-factor authentication.

The system has been built with the following technologies:

- C# and ASP.NET for the server side application logic
- SQL Server for data storage
- Angular and TypeScript for client side User Interface (UI)
- D3 for data visualisations
- AWS for hosting and infrastructure
- Git for source code version control.

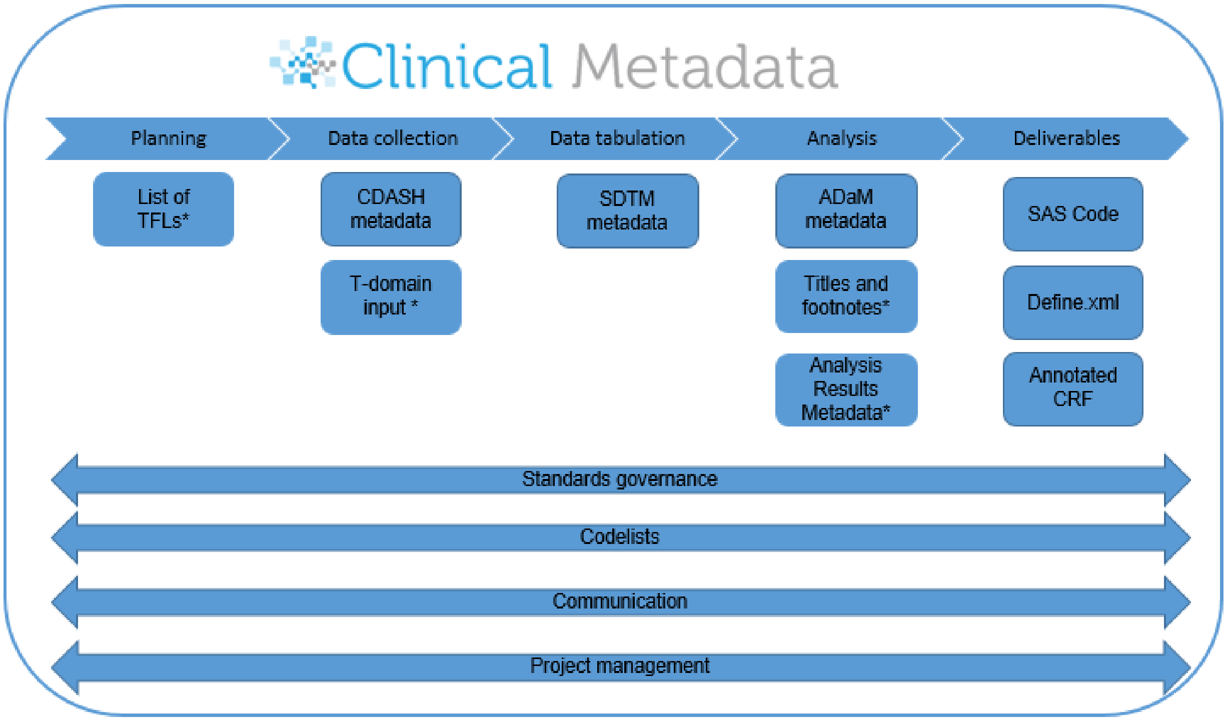
CLINICAL METADATA FEATURES

Clinical Metadata can assist you in managing your metadata and data standards by overcoming the five key metadata and project management challenges identified earlier in the following ways. Through these five points we will outline some of the varied benefits of the tool.

#1 NO MORE METADATA SILOS

Traditional approaches to managing metadata can often result in disconnected metadata silos. Data exchange between these silos either does not happen, relies on human interventions or complicated bespoke programming.

Clinical Metadata is designed to provide a truly end-to-end solution, as shown in the following image[1]:



The system is designed around the principle that metadata should be defined once and used multiple times. Where possible, each metadata element is defined in terms of its predecessors, or imported from a higher level of the standards hierarchy. Built upon relational database principles, the individual properties of an element are inherited rather than duplicated, thus increasing efficiency, reducing the possibility of data entry errors and leading to greater traceability across the entire study metadata lifecycle.

For example, in the disconnected metadata world, SDTM and ADaM metadata might be stored in separate spreadsheets. Metadata for a variable which exists in both SDTM and ADaM would need to be manually entered twice, once in each spreadsheet. This process is error-prone and inefficient. Within Clinical Metadata it is possible to:

- Define metadata once at the SDTM level.
- Use a point and click system to indicate that the ADaM variable is a direct copy of the SDTM variable.

A permanent link is then established between these variables, meaning that any changes to the SDTM metadata for an element will automatically be reflected in the ADaM metadata.

[1] * Functionality to be added quarter 4 2017.

#2 DIFFICULTY IN UNDERSTANDING THE RELATIONSHIP BETWEEN METADATA ELEMENTS

Within the Excel based world, important metadata, for example variable names, is often embedded in human readable text fields which are not machine readable. Each metadata element generally exists in isolation, and there is no way to easily access the metadata of related elements.

Within Clinical Metadata each element is connected to its related elements. Visualization tools help the user to understand these relationships.

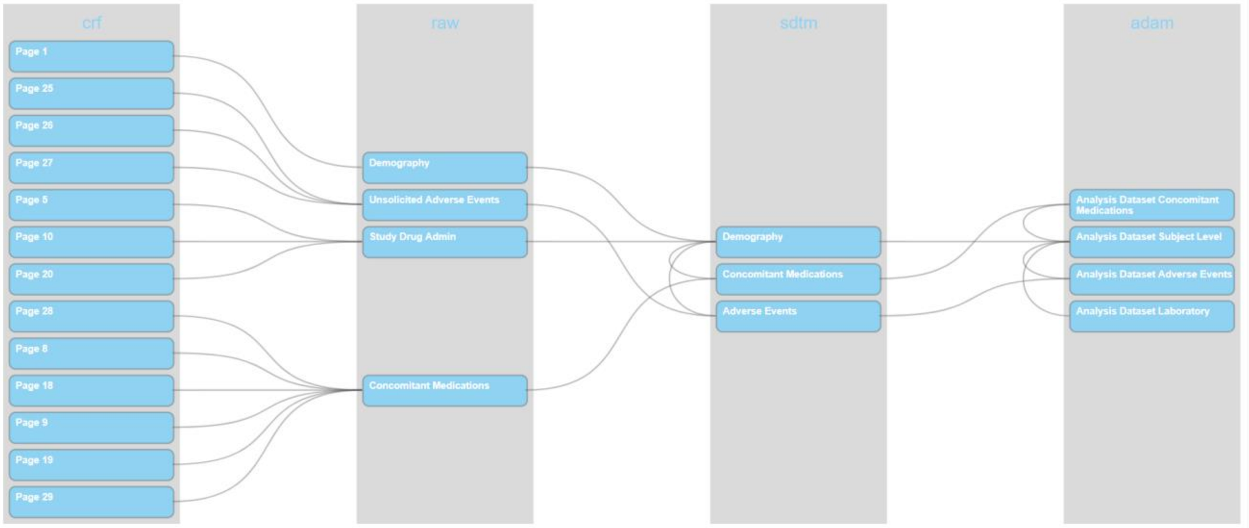
For example, consider a potential derivation for an ADaM variable.

Name	Label	Derivation
ONTRTFL	On Treatment Record Flag	Indication of whether concomitant medication was ongoing during treatment window. Set equal to "Y" if TRTSDT <= ASTDT <= TRTEDT + 28 or ASTDT <= trtsdt AND AENDT >=TRTSDT Otherwise ONTRTFL = ""

A user familiar with CDISC could probably decipher this derivation, however users with limited CDISC knowledge or who are new to the industry would possibly be confused. They may have the following questions:

- Are the source variables TRTSDT, ASTDT, TRTEDT and AENDT taken from ADaM or SDTM?
- In which dataset is each of these variables?
- Where is the metadata of each of the source variables?
- How are the source variables created earlier in the process and how do they link back to the Case Report Form?
- Is the ONTRTFL variable used in any downstream processes, if so which?

Clinical Metadata uses visualisations to ensure that users can easily understand the relationships between metadata elements and to provide end-to-end traceability.



Clinical Metadata's visualisations are built using the latest web technologies. Its drill down functionality allows the user to switch between a high level visualisation, i.e. showing the relationships between datasets, and a lower level, i.e. showing the relationships between variables. Hover over functionality allows the user to see, not only the relationship between variables, but also the derivation rule used to map from one variable to another.

#3 MAKING PROGRAMMATIC USE OF METADATA

Many sponsors moved towards CDISC from a situation where they relied on existing black box software capable of reading input data, performing complex derivations and outputting statistical reports. CDISC standards were often regarded as an undesirable adjunct to this ‘streamlined’ process and little thought was given about the transformational possibilities of true end-to-end metadata management systems, for example that metadata could be used to create code.

Clinical Metadata allows users to generate SAS code and submission deliverables directly from the tool including:

- Variable mappings
- Metadata correct zero observation dataset shells
- Define XML
- Annotated CRFs.

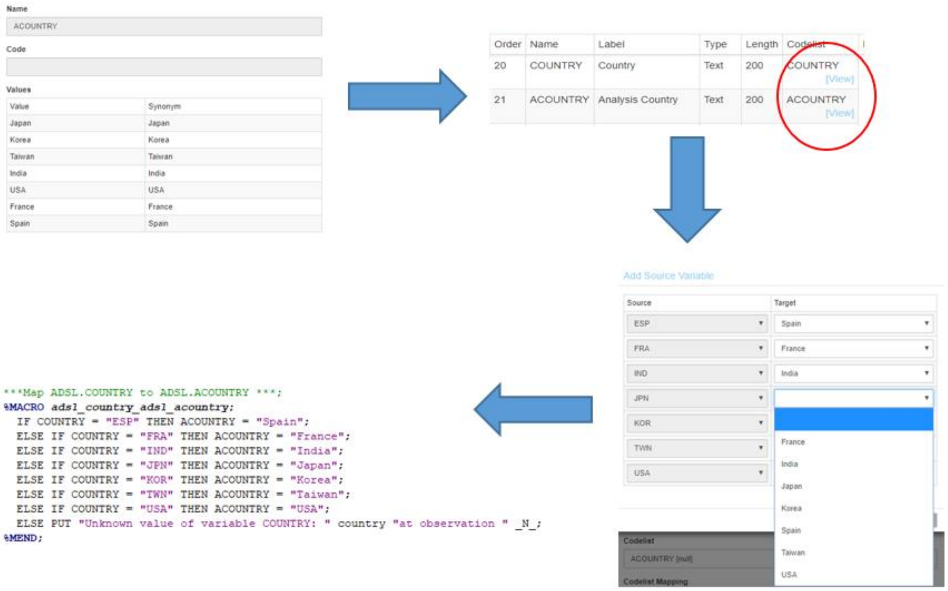
Consider the example of mapping the variable COUNTRY to ACOUNTRY within the ADaM ADSL dataset. Previously users might have:

1. Manually created the COUNTRY codelist.
2. Manually created the ACOUNTRY codelist.
3. Manually created the metadata for the ADSL.COUNTRY variable.
4. Manually created the metadata for the ADSL.ACOUNTRY variable, as well as defined a pseudocode derivation rule.
5. Manually created the SAS code required to perform the mapping.

Within Clinical Metadata the user:

a. Imports the COUNTRY variable from SDTM.DM.
b. Imports the ACOUNTRY variable from a project library.
c. Creates the ACOUNTRY codelist.
d. Uses a point and click system to define the mapping.

The tool will then automatically build the SAS code needed to perform the derivation.



Name: ACOUNTRY

Code:

Values

Value	Synonym
Japan	Japan
Korea	Korea
Taiwan	Taiwan
India	India
USA	USA
France	France
Spain	Spain

Order	Name	Label	Type	Length	Codelist
20	COUNTRY	Country	Text	200	COUNTRY [View]
21	ACOUNTRY	Analysis Country	Text	200	ACOUNTRY [View]

Source	Target
ESP	Spain
FRA	France
IND	India
JPN	
KOR	
TWN	
USA	

Codelist: ACOUNTRY [null]

```
***Map ADSL.COUNTRY to ADSL.ACOUNTRY ***;
%MACRO adsl_country_adsl_acountry;
  IF COUNTRY = "ESP" THEN ACOUNTRY = "Spain";
  ELSE IF COUNTRY = "FRA" THEN ACOUNTRY = "France";
  ELSE IF COUNTRY = "IND" THEN ACOUNTRY = "India";
  ELSE IF COUNTRY = "JPN" THEN ACOUNTRY = "Japan";
  ELSE IF COUNTRY = "KOR" THEN ACOUNTRY = "Korea";
  ELSE IF COUNTRY = "TWN" THEN ACOUNTRY = "Taiwan";
  ELSE IF COUNTRY = "USA" THEN ACOUNTRY = "USA";
  ELSE PUT "Unknown value of variable COUNTRY: " country "at observation " _N_;
%MEND;
```

#4 FACILITATING COLLABORATION BETWEEN MULTIPLE USERS

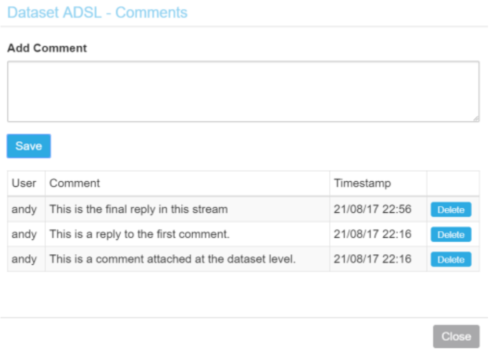
Currently, most approaches to managing metadata do not facilitate collaboration between users. Information is fragmented across multiple non-standardised documents. People working at one stage of the process have little awareness of how their deliverables impact their colleagues working at other stages. This type of approach may have been suitable when projects used waterfall project management methodologies, but as sponsors transition to Agile and modern ways of working, it is no longer appropriate.

Clinical Metadata promotes collaboration between users by:

1. Ensuring that all users always access the latest version of the specifications.
2. Using in-built communication tools, meaning information is exchanged between users via the system, rather than being hidden in emails.
3. Providing project management utilities so that users can clearly understand the status of each stage of the project.
4. Allowing the user to understand how metadata within the system has changed over time.

For example, in existing systems users might share comments about a particular variable by email. Anyone not in the email chain would not be aware of the issues being discussed.

Within Clinical Metadata comments are exchanged within the system, through a chat function, allowing everyone to see what is being discussed, by whom and when.



User	Comment	Timestamp	
andy	This is the final reply in this stream	21/08/17 22:56	Delete
andy	This is a reply to the first comment.	21/08/17 22:16	Delete
andy	This is a comment attached at the dataset level.	21/08/17 22:16	Delete

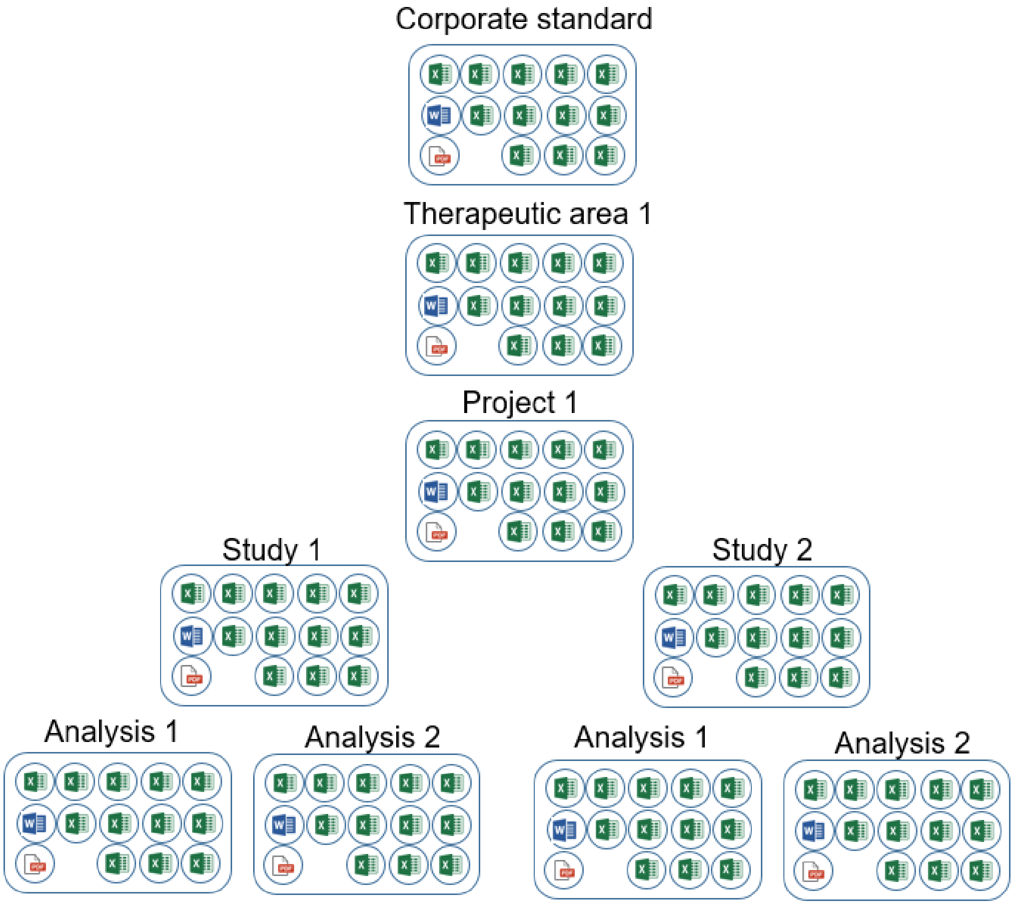
#5 MANAGING STANDARDS AND PROMOTING CDISC COMPLIANCE

In previous disconnected metadata systems, metadata at different levels of the standards hierarchy often existed in separate files. Importing or promoting metadata between levels usually relied on manual copying and pasting information between spreadsheets. Within clinical trials it is often desirable to create standard program code which could be re-used across studies. In order to do this users needed to understand which metadata elements were the same across studies. However, disconnected metadata management systems did not facilitate this sort of comparison.

Within Clinical Metadata, all metadata is held within one connected system built to facilitate the exchange and better understanding of metadata.

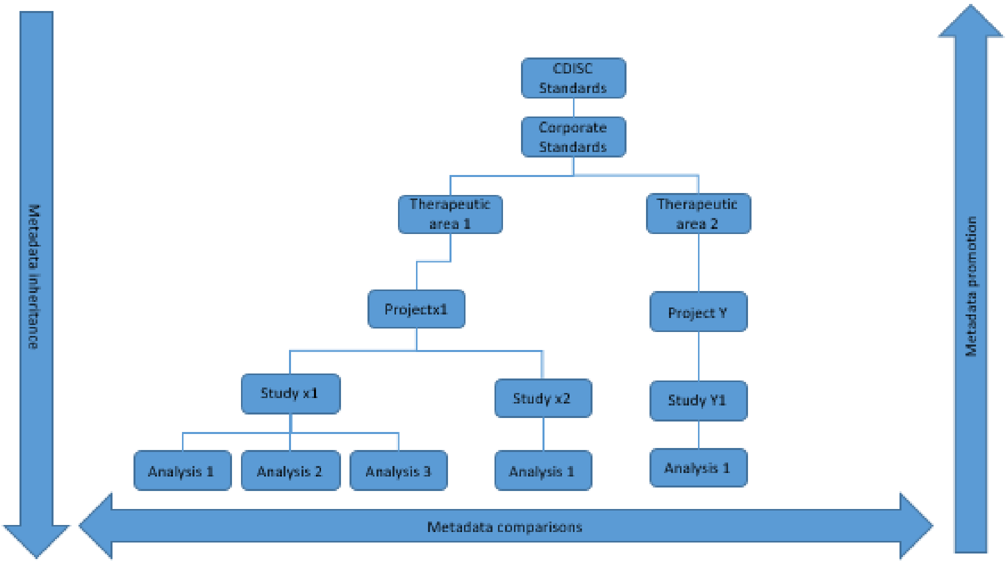
Consider for example a typical data standards implementation. Previously, information held at each level of the hierarchy comprised a multitude of disconnected files. Even before attempting to understand the global metadata environment, considerable time and attention had to be devoted to rudimentary file management.

For example consider the following data standards system for a sponsor, where we assume that at each analysis, study, project therapeutic area or corporate standard contains the files shown in the image on page 3.



It is easy to see how the number of disparate files quickly becomes unmanageable.

Within Clinical Metadata all metadata is held within one system. Metadata can be imported down through the standards hierarchy, or promoted upwards for evaluation through a governance process, increasing both accuracy and efficiency and helping, for example, to avoid the common scenario where different names are given to variables of the same meaning across studies.



CONCLUSION

We have explained how, in recent years, metadata and data standards have become an important part of the clinical trial process. Inefficient metadata management solutions have, however, meant that sponsors failed to realise the transformative possibilities of data standards.

The financial services industry has successfully transitioned from desktop based to web-based applications, and is now realising the benefits of this change in terms of cost, efficiency and simplification of workflows. A similar transition is required in the way sponsors manage their metadata.

In a *Business Case for CDISC Standards (2009)* Rozwell et al[4], the authors set out a convincing argument for CDISC as well as detailing some impressive efficiency savings that can be obtained through adopting CDISC standards. Many sponsors have failed to fully realise these potential efficiencies because they have implemented metadata management processes that are based on outdated technologies, leading to systems that are fragmented, hard to maintain, and with large amounts of data redundancy.

ClinicalMetadata.com is designed to overcome these problems. It leverages the latest technologies to deliver an integrated web-based, metadata management solution that enables the user to fully realise the benefits of CDISC and data standardisation.

REFERENCES

[1] Handelsman D (2005) *An Introduction to CDISC*. Retrieved from:
http://www.lexjansen.com/phuse/2005/cd/cd02.pdf

[2] The FDA (2016) *Study Data Standards, what you need to know.* Retrieved from:
https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM511237.pdf

[3] Winnpenny, G (2014) *What's new in ADaM?* Retrieved from: http://www.lexjansen.com/phuse/2014/cd/CD03.pdf

[4] Rozwell C, Kush R, Helton E, Newby F, Mason T, (2009) *Business Case for CDISC Standards.* Retrieved from:
https://www.cdisc.org/system/files/all/article/application/pdf/businesscasesummarywebmar09.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Author Name: Andrew Ndikom
Company: Clinical Metadata
City: London
Email: andrew.ndikom@clinicalmetadata.com
Web: www.clinicalmetadata.com

Author Name: Liang Wang
Company: Clinical Metadata
City: London
Email: liang.wang@clinicalmetadata.com
Web: www.clinicalmetadata.com