



# E-ARK4ALL

Dr. Sánchez-Royo, Begoña<sup>1</sup> | Prof. Anderson, David<sup>1</sup> | Dr. Kaminski, Jaime<sup>1</sup> | Prof. Anderson, Janet<sup>1</sup>



## Data Quality and Digital Archiving:

### The Intersection of Two Important Data Management Functions in ECIS<sup>2</sup> and Cancer Patient Registries



This poster presents an overview of how E-ARK4ALL can help Cancer Patient Registries in the **long-term sustainability** of their databases and handle the **challenge of Big Data** with interoperable and scalable solutions.

E-ARK4ALL is a service provided by the new eArchiving Building Block, for digital and long-term preservation, in the Connection Europe Facility (CEF) programme.

E-ARK4ALL also draws on the work of the EC to ensure the practical implementation of Directive 2011/24/EU on patient's rights regarding cross-border healthcare and the Standards and Guidelines for Cancer Registration in Europe.<sup>3</sup>

### Introduction

Cancer Patient Registries play a critical role in the fight against cancer. Billions of digital records need to be **preserved over time** in order to produce vital statistics on cancer incidence and survival rates; improve patient care; and develop innovative public health initiatives. Long-term sustainability of digital cancer records is critical for these goals.

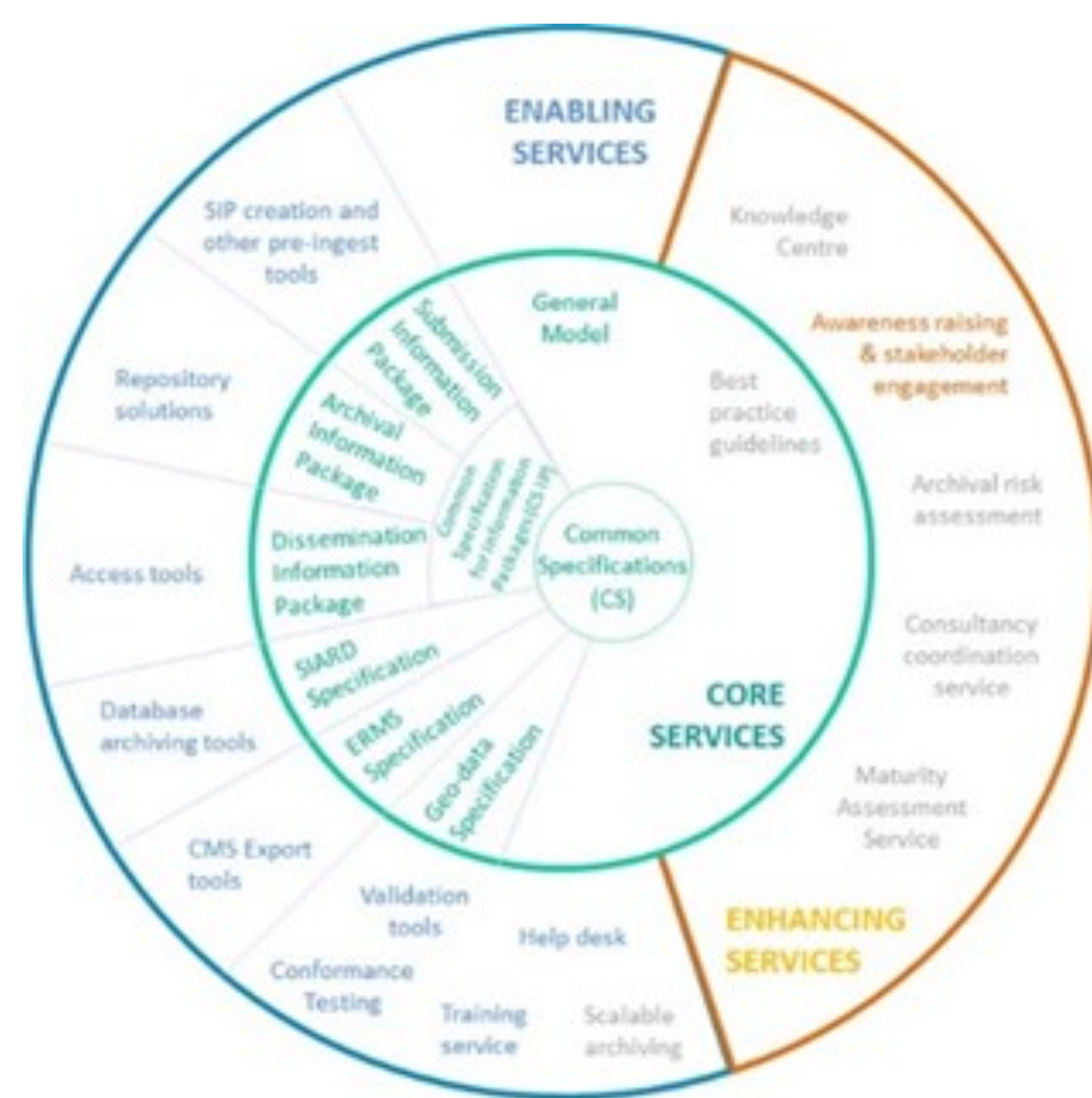
The use of digital archiving standards, interoperability requirements and scalable archival IT tools will be critical if cancer registrars are to better exploit **Big Data** analysis to track and query their data in the future.

With ever-increasing amounts of cancer-related data being generated, the failure to employ best practice in digital preservation and archiving, will have long-term negative repercussions for the health sector.

### What is E-ARK4ALL?

**E-ARK4ALL** is the EC-funded multi-national project tasked with creating the foundation for the new **eArchiving Building Block** in the Connecting Europe Facility (CEF) programme. It is built directly on the results of the EC-funded pilot project E-ARK (European Archival Records and Knowledge Preservation) ([www.eark-project.eu](http://www.eark-project.eu)). Using E-ARK tools, standards and methods it aims to improve the methods and technologies of digital archiving, in order to achieve consistency and interoperability on a Europe-wide scale.

E-ARK4ALL comprises **Core Services** (the specifications); **Enabling Services** (tools, training service, Help Desk and conformance testing); and **Enhancing Services** (awareness raising and stakeholder engagement).



All E-ARK package specifications for database preservation formats and open source software tools are based on well-established core specifications and formats (METS, EAD, PREMIS, SIARD, MoReq2010, GML) that are widely used within the archival and digital preservation sectors.

The E-ARK specifications further refine and specify the use of these underlying core standards. The usability of the E-ARK specifications has been validated in seven pilots conducted in 2016 (at the National Archives of Estonia, Denmark, Norway, Hungary, Portugal and Slovenia, and the Estonian Business Archives).

### How does E-ARK4ALL work?

All E-ARK methods, tools and infrastructure can be used by archival institutions in the health domain for executing production-level digital delivery ingest, long-term preservation and re-use workflows.

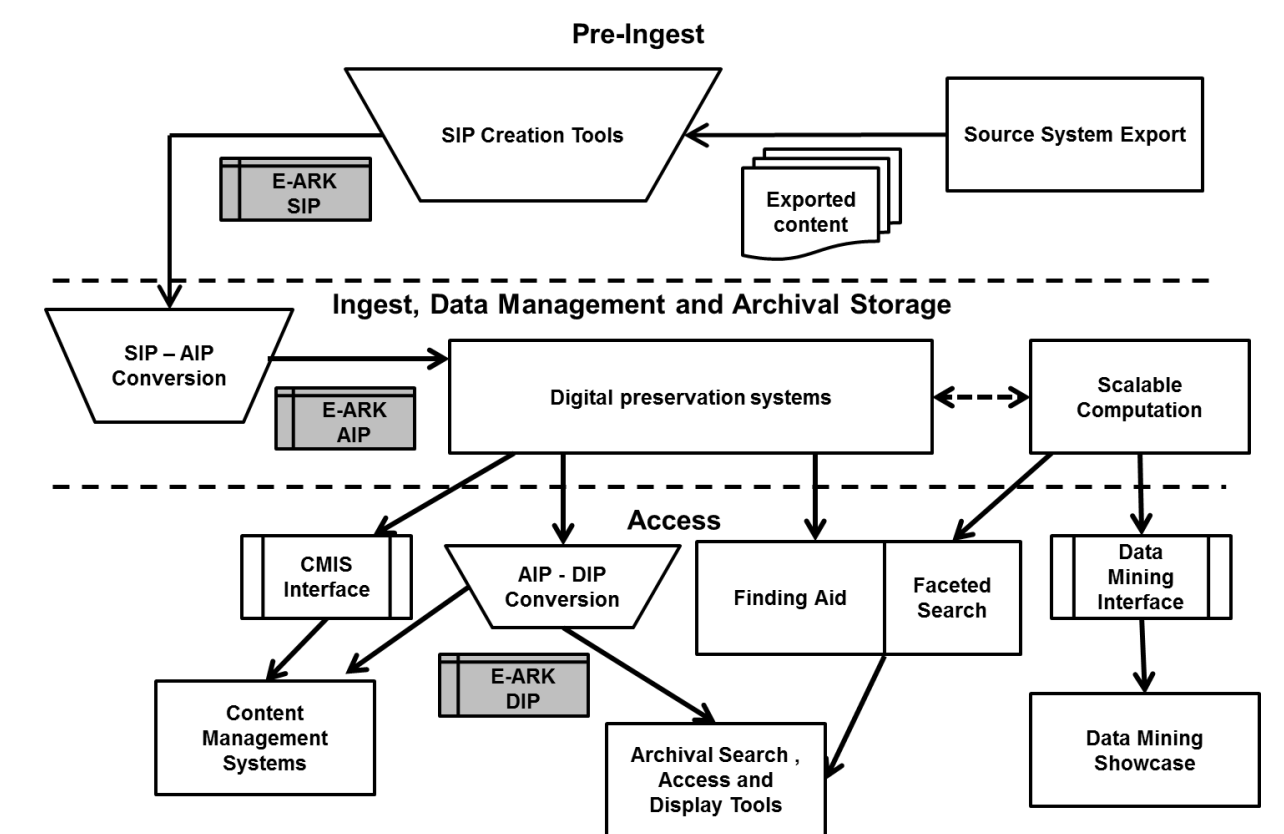
#### The Core Services (specifications)

achieve interoperability and efficiency in digital archiving operations. They detail the structural and metadata requirements for constructing each of the OAIS Information Package types:

- Submission Information Package Specification (SIP)
- Archival Information Package Specification (AIP)
- Dissemination Information Package Specification (DIP)

#### Open source software tools for digital preservation and eArchiving

Pre-Ingest Tools	Migration and Reuse Tools	Scalable Computation Tools
RODA-in;	Database Preservation Toolkit (DBPTK);	E-ARK website;
ETP (data creators);	Database Visualization Toolkit (DBVTK).	Hadoop.
ETA (data repositories).		



### Adding value to data quality management

#### Interoperability

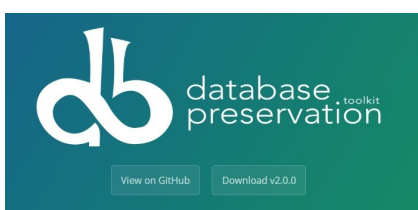
Preserving cancer-related digital content is a **collaborative effort**. Cancer Patient Registries which are running technically heterogeneous repository systems with distinct data models may want to securely share content with selected partners to provide distributed preservation solutions.

#### Interactive Visualizations


In ECIS 'data quality' is focused primarily on assessing the internal quality of the (secondary) data collected from member Cancer Patient Registries. ECIS has little control over the external validity of the original sources and therefore, many issues affecting the conversion of data to a form that is more suitable for long-term retention cannot be resolved at this stage. However, interactive visualization systems using preservation metadata aggregations, categorization and data mining can be used to support large-scale analysis of digital cancer records. This enables a comprehensive understanding of the many information layers within large scale cancer-related collections. The result is high-quality comparability when processing large collections of digital cancer records.

E-ARK4ALL contributes to both goals with two toolkits:

#### Database Preservation Toolkit

 This open source application allows users to access the content of a live database (Oracle, MySQL) to extract its content into the open SIARD 2.0 format.

#### Database Visualization Toolkit

 This open source toolkit allows users to access the content of a stored SIARD 2.0 file without the need to set up a complex management system. It is an interactive visual analytics application that provides access over the long-term in large digital collections.

### Steps forward for E-ARK4ALL

E-ARK4ALL is currently exploring options for providing further services, such as technical and online training related to Digital Preservation and eArchiving.

The project is developing research support services to raise awareness of data preservation and digital archiving to assist Cancer Registrars in finding programs to further their education in the field of health informatics.



<sup>1</sup> E-ARK4ALL | [www.eark-project.eu](http://www.eark-project.eu) | Poster enquiries: [bsanchez13@gmail.com](mailto:bsanchez13@gmail.com)  
<sup>2</sup> ECIS is the European Cancer Information System.  
<sup>3</sup> Standards and guidelines for cancer registration in Europe: the ENCR recommendations /editors, Jerzy E. Tyczynski, Eva Démaret, D. Maxwell Parkin, IARC technical publication; no. 40.