Long-term readability of data: Legacy software virtualization case-study

Natasa Milic-Frayling, Intact Digital Ltd

Extended Abstract **Poster at the 13th European Bioanalysis Forum (EBF) Open Symposium** **19 November 2020**

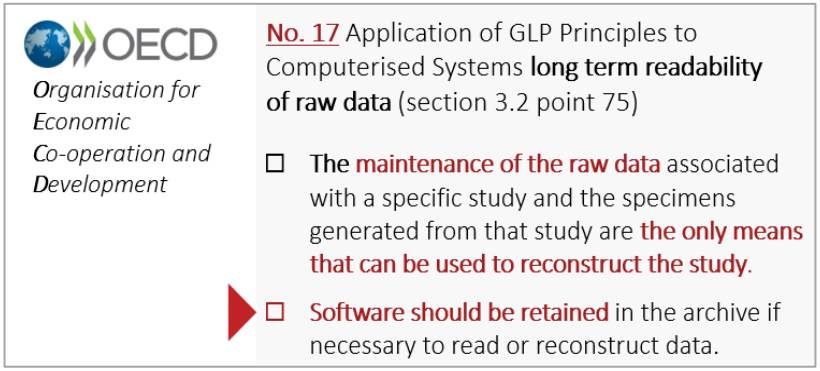
This poster is concerned with the long-term readability of digital data in order to enable reproducibility of research studies. At the EBF Symposium 2019, Carina Ekstroem from Ferring Pharmaceuticals presented results of our joint pilot project on virtualization of legacy software to read archived raw data and reconstruct past studies. We describe the method in detail and use virtualization of ScieX Analyst 1.4.2, 16 year old software as a case study to discuss the approach.

Importance of software. The process of data gathering and analysis normally starts with instruments and specimen processing. The raw data is passed onto specialist software. That software is a critical enabler because it provides the only way for experts to view and reason about collected data. Any changes to the software or the environment within which the software operates may affect the results. For that reason, the technology vendors are concerned with both (1) the implementation of the software and (2) the environment in which the software runs. It is common for vendors to supply a dedicated PC with pre-installed software to be used for processing data in the lab.

Regulations. The importance of raw data and software is emphasized by the GLP principles that apply to computerized systems. OECD guidelines state that

- The raw data must be retained and is the only acceptable means to reconstruct a study
- The software needed to read and interpret the data must be available.

While our focus is on archiving aspects of the bioanalysis practices, we consider essential aspects of the operational activities that are relevant for the transition into archiving: the Data Integrity and the Software Integrity.



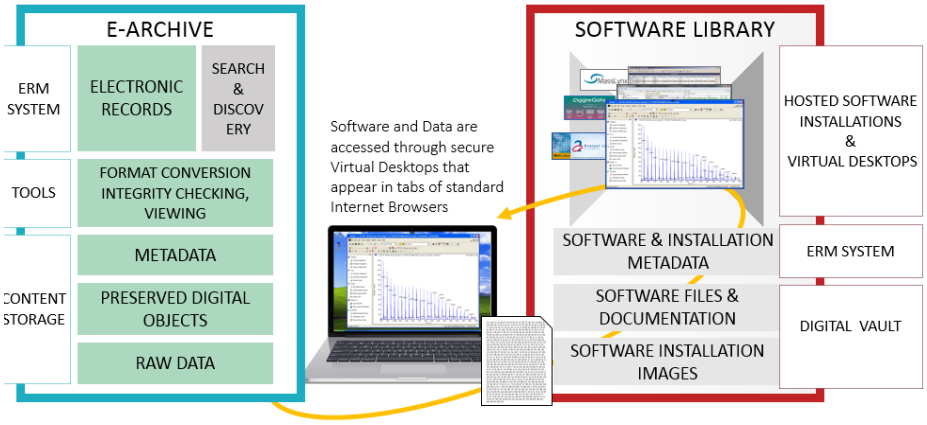
Data Integrity has been an important topic of both this year and last year EBF Open Symposium. There are concerted efforts to improve operational aspects; in particular, to increase the security and interoperability of data and to capture data provenance through digital signatures. It has been proposed to create and adopt an XML-based format for representing raw data and data analysis and enable automated encryption/decryption of data files. That work is ongoing.

Software Integrity, on the other hand, has not been of much concern and discussions. Operations are supported by a careful and comprehensive validation of instruments and software at the time of the technology deployment and upgrades. That ensures that the software stays performant, secure and consistent.

Archiving. Once a study is completed, raw data, analysis data and documentation are placed in e-Archive. Archived data accumulates over time. On the other hand, as new instruments are adopted, the previous ones are decommissioned. The software would normally be decommissioned at the same time but is needed to read the archived data, for decades.

In order to enable data readability, it is common to keep using the original PC with the software installation. However, that approach presents several risks. First, the hardware becomes old, unsupported and may stop working any time. Thus, one needs be prepared to re-install the software. The PC runs an operating system (OS) that becomes unsupported over time; yet it is needed for the software installation. Since the OS presents a security risk, the PC must be isolated from the main communication network. Consequently, the only way to reconstruct a study is to manually copy archived data onto a USB stick and walk to the PC to run the software. Fortunately, neither the hardware nor the security issues are impossible to address, as we show. However, one has to take a principled approach in order to preserve Software Integrity, i.e., to ensure that the new software installations are validated for reading the archived data and reproducing the studies.

Technical solution. In a data centre, we (1) create virtualized installations of software to remove the risks from hardware obsolescence, (2) isolate the installations to manage insecure operating system and (3) enable secure and convenient remote use of software through virtual desktops that can be used on a modern PC. This set up is called Software Library and complements e-Archive: e-Archives stores the data while Software Library runs the software ('data readers').



For a study reconstruction, the archivist creates a copy of the archived data and prepare it (checksum, compress, encrypt) for upload to the Software Library using a secure Transfer Desktop. Users with secure access can then log in from their desktops and complete the study reconstruction tasks, deleting the used copy of the data at the end, as required.

Validation process. Critical aspects of our method are the software installation and validation processes. For Analyst 1.4.2 we followed the best practices adopted for creating operational software installations. First, the IT team gathered (1) software installation files, executables and documentation and (2) original IQ and OQ documentation. We then jointly developed the IQ, OQ and PQ procedures for virtualized installations of Analyst 1.4.2. Since the software is used only for data reading, the OQ procedure was much simplified. The scientists defined the OQ procedure to fit the scope of the study reconstruction. Finally, the PQ procedure is specified and applied before any study reconstruction begins. PQ is used for ongoing testing and maintenance of the installations. All these stages were done very efficiently for ScieX Analyst 1.4.2 since the original software was still available and used for testing and comparison.

In conclusion, the implemented method is effective, fully compliant with organizational policies and aligned with established validation practices. It does not require any changes to the data or software. In fact, it is devised to preserve both Data Integrity and Software Integrity. Going forward, we advise to optimize the process further by adding software to the Software Library at the time it is first deployed and subsequently upgraded. That has two advantages: (1) the validation process need not be performed (again) at the time of decommissioning and (2) the Software Library is always up to date and aligned with the e-Archive data.