iPRES 2021 Beijing • China 19-22 Oct. 2021
17th International Conference on Digital Preservation

Executable Archives

Software integrity for data readability and validation of archived studies

20 October 2021, iPRES 2021

Dr Natasa Milic-Frayling
CEO & Founder, Intact Digital Ltd

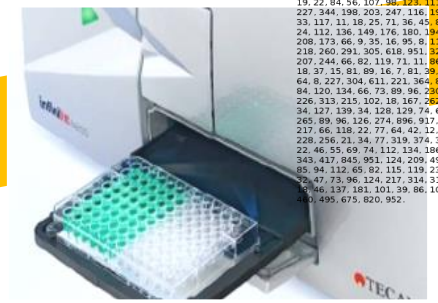
Dr Marija Cubric
University of Hertfordshire, UK

INTACT DIGITAL

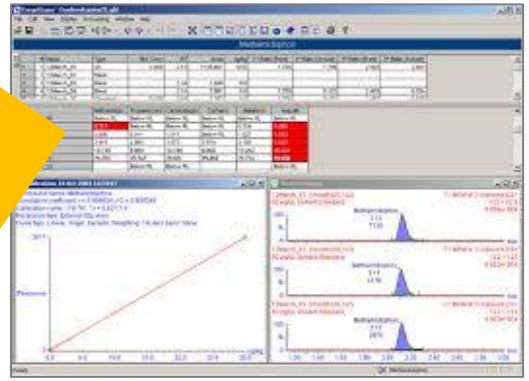
Scientific data archiving: Requirements

- Raw data, collected from instruments must be archived and remain readable for a specified period of time (often decades)
- Research studies must be reproducible directly from archived raw data.

Specimen processing



Raw data



Speciality software for interpreting and analyzing raw data

Regulations on Data integrity

ALCOA+ Data integrity requirements

Attributable
Document must clearly identify who has created and contributed to them, and be protected against falsification or forgery of those details

Legible
Stored documentation must be legible and easy to read

Contemporaneous
Documentation should demonstrate and support contemporaneous record-keeping

Original
Storing original copies of documentation guarantees accuracy and confidentiality

Accurate
The processes and procedures by which companies record and keep their documentation up to date must ensure accuracy and reliability

+Complete
All documentation must have an audit trail to show no data has been deleted or lost

+Consistent
Documentation must be date and time stamped and stored in such a way to prove it has been assembled in the expected sequence

+Enduring
Data must be available for as long as the regulation requires

+Available
Data must not only exist, they must be accessible too, when and where required for reference and auditing purposes

- Archived study data in electronic form must remain immutable, readable and ‘dynamic’, i.e., interactive
- Reproducibility of study results requires software to remain functional and usable—**software integrity**

Challenges

- Software is subject to rapid obsolescence if not regularly updated
- Updated software cannot guarantee the same functionality and output as originally used software version.

Software integrity

- OECD provides guidance on meeting regulatory requirements and adhering to the Principles of Good Laboratory Practices (GLP) when using computerized systems.

***O**rganisation for*
***E**conomic*
***C**o-operation and*
***D**evelopment*

6.16 Archive.

When legacy systems can no longer be supported, consideration should be given to the importance of the data, and if required, to maintaining the software for data accessibility purposes. This may be achieved by maintaining software in a virtual environment.

OECD SERIES ON PRINCIPLES OF GOOD LABORATORY PRACTICE AND COMPLIANCE MONITORING, Number 22. Advisory Document of the Working Party on Good Laboratory Practice on GLP Data.

Digital preservation perspective

- **Entire archived study is a 'digital object' to be preserved.** It includes
  - Raw digital data and derived information
  - Study reports based on the processed data and include some of the visualized information.
- **Significant properties of the digital object** are intrinsically tied to the processing of data and observed through the software use.

Approach: Two preservation principles are considered **sufficient to preserve significant properties of the archived study**:

Stored data integrity – verified through checksum of electronic data files

Software integrity – verified functionality of the legacy software.

Human factors

ARCHIVIST	Study records and raw data are stored and preserved by archivists, following the instructions of scientists.
RESEARCHER	Studies are reconstructed by scientists. Use of the specialty software is outside the scope of archivists' competences.
IT SPECIALIST	Software installation and management is a matter for IT specialists, particularly use of legacy operating systems that present security risks.

How to achieve preservation and reliable use of archived research study, considering

- The dependence on obsolete software and
- Separation of concerns and competencies among three types of specialists?

Research question

What would be an effective design of an **Executable Archive**, a system and services that extend the current electronic data archiving solutions and practices, recognizing

1. The fundamental dependence of electronic data access on software
2. A need for long-term software installation support.

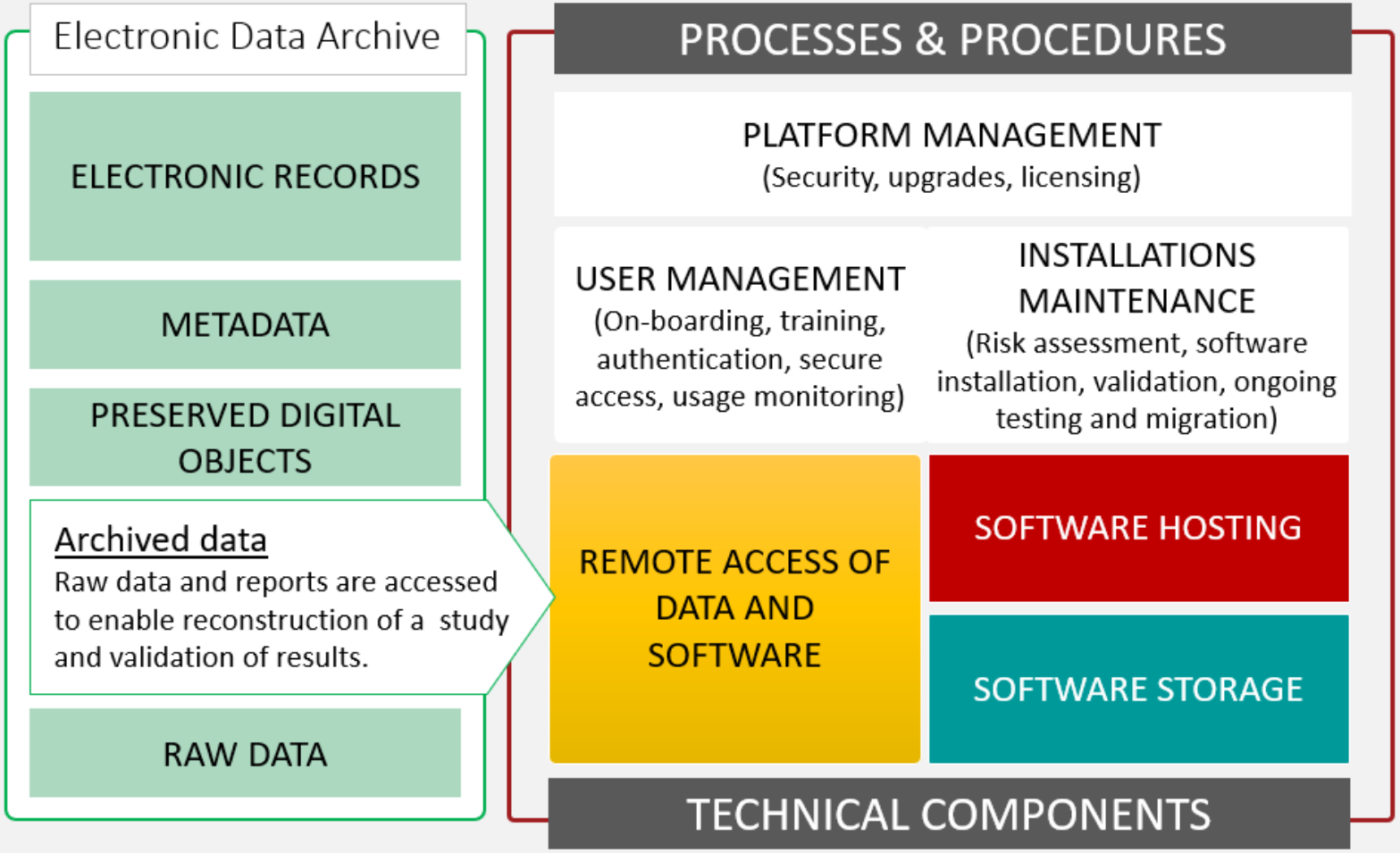
APPROACH

- **Define an Executable Archive Framework**, informed by the software engineering and IT practices and scientific archive requirements
- **Explore designs** of technical infrastructure and services within the framework.

Executable Archive Framework

Electronic data archiving is expended with

- Complementary **technical components** to support software installation, hosting and use
- Corresponding **processes and procedures** for managing installation, validation, maintenance and use of the software.



Case Study: Software Library Design

Executable Archive framework is used to design and implement Software Library with functionality complementary to Electronic Data Archive.

Electronic Data Archive + Software Library → Executable Archive

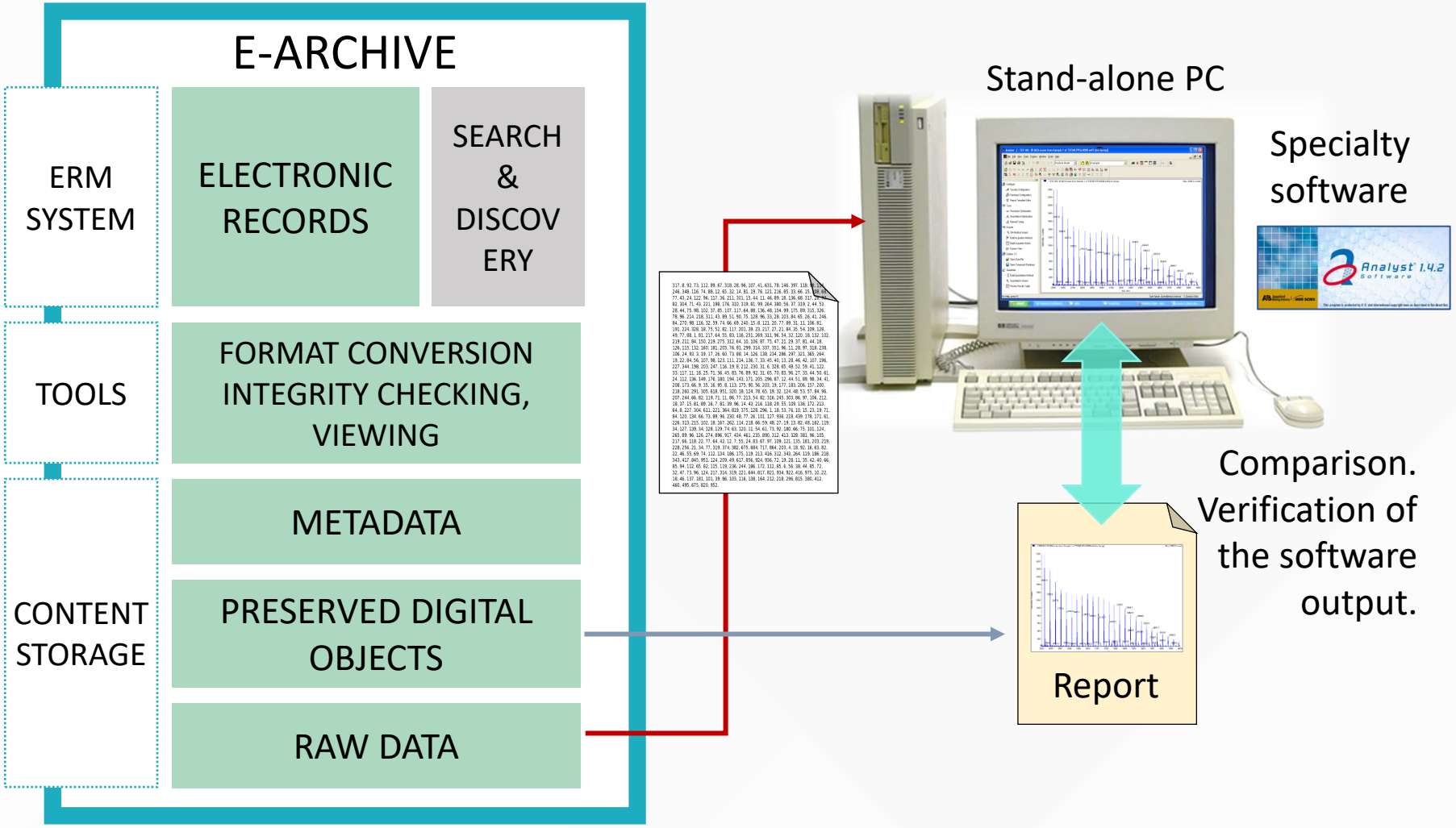
Software Library platform and services

- Host the collection of validated software installations
- Provide secure connections to data repositories
- Enable reliable use of software and data.

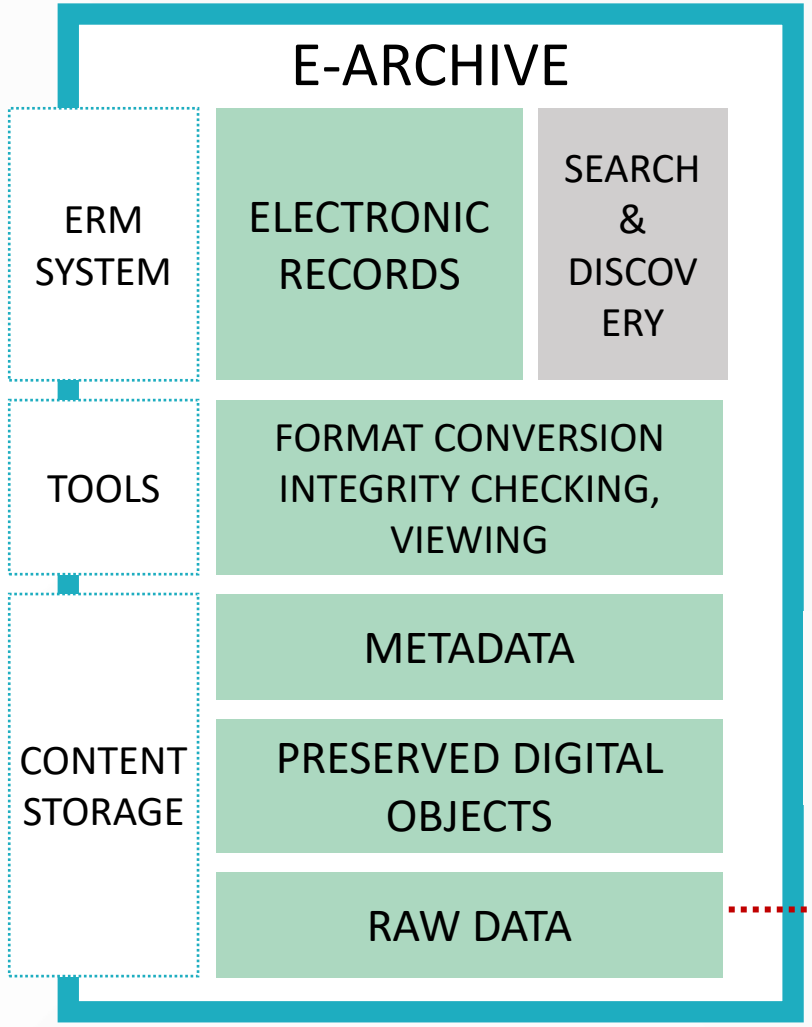
A transition of software into the Software Library is a standard procedure that involves:

- Software installation
- Software validation, in the context of a task (e.g., a study reconstruction)
- Software maintenance.

Requirements: Current practice



Technology obsolescence impact



Physical transfer of data files

Isolation, to mediate security risks

Restriction to physical access only.

Technology obsolescence

- Unsupported PC hardware
- Unsupported operating system (Windows XP, Windows 7)
- Unsupported application
- Compatible software may not be available.

Software virtualization

E-ARCHIVE

ERM SYSTEM	ELECTRONIC RECORDS	SEARCH & DISCOVERY
TOOLS	FORMAT CONVERSION INTEGRITY CHECKING, VIEWING	
CONTENT STORAGE	METADATA	
	PRESERVED DIGITAL OBJECTS	
	RAW DATA	

Stand-alone PC

VIRTUALIZATION

SOFTWARE LIBRARY

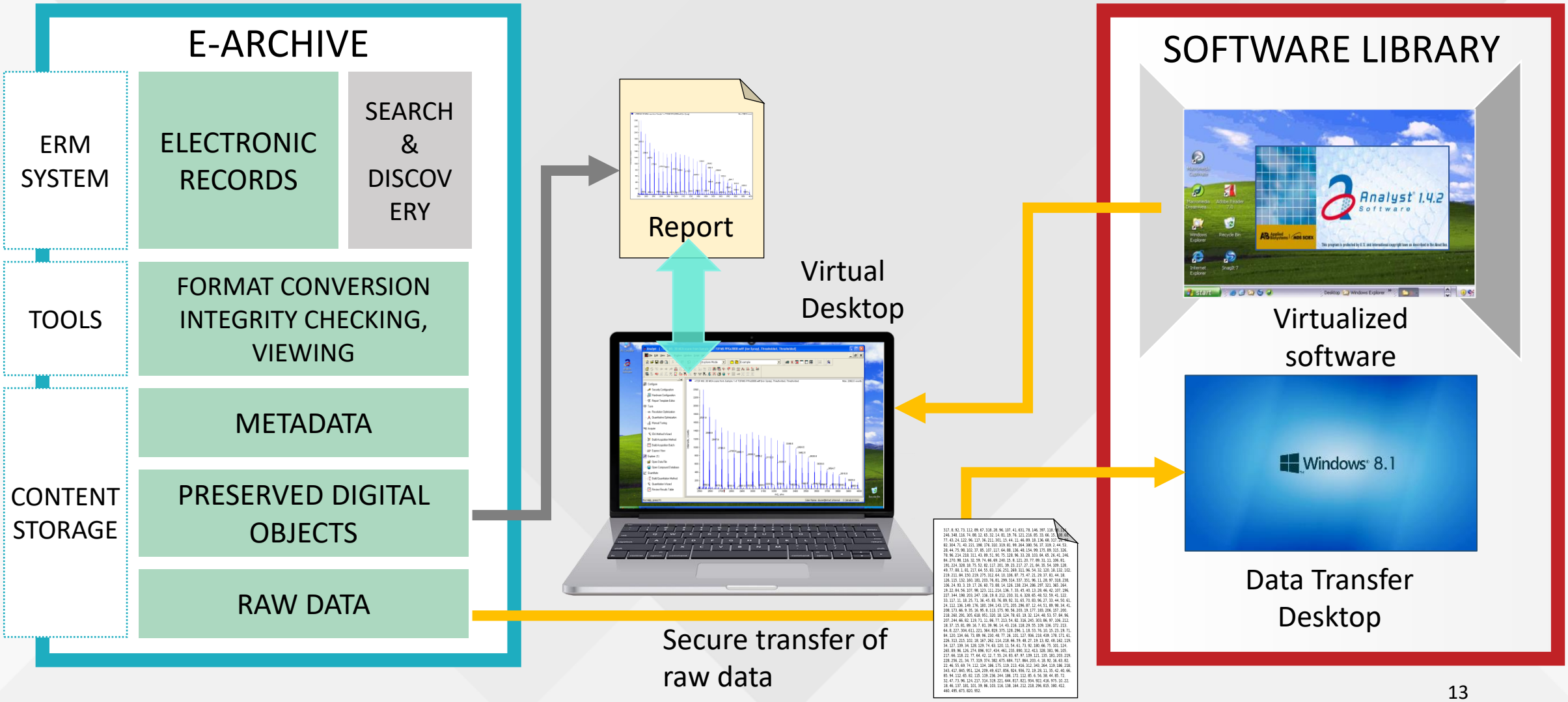
Virtualized software

SECURE DATA TRANSFER

- Data is transferred secure for study reconstruction
- Data is removed from the Software Library once the task is completed.

- Software hosting with no risk from hardware obsolescence
- Validated software installations
- Secure access to VMs with legacy OS through Virtual Desktops from standard browsers.

Study reconstruction in Software Library



Software Library technical components

EXECUTABLE ARCHIVE FRAMEWORK

Electronic Data Archive

- ELECTRONIC RECORDS
- METADATA
- PRESERVED DIGITAL OBJECTS
- **Archived data**
  Raw data and reports are accessed to enable reconstruction of a study and validation of results.
- RAW DATA

PROCESSES & PROCEDURES

PLATFORM MANAGEMENT
(Security, upgrades, licensing)

USER MANAGEMENT
(On-boarding, training, authentication, secure access, usage monitoring)

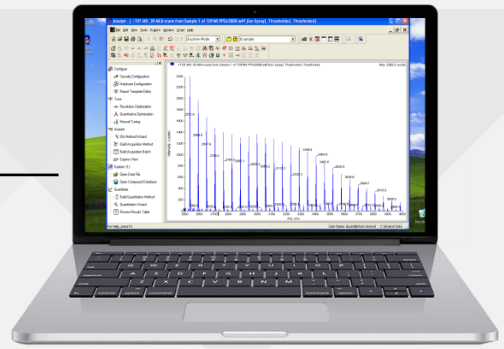
INSTALLATIONS MAINTENANCE
(Risk assessment, software installation, validation, ongoing testing and migration)

REMOTE ACCESS OF DATA AND SOFTWARE

SOFTWARE HOSTING

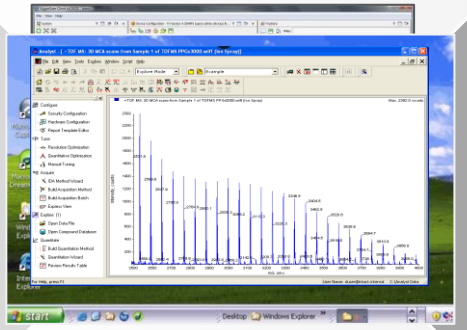
SOFTWARE STORAGE

TECHNICAL COMPONENTS



VIRTUAL DESKTOPS

SOFTWARE LIBRARY

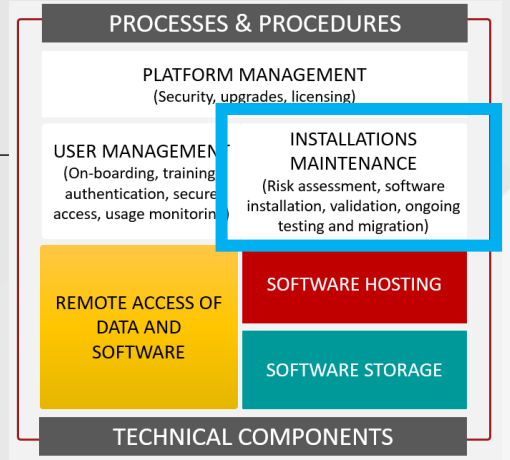


SOFTWARE LIBRARY MANAGEMENT PORTAL

SOFTWARE HOSTING PORTAL

DIGITAL VAULT

Procedures: Installation management



INSTALLATION QUALIFICATION (IQ → SL-IQ)

- **IT** Create a sandboxed VM environment for software installation
- **IT** Upload of software and software installation documentation
- **IT** **Document the process of installing the software in the VM**
  - Follow the original **Installation Qualification** (IQ) used to create installations in the Lab.
  - Create **Software Library IQ** (SL-IQ) for the specific software.

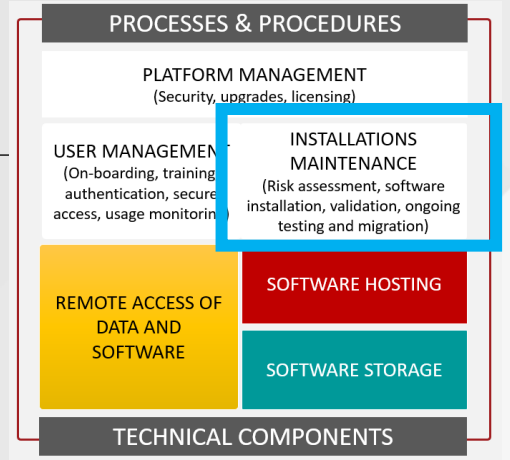
OPERATIONAL QUALIFICATION (OQ → SL-OQ)

- **IT** Configure Virtual Desktops (VD) to support data transfer and software use.
- **RS** Review the Operational Qualification (OQ) of the original software installed in the Lab
- **RS** **Document the testing of the virtualized software installation**
  - Select and test software features that support the study reconstruction task
  - Document the operational qualification process SL-OQ for virtualized software installations.

Roles

- **IT** IT Specialist
- **RS** Researcher
- **AR** Archivist

Procedures: Software integrity checks



PERFORMANCE QUALIFICATION (PQ → SL-PQ)

RS Create a **Performance Qualification (PQ)** test and select representative test data

- Select a minimal set of steps to establish the **software integrity**
- Perform the test with the original software (in the Lab) and virtualized software

RS **Document the PQ procedure for the virtualized installation to create SL-PQ**

- Apply SL-PQ test before importing the study data.

STUDY RECONSTRUCTION

AR Use Transfer Desktop to transfer data into the Software Library environment

- Data is accessible to all the VMs available to the user

RS Activate the desktop with specific software and gain access to data

- Move the data to the VM if it needs to be used locally by the software

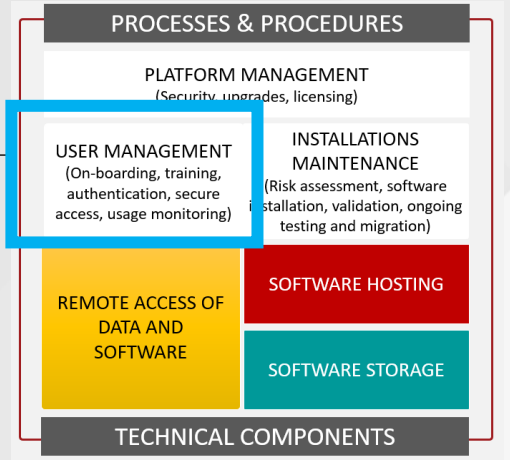
AR Access the study documentation and reconstruction process.

RS Reproduce the results and document the study reconstruction process.

Roles

- **IT** IT Specialist
- **RS** Researcher
- **AR** Archivist

Human factor management



Software Library supports a systematic capture of knowledge about software installation, maintenance and use through

- Regular user training
  - Refresher course every 6 months
  - Involvement in internal compliance audits and software verification annually or semi-annually and in regulatory inspections (every 2 years)
- Detailed documentation
  - Software management is documented by describing steps and capturing screenshots.

Concluding remarks

- Archiving of digital content cannot be separated from long-term care of software

Data integrity + Software integrity → Preserved content
(significant properties)

- Executable Archive Framework supports a design of software management systems that complement data archives.

Electronic Data Archive + Software Library → Executable Archive

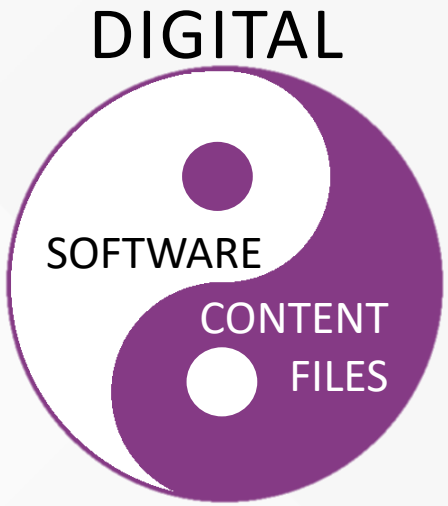
- Software Library platform and services support activities of three expert roles: IT staff, archivists and researchers.

Optimization: Keep the Software Library ‘aligned’ with the stored data.

iPRES 2021
17th International Conference on Digital Preservation

INTACT DIGITAL

Thank you!



Software is the Yang of Digital.

Interacting with stored content files, the Yin of Digital, it brings to life ephemeral existence of digital creations.

Dr Natasa Milic-Frayling

natasamf@intact.digital

www.intact.digital