PhUSE 2014

Paper DH07

Study Data Handling: A CRO Experience

Paul Crean, ICON, Dublin, Ireland

ABSTRACT

As a CRO ICON receives data from multiple systems (internal and external) and vendors, processes it, and delivers it in a variety of formats. In recent years, ICON is delivering outputs to data consumers earlier in the study, more frequently, and in a greater variety of formats.

This paper will describe how the process of data handling at ICON has changed in light of these requirements, and the tools implemented to enable study teams to have a clearer picture of data processing in a study and to alert them of any potential data processing issues.

INTRODUCTION

In an effort to improve the safety and efficiency of the process of clinical research, approaches such as Risk Based Monitoring (RBM), Quality by Design (QbD) and Adaptive Clinical Trials are being used increasingly. The success of these approaches depends on earlier and more frequent data analysis than previously required.

A key enabler for these approaches has been the availability of new technology solutions to collect data in a timely manner, to integrate and consolidate data from diverse systems and data structures, and to deliver data through a variety of presentation tools. In addition to implementing new technology, the programming teams and processes for data handling need to become more efficient as the complexity, volume and rate of transfers are increasing.

RESTRUCTURED PROGRAMMING TEAMS

In the past the Data Management programming team was made up of DBAs – SAS® programmers who could also build collection systems.

A lead DBA was assigned to a study for the duration of the trial and was responsible for building the collection system and for programming all edit checks, listings, reports and transfers. The DBAs had detailed knowledge of the study data sources and outputs so that if, for example, a protocol amendment was needed they could easily assess the impact of the change on any study output.

With the introduction of more advanced technology and a wider range of data collection systems (due to the uptake of electronic data capture systems) plus the requirement for more complex and specialist outputs, the use of generalist DBAs was no longer viable. In light of this, the Data Management programming team was reorganized into specialist teams.

- **Clinical Data Management Systems (CDMS):** Responsible for database builds and validation for all studies; comprises pools of resources with in-depth knowledge of each EDC platform.
- **Clinical Data Services (CDS):** Responsible for the production of all clinical data outputs, including SDTM, sponsor defined data standards, listings, reports, visualisations and the continued development of ICON's clinical data warehouse.

Now, on award of a new Data Management study dedicated CDMS and CDS leads are assigned. The leads work directly with the study team and are responsible for coordinating all build and delivery related activities for the duration of the trial.

CONSISTENT DATA FLOW

As the scope and complexity of requirements grew the DM programming team reacted on a study-by-study basis, over time this ad-hoc approach caused inconsistencies in how programs, outputs and supporting documentation were managed across studies. Regional variations were also common with different programming approaches being used in each geographical location. These inconsistencies limited the opportunity for increasing efficiency through reuse of programming code.

To overcome this, a new global data flow and standard directory structure has been implemented that enforces consistent global processes for storing, accessing and delivering study data and outputs. Data and outputs for new studies are processed as follows:

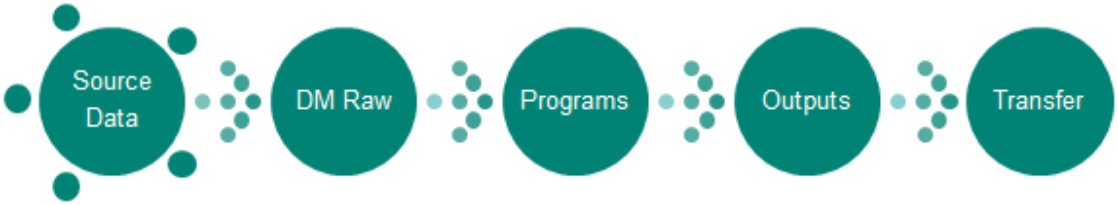


Figure 1: Consistent Data Flow

The directory structure is created from a template at the beginning of a study and additions to the structure can only be made with management-level approval.

Development and production-ready data, programs and outputs are stored separately in the folder structure. This separation acts as a safety net to prevent untested code and data from being mistakenly used as production standard.

The template consists of:

Folder	Content
0_docs\	This folder contains any study documents related to programs, including program sources, outputs and transfers. Documents are organized according to group (study team, CDMS, SDTM, ICONiK®, Listings etc.).
1_source_data\	All source data as it acquired from the study source systems, and associated extract, medical coding and data verification programs.
2_dmraw\	Contains the most recent collection of verified study data for data management use and is the source data folder for listings, SDTM, data transfers and ICONiK deliverables.
3_progs\	All programs, macros and associated logs for data management deliverables. Programs are organized according to type of deliverable.
4_outputs\	Study programming outputs organized according to type of deliverable.
5_transfer\	Historical copies of data transfers to internal and external users

SOURCE DATA HANDLING

The verification of all source data transfers has traditionally been a manual process in which the assigned programmer inspected the transferred files to confirm they were compliant with the agreed source data specification document. The programmer verified, at a minimum that:

- The number of files transferred was the expected number.
- The structure of datasets and the format of variables within matched the data specification.
- The observation count for each dataset was within expected limits.

If any issues were found, the DMPM for the study was notified and the data was rejected.

The manual process is no longer viable because of increased complexity and variety in data collections systems and the considerable increase in data transfer frequency required for near real-time analysis. In view of this, a new process has been developed that programmatically evaluates source data before release for use by Data Management and programming teams.

SOURCE DATA HANDLING SETUP

The DMPM is responsible for ensuring that all source data providers supply a valid source data specification document.

The specifications document details the structure of expected files, including variables, variable formats, variable descriptions and key/header variables. The specification is manually compared to a test transfer from the data provider, and if the transfer complies the specification document is signed and versioned.

The test transfer is given to the assigned CDS lead who programmatically generates a versioned source data definition file. The definition file is a metadata description of the test transfer.

A source data definition file is required for each study source (EDC, PK, lab, etc.).

SOURCE DATA HANDLING PROCESS

The operational process involves three stages:

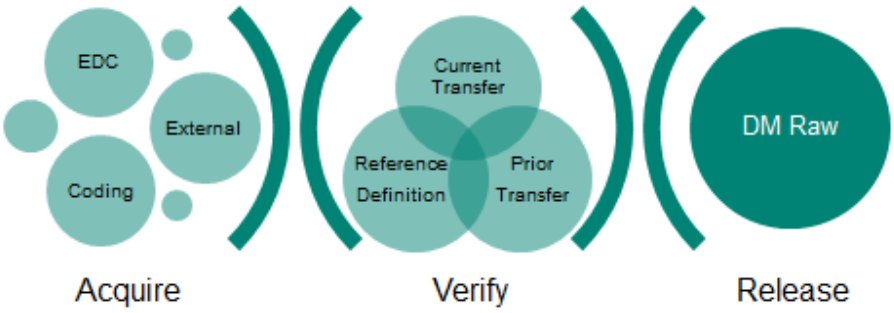
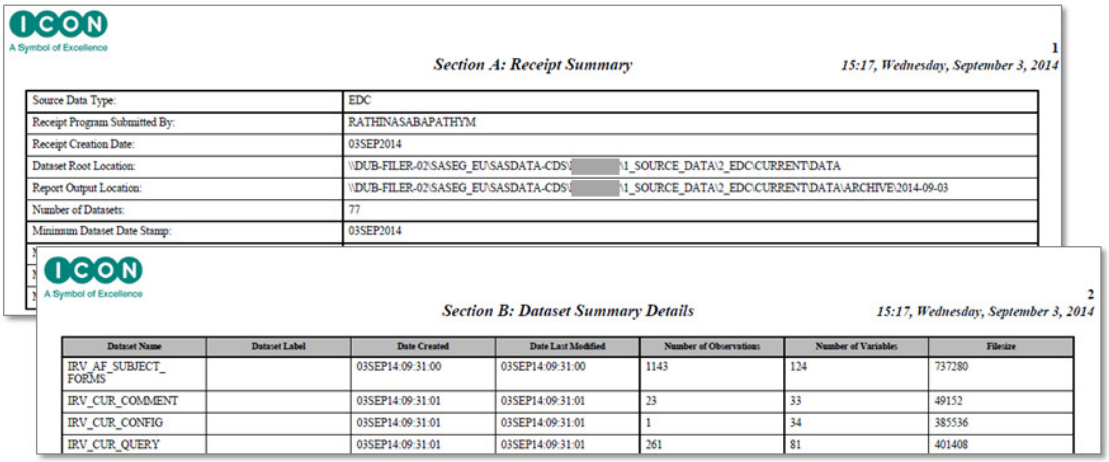


Figure 2: Source Data Handling Steps

- **Acquire Data:** The transfer file is saved to a dated archive folder, and if necessary data is extracted and converted to SAS dataset(s) in the archive folder.

  A data transfer receipt is programmatically published to PDF and the associated metadata is written to a dataset in the dated archive.



ICON
A Symbol of Excellence

Section A: Receipt Summary — 1 — *15:17, Wednesday, September 3, 2014*

Source Data Type:	EDC
Receipt Program Submitted By:	RATHINASABAPATHYM
Receipt Creation Date:	03SEP2014
Dataset Root Location:	\\DUB-FILER-02\SASEG_EU\SASDATA-CDS\[redacted]\1_SOURCE_DATA\2_EDC\CURRENT\DATA
Report Output Location:	\\DUB-FILER-02\SASEG_EU\SASDATA-CDS\[redacted]\1_SOURCE_DATA\2_EDC\CURRENT\DATA\ARCHIVE\2014-09-03
Number of Datasets:	77
Minimum Dataset Date Stamp:	03SEP2014

ICON
A Symbol of Excellence

Section B: Dataset Summary Details — 2 — *15:17, Wednesday, September 3, 2014*

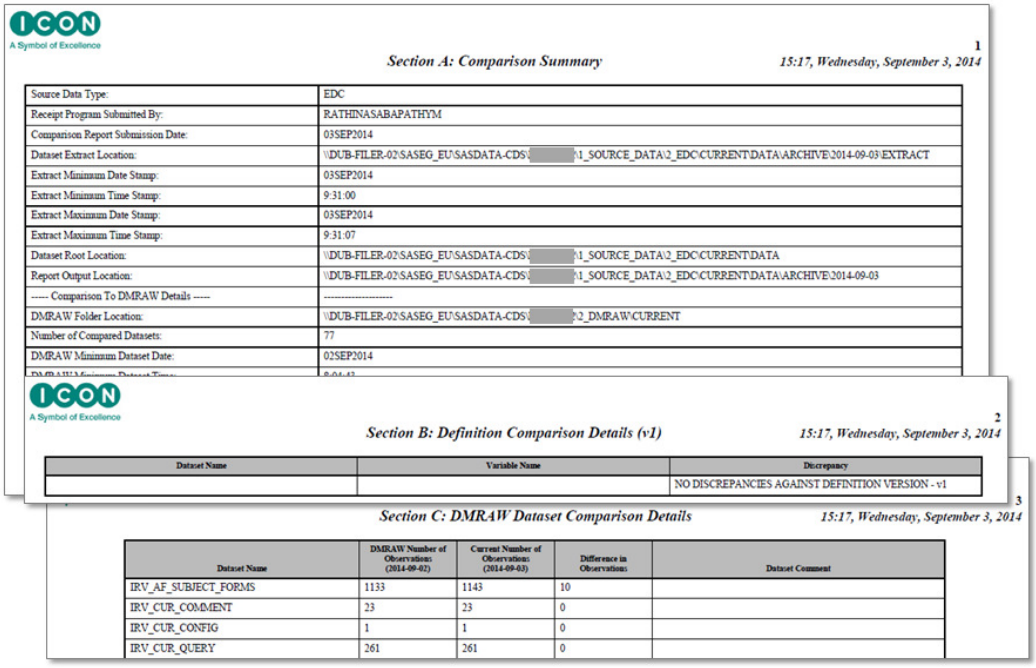
Dataset Name	Dataset Label	Date Created	Date Last Modified	Number of Observations	Number of Variables	Filesize
IRV_AF_SUBJECT_FORMS		03SEP14:09:31:00	03SEP14:09:31:00	1143	124	737280
IRV_CUR_COMMENT		03SEP14:09:31:01	03SEP14:09:31:01	23	33	49152
IRV_CUR_CONFIG		03SEP14:09:31:01	03SEP14:09:31:01	1	34	385536
IRV_CUR_QUERY		03SEP14:09:31:01	03SEP14:09:31:01	261	81	401408

Figure 3: Data Transfer Receipt

- **Verify Data:** The validity of the source data transfer is evaluated and a comparison report is generated describing the outcome of the verification.
  - **Data Structural Checks:** Metadata describing the transferred datasets is programmatically compared to the current source data definition to confirm that the data structure matches the agreed data specification.

    If the dataset metadata does not match the reference definition or no reference definition is found the transfer fails verification; data is not released for use and the reasons for the verification failure are documented in the comparison report (Figure 4).
  - **Observation Counts:** Observation counts for the transfer are compared to the observation counts for the most recent release of the current source type (e.g., EDC) to DM Raw, and if observations counts have dropped for any dataset the drop is noted in the comparison summary.

  The comparison summary is programmatically published to PDF and the associated metadata is stored in the dated archive. If the transfer passes the verification process, the extracted datasets are moved to the root of the relevant source data folder in preparation for release to DM Raw.



Section A: Comparison Summary — 15:17, Wednesday, September 3, 2014

Source Data Type:	EDC
Receipt Program Submitted By:	RATHINASABAPATHYM
Comparison Report Submission Date:	03SEP2014
Dataset Extract Location:	\\DUB-FILER-02\SASEG_EU\SASDATA-CDS\[redacted]\1_SOURCE_DATA\2_EDC\CURRENT\DATA\ARCHIVE\2014-09-03\EXTRACT
Extract Minimum Date Stamp:	03SEP2014
Extract Minimum Time Stamp:	9:31:00
Extract Maximum Date Stamp:	03SEP2014
Extract Maximum Time Stamp:	9:31:07
Dataset Root Location:	\\DUB-FILER-02\SASEG_EU\SASDATA-CDS\[redacted]\1_SOURCE_DATA\2_EDC\CURRENT\DATA
Report Output Location:	\\DUB-FILER-02\SASEG_EU\SASDATA-CDS\[redacted]\1_SOURCE_DATA\2_EDC\CURRENT\DATA\ARCHIVE\2014-09-03
----- Comparison To DMRAW Details -----	----------
DMRAW Folder Location:	\\DUB-FILER-02\SASEG_EU\SASDATA-CDS\[redacted]\2_DMRAW\CURRENT
Number of Compared Datasets:	77
DMRAW Minimum Dataset Date:	02SEP2014

Section B: Definition Comparison Details (v1) — 15:17, Wednesday, September 3, 2014

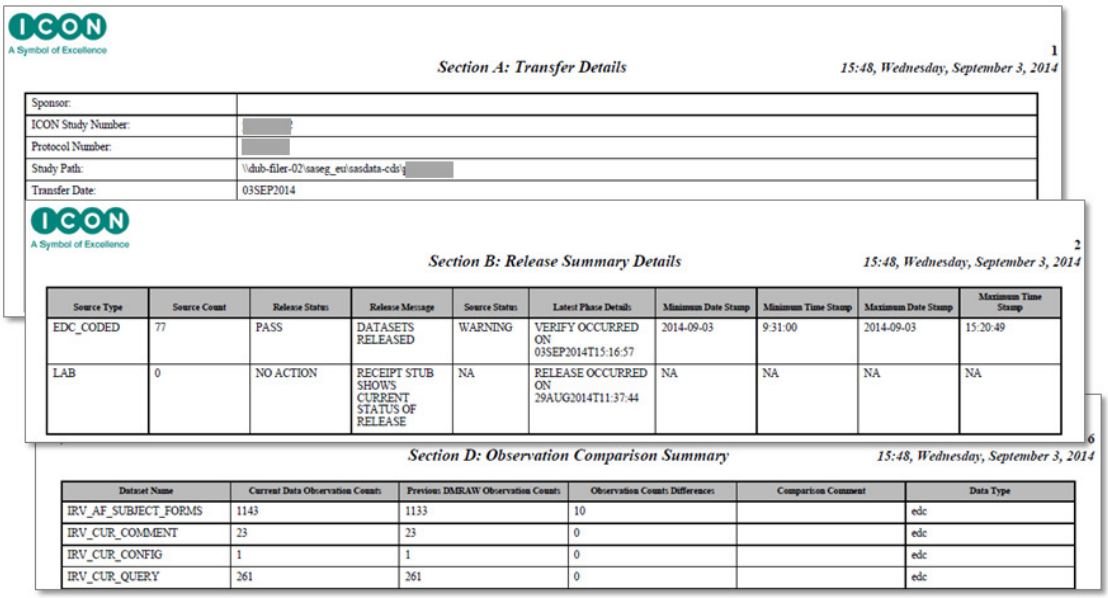
Dataset Name	Variable Name	Discrepancy
		NO DISCREPANCIES AGAINST DEFINITION VERSION - v1

Section C: DMRAW Dataset Comparison Details — 15:17, Wednesday, September 3, 2014

Dataset Name	DMRAW Number of Observations (2014-09-02)	Current Number of Observations (2014-09-03)	Difference in Observations	Dataset Comment
IRV_AF_SUBJECT_FORMS	1133	1143	10	
IRV_CUR_COMMENT	23	23	0	
IRV_CUR_CONFIG	1	1	0	
IRV_CUR_QUERY	261	261	0	

Figure 4: Comparison Report

- **Release Data:** An automated process loops through each source data folder searching for data that has not previously been released. If un-released data with matching verification metadata is found, the datasets are moved to the DM Raw folder. If no matching verification metadata is found, the data is not released.



Section A: Transfer Details — 15:48, Wednesday, September 3, 2014

Sponsor:	
ICON Study Number:	[redacted]
Protocol Number:	[redacted]
Study Path:	\\dub-filer-02\saseg_eu\sasdata-cds\[redacted]
Transfer Date:	03SEP2014

Section B: Release Summary Details — 15:48, Wednesday, September 3, 2014

Source Type	Source Count	Release Status	Release Message	Source Status	Latest Phase Details	Minimum Date Stamp	Minimum Time Stamp	Maximum Date Stamp	Maximum Time Stamp
EDC_CODED	77	PASS	DATASETS RELEASED	WARNING	VERIFY OCCURRED ON 03SEP2014T15:16:57	2014-09-03	9:31:00	2014-09-03	15:20:49
LAB	0	NO ACTION	RECEIPT STUB SHOWS CURRENT STATUS OF RELEASE	NA	RELEASE OCCURRED ON 29AUG2014T11:37:44	NA	NA	NA	NA

Section D: Observation Comparison Summary — 15:48, Wednesday, September 3, 2014

Dataset Name	Current Data Observation Counts	Previous DMRAW Observation Counts	Observation Count Differences	Comparison Comment	Data Type
IRV_AF_SUBJECT_FORMS	1143	1133	10		edc
IRV_CUR_COMMENT	23	23	0		edc
IRV_CUR_CONFIG	1	1	0		edc
IRV_CUR_QUERY	261	261	0		edc

Figure 5: Release Report

After looping through all source data folders, the release routine publishes a release report to PDF (Figure 5) containing details of datasets released, datasets not released and any warnings encountered.

COMMUNICATION AND TRANSPARENCY

At all stages in the process, versioned PDFs and metadata are created by the process describing what data has been processed, when it was processed, the quantity of data, the validity of the data and the data structures. This information is retained for the duration of the study and is available in the study folder for review by all study and programming team members.

A Windows PowerShell® script harvests the study-level metadata daily and consolidates it in a central repository for publication. We are in the process of developing a data processing dashboard published through TIBCO Spotfire®

to allow for centralized monitoring of data processing issues. The initial release of the dashboard focuses on the final release process, tracking data growth by study and dataset.

IMPACT ANALYSIS OF POST PRODUCTION CHANGES

Post production changes, including protocol amendments, are a fact of life in clinical trials and despite the fact that the introduction of EDC systems has simplified the process of implementing changes from a deployment perspective, for data consumers there are costs to be considered as these changes inevitably necessitate further changes in downstream systems. This cost rises the later in the trial the change occurs.

Within the new standard structure, test and production ("current") data are stored separately in the source data folder (Figure 6) with the underlying folders being structural mirrors. This allows us to leverage some of the functionality of the source data handling programs to identify the actual structural changes that a post production change will cause and assess impact accordingly.

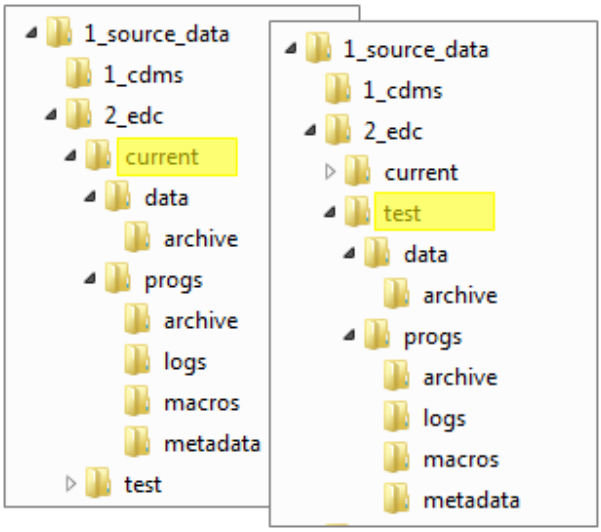


Figure 6: Source Data Folder

To analyze the differences, the programmer saves the post-amendment test data to the test folder and runs the "acquire" and "verify" steps of the data handling process in the test location. The verify fails because of structural changes and the generated comparison report will list the structural differences (Figure 7).

ICON
A Symbol of Excellence

Section B: Definition Comparison Details (v2) *10:36, Thursday, May 15, 2014*

Dataset Name	Variable Name	Discrepancy
IE_RAND	IE_RAND2_STD	STRUCTURE DIFFERS TO DEFINITION
IE_RUNIN	IE_RUNIN3	STRUCTURE DIFFERS TO DEFINITION
IE_RUNIN	IE_RUNIN3_STD	EXISTS IN DEFINITION, NOT IN CURRENT
KIT_ASSIGN	KIT_ASSIGN3	STRUCTURE DIFFERS TO DEFINITION

Figure 7: Comparison Report with Amendment

This report is distributed to downstream data consumers to allow them to prepare for the post production change.

The programmer creates an updated source data definition file, and when the change goes to production this updated definition file is used to confirm that the actual change implemented matches the change that was communicated as part of the impact assessment.

CHANGES IN DEVELOPMENT

The source data handling process described in this paper is implemented on new studies that require daily refreshes of source data, it is expected that the process will be further refined as it is extended to other studies. Besides these refinements development has started on other enhancements to the data handling process.

TRANSFER DATA RELEASE PROCESS

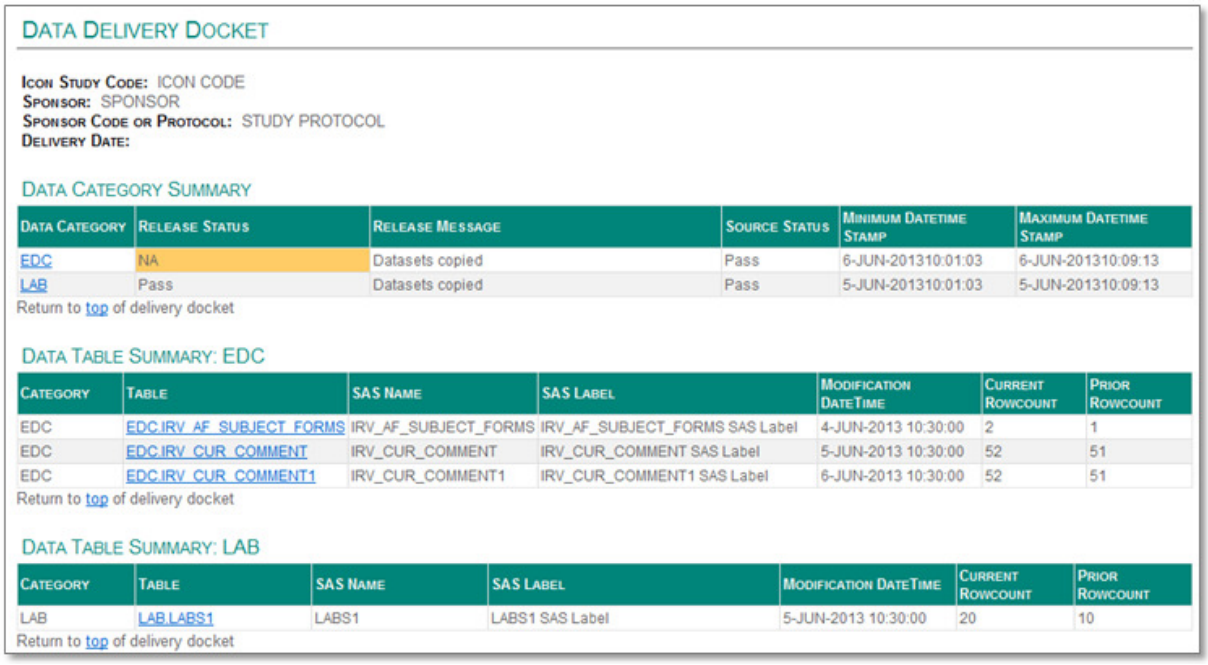
There are a number of study-specific programs being used that generate data transfer specifications before data is distributed to non-DM users. To replace these programs, we are developing a standard automated process for the verification of outputs before release to the "transfer" folder.

Functionally this is very similar to the source data verification process.

SWITCH TO XML FOR DATA HANDLING ARTIFACTS

The source data handling process as it is currently implemented stores metadata in SAS datasets and CSV files, and reports are published through PDF. The next release of the process will use XML for metadata storage and an XSL will be used for presenting the summary.

Figure 8 contains a screenshot of the latest draft of transfer data delivery docket that is in development.



DATA DELIVERY DOCKET

ICON STUDY CODE: ICON CODE
SPONSOR: SPONSOR
SPONSOR CODE OR PROTOCOL: STUDY PROTOCOL
DELIVERY DATE:

DATA CATEGORY SUMMARY

DATA CATEGORY	RELEASE STATUS	RELEASE MESSAGE	SOURCE STATUS	MINIMUM DATETIME STAMP	MAXIMUM DATETIME STAMP
EDC	NA	Datasets copied	Pass	6-JUN-201310:01:03	6-JUN-201310:09:13
LAB	Pass	Datasets copied	Pass	5-JUN-201310:01:03	5-JUN-201310:09:13

Return to top of delivery docket

DATA TABLE SUMMARY: EDC

CATEGORY	TABLE	SAS NAME	SAS LABEL	MODIFICATION DATETIME	CURRENT ROWCOUNT	PRIOR ROWCOUNT
EDC	EDC.IRV_AF_SUBJECT_FORMS	IRV_AF_SUBJECT_FORMS	IRV_AF_SUBJECT_FORMS SAS Label	4-JUN-2013 10:30:00	2	1
EDC	EDC.IRV_CUR_COMMENT	IRV_CUR_COMMENT	IRV_CUR_COMMENT SAS Label	5-JUN-2013 10:30:00	52	51
EDC	EDC.IRV_CUR_COMMENT1	IRV_CUR_COMMENT1	IRV_CUR_COMMENT1 SAS Label	6-JUN-2013 10:30:00	52	51

Return to top of delivery docket

DATA TABLE SUMMARY: LAB

CATEGORY	TABLE	SAS NAME	SAS LABEL	MODIFICATION DATETIME	CURRENT ROWCOUNT	PRIOR ROWCOUNT
LAB	LAB.LABS1	LABS1	LABS1 SAS Label	5-JUN-2013 10:30:00	20	10

Return to top of delivery docket

Figure 8: Data Delivery Docket.XML (In Development)

CONCLUSION

The push to near-real time analysis of trial data initially raised some challenges, but ultimately it has resulted in the introduction of more efficient processes within the Data Management programming team. It has also lead to an increased focus on transparency and communication so that we can give study teams a more complete picture of the status of data processing in a study and alert them to any potential data handling issues.

ACKNOWLEDGEMENTS

The development of the source data handling process and associated automations was a team effort. It would not have been successful without the input of the following people.

- Stephen Robertson (Senior Clinical Data Developer): Lead developer for source data handling automation.
- Margaret Skrzypczak (Clinical Data Developer): Lead tester for source data handling automation.
- Louise Kenny (Senior Clinical Data Delivery Lead): Design of impact analysis process.
- Niall AlRubeai (Clinical Data Developer): Designer of data processing dashboard.

My sincere thanks to each of them for their help.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author(s) at:

Paul Crean
ICON
South County Business Park
Leopardstown, Dublin 18, Ireland
Work Phone: +353 (0) 1 291-2000
Fax: +353 (0) 1 291-2200
Email: paul.crean@iconplc.com
Web: www.iconplc.com