Specification for the E-ARK Content Information Type Specification for eHealth2 (CITS eHealth2)

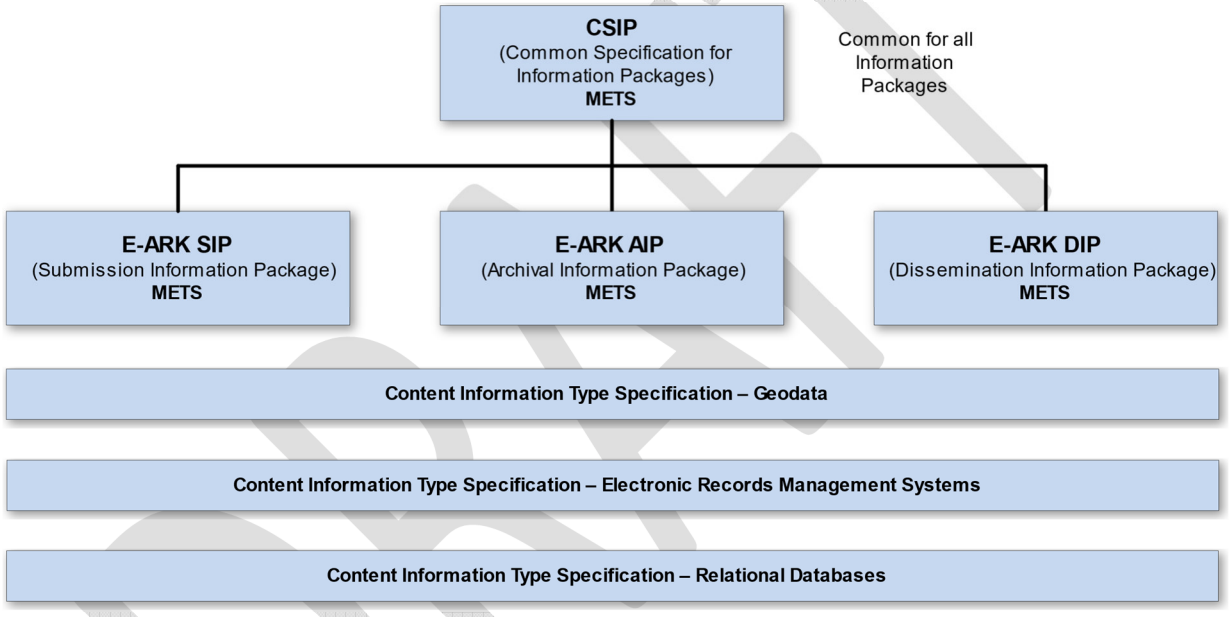
A proper front page will be created for the publication occurring after implementation of review comments.

A first page is generated with correct color and logos and all.

1 Preface

1.1 Aim of the specification

This E-ARK specification is part of a family of specifications that provide a common set of requirements for packaging digital information. These specifications are based on common, international standards for transmitting, describing and preserving digital data. They have been produced to help data creators, software developers, and digital archives tackle the challenge of short-, medium- and long-term data management and reuse in a sustainable, authentic, cost-efficient, manageable and interoperable way.
The foundation for these specifications is the Reference Model for an Open Archival Information System (OAIS) which has Information Packages at its core. Familiarity with the core functional entities of OAIS is a prerequisite for understanding the specifications. A visualisation of the current specification network can be seen here:



The E-ARK specification dependency hierarchy

Specification	Aim and Goals
Common Specification for Information Packages	This document introduces the concept of a Common Specification for Information Packages (CSIP). Its three main purposes are to: • Establish a common understanding of the requirements which need to be met in order to achieve interoperability of Information Packages. • Establish a common base for the development of more specific Information Package definitions and tools within the digital preservation community. • Propose the details of an XML-based implementation of the requirements using, to the largest possible extent, standards which are widely used in international digital preservation. Ultimately the goal of the Common Specification is to reach a level of interoperability between all Information Packages so that tools implementing the Common Specification can be adopted by institutions without the need for further modifications or adaptations.
E-ARK SIP	The main aims of this specification are to:

	• Define a general structure for a Submission Information Package format suitable for a wide variety of archival scenarios, e.g. document and image collections, databases or geographical data. • Enhance interoperability between Producers and Archives. • Recommend best practices regarding metadata, content and structure of Submission Information Packages.
E-ARK AIP	The main aims of this specification are to: • Define a generic structure of the AIP format suitable for a wide variety of data types, such as document and image collections, archival records, databases or geographical data. • Recommend a set of metadata related to the structural and the preservation aspects of the AIP as implemented by the reference implementation eArchiving ToolBox (formerly earkweb). • Ensure the format is suitable to store large quantities of data.
E-ARK DIP	The main aims of this specification are to: • Define a generic structure of the DIP format suitable for a wide variety of archival records, such as document and image collections, databases or geographical data. • Recommend a set of metadata related to the structural and access aspects of the DIP.
Content Information Type Specifications	The main aim and goal of a Content Information Type Specification is to: • Define, in technical terms, how data and metadata must be formatted and placed within a CSIP Information Package in order to achieve interoperability in exchanging specific Content Information. The number of possible Content Information Type Specifications is unlimited.

1.2 Organisational support

This specification is maintained by the Digital Information LifeCycle Interoperability Standards Board (DILCIS Board). The DILCIS Board (http://dilcis.eu/) was created to enhance and maintain the draft specifications developed in the European Archival Records and Knowledge Preservation Project (E-ARK project) which concluded in January 2017 (http://eark-project.com/). The Board consists of eight members, but there is no restriction on the number of participants in the work. All Board documents and specifications are stored in GitHub (https://github.com/DILCISBoard) while published versions are made available on the Board webpage. Since 2018 the DILCIS Board has been responsible for the core specifications in the Connecting Europe Facility eArchiving Building Block (https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eArchiving).

1.3 Authors

A full list of contributors to this specification, as well as the revision history can be found in Appendix 1.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

2 Glossary

Table 2: Glossary

Name	Description
Aggregator	Person or organisation that requires data from the cancer registry.
Archive	A state agency or organisational unit of the organisation responsible for the long-term preservation of data.
Cancer case data	Data gathered as a result of the cancer registry activity and prepared for the particular export.
Cancer registry (CR)	An institution that systematically receives and collects data about cancer patients.
Cancer registry data	Data that is created as a result of the cancer registry's activities.
Cancer registry export	Data set and accompanied documentation exported from the cancer registry based on an aggregator's specifications.
CONCORD Programme	CONCORD is the programme for worldwide surveillance of cancer survival trends, led by the London School of Hygiene and Tropical Medicine.

ENCR	European Network of Cancer Registries promotes collaboration between cancer registries, defines data collection standards, provides training for cancer registry personnel and regularly disseminates information on incidence and mortality from cancer in the European Union and Europe. Its secretariat is hosted at the European Commission's Joint Research Centre (JRC).
Export definitions	Documentation that defines the content and structure of the cancer registry export agreed between the cancer registry and the aggregator.
Exportation documentation	Documentation created during the export and during the transfer to the aggregator.
Incidence data	Absolute number of all cancer cases who were newly diagnosed in a defined population in one calendar year reported by gender and age at diagnosis.
JRC	European Commission's Joint Research Centre.
Life table	A table which shows, for each age, what the probability is that a person of that age will die before their next birthday.
Mortality data	Absolute number of all persons who died because of a certain disease in a defined population in one calendar year reported by gender and age at death.
Population data	Population data is provided from official censuses or other official sources. General population data is provided by sex, age and calendar year for the area covered by the cancer registry (e.g. the population the cancer cases come from).
TNM	The TNM Classification of Malignant Tumors is a globally recognised standard for classifying the extent of the spread of cancer. TNM stands for tumour (T), nodes (N), and metastases (M).

3 Introduction

3.1 Purpose

The purpose of this document is to describe the Content Information Type Specification for eHealth2 (exported cancer registry data sets). The specification is designed to archive exports from cancer registries and exchange data between cancer registries and different users. This specification is supported by an XML-schema and a Schematron document described in CSIP (the E-ARK Common Specification for Information Packages) and E-ARK SIP (the E-ARK Specification for Submission Information Packages).

The specification is built on work by the Archives of the Republic of Slovenia, the Joint Research Center and the Slovenian Cancer Registry.

3.2 Aim

This specification aims to define the structure and content of cancer registry data sets exported by cancer registries and aggregated by different aggregators and need to be archived. This specification makes the following assumptions:

- exports (content and structure) reflect data in cancer registries
- international aggregators define the content of the export with specifications based on accepted international health standards
- the specification is included in a data call
- different aggregators can use the same health standards and similar manuals for coding; therefore, exports usually contain a certain amount of identical types of data
- researchers require the same type of data as international aggregators but are usually focused on a narrower subset of data and different set of variables
- the version of the health standards used must be defined
- data is collected for a certain time period
- cancer registries need to acquire additional data from outside sources for proper cancer burden estimation
- aggregators can use software tools to check consistency and/or correctness of data

Since cancer is a major public health and economic issue, many international organisations and projects aggregate data and report the cancer burden on a wider scale and undertake additional analyses and comparisons between countries. For this specification, representative cases for aggregating and processing data were provided by the ENCR-JRC and CONCORD Programme.

4 Context

4.1 Cancer registry and cancer registration

Cancer registration is a process of continual systematic collection of data on the occurrence, characteristics, and outcome of reportable neoplasms to assess and control the impact of malignant disease in the community.

The data on cancer collected by the cancer registry, incidence, survival and prevalence, serve together with mortality data as the basis for assessing the cancer burden in the geographical area covered by cancer registries. The data is vital for the evaluation of cancer burden and planning in the field of primary and secondary prevention, diagnostics, treatment and rehabilitation, for planning facilities and funding needed for cancer control (personnel, equipment and hospital capacities) as well as for clinical and epidemiological research and the evaluation of the effectiveness of cancer screening programs.

For any cancer registry, it is essential to define the population it comprises, the territory it covers, and the way data is collected and updated (continuous/periodical; active registration/passive registration, etc.).

Cancer registries collect and process personal data. This data's use and distribution are regulated with European and national legislation, which has to be considered when archiving cancer registry's data.

4.2 Cancer registries data use

Data gathered by cancer registries are used by cancer registries themselves and numerous other institutions (international, European and state agencies, academic institutes) and individual researchers. It influences scientific progress, new cancer treatment methods, health and environmental policies.

Aggregations and data exports

The cancer registries' **Annual Reports** are a key means of disseminating information about cancer among professionals and other interested groups. As cancer registries are limited by data collection technology from a variety of sources, data collection, linking, and publication are time-consuming (here and elsewhere) and usually takes several years to publish a report. Besides annual reports, cancer registries publish their findings and data in other **special publications** and **scientific papers.**

Some cancer registries have established **interactive web portals** that offer online access to cancer registry data and search facilities (e.g. SLORA and ELVIS).

Cancer registries publish their data in **international databases and projects**. Data has been published in volumes of the Cancer Incidence in Five Continents issued by the International Agency for Research on Cancer. The data are regularly included in international databases ECIS (European Cancer Information System), GLOBOCAN database (Global Cancer Observatory) and ACCIS (Automated Childhood Cancer Information System). The survival of cancer patients has been analysed in international studies (e.g. EUROCARE, CONCORD and RARECARE). Cancer registries also contribute their data to other international projects.

Cancer registries enable access to their data for scientific research. When researchers have **special requirements**, they can apply for specific data directly from the cancer registries, where a designated body considers the application, especially regarding regulations on data protection and research aims and ethics. To properly analyse and interpret data, the context of how cancer registry data was captured and processed has to be preserved and clearly reported.

When cancer registries share their data with other **national institutions** (e.g. national health institute, statistical office, etc.), such use is regulated by national legislation. Aggregations can be classified according to different criteria into several groups. The classification criteria are:

- one-off / periodic (key information: date or frequency; in case of periodic possible differences in the structure of exports)
- determined by regulations that are based on formal agreement (key information: the formal agreement has to be preserved together with the aggregation)
- purpose: preparation of a one-off study/statistical comparison/preparation of strategies and policies
- international (involvement of several cancer registries) / national - regional (only one cancer registry) (key information is whether the results are / will be based on only one CR data or more)

Those criteria affect the retention period of the aggregation and the decision of whether aggregation needs to be archived or not.

5 Cancer registry data

Population-based cancer registries worldwide systematically gather data on the occurrence, characteristics, and outcome of cancer patients in the underlying population. The main purpose of cancer registries is to monitor the community's cancer burden for public health and clinical implications.

5.1 Data sources

The main sources of data in cancer registries are notifications of cancer, gathered from all hospitals and diagnostic centres, exceptionally also from primary health care centres if the patient has not been referred for further diagnostic investigations and/or treatment. The notification process is different from country to country and among data providers within the country and depends strongly on individual agreements and IT solutions. The collection of data is continuous for the majority of cancer registries.

Data that enters the cancer registry's database includes not only raw data accumulated from several sources but also data that has already been processed by cancer registry experts.

The data gathered with notification forms are patient identification data, data on the cancer disease (anatomic cancer site, histological or cytological diagnosis and stage by TNM (Tumour, Node, Metastasis) or other clinical classification described in further sections) and a general description of the first treatment (optional).

Other relevant databases are used to obtain more information on registered cancer cases and gather cases that might otherwise slip through the registration process (e,g, screening programme's registries, population registry, mortality registry, etc.). These are so-called complementary data sources, which are defined as other health data sources and administrative data sources. Both are crucial for improving the data quality.

5.2 Data quality and completeness

Data about the patient is collected and continuously updated for the duration of medical treatment or until the patient's death. The time span of such a collection can last several years or even decades.

Cancer registries usually publish annual reports with a few years delay to assure maximum completeness of the gathered data. Even so, cancer registries have live databases, meaning data is added and corrected all the time, resulting in the registration of several percentage points of more cancer cases after the publication of annual reports, depending on the country and publication delay.

5.3 Health classifications and standards in cancer registries

Part of cancer registry data is defined with international classifications and standards. Those classifications and standards exist in several versions, which can be used simultaneously.

Cancer registries code the data retrieved from different health data sources following the international and internal guidelines, which are regularly updated. The use of specific classification (and updates) depends upon the cancer registry's decision, although international organisations promote the latest versions of the classifications.

For coding, most European cancer registries use the standards described in the following paragraphs.

Cancer sites are coded according to the International Statistical Classification of Diseases and Related Health Problems (ICD) (available at https://www.who.int/classifications/icd/en/). When classifying tumours as malignant, the behaviour digit of the morphology code of the third edition of International Classification of Diseases for Oncology (ICD-O) (available at http://www.iacr.com.fr/index.php?option=com_content&view=category&layout=blog&id=100&Itemid=577) is used. Stage definition generally follows the TNM classification, where the T category describes the primary tumour site, the N category describes the regional lymph node involvement, and the M category describes the presence or otherwise of distant metastatic spread (available at https://www.uicc.org/resources/tnm).

For specific cancer sites, there are also more specific classifications in use. Gynaecological tumours are also defined according to the FIGO classification (available at https://screening.iarc.fr/viaviliappendix1.php) and in skin melanoma size according to Clark's level and Breslow's thickness is registered (available at https://www.curemelanoma.org/about-melanoma/melanoma-staging/breslow-depth-and-clark-level/) Malignant lymphomas are classified according to the Ann-Arbor System (available at https://radiopaedia.org/articles/ann-arbor-staging-system).

5.4 Data coding

The cancer registry undertakes coding according to health standards that were valid at the time of coding. Data aggregators define which version of the particular health standard must be used for the export and, in some cases, offer tools for transformation from one standard to another. Cancer registries can do transformation manually, depending on a data call. Such transformation is normally more accurate than automatic transformation, however, they are not achievable for large amounts of data. For some elements, special non-standardised rules apply. The rules are set within the data call and can vary from aggregator to aggregator.

Therefore, when archiving exports, documentation that defines the standards and rules used for the particular export must be preserved.

5.5 Cancer case data

Cancer case data are data in the cancer registry related to cases of cancer diseases. Those data can be arranged in the following categories:

a) Registry

The registry data aims to identify the country and cancer registry. They are usually included in exports gathered by international aggregators.

b) Patient

The patient data usually contains metadata on the patient's identification, date of birth, sex, region or address of living and race/ethnicity.

c) Tumour

The tumour data contains metadata (cancer case ID) to refer to each registered tumour.

d) Incidence

The incidence data assemble metadata on full and accurate date of incidence or patient's age when date of birth and/or incidence are not available. Cancer registries usually follow ENCR rules for determining the date of incidence. The incidence data may also contain metadata on the source of information the incidence is based on (e.g. the incidental finding of cancer in an autopsy, the death certificate being the only source of information, registries of screening programmes).

e) Basis of diagnosis

The basis of diagnosis data contains metadata that indicates the degree of certainty with which a cancer diagnosis has been established.

f) Topography

The topography data consists of metadata describing the anatomic site of the tumour. The description may be based on the use of international standards.

g) Morphology

The morphology data contains metadata describing the type of cell that has become neoplastic. The description may be based on the use of international standards.

h) Behaviour

The behaviour data contains metadata that indicates whether a tumour is malignant, benign, in situ, or of uncertain behaviour. The metadata may be coded according to international standards.

i) Grade

The grade describes how much or how little a tumour resembles the normal tissues from which it arose. The metadata may be coded according to international standards.

j) Vital status

The vital status data usually contains metadata on the patient's last known vital status or duration of survival after the cancer diagnosis (in days or months) when the complete date of incidence and/or date of last known vital status cannot be provided due to personal data protection.

k) Other data

The exports may contain several additional types of cancer registry data related to cases of cancer (e.g. laterality of paired organs, date of case registration, the official underlying cause of death, TNM stage, summary extent of disease, tumour size, number of examined and of metastatic nodes, data on surgery, systemic anti-cancer therapy, radiotherapy and/or bone marrow transplantation). The metadata may be coded according to international standards.

5.6 Cancer case-related data

a) Mortality data

Mortality data is essential for assuring the completeness of cancer registration. It is based on the official cancer mortality data obtained from the vital statistics department, or equivalent, and is based on death certificates. Cancer registry personnel use this information for querying regarding patients for whom there is no information on cancer in the cancer registry's database, but cancer is stated as the cause of death on the death certificate. In case no additional information is obtained based on the cancer registry's queries, the cancer cases are registered as so-called "death certificate only".

When there is no linkage with the official population registry or other data sources of population statistics, this data source is also used to update information on vital status and date of death for the deceased. Another cancer burden measure, prevalence, cannot be assessed without information on cancer patients' vital status.

Mortality data is the cancer burden measure per se and can be stratified by gender, topography, etc. It is also used to calculate the survival estimates or simply a measure of the ratio of mortality to incidence.

b) Population data

Population data is essential for calculating several cancer burden measures such as crude rates, age-standardised rates and survival estimates. Population data is data in cancer registries related to cases of cancer diseases and is provided from official censuses or other official sources and describes the general underlying population by sex, age and calendar year. The population data should cover the same people and geographic area as the cancer cases.

c) Life tables

Life tables are the oldest demographic tool and still among the most important instruments for mortality analysis and other investigations concerning the length of life. Life tables (i.e. the background mortality in the general population of the administrative territory covered by the cancer registry must be provided by registries participating in survival studies).

6 Elements of the eHealth2 archival information package

The eHealth2 specification conforms and extends the Common Specification for Information Packages (CSIP) and the Specification for Submission Information Packages (E-ARK SIP, E-ARK DIP and E-ARK AIP). The eHealth2 specification defines an information package (Figure 1) that aims to ensure long-term usability and authentic interpretation of the content and context of the export created when the aggregator (international, national, researchers, etc.) requests data from the cancer registry.

6.1 Cancer registry exports

Cancer registry export is a dataset containing at least one file with data from the cancer registry database (for example, in CSV format) and other files that help contextualise and interpret data from the cancer registry database. These are usually text-based (agreements, code lists, reports, etc.) but can be in other formats (audio, interviews, images). The database export must consist of cancer case data, and possibly mortality, population data and life tables. The cancer case data is purely a health record, while the other three can provide context. The complexity of structure, the quantity of data and the retention period of the export depending on the user or the purpose (as described in chapter 1.4) of the export. For example, international or national aggregators will usually need more complex exports with possible bigger implications in health policies than an academic or local cancer registry research team. If data is gathered by an international aggregator (e.g. JRC, CONCORD), whose aim is to compare cancer burden across several countries, a questionnaire, which provides data about the cancer registry and metadata on cancer registration process and tools, is an essential part of the data submission. Documentation that describes why, how and who the export was created for must be presented in the SIP regardless of the export user.

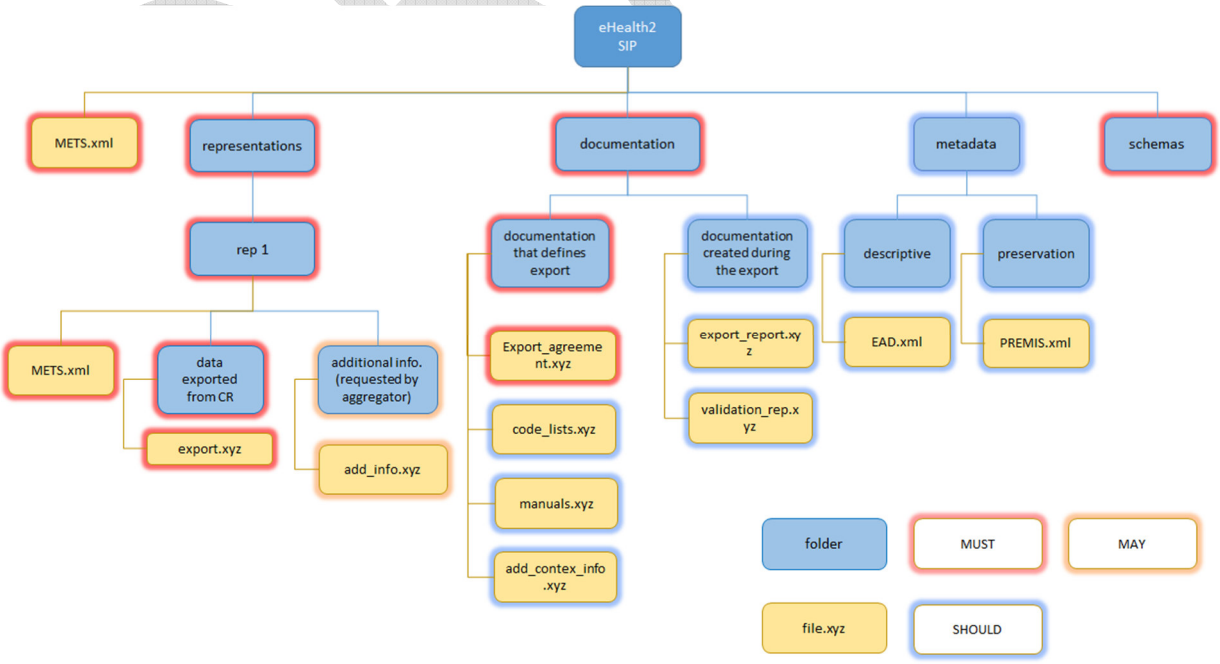


Figure 1: eHealth2 archival information package structure

6.1.1 Representation folder

Representation folder contains:

- the data file exported from the cancer registry
- additional information requested by an aggregator

a) Data exported from cancer registries

For the eHealth2 specification to be CSIP conformant, the IP (information package) has to have at least one representation folder with a data folder that contains the file that represents the export from the Cancer registry (cancer case data). The data exported from the cancer registry is the core data of the specification and based on the agreement between the cancer registry and aggregator. When an export results from an international data call, the structure and content of the export are defined by the data call. The export must contain cancer case data and could contain several other files that help interpret the cancer registry's data (mortality data, population data and life tables).

b) Additional information requested by an aggregator

If the aggregator asks for any other data besides the files exported from the cancer registry database, this data must be put in the additional information folder, enabling proper interpretation of cancer case data. This folder usually stores a questionnaire, which is a file required by international aggregators, that collects data from several cancer registries and needs an identification file for each cancer registry. In the questionnaire, aggregators collect basic data about cancer registry, persons responsible for the export, other data about particular export and the cancer registry's policies on data gathering and data processing.

6.1.2 Documentation folder

The documentation folder contains:

- export definitions
- exportation documentation

a) Export definitions

Exports from the cancer registry must be prepared based on the agreement between the aggregator and the cancer registry. This agreement must be placed in a documentation folder of the information package. Agreement always contains information on the data-set definition (scope and timeframe), framework (legal or any other), and purpose for gathering and reusing the cancer registry data. When the agreement contains the definition of the structure of the data, use of standards and code-lists, validation schemas, these must be included in the "documentation that defines export" folder. In the case of international aggregators, the data call and accompanied correspondence take the role of agreement. Other aggregators can store any documentation created in collaboration with the cancer registry when defining the export that contains the definition of the content of exported data, formats, used standards and other code lists, any re-use restrictions in this folder.

In this folder, all documentation that supports the interpretation of the export (description of standards and description of codes, manuals and other documentation) should be put.

If in the process of negotiating an agreement or executing export, other documentation that significantly amends the agreement is created and not included in the export agreement (e.g. correspondence between the cancer registry and aggregator), this documentation should also be included in this folder.

b) Exportation documentation

If the export process is documented (e.g. export reports, validation reports, delivery confirmation), it is strongly recommended that this documentation is included in the "exportation documentation" subfolder of the documentation folder. This documentation is important for preserving the completeness, authenticity and (re)usability of the export content. It can be created automatically by the software used for exporting or manually by the person conducting the export (e.g. print screens, notes). The documentation can be in the form of translation files (in case of translation of data from one version of the standard to another), validation reports (if validation software is used), ingest report (if aggregator provides software for ingest), confirmation report (a statement that the aggregator received the export as it was agreed), etc.

6.1.3 Metadata folder:

Metadata folder contains:

- descriptive metadata
- preservation metadata

a) Descriptive metadata

The SIP package must be described if or when the archive requires it. Metadata for the description of a SIP package can be based on EAD standard. Tools, available in the eArchiving Building Block, support description in conformance with EAD standard.

b) Preservation metadata

SIP package creation must be documented if/when the archive requires it. Metadata for documentation of SIP package creation can be based on PREMIS standard. Tools, available in the eArchiving Building Block, support documentation in conformance with PREMIS standard.

7 Postface

AUTHOR(S)	
Name(s)	Organisation(s)
Anja Paulič	Archives of the Republic of Slovenia
Jože Škofljanec	Archives of the Republic of Slovenia
Sonja Tomšič	Institute of Oncology Ljubljana, Cancer Registry of Republic of Slovenia
Tina Žagar	Institute of Oncology Ljubljana, Cancer Registry of Republic of Slovenia
Vesna Zadnik	Institute of Oncology Ljubljana, Cancer Registry of Republic of Slovenia

REVIEWER(S)	
Name(s)	Organisation(s)
Karin Bredenberg	Kommunalförbundet Sydarkivera
Jaime Kaminski	Highbury R&D
Begoña Sanchez Royo	Highbury R&D

Project co-funded by the European Commission within the ICT Policy Support Programme Dissemination Level		
P	**Public**	x
C	**Confidential, only for members of the Consortium and the Commission Services**	

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Submitted Revisions History

Revision No.	Date	Authors(s)	Organisation	Description
0.1.0	9th of April 2021	Anja Paulič	Archives of the Republic of Slovenia	Draft for internal review

Statement of originality:
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.