**Paper PP22**

Annotate CRFs and Keep Track of them

Rohit Banga, Biometrical Practice BIOP AG

ABSTRACT

Keeping track of CRFs and managing Data entry is a daunting task. It requires considerable man power and often there is a danger of CRF pages not getting entered in the database. With this paper I wish to demonstrate a novel way of annotating and indexing CRFs and then building a database of pages which have been annotated. This database when combined with data extracts can produce status reports which reflect an accurate status of CRFs received and CRFs entered.
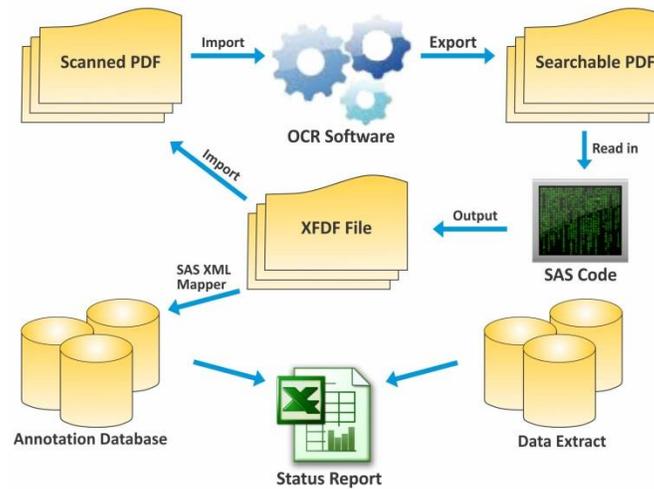
The process is as follows. The CRF page is first scanned and passed through OCR software. The OCR software converts the scanned CRF into a text file. This text file is then read in by SAS and an algorithm determines which scanned PDF page belongs to which panel/visit/page. Using this information an XML Forms Document file (XFDF) file is programmed. This XFDF file can be read in by Adobe Acrobat software and annotations are imported inside the PDF. The XFDF file, essentially being an XML file, can also also be read in by SAS XML Mapper and can populate a database of pages that have been received.

When this database is combined with data extracts from Clinical database, a detailed status report emerges. It has details of pages that have been received by the Data Management department and pages that entered in the database. One can drill down to country/site/patient level and know the exact status of Data Entry.

INTRODUCTION

Maintaining track of CRF pages that have received by the CRO and entered in the DB System is traditionally done manually. In this paper we present a way to automate the process that has been implemented by us. The different steps are outlined below

STEPS TO ANNOTATE AND TRACK CRFs

STEP 1 – SCAN THE CRF
When a CRF is received, it is scanned and converted into a PDF file [1]. The PDF file is then stored in a central location in a folder named after the subject identifier. All files related to a particular subject such as CRFs, DCFs (Data clarification forms) or questionnaires are stored in the same location. Storing PDF files of a particular subject together helps organize the files in a better way.
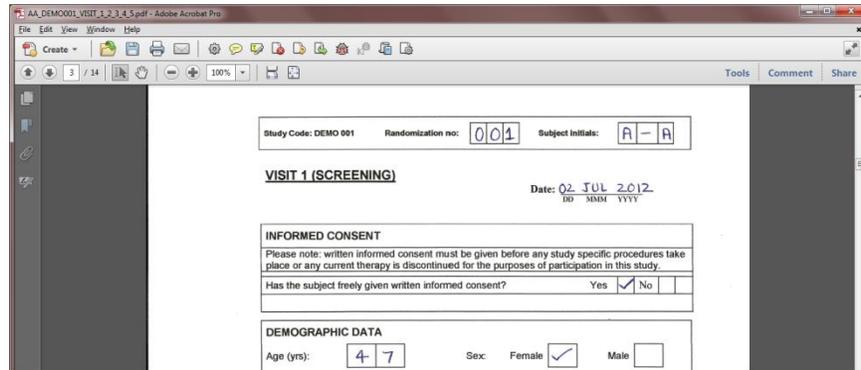
STEP 2 – PASS the PDF through OCR (Make them SAS Readable)
PDF is proprietary software from Adobe and a file in PDF format cannot be directly read by SAS. Moreover the scanned CRF is stored as an image and not as text. Therefore, we need optical character recognition (OCR) software to convert the scanned image into a format that can be read and interpreted by SAS. We prefer to convert the PDF file into an accessible text (text file with line breaks, *.txt) format.
There are many OCR software available in the market, some free and some paid. However, in our opinion one should go for an option in which the conversion to text can be automated (or run as a batch job) so the multiple PDF files can be processed at once. Deciding on good OCR software is a very important part of the process; if the character recognition from image to text is not good it will be difficult to write a SAS program to interpret it. In our organization we use OmniPage OCR software from Nuance [2]

STEP 3- Read the text file into SAS
The next step is to read the text file created in step 2 in SAS. This is done to interpret the contents of the scanned PDF File. The unfortunate reality is that no OCR software is perfect. The conversion of an image into a text file perfectly all the time is an impossible task. If the investigator crosses a part of the text, overwrites something or in general doesn't write nicely then the text is garbled. There are blocks of readable text intermixed with gibberish.

To overcome this challenge we developed a lookup table which had keywords corresponding to a combination of CRF Panel, Visit Numbers and Page numbers. It can look something like this –

CRF Panel	CRF Page Number	CRF Visit Number	CRF Visit Name	Key Words
DEMOGRAPHICS - DM	3	1	VISIT 1 (SCREENING)	DEMOGRAPHIC DATA ## INFORMED CONSENT ## written informed ## consent must be ## purposes of participation ## freely given written ## Age ## Male ## Body Mass Index ## SMOKING HABITS ## cigarettes ## ALCOHOL CONSUMPTION ## MEDICATIONS TAKEN
MEDICAL HISTORY - MH	4	1	VISIT 1 (SCREENING)	PREVIOUS MEDICAL HISTORY ## relevant medical ## Cardiovascular ## Respiratory ## Endocrine ## Haematological ## Neoplasia ## Neurological ## Immunological ## Allergies ## If giving details of surgery

The SAS program would read the text file, see the panel in which maximum keywords fit and store the result in a SAS Dataset. The keywords are carefully chosen. An effort is made to record as many unique keywords in a particular CRF page as possible. It is not a problem even if a keyword repeats for different panels. The SAS program selects the panel in which maximum number successful keyword matches are found.

STEP 4 – Create XFDF file. Make Annotations.
The SAS dataset created in the above step is used to create an annotation file (*.xfdf) for adobe acrobat. An XFDF is technically an Acrobat Forms Document file. For all practical purposes a XFDF file is essentially a XML file. When this XFDF file is exported inside adobe annotations start to appear in the PDF file. Keep in mind that only Adobe Professional can import/export XFDF file and the freely available adobe reader doesn't have this functionality.
To know how an XFDF file is organized, you can go through the document referenced in point [3] at the end of the paper.
If you do not have enough time or patience to go through the entire XFDF specification, you can take the smart way out. You can manually annotate the PDF file (look in the top left hand side in the example below). Then export the annotation using the option in comments section – 'Export all to data file'.
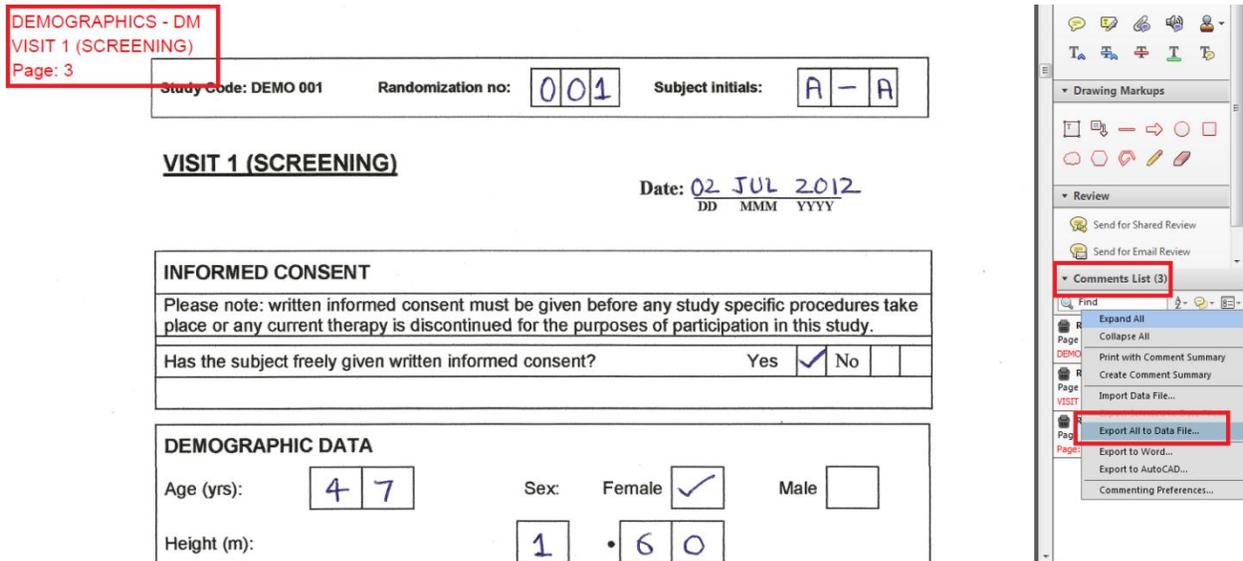
Once you understand how the XFDF file is organized and how it works for one annotation file, you can program it to put annotation on the entire PDF. The content of the XFDF file is determined by the text interpreted by the SAS program in the previous step. When the programmed XFDF file is ready, it can be imported inside the PDF document using the 'Import Data File' option in the comment section of Adobe Professional.

STEP 5 – Begin to make Status Report. Read XFDF Files
The XFDF files created in the above step serve a dual purpose. They are not only read in by Adobe to put annotations on PDF files, but they can also be imported by SAS to create a database of all indexed files.
The XFDF files can be read by SAS just like a regular xml file. SAS XML mapper is used to create a map file which maps the xml tags to SAS datasets and variables. All the XFDF files can be imported using a SAS macro and a database of the indexed CRF Files is created.
This is the first half of making a report to track CRFs. Till now we have made a list of all PDF files which have been received, scanned & annotated by our organization.

STEP 6 – Extract data from Database System. Create Status Report.
The next step is to extract data from DB Repository. A list of forms and pages that have been entered in the clinical database are extracted. In our case, the DB system was OpenClinica and it was quite straightforward to extract list of all forms that were entered. This data was merged with the indexed database created in the previous step and Status Reports were created.

From the Status Reports it became clear which CRF pages were indexed and what was the Data Entry Status of a particular page. We were also able to tell whether a CRF page had not been entered within 48 hours and we could take appropriate action with the data entry team.
An example of Status Report is shown below -

DEMO 001
CRF Life Cycle of SubjectID - DEMO-001

Patient number	Visit Name	Visit number	CRF Page Number	Panel	Date Indexed	Data Entry Status	Not Entered in 48 hour
DEMO-001	SCREENING VISIT (Visit 1)	1	Page: 2	Demographics - DM	1-Jan-13	data entry complete	
DEMO-001	SCREENING VISIT (Visit 1)	1	Page: 3	Medical History - MH	1-Jan-13	initial data entry	
DEMO-001	SUMMARY VISIT	888	Page: 4	Adverse Events - AE	2-Feb-13		x
DEMO-001	VISIT 2	2	Page: 12	Vital Signs - VS	10-Feb-13		x
DEMO-001	VISIT 2	2	Page: 12	Laboratory Tests Results - LB	10-Feb-13	data entry complete	
DEMO-001	SUMMARY VISIT	888	Page: 5	Concomitant Medications - CM	14-Mar-13	data entry complete	
DEMO-001	VISIT 3	3	Page: 15	Laboratory Tests Results - LB	15-Mar-13	data entry complete	

Reference:
[1] PDF files, Adobe Portable Document Format - http://www.adobe.com/products/acrobat/adobepdf.html
[2] Omnipage - Document Conversion Software from Nuance http://www.nuance.com/for-business/by-product/omnipage/index.htm
[3] XML Forms Data Format (XFDF) Specification -
http://partners.adobe.com/public/developer/en/xml/XFDF_Spec_3.0.pdf

CONTACT INFORMATION

Rohit Banga
Biometrical Practice BIOP AG | Biometrical Practice India Pvt. Ltd
Centralbahnstrasse 9
CH–4051 Basel, Switzreland
Email – rohit.banga@biop.ch
www.biop.ch | www.biop.in

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.