_____

Guideline for the E-ARK Content Information Type Specification for Patient Medical Records (CITS eHealth1)

_____

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:
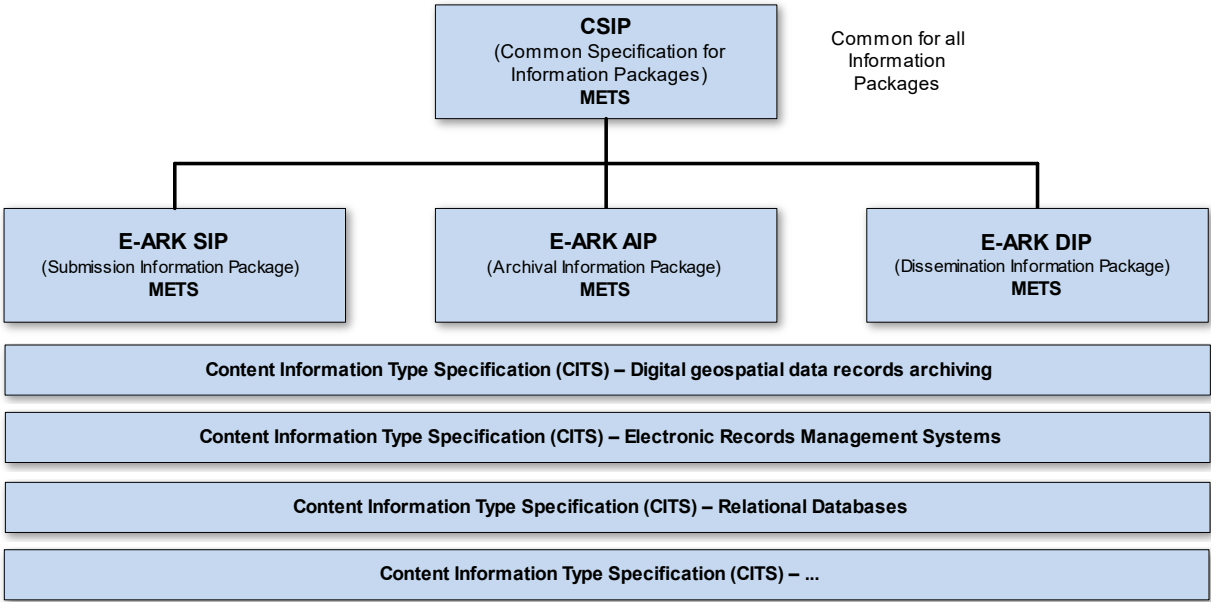

Figure 1: E-ARK specification dependency hierarchy

Table 1: Specification hierarchy aims and goals

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.

_____

_____

Specification	Aim and Goals
	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, http://dilcis.eu/). The role of the DILCIS Board is to enhance and maintain the draft specifications developed in the European Archival Records and Knowledge Preservation Project (E-ARK project, http://eark-project.com/), which concluded in January 2017. The Board consists of eight members, but no restriction is placed on the number of participants taking part 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. The DILCIS Board have been responsible for providing the core specifications to 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 2.

_____

TABLE OF CONTENTS

_____

_____

_____

_____

TABLE OF FIGURES

LIST OF TABLES

_____

_____

2 Introduction

2.1 Purpose

The purpose of this document is to provide background and context for the Content Information Type Specification (CITS) for Patient Medical Records (eHealth1). The specification is further supported by METS profiles for the Root and Representation METS files. The initial development of the specification was based on work done at the Norwegian National Health Archive (NHA) and is being enhanced and developed further through community feedback.

2.2 Scope

This specification makes the following assumptions:

- A business case for the creation of an eHealth archive includes the incorporation of a backlog of physical and digital Patient records.

- An eHealth archive concerns the Complete Patient Medical Records for deceased Patients within the jurisdiction. Note that the term 'jurisdiction' does not imply that a Central Health Archive must be at a national or federal level. Many health administrations are organised at a state or region level, and the specification is equally valid for this scenario. Note also that there are significant potential benefits for the use of the standard for archiving of Patient Medical Records if complied with by all regional administrations within a federation. This can also apply to environments where there are private healthcare providers, and a Central Health Archive is being created by a controlling administration.

- Implementation of Electronic Health Record (EHR) systems is not widespread, and the creation of an eHealth archive that aggregates information from both EMR and EHR systems is considered to be a special case that can be considered within future iterations of this specification (see section 4 for how the specification defines EHR and EMR systems).

- The use cases for an eHealth archive are described in section 4.3.

There are two options for extracting Patient records from an EMR or EHR system which can be dependent to a certain extent on the source system data structure:

2.2.1 Extracting data in a relational database structure

If the structure of the source EHR/EMR system is largely or wholly a relational database, then the extraction of selected records can be made into a long-term database preservation format (SIARD) that preserves the properties of the database such that the data can be imported into a relational database management system (RDBMS) at the time of access. Access can happen through database queries or a search field.

Further information on the limitations of this approach, particularly for the use cases behind the eHealth1 CITS is given in this document.

The SIARD specification, together with a Content Information Type Specification for SIARD, represents the SIP profile for the relational databases content type. More information can be found at https://dilcis.eu/content-types/siard.

Extractions can be made from such relational database systems programmatically that create the aggregated structure described below which conforms to that seen in traditional EMR systems and physical Patient record archives. For the use cases described in this Content Information Specification it is recommended that this approach is followed.

_____

_____

2.2.2 Extracting data and metadata as aggregations of Patient medical records

Digitisation of physical Patient Medical Records or extraction of electronic records from more traditional EMR systems produces a case type structure of files and accompanying metadata as described in section 6. Being extracted in this manner makes them directly accessible for validation, data management, indexing and searching. The structured semantic metadata description is explicit rather than hidden inside an RDBMS. This methodology also supports the incremental extraction of records over time (such as when Patients are deceased) for submission to the Archive, and in addition:

 – Records from different sources can be merged (complete Patient Medical Records can be synthesised from multiple submissions)
 – Search and access is possible across all records and sources
 – Records can be managed individually and uniformly
 – The original EMR/EHR system software does not need to be licensed or preserved

The specification considers this particular extraction method within the context of the use cases as described in section 4.

2.3 Layered Data Model

This section introduces the role of the CITS eHealth1 and its dependencies on the basic structures of the Information Package.

> This specification is created based on the requirements of the Common Specification for Information Packages (CSIP), the specification for Submission Information Packages (E-ARK SIP) and the specification for Archival Information Packages (E-ARK AIP). To fully understand its requirements, we highly recommend that users review the requirements and the terminology of the source documents, before using this specification.

The data model structure is based on a layered approach for information package definitions (Figure 2). The Common Specification for Information Packages (CSIP) forms the outermost layer. The general SIP, AIP and DIP specifications add respectively, submission, archiving and dissemination information to the CSIP specification. The third layer of the model represents specific content information type specifications, such as the eHealth1 specification. Additional layers for business-specific specifications and local variant implementations of any specification can be added to suit the needs of the organisation.

_____
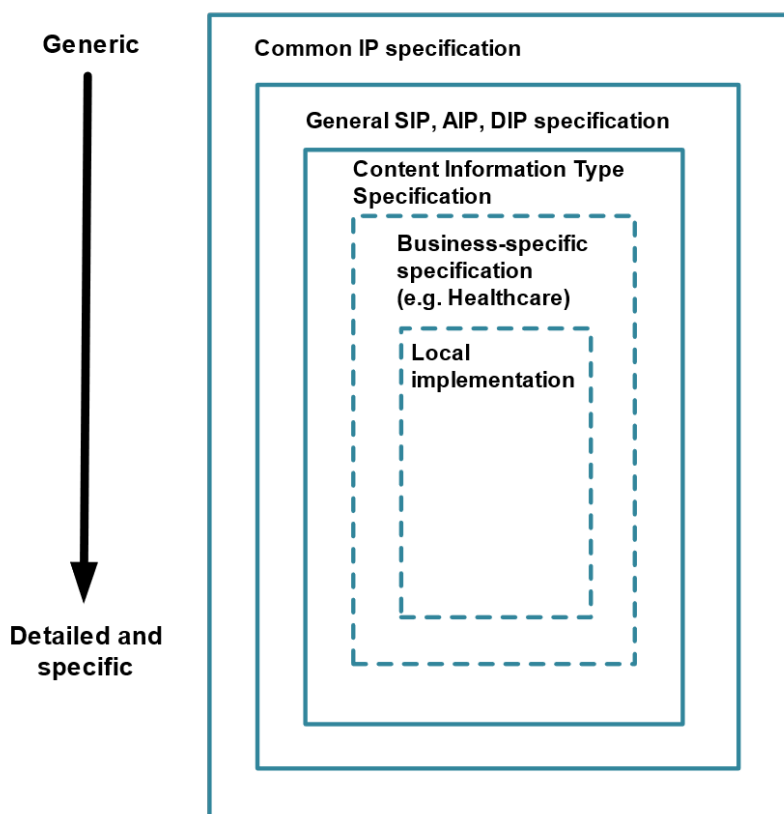
Figure 2: Data Model Structure

Every level in the data model structure inherits metadata entities and elements from the higher levels. In order to increase adoption, a flexible schema has been developed. This will allow for extension points where the schema in each layer can be extended to accommodate additional information on the next specific layer until, finally, the local implementation can add specific entities or metadata elements to satisfy specific local needs. Extension points can be implemented by:

- Embedding foreign extension schemas (in the same way as supported by METS [http://www.loc.gov/standards/mets/] and PREMIS [http://www.loc.gov/standards/premis/]). These both support increasing the granularity of existing metadata elements by using more detailed data structures as well as adding new types of metadata.

- Substituting metadata schemas for standards more appropriate for the local implementation.

The structure allows the addition of more detailed requirements for metadata entities, for example, by:

- Increasing the granularity of metadata elements by using more detailed data structures, or

- Adding local controlled vocabularies.

For consistency, design principles are reused between layers as much as possible.

3 CITS eHealth1 Specification Requirements Structure

The Content Information Type Specification for Patient Medical Records (CITS eHealth1) aims to define the necessary elements required to preserve the accessibility and authenticity of Patient Medical Records over time and across changing technical environments. The specification elevates the level (and

_____

_____

adjusts the cardinality) of some of the requirements set out in the Common Specification (CSIP) and package specifications (namely SIP) and adds new requirements for the package structure, descriptive metadata and accompanying METS files. The specification sets out general principals that underpin the specific requirements and further context for the requirements and principals can be found in this Guideline document.

4 Elements of an eHealth Archive

4.1 Physical and Electronic Patient Records

A Patient Medical Record can be defined as: "a collection or compilation of recorded information about a Patient in connection with healthcare; the Patient record is the principal repository for information concerning a Patient's health care."[1] Prior to the widespread implementation of Electronic Medical Record (EMR) systems, the recording of Patient health records was paper and film-based (plus additional materials which could be images, video, audio).

Electronic Medical Records (EMRs) are digital versions of paper or film records. A healthcare provider may have a single EMR system for all of its Patient records or for larger organisations; there can be fragmentation because of specialisation or organisational sub-division, and a Patient's total medical record at that organisation may be constituted from many subsidiary systems. A considerable volume of these Patient records exist at healthcare providers and within centralised organisations because of legal remits to store the records for extended periods.

A Complete Patient Medical Record may contain information that is sourced from several different organisations' systems (e.g. different hospitals, specialist healthcare providers, primary healthcare providers) and viewed from an archive/academic perspective; the information in each of these organisations constitutes an archive (or several archives). In creating a Central Health Archive, it is necessary for a healthcare provider to make separate extractions from each system, for each Patient to be included in a delivery and to aggregate them before submission to the central archive.

The creation of a Central Health Archive can encompass the digitisation and preservation of physical records as well as the collection and preservation of electronic records from EMR systems. In general, as a Patient's aggregated medical record is not complete until there are no new additions to it (i.e. when the individual has died), then a health archive will consist only of records for Patients who are known or who are believed to be deceased.

4.2 Electronic Medical Record and Health Record Systems

The terms "electronic medical record" and "electronic health record" (or "EMR" and "EHR") can be used interchangeably. However, the difference between the two terms is quite significant and particularly so in the context of archiving standards.

EMR is the older term, and early EMRs were medical in nature; they were for use by clinicians mostly for diagnosis and treatment. Because of a lack of available standards when EMR systems were first developed, the information in EMRs does not travel easily out of a healthcare provider. In fact, the

---

[1] Institute of Medicine (US) Committee on Improving the Patient Record; Dick RS, Steen EB, Detmer DE, editors. The Computer-Based Patient Record: Revised Edition: An Essential Technology for Health Care. Washington (DC): National Academies Press (US); 1997. 1, Introduction. Available from: https://www.ncbi.nlm.nih.gov/books/NBK233055/

_____

_____

Patient's record might have to be printed out and delivered by mail to specialists or other members of the care team. In that regard, EMRs are not much better than paper records.

Electronic health records (EHRs) focus on the total health of the Patient—going beyond standard clinical data collected in the provider's office and inclusive of a broader view on a Patient's care. EHRs are designed to reach out beyond the health organisation that originally collects and compiles the information. They are built to share information with other healthcare providers, such as laboratories and specialists, so they contain information from all the clinicians involved in the Patient's care. The National Alliance for Health Information Technology stated that EHR data "can be created, managed, and consulted by authorised clinicians and staff across more than one healthcare organisation."[2]

The information moves with the Patient—to the specialist, the hospital, the nursing home, or even across a region or country. In comparing the differences between record types, HIMSS[3] Analytics stated that "the EHR represents the ability to easily share medical information among stakeholders and to have a Patient's information follow him or her through the various modalities of care engaged by that individual." EHRs are designed to be accessed by all people involved in the Patient's care—including the Patients themselves. Indeed, that is an explicit expectation in the Stage 1 definition of "meaningful use" of EHRs.

The benefits of EHR systems to Patient care mean that the trajectory for healthcare worldwide is towards national or regional EHR systems. The complexity and lack of standards in existing systems mean that realisation is difficult and expensive. Adoption is hence not yet widespread. Implementations of EHR systems can also rely on summary Patient data gathered by means of standardised clinical documents (such as HL7 CDAs). This means that extractions from EHR systems can sometimes only yield Patient summary data and not the Complete Patient Medical Record.

The development of standards and technology that make EHR systems possible (such as the encoding of key clinical data, medical data interoperability standards such ICD, DICOM, SNOMED and HL7 FHIR) makes the future scope of a national health archive a different proposition; systems will exist containing a Patient's total health history, richly encoded and ideally suited to analytical techniques for 'big data'. Systems in principle will be able to grow over time containing records from both live and deceased Patients.

4.3 Use Cases for a Central Health Archive

The eHealth1 Content Information Specification (CITS) was developed as a result of creation of a centralised Patient Medical Record archive in Norway. According to the health archive regulation, the mission of the Norwegian National Health Archive (NHA)[4] is to:

a) receive and preserve Patient archives from public and private hospitals, and

b) to disseminate health information for researchers and the Patients next of kin in compliance with regulations and confidentiality acts.

There is no limit to the age of the records to be presented to the NHA from hospitals and so consist of physical and electronic Patient records.

_____

[2] https://www.healthcareusability.com/article/terminology-hit-emr-ehr
[3] HIMSS Analytics himssanalytics.org
[4] https://ehandbok.arkivverket.no/folder/92

_____

The Norwegian regulation envisions two possible use cases for the archive when built, which are:

a) To provide records to next of kin in compliance with open information regulation;

b) To harvest the vast amount of historical healthcare-related data within the archive for medical research.

In order to achieve use case a, it is necessary to ensure that the specification allows for access to all of the records pertaining to a single Patient, regardless of the submitting institution.

Use case b. requires that the specification allows for ingestion of digitised records and the ingestion of extracts from EMR systems for all deceased Patients and that sufficient metadata is provided to enable searches across the archive to create cohorts to support medical research. Metadata regarding Patient personal information and Patient clinical information may be encoded in EMR systems or may have to be entered at a digitisation stage. The scope of the metadata to be included in the archive is therefore very much a determination for the local and national organisations based on the existing records, resources available, standards, etc.

5 Standards

5.1 eHealth standards and use in the eHealth1 specification

Controlled vocabularies and coding provide a standardised way for the unambiguous recording of health data. Most EMR and all EHR systems will hold coded data concerning Patient Cases that can be extracted as metadata for the Patient Medical Record and will use an international standard such as ICD or SNOMED. Data can be recorded in a standardised (such as OpenEHR) formats or to a local format which is specified by the health archive and referenced within a Submission Agreement.

5.2 HL7 FHIR5

Fast Healthcare Interoperability Resources (FHIR, pronounced "fire") is a standard describing data formats and elements (known as 'resources') and an application programming interface (API) for exchanging electronic health records (EHR). The standard was created by the Health Level Seven International (HL7) healthcare standards organisation.

Its goals are to facilitate interoperation between legacy healthcare systems, to make it easy to provide healthcare information to healthcare providers and individuals on a wide variety of devices from computers to tablets to mobile phones and to allow third-party application developers to provide medical applications which can be easily integrated into existing systems.[6]

FHIR provides resources that can be used for the standardised description of Patient Personal Information and Patient Clinical Information, which reference controlled vocabulary and coding standards such as ICD and SNOMED. The use of FHIR is suggested within eHealth1, but local standards for encoding metadata are allowable if specified elsewhere and referenced within a Submission Agreement.

---

[5] https://www.hl7.org/fhir/summary.html
[6] https://en.wikipedia.org/wiki/Fast_Healthcare_Interoperability_Resources

_____

5.3 HL7 Clinical Document Architecture and CDA R2 International Patient Summary[7]

HL7 CDA provides a standard for the organisation of material within clinical documents for exchange between systems. By using XML, the HL7 v3 standard and coded vocabularies, the CDA facilitates the exchange of both machine and human-readable documents, enabling electronic processing for decision support, etc., whilst being easily retrieved and used by the people who need them (HL7 UK, 2018).

According to HL7 : "An International Patient Summary (IPS) document is an electronic health record extract containing essential healthcare information intended for use in the unscheduled, cross-border care scenario, comprising at least the required elements of the IPS dataset. The IPS dataset is a minimal and non-exhaustive Patient summary dataset, specialty agnostic, condition-independent, but readily usable by clinicians for the cross-border unscheduled care of a Patient."

5.4 ICD[8]

The International Classification of Diseases is the foundation for the identification of health trends and statistics globally and the international standard for reporting diseases and health conditions. It is the diagnostic classification standard for all clinical and research purposes. ICD defines the universe of diseases, disorders, injuries and other related health conditions, listed in a comprehensive, hierarchical fashion that allows for:

- easy storage, retrieval and analysis of health information for evidence-based decision making;

- sharing and comparing health information between hospitals, regions, settings and countries; and

- data comparisons in the same location across different time periods.

ICD is mapped from other standards such as HL7 FHIR and will be part of the process used by many institutions to record Patient Clinical Information. The use of international standards such as ICD within supplied clinical metadata is encouraged but will be limited by their use within the source EMR or EHR system.

5.5 SNOMED9

SNOMED CT or SNOMED Clinical Terms is a systematically organised computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. SNOMED CT is considered the most comprehensive, multilingual clinical healthcare terminology in the world. The primary purpose of SNOMED CT is to encode the meanings that are used in health information and to support the effective clinical recording of data to improve Patient care. SNOMED CT provides the general core terminology for electronic health records.[10]

SNOMED CT is mapped from other standards such as HL7 FHIR and will be part of the process used by many institutions to record Patient Clinical Information. The use of international standards such as SNOMED CT within supplied clinical metadata is encouraged but will be limited by their use within the source EMR or EHR system.

_____

[7] http://www.hl7.org/implement/standards/product_brief.cfm?product_id=483

[8] https://www.who.int/classifications/icd/en/

[9] http://www.snomed.org

[10]

https://en.wikipedia.org/wiki/SNOMED_CT#:~:text=SNOMED%20CT%20or%20SNOMED%20Clinical,in%20clinical %20documentation%20and%20reporting.

_____

5.6 DICOM[11]

Digital Imaging and Communications in Medicine (DICOM) is the standard for the communication and management of medical imaging information and related data.

A DICOM file is a file that encapsulates attributes and bit streams (image, video, etc.) and has embedded Patient Personal Information and IDs. DICOM files have a recognised MIME file type. Extraction of DICOM files from specialised EMR systems for inclusion in Patient Medical Records should present no problem, but it should be ensured that Patient IDs in DICOM files match those in archival package Patient Personal Information.

5.7 eHealth DSI (eHealth Digital Service Infrastructure)[12]

The eHealth Digital Service Infrastructure (eHDSI or eHealth DSI) is the initial deployment and operation of services for cross-border health data exchange under the Connecting Europe Facility (CEF). It defines a document framework or Clinical Document Architecture (CDA) for sharing medical data across borders (Patient Summary). As E-ARK eHealth1 considers the totality of a Patient Medical Record, the eHDSI is too limited in scope to be useful in this context, eHDSI aims to specify an interchangeable derivation and extract of a Patient Medical Record, whereas the E-ARK eHealth1 CITs aims to preserve the Patient Medical Record in its entirety.

6 Data Structure and Aggregations

6.1 Case Structure

The names of aggregation levels within an archive and represented within an archival package (IP) will depend on the agreements between data producers (Creators) and archives. EAD3 has defined a set of values (class, collection, file, fonds, item, otherlevel, recordgrp, series, subfonds, subgrp, subseries) for that purpose, and it allows other values to be used in addition if they are defined as "otherlevel". However, even though the aggregation levels in this context could be described in this way, the EAD template for archival description is considered broadly unsuitable for a Patient Health Archive.

A Central Patient Health Archive has a single purpose and will most probably (due to security constraints) be instituted as a stand-alone entity or as a sub-entity within a larger institution (e.g. National Archive or Health Authority). The overall aggregation of a health archive is therefore implicit (it is an aggregation of Patient Medical Records), and further aggregation levels must be defined that suit the use cases for navigation within the archive and for the way in which the archive is populated.

Patient data will most likely be submitted by hospitals or other healthcare providers in periodic batches, consisting of multiple Patient records. Patient Medical Records will be submitted to a Central Health Archive either when a Patient is known to have died or after a period of time when it is not feasible that a Patient is still alive (determined through regulations). Depending upon the availability of a National Death Register, the accessibility and responsiveness to such a register and the periodic batching of archival extracts at healthcare providers, it cannot be expected that individual Patient submissions from multiple creators will be at all coordinated. Aggregation of a Complete Patient Record at the archive prior to submission into the preservation system is therefore deemed in this specification to be unlikely to be practical.

---

[11] https://www.dicomstandard.org/current/
[12] https://ec.europa.eu/cefdigital/wiki/display/EHOPERATIONS/eHealth+DSI+Operations+Home

_____

The proposed data structure for the aggregations of the submissions of Patient Medical Records is as shown in the data model in Figure 3. As Patient data is likely to be submitted in batches, each submission package will contain information from multiple Patients, and it is likely that these submissions will be split by the archive on receipt to create Patient-specific archival information packages (AIPs) in order to simplify the dissemination process. In this context, the submission package could be considered as a submission information collection (SIC) or collation of SIPs which is compiled to simplify extraction and transmission. However, for the purposes of this specification, the term SIP is used to mean both a submission package for a single Patient record or a submission package containing multiple Patient records.

The levels of the aggregation in an eHealth1 package are as follows:

- Patient: An individual who has received healthcare at any number of healthcare providers and who is described by Patient Personal Information (see 8.4.2). Each Patient will be identified by means of a unique identifier (ID) which is provided from the source EMR system. This unique ID connects the Patient Personal Information and the Patient Medical Record in the information package.

- Case: A Patient Medical Record can be structured in various ways, which may be dictated by national standards, guidance or local practice. A Patient's Complete Medical Record will consist of multiple individual thematic Cases which may be concerned with particular medical conditions, periods or treatments. The proposed aggregation allows for flexibility in this grouping. These cases will be held in a healthcare provider's local archive and may contain a number of Sub-cases and/or Documents with associated Data Files.

- Sub-case: A Sub-case is an allowable type of component consisting of a set of Documents and Data Files that is nested below a Case. Sub-cases may originate in departments within a large hospital or may be related to a different diagnosis to other Sub-cases. A Sub-case may have common (to the Case) or specific metadata.

- Document: A Document is a component that may consist of multiple related Data Files with common metadata; for example, a document may be a PDF file together with associated attachments, or there may be a document and a separate signature sheet. A document can be considered to be an entity that is approved/signed as a whole.

- Data File: A Data file is a component that contains data and has an associated MIME file type. A Data File can be a single bit stream or can encapsulate bit streams and attributes according to a standard such as a DICOM or MP4, in which case it will have a recognised MIME file type. A Data File which is a container for multiple byte streams and metadata, can be included in the package as a Data File or can be unpacked and included as separate Byte Streams and metadata in METS. It is expected that containers such as DICOM and MP4 files will be submitted unaltered in Submission Information Packages (SIPs) and that any decision to unpack them is part of a preservation plan at the archive.

- Byte Stream: A Byte Stream is a component that contains data, has an associated MIME file type and is encapsulated in a container such as MP4, DICOM or Matroska. Each Byte Stream has its own associated metadata, such as technical metadata which is generally only accessible with specialised tools (such as ffprobe for video container formats).

_____
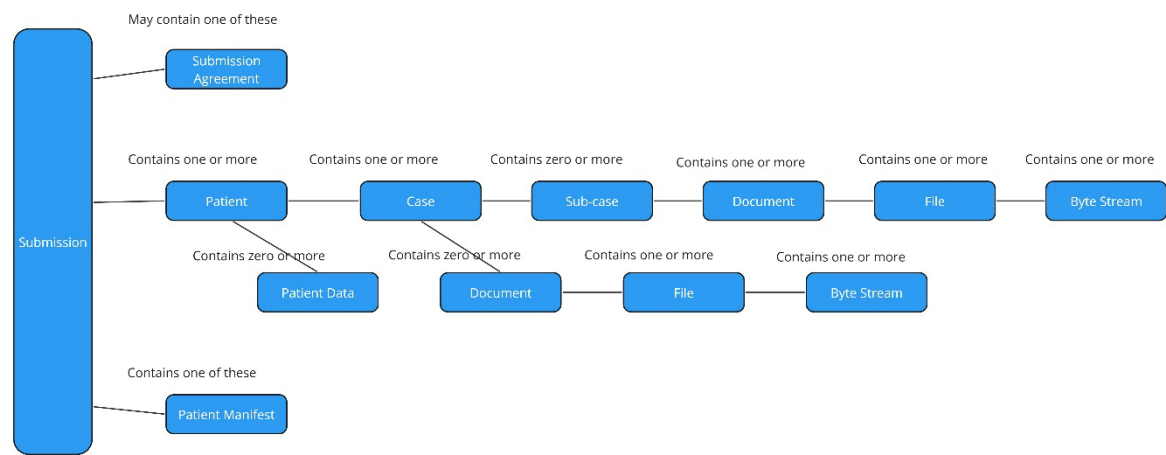
Figure 3: eHealth1 SIP Data Model

6.2 Examples of Different Patient Record Submissions

With the flexibility of the structure of the eHealth1 archival package and the differences that are likely to be found in making Patient Medical Record extractions from disparate EMR systems, there can be expected to be different cases for the extraction of records.

6.2.1 Example 1: The entire archived Patient Medical Record as one file (document)

In this example, the extraction of a Patient's Medical Record consists of one unstructured file in, for example, PDF format, which contains a complete extract from an EMR system. In such a case, an archived Patient Medical Record will consist of one Case containing one Document and one Data File (see Figure 4).
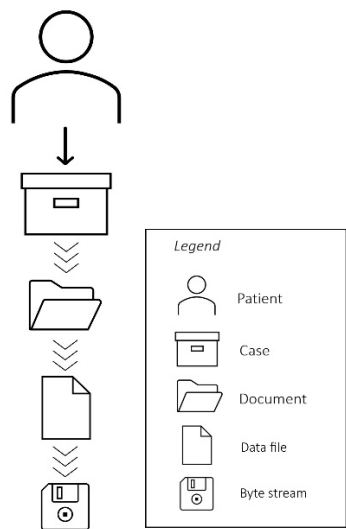
Figure 4: Archived Patient Medical Record as one file

_____

6.2.2 Example 2: The archived Patient Medical Record as a set of thematic files (documents)

In this example, extraction of the Patient's Medical Record consists of a set of unstructured files, typically PDF documents, where each file includes all of the information within a subject/theme that reflects the organisation of information in the current system. In this example, an Archived Patient Medical Record would consist of a number of Cases, each containing one Document, each containing one Data File (see Figure 5).
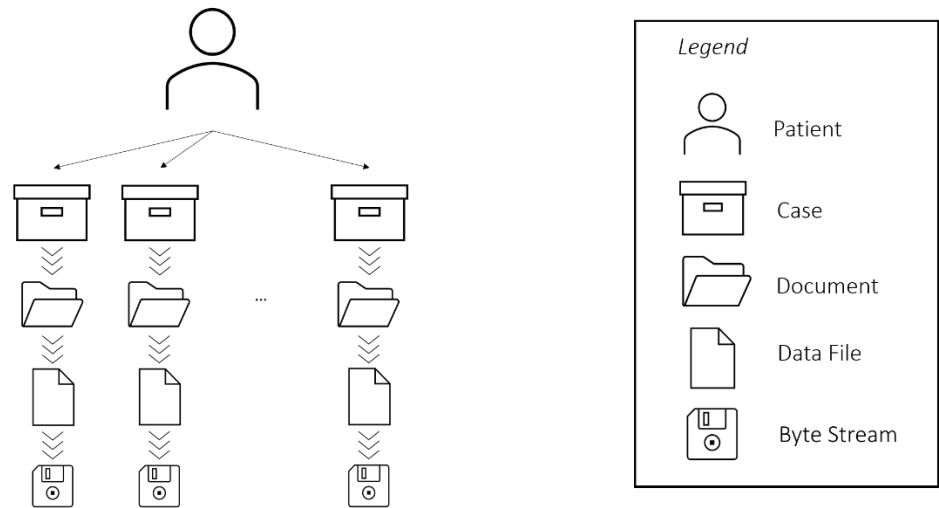
Figure 5: Archived Patient Medical Record as a set of thematic files

6.2.3 Example 3: The archived Patient Medical Record as a set of Documents per Case

In this example, extraction of the Patient's Medical Record consists of a set of unstructured files which can be documents, images, videos, DICOM files, etc., and where each Data File may be related to other Data Files within a Document which can be related to each other within a Case or a Sub-case (see Figure 6).

_____

_____
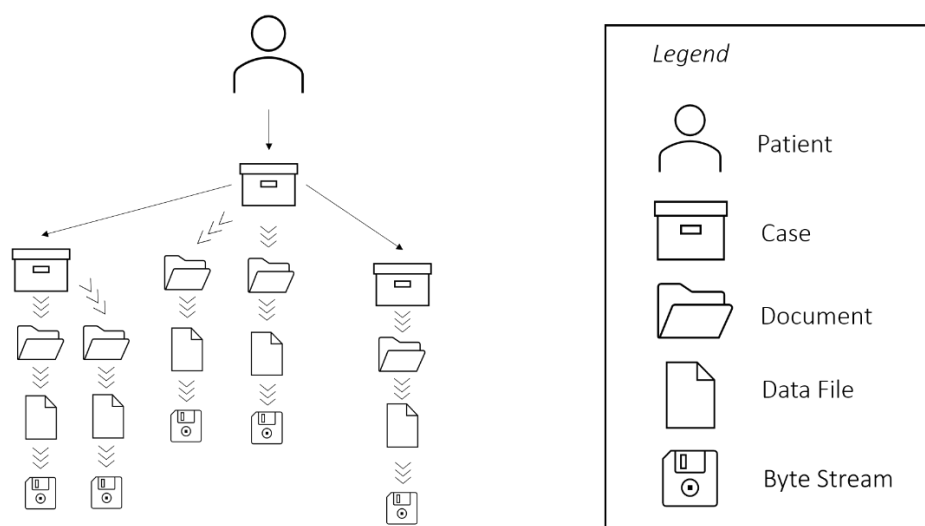
Figure 6: Archived Patient Medical Record as a set of Documents per Case or Sub-case

6.3 Using the eHealth1 specification together with the Common Specification for Information Packages (CSIP)

The eHealth1 specification conforms to and extends the Common Specification for Information Packages (CSIP) and the specification for Submission Information Packages (E-ARK SIP). When extractions are made from EMR systems according to the structure described, they can be transmitted in a package following the principles described in the CSIP and IP specifications.

6.4 Placement of data in an eHealth1 Information Package

As described in 6.1, Patient data as submitted by hospitals or healthcare providers are likely to be periodically extracted from source systems and sent in batches. The eHealth1 specification allows for the inclusion of multiple Patients per package, and so these batches can be transmitted in a single submission. The number of Patients included in each AIP is then a matter for local implementation, although the decision in Norway at the National Health Archive was for each AIP to consist of data from a single Patient and from a single Submitting Organisation..

Patient Medical Records are placed in a single representation within the representations folder of the package. The ID of the representation should have a name that follows the requirements of CSIP.

The Patient Medical Record representation should contain a METS file at its root (Representation METS), and the folder structure must follow that defined by the CSIP including containing a data folder. Multiple Patient Records are organised within the representation data folder in individual folders that have names that must contain the Patient unique identifier.

It is recommended but not mandated that within each Patient Record folder that there are further folders that physically represent the Case, Sub-case, Document structure to aid human readability and navigation of the archive. If Patient Administrative and/or Patient Clinical Information is provided, then this is held at the appropriate level within the Case structure in the Patient Medical Record. Figure 7 shows an example of a folder structure for a representation where there are multiple Patient submissions and clinical metadata included.

The package should contain a Patient Administrative Information or manifest file within the root metadata/descriptive folder that at minimum contains the names of the Patients whose records are contained in the package and a reference to their Patient ID.

_____

_____
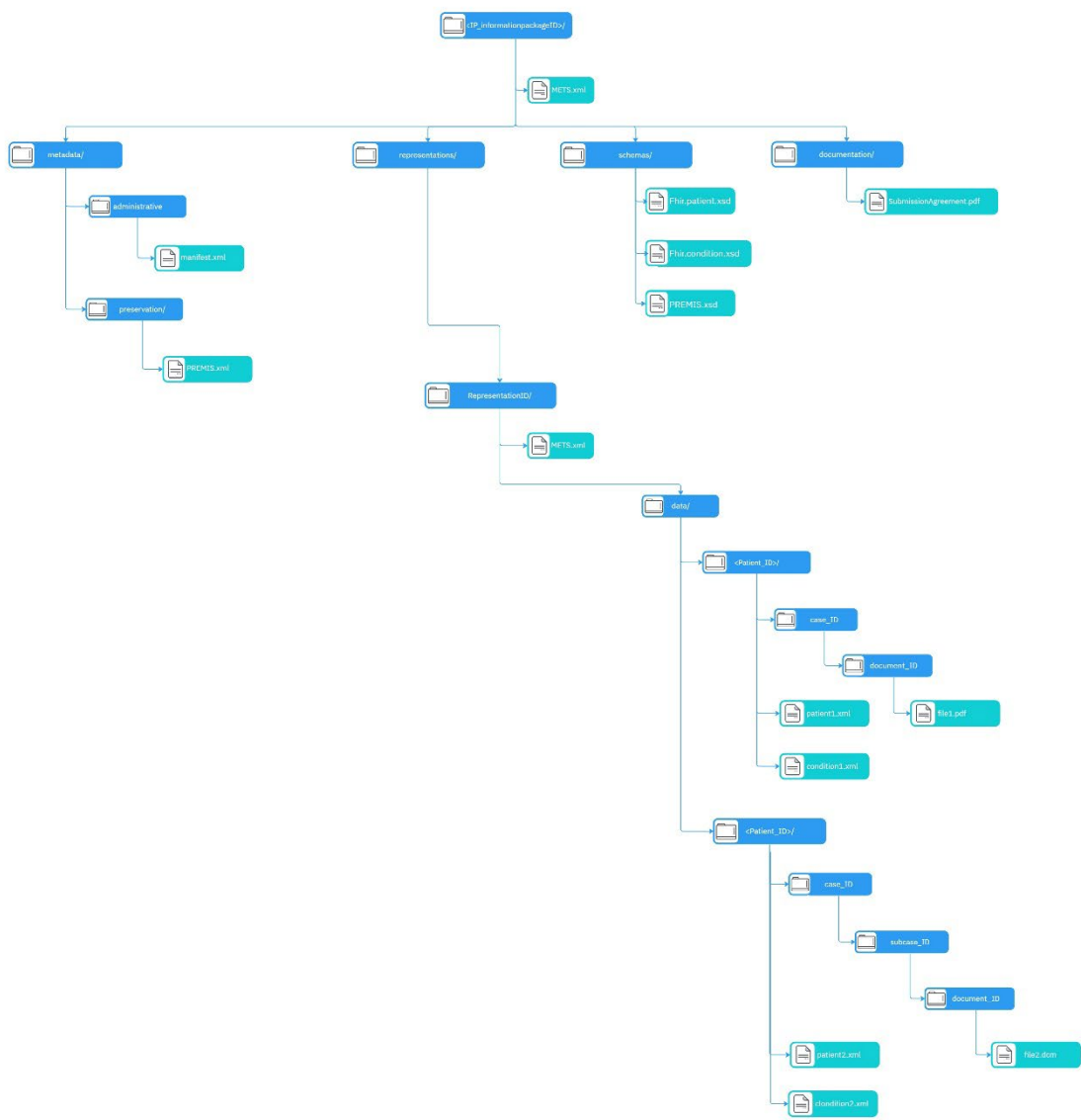
Figure 7: Example of Package Folder Structure with Multiple Patient Submissions and Case Structure

6.5 Archival Package (AIP) Representations

The CSIP and SIP specifications allow that packages contain multiple representations of data that form a single intellectual entity, be this an aggregation within an archival taxonomy, geophysical data with a given boundry or a complete archived relational database. Representations allow the same intellectual entity to be represented in different formats for example for long-term preservation or for access purposes.

In this version 2 of the eHealth1 specification Patient Medical Records have been organised into a single representation of a Submission Package such that multiple representations of data can be created within Archival Packages in order to aid preservation. This is a different organisation to version 1 of the specification, but allows the archiving of packages containing multiple, different Patient Records per AIP and for the for generation of additional representations of the data over time.

_____

_____

7 General Requirements & Rationales

EHGR1 – submission packages MUST contain at least one representation containing data from one or more Patients.

Rationale – the eHealth1 SIP is structured to allow aggregations of multiple Patient submissions into a single package, within a minimum single representation.

EHGR2 – data from multiple Patients if present MUST be divided into separate Patient Record folders in the data folder of the representation.

Rationale – aggregations of multiple Patient's data into a single SIP eases and simplifies the task for data producers to submit many records in batch submisions and subsequent processing at the archive. Organisation of the Patient Records into a single representation in the SIP allows for creating multiple Patient Recoprd AIPs and for generation of additional representations over time.

EHGR3 – Patient data in a Patient record SHOULD follow a Case/Document/File or Case/Sub-case/Document/File structure.

Rationale – the Case/Sub-case/Document/File structure in eHealth1 follows a typical Patient Medical Record structure found in physical and EMR systems. It allows both digitised and born digital records to co-exit in the same archive and the Patient centricity allows secure and simple extraction of all data for a single Patient.

EHGR4 – each submission package SHOULD contain a submission agreement in the root /documentation folder.

Rationale – a submission agreement between producer and archive details the agreement reached between the archive and producer on submission formats and other submission conditions or arrangements. A machine-readable format is recommended sych as that developed by Docuteam GmbH at: http://www.loc.gov/standards/mets/profiles/00000041.xml

EHGR5 – there MUST be a Patient manifest or Patient Administrative Information file located in the root /metadata/descriptive folder that at a minimum contains Patient names and unique identifiers. The Patient Administrative Information file MAY contain personal, demographic and clinical information such as to aid searches for next of kin and research cohorts.

Rationale – metadata in the root of the package should facilitate search and location of individual Patient Records such as to facilitate location of all data associated with a single Patient. It can also contain information such as to support searches based on demographic or clinical information in order to build cohorts of data for researchers.

EHGR6 – each Patient Record SHOULD contain additional Patient Administrative Information and Clinical Information file(s).

Rationale –data/metadata located in the individual Patient Records can be extracted at the archive and used to build search indexes and databases such as to meet use cases for access such as next of kin information requests and creation of cohorts of data for research purposes

_____

_____

8 Metadata

8.1 Use of METS in eHealth1

CSIP specifies that METS files be located at the root of the package folder structure (Root METS) and optionally in each of the Representations within its respective root folder (Representation METS). As has been described previously, the eHealth1 CITS defines a package that has been submitted by a single institution and may contain information concerning either single or multiple Patients. In the case of a multiple Patient submission, there will be multiple Patient Records within a data folder in a single Representation, with its own Representation METS file.

8.2 Root METS File

The root METS file must adhere to the requirements of the CSIP and Information Package specifications. In addition, there are specific requirements for the eHealth1 CITS, and in some cases, the level of the CSIP or package requirements have been increased (but never decreased). Detailed requirements for the root METS file are found in the specification document.

8.3 Representation METS File

The Representation METS file is used to describe the data structure of the Patient Medical Records held in the data folder of the Representation via the structMap element and to reference any additional technical metadata. Details of requirements for the Representation METS can be found in the specification document.

8.4 Use of Descriptive Metadata in eHealth1

8.4.1 Patient Identifiers

Patients must have a nationally unique identifier that is referenced within the source EMR system, such as a Social Security or other unique individual identifier.

8.4.2 Patient Personal Information

Patient Personal Information should, wherever possible conform to an international or national standard for describing Patient information within EMR of EHR systems (e.g. HL7 FHIR contains a well-specified resource for Patient Personal Information and provides schemas in a number of formats).[13] At a minimum the file located in the root/metadata/descriptive folder must contain Patient names and unique identifiers and may contain personal and demographic information.

8.4.3 Patient Clinical Information

Structured Patient Clinical Information such as diagnoses, procedures, medication, allergies, etc., can add significant value to the Health Archive and, in particular, to the research use cases as described above. Patient Clinical Information associated with the Patient or Patient Cases can be added to the individual Patient Records either at Patient or appropriate Case, Sub-case or Document level. Patient Clinical Information should, wherever possible, conform to an international or national standard for describing Patient Clinical Information with EMR or EHR systems (e.g. HL7 FHIR contains well-specified resources for clinical, diagnostic and medication data and provides schemas in a number of formats).[14] Clinical Information should use recognised vocabularies and coding such as ICD and SNOMED.

---

[13] https://www.hl7.org/fhir/Patient.html

[14] https://www.hl7.org/fhir/clinicalsummary-module.html, https://www.hl7.org/fhir/diagnostics-module.html, https://www.hl7.org/fhir/medications-module.html

_____

_____

9 Examples of application of the eHealth1 Specification

9.1   Example 1 – the Piql eHealth SIP Creator (piqlIngest)

9.1.1 Introduction

The Piql EHealth SIP Creator was developed as part of the E-ARK3 project in 2021 by Piql AS of Norway who were also the lead authors of the eHealth1 CITS as part of the project consortium. The objective of the development was to accompany the creation of the eHealth1 specification with: "..new supporting software services ..." and specifically "inclusion of an eHealth1 SIP creation component into the Sample Software Portfolio." Consequently the Piql activity focused on the creation of an eHealth SIP creation software tool.

The eHealth1 specification builds on work done in Norway for standardisation of submissions to the National Health Archive (NHA), specifically via the EPJ (Electronic Patient Journal) specification for extraction of records from hospital Electronic Medical Record systems (EMR)[15]. The Norwegian specification assumes that EMR system vendors will adapt their software to produce compliant submissions (as is required of the hospital by law). It is anticipated that this could be problematic in any jurisdiction, as there are few incentives for EMR system vendors to do this work. In general, any data producers requirement will be for simplicity in the production of compliant SIPs, i.e. using the standard database structures of the host EMR, extraction using easily available tools on standard platforms, consistency in processes and efficient production with common IT skills.

The Piql SIP Creator specification described an approach that required a lower level of bespoke work at system vendor or hospital level and proposed a SIP creation tool that could universally produce compliant submissions (to eHealth v1.0) for data exported from source EMR systems through SQL (or similar) queries. In addition the project produced a User Guide detailing the requirements for the submitting organisation.

Any central health archive may introduce additional requirements for the SIPs which extend or make mandatory requirements within the existing specification (e.g. metadata standards, submission agreements). The tool is open-source and available at GitHub at: https://github.com/E-ARK-Software/piql_ingest. Any Central Health Archive or data producer can modify the code to suit individual requirements. Note that the tool was written to support v1 of the specification and there is no current plan for it to be adapted to support v2.

9.1.2 SIP Creator Tool

The Piql E-ARK SIP Creator is a version of an existing Windows desktop ingest tool (PiqlIngest) that was modified by Piql to create conformant E-ARK eHealth1 SIPs (to v1.0) from exports of Electronic Medical Record (EMR) systems structured in a prescribed manner. The tool can also automate bespoke scripts that transform non conformant into conformant exports by mapping and renaming directory structures and by mapping metadata files.

The eHealth1 SIP Creator is delivered as a self-contained zipped directory structure which contains executable code, dependencies, reference metadata schemas and sample data. The code is available from GitHub at: https://github.com/E-ARK-Software/piql_ingest. The contents of the SIP Creator package are shown in Figure 8 and the GUI is shown in Figure 9.
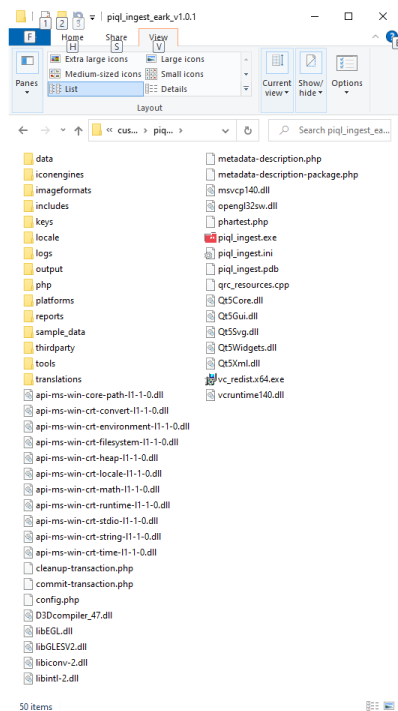
_____

[15] https://www.ehelse.no/standarder/epj-standard-del-5-arkivuttrekk

_____

_____

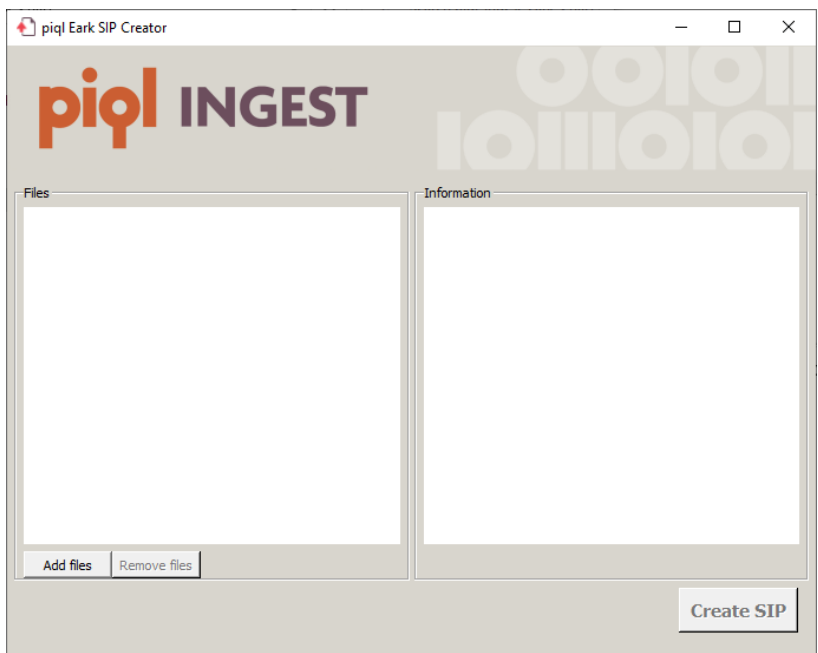**Figure 8: Piql SIP Creator Package Contents**

Figure 9: SIP Creator Window

9.1.3 Descriptive Metadata

The eHealth1 CITS Specification requires that Patient Administrative Information must be included in the information package at the root level and that Patient Clinical Information should be included within

_____

_____

each Patient Medical Record.The FHIR Patient and Condition resources are recommended but not mandated and the SIP Creator has been configured to use these schemas as standard. At a minimum the Patient Administrative Information should be a simple Patient manifest with Patient names and references to unque identifiers.

eHealth1 SIPs should include all necessary xml schemas and the SIP Creator has been profiled to include schemas for METS, METS extensions, xlink, xml and the fhir base, Patient and condition schemas.

9.1.4 Extracting Patient Records from EMR Systems

It is assumed that with the skills of a local DBA and basic scripting skills that organisations will be able to:

- Extract Patient administrative metadata together with unique Patient IDs and save to an xml file which will comply to either: the fhir-Patient resource, or a local xml schema. This can be for multiple Patients for each submission.

- Extract Patient medical record information (folders, documents, files) as separate Patient records each organised according to one of the case structures as described in the eHealth1 specification.

- Extract minimum viable Patient case clinical data that conforms to a recognised vocabulary (such as ICD), is linked to Patient cases in one of the above structures, is saved to xml files that comply with either: fhir-condition resource, any other fhir clinical resource, or to a local xml schema. There should be references for each record to the relevant Patient cases.

Patient ID and Case IDs should be included in the Case folder names.

9.1.5 Using the SIP Creator
Submission packages are ingested into the tool by dragging and dropping from the top level folder of the submission.

The tool presents a metadata editing template with four tabs as follows:

- Software Version – this is pre-filled

- Submission Agreements

- Path or URL to submission agreement

- An identifier or reference code for the submission agreement

- Path or URL to the previous submission agreement

- An identifier or reference code for the previous submission agreement

- Agents

    o Creator organization name

    o Creator organization identifier

    o Archive organization name

    o Archive organization identifier

_____

_____

- o   Preservation organization name

- o   Preservation organization identifier

- o   Submitter name (individual)

- o   Submitter details (e.g. email address)

- • Metadata schemas

- o   Patient personal information schema name

- o   Path or URL to Patient personal information schema

- o   Patient clinical information schema name

- o   Patient clinical information schema location

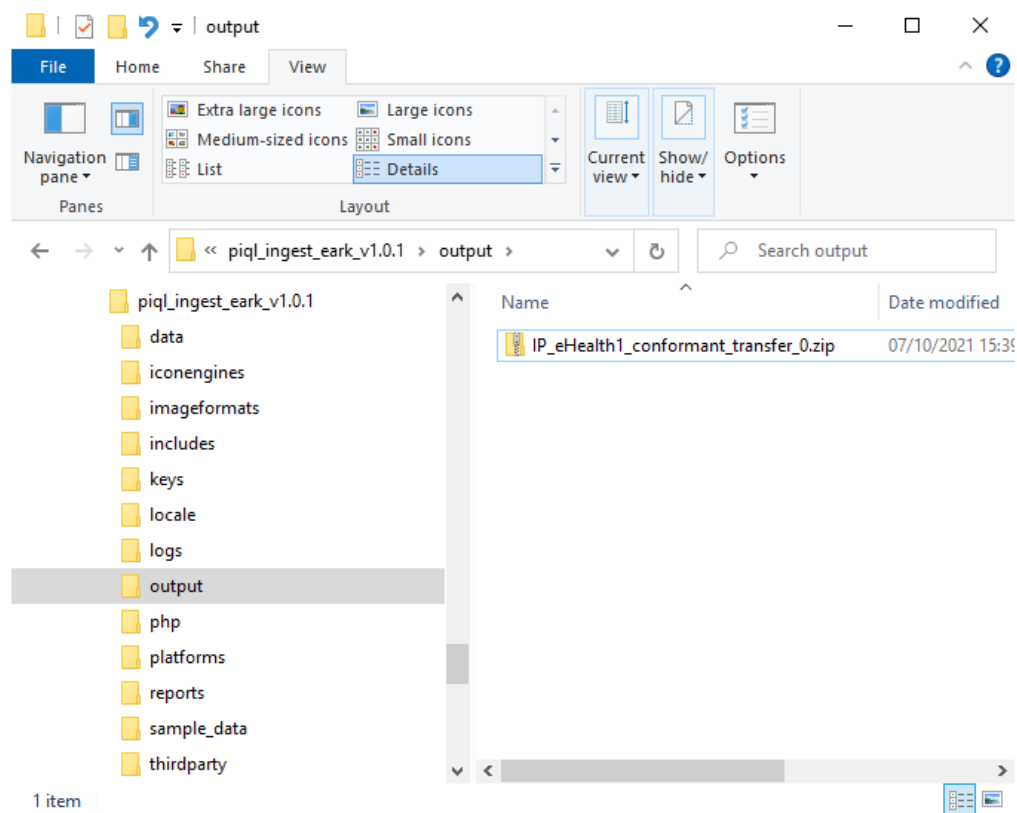The output of the process is placed in the /outputs folder of the package as seen in Figure 10.

Figure 10: SIP in output folder

_____

10   Glossary

Table 2: Glossary

Term	Description
Archival Creator	Organisation unit or individual that creates records and/or manages records during their active use.
Archival Information Package (AIP)	An information package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an Open Archival Information System (OAIS).
Cardinality	The term describes the possle number of occurrences for elements in a set. The numbers have the following meanings: (1..1) – in each set, there is exactly 1 such element present (0..1) – the set can contain from 0 to 1 of such elements (1..n) – the set contains at least one element (0..n) – the set can contain up to n of such elements, but it is not mandatory (0..0) – the element is prohibited to use
Case or Patient Case	Type of component consisting of a set of objects and/or sub-cases. This is represented in the specification as a directory that sits within the data directory of a representation (which in this case is a Patient's Medical Record). A Case is an aggregation of individual records related to one Patient and which are related in a way that is defined by national standards, guidance or local practice. A Patient's Medical Record will consist of multiple individual thematic Cases which may be concerned with particular medical conditions, periods or treatments.
Central Health Archive	An organisation within a national or regional jurisdiction with a (usually legal) remit to create an archive of Patient Medical Records for people who have received primary or secondary healthcare in the jurisdiction. The Central Health Archive will be populated with Patient Medical Records from multiple healthcare providers in the jurisdiction, which will be drawn from Local Patient Health Archives (e.g. a hospital archive).

_____

Component	In this standard: meaningful, logically delimited, and uniquely identifiable information that may be subject to treatment in manual and/or automated processes. This standard operates with four generic types of components: Case, Document, Data File and Byte Stream.
Complete Patient Medical Record	The sum of the submissions of Patient Records made for an individual.
Content Data Object	The Data Object, that together with associated Representation Information comprises the Content Informartion (Source OAISA – ISO 14721:2012)
Content Information	A set of information that is the original target of preservation or includes part or all of that information. It is an Information Object composed of its Content Data Object and its Representation Information. (Source OAIS – ISO 14721:2012)
Data File	A component which contains data and has an associated MIME file type. A Data File can encapsulate multiple bit streams and metadata according to a standard such as a DICOM but must have a recognised MIME file type. A Data File may comprise one or more subsidiary Byte Streams; for example, an MP4 file might contain separate audio and video streams, each of which has its own associated metadata.
Death Register	National system which records deaths within the jurisdiction.
Dissemenation Information Package (DIP)	An Information Package, derived from one or more AIPs and sent by Archives to the Consumer in response to a request to the OAIS.
Document	A single or group of related Data Files with common metadata. For example, a Document may consist of a PDF file together with associated attachments or a word file with a separate image signature sheet. A document can be considered to be an entity that is approved/signed as a whole by a practitioner.
General EMR System	Electronic Medical Record system intended for documentation of all forms of healthcare. Note: large scale healthcare providers may have a main general-purpose EMR system but can also have a number of distributed general-purpose EMR systems serving parts of the organisation that operate as separate sub-services.
Healthcare Provider	An organisation providing primary or secondary healthcare. Can be general in scope or specialised, public or private.
Information Package	A logical container composed of optional Content Information and optional associated Preservation Description Information used to delimit and identify the Content Information and Package Description information used to facilitate searches for the Content Information.
Internal Archival Long Term Preservation guidelines	This type of guideline can have different names depending on the creator. Generally, archives specify technical guidelines and/or regulations for formats, specifying what they will accept

_____

_____

	and maintain for the long term/ Depending on the archive and available technical resources, the criteria for the selected formats can differ from archive to archive.
Level	The level of requirements of the element following RFC 2119 http://www.ietf.org/rfc/rfc2119.txt **MUST –** this means that the definition is an absolute requirement **SHOULD** – this means that in particular circumstances, valid reasons may exist to ignore the requirement, but the full implications must be understood and carefully weighed before choosing a different course. http://www.ietf.org/rfc/rfc2119.txt **MUST NOT** – this means that the prohibition described in the requirement is an absolute prohibition of the use of the element. **SHOULD NOT** – this means that in particular circumstances, violating the prohibition described in the requirement is acceptable or even useful, but the full implications should be understood and the case carefully weighed before doing so. The requirement text should clarify such circumstances. **MAY** – means that a requirement is entirely optional.
Local Patient Health Archive	An archive of physical or electronic Patient Medical Records within a Healthcare Provider or group of Healthcare Providers. A Patient Medical Record will normally be expected to be transferred to an archive either when the Patient is known to have died, or after a number of years have passed since its creation that exceeds normal life expectancy.
Open Archival Information System (OAIS)	An Archive consisting of an organisation, which may be part of a larger organisation, of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community. It meets a set of responsibilities that allows an OAIS Archive to be distinguished from other uses of the term 'Archive'.
Patient	A person who has received medical treatment
Patient Clinical Information	Structured Patient clinical data related to Cases such as diagnoses, procedures, medication, allergies, etc.
Patient Manifest	Structured manifest containing at minimum the full names of the each Patient who has records in the package together with a unique ID (such as a social security or health number).
Patient Medical Record	Collection or compilation of recorded information about a Patient in connection with healthcare.

_____

_____

	Note: a Patient Medical Record may contain information in digital form and/or information recorded on other types of media such as paper or film. For the purposes of this specification, Patient Medical Records are assumed to be digital where the content may be born digital and/or digitised from physical records.
Patient Medical Record Extraction	Extract from a Local Health Archive for the purposes of handing off to the Central Health Archive. All Patient Medical Record Extractions should be under a Submission Agreement.
Patient Administrative Information	Demographics and other administrative information about an individual receiving care or other health-related services. For example, as can be described using the resource FHIR.Patient. Information will include but not be limited to name, Patient ID(s), administrative gender, date of birth, date of death, address(es).
RDBMS	Relational Database Management System
Representation	A Representation within an Infiormation Package contains archival data. If an Information Package contains the same data in two or more different formats (i.e. an original and a long term preservation format) or in different types of organisations (arrangements), the are placed within two or more separate Representations within the Representations folder of the Information Package of the Information Package.
Representation Information	The Representatiuon Information must enable or allow the re-creation of the significant properties of the original data object.
Specialised EMR System	Electronic Medical Record system specially adapted for documentation of a type of specialised healthcare or integrated with a specialised device. Examples: food/maternity system, gastrosystem, laboratory system, etc.
Standardised Machine-readable Documentation	A standardised machine-readable document is a document whose content can be readily processed by computers and is based on a commonly accepted standard. Such documents are distinguished from machine-readable data by virtue of having sufficient structure to provide the necessary context to support the business processes for which they are created.
Sub-case	Type of component consisting of a set of thematically related Data Files which are also related to a Case. Sub-cases are represented in the specification as folders that sit within a Case.
Submission Agreement	The agreement reached between an archive and the submission producer that specifies a submission format (eHealth1 CITS), and any other arrangements needed, for the data submission session. Any special conditions on Patient confidentiality could be specified in the submission agreement.

_____

_____

Submission Information Package (SIP)	An Information Package that is delivered by the Producer to the OAIS for use in the construction or update of one or more AIPs and/or the associated Descriptive Information.
Submitting Organisation	Name of the organisation submitting the package to the archive.

Appendix 1 – Extended Vocabulary

Vocabulary extensions are provided in the Vocabulary EHealth1 for the content information type and the different divisions present in the structural map.

<div align="center">**Table 3: Extended Vocabulary**</div>

Value	Vocabulary and Context
citsehpj_v2_0	VocabularyeHealth1 Value for content information type: `@csip:CONTENTINFORMATIONTYPE`
Patient Medical Records	VocabularyeHealth1 Value for other content category: `@csip:OTHERTYPE`
eHealth1	VocabularyeHealth1 Value for representation structural description label: `div/@LABEL`
Data	VocabularyeHealth1 Value for data division structural division label: `div/div/@LABEL`
Patient Record	VocabularyeHealth1 Value for Patient record structural division label: `div/div/div/@LABEL`
Case	VocabuilaryeHealth1 Value for structural map case division label: `div/div/div/div/@LABEL`
Subcase	VocabularyeHealth1 Value for structural map subcase division label: `div/div/div/div/div/@LABEL`

Document	VocabularyeHealth1
	
	Value for structural map document division label:
	
	`div/div/div/div/div/@LABEL, or`
	`div/div/div/div/div/div/@LABEL`

_____

Appendix 2 - Postface

AUTHOR(S)	
Name(s)	Organisation(s)
Stephen Mackey	Penwern Limited

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

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
v1.0.0	31st August 2021	Stephen Mackey	piqlAS	Published
V2.0.0	17th May, 2024	Stephen Mackey	Penwern Limited	First major revision to amend structure of aggregations for compliance with CSIP and correct identified errors

> **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.