

A simple tool to catalogue statistical outputs developed for submission by linking two in-house systems – experience from a submission project

Katja Diezel

Clinical Data Review & Reporting, Novartis Pharma AG, Basel, Switzerland

Introduction

During the submission process for an in-licensed compound a large quantity of tables, listings, figures and reports were produced. As the project grew the team felt the need for a user-friendly repository for such outputs.

Different documents are referenced in order to gain an overview of the project:

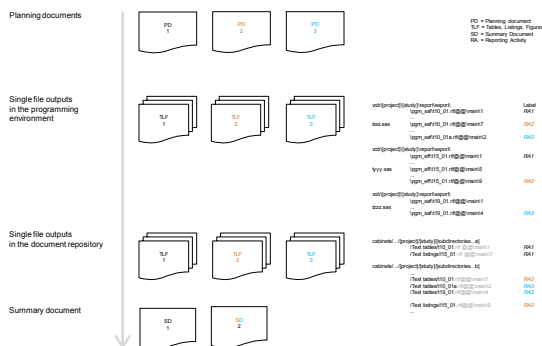
Planning documents (e.g. Statistical Analysis Plan and associated output shells, tracking documents of SAS programs) cover planned outputs per reporting activity. Those planning documents list planned outputs, but outputs might not have been programmed yet at a given point in time when the reporting is still ongoing. Information is spread out over multiple documents: one set of planning documents is available per reporting activity (CSR, SCS, SCE, 120 Day update, etc.).

Single file outputs in the programming environment are available once programmed and validated. The same program might be used for several reporting activities. In this situation the program would run on different versions of the datasets and for each set of data an output file is created. These output files might have identical file names and will differ in file version number, label and content.

For each reporting activity, once an output has been approved by the statistician the output is transferred from the programming environment to the document repository.

Single file outputs in the document repository are stored in different directories depending on the reporting activity and publishing needs. The document repository has a company wide pre-defined structure.

Summary documents (e.g. Clinical Study Report, Summary of Clinical Safety, Summary of Clinical Efficacy, etc.) will contain all single file outputs programmed, compiled into a report. They become available only some time after the production of the outputs due to time needed by Medical Writing and Publishing group to generate these documents. Tables, listings and figures produced may spread across multiple summary documents.



To bridge the gap that results from searching for outputs in the different referencing documents, the team was looking for a tool that:

- lists all outputs that have already been produced and validated
- flattens out the vertical structure of the programming environment
- pulls together files spread out across multiple subdirectories in the document repository

Output catalogue [Information collected from Programming environment / Document repository]

RA1	cabinets/_/project/study/subdirectories...a)Text tables	110_01_rf@@main1	<hyperlink>	Table 10.1	Summary table
RA1	cabinets/_/project/study/subdirectories...a)Text listings	110_01_rf@@main7	<hyperlink>	Listing 15.1	Data listing
RA2	cabinets/_/project/study/subdirectories...b)Text tables	110_01_rf@@main2	<hyperlink>	Table 10.1a	Summary table
RA3	cabinets/_/project/study/subdirectories...b)Text tables	110_01_rf@@main2	<hyperlink>	Table 10.1a	Summary table - Subgroup
RA3	cabinets/_/project/study/subdirectories...b)Text tables	110_01_rf@@main4	<hyperlink>	Table 18.1	Frequency table
RA2	cabinets/_/project/study/subdirectories...b)Text listings	110_01_rf@@main9	<hyperlink>	Listing 15.1	Data listing

In short, the team was looking for a simple and easy to update tool that can be used by programming and non-programming team members which serves as a catalogue of outputs produced during the submission.

Background - Systems

- GPS II (Global Programming and Statistics) is the standard in-house programming environment
- used as a repository for data imported from in-house or external sources and for generating outputs (e.g. tables, listings)
- GPS II environment incorporates Version Control using Rational ClearCase software which records file history and provides traceability to guarantee 21CFR Part 11 compliance



- **Version Control in a nut-shell**
 - each ClearCase VOB stores version-controlled file system objects (elements)
 - an element is a file or directory for which ClearCase maintains multiple versions
 - each version of an element has a unique version-id
 - any version of an element can be uniquely specified by appending its version-ID to its standard pathname

Document repository

- CREDI (Clinical Research Documentation and Information System) is a global document management system
- it is a Documentum 6.5 Docbase with a web-based client interface for managing documents
- users can access basic services, such as creating documents and participating in lifecycles and workflows using EMC Documentum Webtop

- **Content Management in a nut-shell**
 - a document management system such as Documentum coordinates the changes and availability of business critical documents on a global basis
 - two kinds of information for a document are stored: Content (e.g. Text, Graph) = information in the document Properties (e.g. Title, Owner, Status) = information about the document

Technical workflow

- 1) Identify project directories that are active and/or part of the submission project.
- 2) Retrieve detailed information on upload to document repository (CREDI) by screening through utility files and all their historic versions.

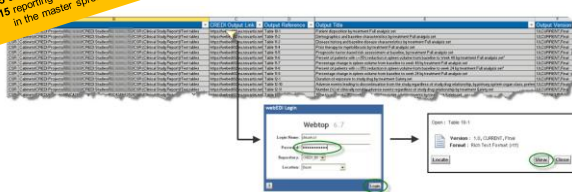
- 3) Obtain information from document repository (CREDI):
 - 1) Extract file properties of documents in relevant subdirectories in the document repository as identified in previous step.

- 4) Combine information from programming environment (GPS) and document repository (CREDI):
 - 1) Read in data from both sources (*.txt, *.csv)
 - 2) Combine information and create SAS dataset (CREDI information on directory and file properties, GPS upload information)
 - 3) Manipulate data
 - (Create hyperlinks, prepare master spreadsheet, prepare summary reports, etc.)
 - 4) Create reports using SAS ExcelXP tagset

Benefits

Line Function	Information provided	Information used for	Impact
Statistical Reporting - Project level	Access to all information extracted from both systems. Access to summary reports on process adherence. Access to master spreadsheet	- project status and process adherence checks can be automated - monitor process adherence and project status done more consistently across project	- quality, consistency of SR deliverables - process awareness and training within the team
Statistical Reporting - Reporting activity level	Access to master spreadsheet	- checking if a newly requested output is truly new - identify programs/macros/outputs to be reused within and across reporting activities - ensure consistency of outputs by referring to tables already done	- increased speed - increased consistency - increase quality due to re-use of program and macros code

Output catalogue: Master spreadsheet



Line Function	Information provided	Information used for	Impact
Biostatistics	Access to master spreadsheet	- rapid access & overview of already delivered outputs - check if an output request is truly new - verification of outputs across different reporting events - draw shells based on existing outputs	- speedier cross verification - quicker generation of shells - consistency of shells across project - consistency of analysis across project
Outside BIO&SR	Information of master spreadsheet embedded in communication, e.g. subset of spreadsheet, hyperlinks (to avoid misleading conclusions outputs are presented in context)	- identify existing outputs that may answer a question - identify new outputs to be generated - reference outputs by hyperlinks instead of distribution of physical files	- quicker access to outputs produced - shorter turn-around time for newly requested outputs - easier use of document repository

Conclusion

Programming does not stop at the clinical data level !

Thinking outside the clinical data reporting box, SAS programmers can create new opportunities to work together and ease the daily burden for everyone on the team.

Joining information from two in-house systems using SAS, a simple tool was created to allow for the referencing of outputs and relevant information, thus increasing the speed, productivity and quality of output production.