

A metadata-driven system to generate annotated CRF automatically

Guojun Liufu, Jun Wu, Li Chen, Zhen Peng, Fushan Lai, Hanjia Zheng, Clinical Trial Service (Guangzhou) Co.,Ltd

ABSTRACT

Annotated CRF (aCRF) plays a pivotal role in Study Data Tabulation Model (SDTM) generation and supporting regulatory submissions. However, aCRF generation can be work-intensive in mapping domains and variables to CRF contents, manually adding annotations, and adjusting formatting. Especially when CRF version updates, it has to spend repeated work for adjustments. It will make the generation process too inefficient, with a high error rate, inconsistency among projects, and the inability to reuse annotated contents.

We have developed a metadata-driven and automatic aCRF generation system. The system adapts to mainstream Electronic Data Capture (EDC) systems, and could efficiently generate aCRF by matching EDC library information with backend annotation database. The system is not limited to SDTM versions, languages, etc. As annotation libraries accumulate, accuracy can be continuously improved, greatly enhancing efficiency, accuracy, and reusability. In addition, SDTM specifications and SDTM programs can be efficiently generated by utilizing the metadata generated during the annotation process, establishing a solid foundation for SDTM automation.

INTRODUCTION

aCRF plays a pivotal role in SDTM generation and supporting regulatory submissions.

Traditional methods of creating aCRFs involve manual annotation, which is inefficient and prone to errors. Some companies have sought more efficient methods, including: adding annotations to library files, using Python packages to store CRFs in XML format, merging annotations with XML to create XFDF files, and importing these files into CRFs; using R's tm package, PyMuPDF package, or other tools to extract text information from CRFs, obtaining its coordinates in PDFs, and adding annotations based on these coordinates; storing PDF information in hash trees, converting it to JSON format, and adding annotations based on this format before replacing it in CRFs.

These tools have addressed the issue of manually adding annotation text boxes, particularly reducing the time spent on format adjustments. However, they still rely on manual annotation and cannot automatically match appropriate annotations based on CRF field information. Moreover, they require the use of multiple software languages, lacking an advantage in user experience.

To address these issues, we have developed a metadata-driven automated aCRF generation system with the following advantages:

- **High Automation:** Efficiently generates aCRFs by matching EDC library information with a backend annotation database.
- **Self-iterative Optimization:** As the annotation database grows, the accuracy of annotations can continuously improve, allowing for internal self-optimization of the system.
- **Flexibility:** Not limited by EDC systems, SDTM versions, or languages.
- **Version Management:** When CRF versions are updated, the system can inherit annotations from previous aCRF versions, automatically matching them to the new CRF version, thus enabling quick management of version updates.
- **Team Collaboration:** The system is configured for project personnel and their permission settings, facilitating efficient multi-person collaboration.
- **User-friendly Interaction:** Based on a web system, the user interface is friendly, and all file operations (editing, viewing, updating) can be completed online.

- **Metadata Reusability:** Utilizes metadata generated during the annotation process to efficiently produce SDTM specifications and SDTM programs, laying a solid foundation for the automated output of SDTM datasets.

WORKFLOW OF THE SYSTEM

The primary inputs for the system are the EDC library information and blank CRF. Due to the varying structures of library information across different EDC systems, this information must be converted into a standard format. By matching the standardized library information with the backend annotation database, the system generates aCRF, which users can edit, view, and download online. To ensure consistency in the aCRF style, the annotations in the database are standardized.

Considering the differing eCRF page styles of various EDC systems, the system has been designed to accommodate and adapt to mainstream EDC systems such as RAVE and eCollect. This design significantly enhances the automation and accuracy of the aCRF generation process.

The metadata generated during the annotation process is utilized to continuously build and refine the annotation database. This self-iterative optimization model improves the precision of annotations over time. Additionally, using this metadata, the system can efficiently produce SDTM specifications and programs, laying a solid foundation for the automated output of SDTM datasets.

Figure 1 illustrates the workflow of the aCRF automatic generation system, including the aCRF production process and the system's self-iterative optimization process. The automated output process for SDTM (indicated by the dashed line) is a module under consideration for future development.

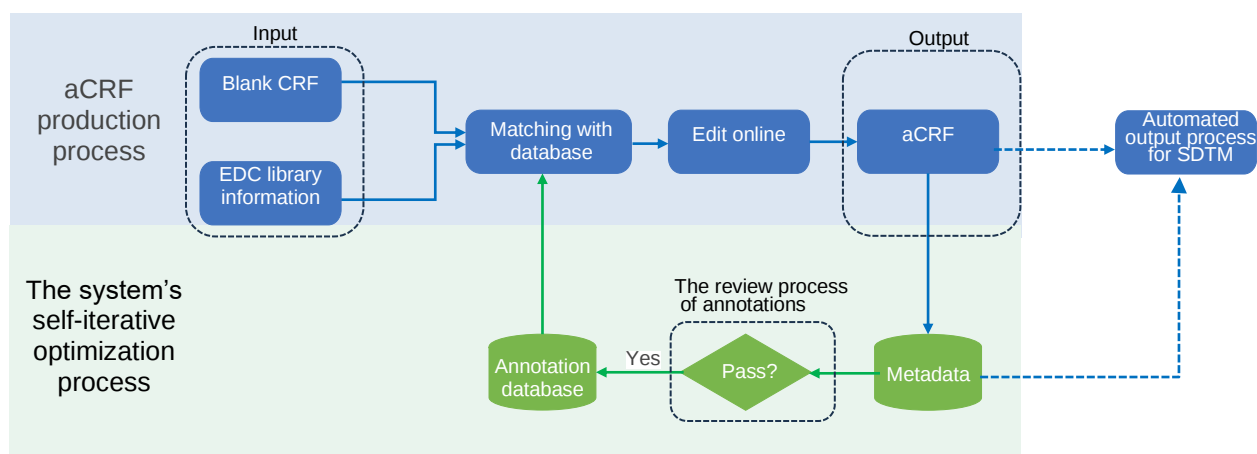


Figure 1. workflow of the aCRF automatic generation system

PRIMARY FEATURES OF THE SYSTEM

This section primarily outlines the functionalities of the aCRF automatic generation system and demonstrates the methods and effects of using these features through examples.

• Project Creation

When creating a project, users can specify parameters such as language, SDTM Implementation Guide (SDTM IG) version, and the corresponding EDC system. These details significantly enhance the accuracy of the system's matching with the annotation database. The system is configured for project personnel and their permission settings, facilitating efficient multi-person collaboration. In case of changes to project personnel or permissions, the system retains a record of these modifications. Figure 2 illustrates the interface for creating a project.

* 项目名称 test

* 语言 CN

* 版本 V3.2

* 项目成员 liufuguojun@citinc.org

* 对应系统

Figure 2. the interface for creating a project

- **Version Management**

Once a project is created, users can create versions of the aCRFs. This involves inputting the EDC library information and the blank CRF. In the event of a CRF version update or the need to upgrade the aCRF due to other reasons, users can choose to inherit from a previous version. The system automatically matches the annotation content from the inherited version to the new version of the CRF, enabling quick and efficient management of version updates.

Figure 3 and figure 4 illustrate the interfaces for creating and updating aCRF versions, respectively.

创建版本

* 版本名称: 1.0 3/200

* DBS文件 上传文件

* 空白CRF文件 上传文件

Figure 3. the interface for creating aCRF version

Figure 4. the interface for updating aCRF version

- **Generating aCRF automatically**

The system automatically generates annotation boxes based on the annotation database and matching rules. Users can preview the display style of the annotation boxes in real-time and also manually adjust the position, size, and other attributes of these boxes as needed. Developed as a web-based application, the system offers a user-friendly interface, allowing for all file operations (editing, viewing, updating) to be conducted online. This accessibility enhances user convenience and streamlines the workflow.

Figure 5 displays an example of an aCRF page automatically generated by the system.

Figure 5. an example of an aCRF page automatically generated by the system

If the generated annotation content is not suitable, users have the option to click on "Other Annotations" to select or manually add alternative annotation content, as illustrated in figure 6. Users can choose from other annotations matched by the system, search for additional annotations in the annotation database, or manually input new annotation content. After making these adjustments, users can instantly preview the effects of the modifications on the aCRF.

字段注释

注释库已有注释：

民族 其他民族

操作	字段名称	注释文本	域	类别	引用次数	备注
添加	其他民族	ETHNIOTH in SUPPDM	DM	1	3	/
添加	民族	ETHNICC in SUPPDM	DM	1	2	/
添加	民族	ETHNICC in SUPPDM	DM	2	2	/
		ETHNIOTH in SUPPDM				/

手动添加：

请输入注释内容

请选择

实践框

添加

ETHNICC in SUPPDM

取消

确定

Figure 6. an example of "Other Annotations" to select or manually add alternative annotation content

The system offers additional functionalities to further enhance the aCRF generation process:

1. **Annotation Box Types for Fields:** Users can select between different types of annotation boxes for fields, such as solid-line boxes and dashed-line boxes. This feature allows for customization of the aCRF based on specific annotation requirements.
2. **Submission Indication:** The system can indicate whether specific fields or annotations require submission.
3. **Summary of Unrecognized Content:** The system provides a summary of content that could not be automatically recognized or annotated. This feature aids in identifying and addressing any potential issues or missing annotations.
4. **Exporting aCRF:** Users can export the aCRF in PDF format.

These additional features contribute to the system's versatility and user-friendliness, ensuring that it meets a wide range of needs in the clinical research and pharmaceutical industries.

- **Metadata management**

The system is capable of exporting metadata generated during the annotation process. Utilizing this metadata, users can efficiently produce SDTM specifications and programs. This functionality significantly contributes to the automation of SDTM dataset output, streamlining the data tabulation process in clinical research.

Furthermore, the system incorporates an annotation review feature. If an annotation passes the review process, it can be added to the annotation database. This continuous process of review and addition enhances the precision and accuracy of annotations over time.

Figure 7 illustrates the interface for the annotation review feature.

项目:	请选择	状态:	待审核				
项目名称	表单	字段	字段注释	域	状态	注册人	操作
test	死亡记录	不良事件编号	DDAENUM in SUPPDO	DD	待审核	liufuguojun@cltinc.org	项目信息 入库 忽略
test	死亡记录	死亡日期	AA	DD	待审核	liufuguojun@cltinc.org	项目信息 入库 忽略

Figure 7. the interface for the annotation review feature

CONCLUSION

The core concept of the automated aCRF generation system is metadata-driven, utilizing the annotation database from previous studies to efficiently produce aCRF. As the annotation database grows, the accuracy of annotations can continuously improve, significantly reducing workload and minimizing errors. Additionally, the metadata generated during the annotation process can be effectively used to produce SDTM specifications and programs, laying a solid foundation for the automated output of SDTM datasets.

Currently, the annotation database of the system is still in the process of expansion. Future efforts will focus on continuously accumulating data and further optimizing matching rules to enhance the accuracy of annotations. Moreover, integrating the metadata generated by the system with subsequent SDTM production processes will be crucial to achieve the automation of SDTM dataset output.

REFERENCES

Pengpeng Li. 2023. "Automatic CRF Annotations Using Hash Tree Algorithm." PharmaSUG China 2023 - Paper AP-117.

Peng Wang, Yiqun Wang. 2021. "Automation for CRF Annotation and Bookmark Generation via Python." PharmaSUG China 2021 - Paper AD-040.

Teresa Tang. 2021. "Auto-annotation in SDTM aCRF Using R Package and SAS." PharmaSUG China 2021 - Paper AD-11.

ACKNOWLEDGMENTS

Sincere appreciation is extended to all individuals involved in the design, development, and testing of the system for their diligent efforts and contributions.

CONTACT INFORMATION

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

Guojun Liufu
Clinical Trial Service (Guangzhou) Co.,Ltd
liufuguojun.cltinc.org

Jun Wu
Clinical Trial Service (Guangzhou) Co.,Ltd
wujun.cltinc.org

Li Chen
Clinical Trial Service (Guangzhou) Co.,Ltd
chenli.cltinc.org

Zhen Peng
Clinical Trial Service (Guangzhou) Co.,Ltd
pengzhen.cltinc.org

Fushan Lai
Clinical Trial Service (Guangzhou) Co.,Ltd
laifushan.cltinc.org

Hanjia Zheng
Clinical Trial Service (Guangzhou) Co.,Ltd
zhenghanjia.cltinc.org

Any brand and product names are trademarks of their respective companies.