Guideline for the E-ARK Content Information Type Specification for Cancer Registry Exports (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.

DRAFT

1 Preface

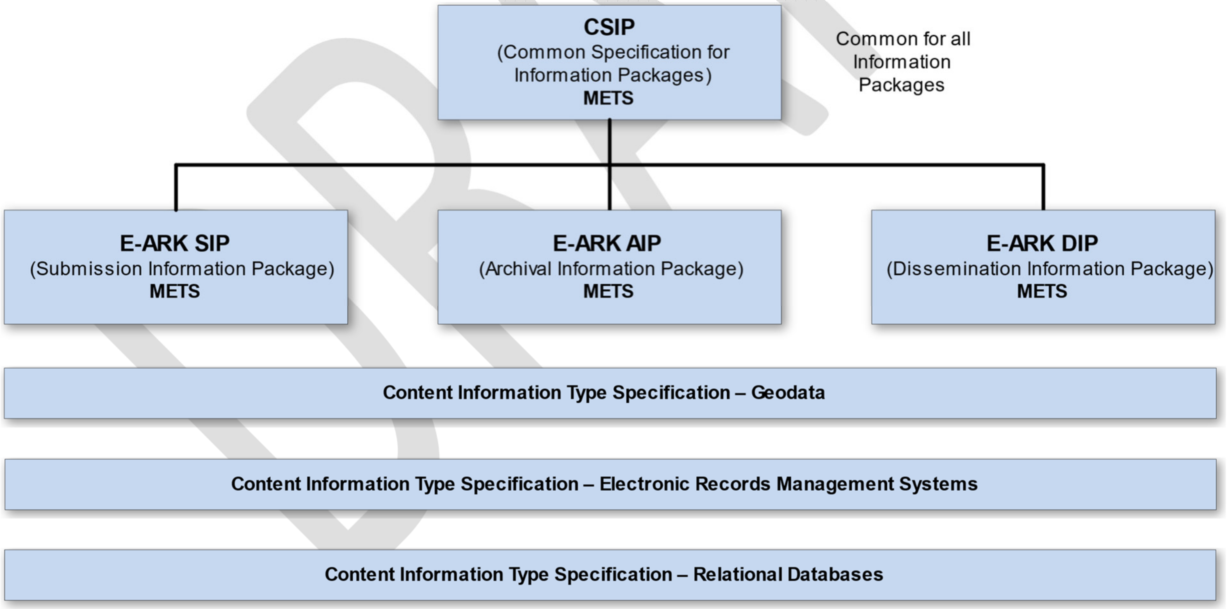
The correct preface will be inserted for the publication occurring after implementation of review comments.

DRAFT

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:


CSIP
(Common Specification for Information Packages)
METS
Common for all Information Packages
E-ARK SIP
(Submission Information Package)
METS
E-ARK AIP
(Archival Information Package)
METS
E-ARK DIP
(Dissemination Information Package)
METS
Content Information Type Specification – Geodata
Content Information Type Specification – Electronic Records Management Systems
Content Information Type Specification – Relational Databases

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.

DRAFT

TABLE OF CONTENT

DRAFT

2 Context

2.1 Purpose

The purpose of this guideline is to further explain and describe the eHealth2 Content Information Type Specification to stakeholders that are interested in keeping and preserving cancer registry exports. These can be for example cancer registries, international aggregators, research organisation/academia, archives.

The eHealth2 specification is a document that describes which information, documents, records have to be preserved to make content of CR exports available and reusable for long-term. It defines how can user of specification organise their data in a submission information package (package that enables maintenance of CR export content), what kind of data should be included in the package (what are mandatory and optional elements).

2.2 Scope

This guideline will provide further information and insights as to how to preserve exports from the cancer registry environment. This guideline will also provide further information and insights to the information packages which are not covered in the CSIP and E-ARK SIP.

2.3 Structure of the document

Aim of section 3 is to get users familiarised with the cancer registries and the data they are gathering. It explains the concept of CR Data and its Export in general and the digital preservation of the CR Data and its Export. It also describes the needs of different types of stakeholders and explains the use of standards in a cancer registry setting.

Aim of section 4 is to explain to the specification users where in the workflow of preparation of cancer registry exports, this specification can be used and what documentation should be preserved/prepared to be later used as a content of the SIP package.

Section 5 provides a rationale for each of the requirements found in the eHealth2 CITS. This is meant to provide a better basis for understanding the reasons behind the requirements. Each rationale contains: requirement, description, rationale and also an example from the actual cancer registry export.

Section 6 is a Glossary of less known terms used in the guidelines.

Section 7 provides a full example of the information package specified in the eHealth2 CITS.

DRAFT

3 CR Exports – Introduction

Aim of this section is to familiarise users of the specification with cancer registries and their activity. Users from the cancer registries, who are familiar with the role and activities of CRs, can find terminology of specification explained here.

3.1 Cancer registry and Cancer data

This section will be extended for the publication occurring after implementation of review comments.

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.

To fulfil that mission cancer registries gather usually in databases next main categories of data and metadata related to cases of cancer diseases: patient, tumour case, incidence, basis of diagnosis, topography, morphology, behaviour, grade, vital status of patient and other data.

3.2 Cancer Registry Export

This section will be extended for the publication occurring after implementation of review comments.

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.

3.3 Stakeholders and their needs

This section will be extended for the publication occurring after implementation of review comments.

The complexity of structure, the quantity of data and the retention period of the export depending on the user or the purpose of the export. For example, international or national

DRAFT

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. ENCR - 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. On the other hand individual researchers may have special requirements and they apply for specific data directly from the cancer registries. Cancer registries share their data also with other national institutions (e.g. national health institute, statistical office, etc.). Such use is usually regulated by national legislation.

DRAFT

3.4 Use of standards

This section will be extended for the publication occurring after implementation of review comments.

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 (aggregators) promote the latest versions of the classifications.

4 How to use the Specification (step by step)

Users of the eHealth2 specification will tailor the use of specification for their use-case. Some will be able to use it at the beginning of the creating export (e. g. those requesting data from the cancer registry), some will need it after the end of the export life cycle (long-term record keepers, archives). This chapter brings attention to those steps of the cancer registry export preparation that produce the documentation that has to be included in the SIP. All descriptions in this section have a reference to requirements in section 5.

4.1 Data export definition documentation (Negotiation and Agreement/Data call)

This section will be extended for the publication occurring after implementation of review comments.

When defining the content of cancer registry export, different documentation can be produced. The type and quantity of it will depend on the type of agreement between cancer registry and aggregator. For example, data calls of international aggregators will produce different documentation, then smaller requests from local research institutions. However, in this step documentation that describes the content and the use of the cancer registry export is produced. Such documentation is important for further use of the export and key for proper future interpretation of the export. Part of documentation that defines export are any amendments and/or clarifications of the content of the cancer registry export. This documentation should also be preserved.

4.2 Documentation about exporting data from CR

This section will be extended for the publication occurring after implementation of review comments.

Especially with data calls from international aggregator’s cancer registry data has to be restructured, selected or otherwise manipulated to conform with the aggregators requirements. If documentation that describes manipulation with cancer registry data so it conforms to the aggregator's requirements is available, it should be included in the SIP.

4.3 Data validation documentation

This section will be extended for the publication occurring after implementation of review comments.

In some cases software validation of the cancer registry export content is possible. Including validation reports in the SIP is strongly suggested. If they are not automatically generated, the print screen of completed validation can be used as such a documentation.

4.4 Export submission documentation

This section will be extended for the publication occurring after implementation of review comments.

If available, documentation that demonstrates successful handover of the cancer registry export should be included in the information package. Later on it can serve as a demonstration of the exported content, adherence to what was agreed/required and possibly changed or added during the export preparation.

DRAFT

4.5 Creating Information Package

This section will be extended for the publication occurring after implementation of review comments.

After submitting the cancer registry export, SIP package can be created. Besides the documentation, described in previous sections, the SIP package should have enough metadata, describing IP content and assure the package's authenticity and completeness. Creating a specification compliant SIP is possible with software tools available within eArchiving Building Block.

4.6 Keeping and maintaining of the Export IP

This section will be extended for the publication occurring after implementation of review comments.

After ingestion in the preservation information system, the SIP package is transformed in the AIP package. Content of the AIP package must be regularly monitored to preserve the package's long term usability. Doing preservation actions (e. g. format migrations) ensures AIP’s long-term usability.

5 Rationale for requirements in eHealth2 CITS

This section will be extended for the publication occurring after implementation of review comments.

This chapter systematically describes all requirements from the eHealth2 specification. All requirements are listed, described and have a rationale, which explains why a requirement should be met, added. Every requirement is illustrated with an example.

DRAFT

6 Glossary

This section will be extended for the publication occurring after implementation of review comments.

A glossary with terms mainly from the OAIS which are described here for understandability of the specifications and the guideline.

Table 1: Glossary

Name	Description
Aggregator	Person or organisation that requires data from the cancer registry.
Archives	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 Tumours is a globally recognised standard for classifying the extent of the spread of cancer. TNM stands for tumour (T), nodes (N), and metastases (M).

DRAFT

7 Example following CITS eHealth2

This section will be extended for the publication occurring after implementation of review comments.

In this section an XML document built upon the specification and the METS Profile for CSIP is described.

[The full example in XML]

DRAFT

8 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

REVIEWER(S)	
Name(s)	Organisation(s)
Karin Bredenberg	Kommunalförbundet Sydarkivera
Jaime Kaminski	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
V 1.0	2021-04-30	eHealth2 team	Various	Draft for 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.

DRAFT