MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220025

How to Optimize Connection Between PACS and Clinical Data Warehouse: A Web Service Approach Based on Full Metadata Integration

Pierre Lemordant[a,b,1], Guillaume Bouzille[a], Romain Mathieu[ce], Ronan Thenault[ce], Bernard Gibaud[a], Cyril Garde[b], Boris Campillo-Gimenez[f], Didier Goudet[d], Sébastien Delarche[c], Yann Roland[f], Marc Cuggia[a]

[a] Univ Rennes, CHU Rennes, Inserm, LTSI - UMR 1099, F-35000 Rennes, France; [b] Enovacom, Marseille, France
[c] CHU Pontchaillou, F-35000 Rennes, France; [d] CLCC Eugène Marquis, F-35000 Rennes, France
[e] Univ Rennes, Inserm, EHESP, Irset – UMR_S 1085, F-35000 Rennes, France
[f] Univ Rennes, CLCC Eugène Marquis, Inserm, LTSI - UMR 1099, F-35000 Rennes, France

Abstract

Clinical image data analysis is an active area of research. Integrating such data in a Clinical Data Warehouse (CDW) implies to unlock the PACS and RIS and to address interoperability and semantics issues. Based on specific functional and technical requirements, our goal was to propose a web service (I4DW) that allows users to query and access pixel data from a CDW by fully integrating and indexing imaging metadata. Here, we present the technical implementation of this workflow as well as the evaluation we carried out using a prostate cancer cohort use case. The query mechanism relies on a Dicom metadata hierarchy dynamically generated during the ETL Process. We evaluated the Dicom data transfer performance of I4DW, and found mean retrieval times of 5.94 seconds and 0.9 seconds to retrieve a complete DICOM series from the PACS and all metadata of a series. We could retrieve all patients and imaging tests of the prostate cancer cohort with a precision of 0.95 and a recall of 1. By leveraging the CMOVE method, our approach based on the Dicom protocol is scalable and domain-neutral. Future improvement will focus on performance optimization and de identification.

Keywords:

Medical Imaging, Clinical Data Warehouse, Health Information System Interoperability

Introduction

Clinical images data analysis, especially with artificial intelligence methods, is an active area of research that holds promise for disease characterization, precision medicine, and early assessment of treatment response. However, a key challenge needs to be overcome: unlocking hospital imaging software components (i.e. Picture Archiving and Communication Systems, PACS and Radiology Information Systems, RIS) to integrate imaging data with the other patient clinical data into Clinical Data Warehouses (CDW) or data lakes for research reuse [1,2]. Researchers have developed several solutions for reusing imaging data, such as Research PACS [3,4] that manages the whole process for imaging-oriented clinical trials. Other solutions exist to carry out imaging studies, but they are fully imaging oriented [5,6], or extract data from the PACS for a given predefined cohort and allow crossing imaging data with clinical data a posteriori [7]. Kaspar et al [8] described the implementation of a technical component to integrate imaging metadata from the clinical PACS into a CDW. They proved the feasibility of routine feeding via basic metadata queries (CFIND), or queries on the first image of a series (CMOVE) to retrieve all metadata for the identified patient subsets.

However, the metadata recovered via CFIND queries are very limited. To tackle this issue, we developed a prototype (Images for Data Warehouse, I4DW) that fully captures the semantics around the image and efficiently connects a PACS to a CDW. Here, we present the technical implementation of this workflow, and its evaluation using a prostate cancer cohort use case.

Material and Methods

Functional and technical requirements: To prioritize the R&D tasks, we interviewed a group of radiologists, clinicians and data scientists at our hospital to define the following list of functional and technical requirements:

- secondary reuse of data for the widest choice of purposes;
- combining queries on imaging data and other clinical data using the same interface;
- possibility to connect to the PACS and perform queries without jeopardizing the care processes; no duplication of the PACS data due to safety, GPRD, and IT resources reasons;
- possibility to browse and view clinical images on the CDW graphical user interface (GUI) to perform pre-screening and/or for data quality control;
- leveraging the Digital Imaging and Communication in Medicine (DICOM) and terminology standards used in Dicom metadata (e.g., SNOMED, LOINC) to

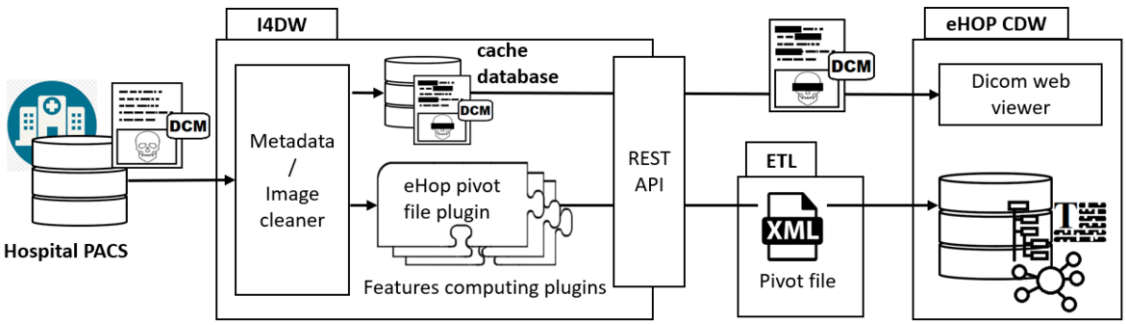

Figure 1 - Global architecture: Flow of the imaging data through the I4DW server

share data between CDWs and perform multi-center studies;

- possibility to enrich metadata during the Extraction, Transformation, Loading (ETL) process to improve image indexing and retrieval.

System design: We designed a stand-alone JAVA server component called I4DW to connect the ETL component with the PACS. Figure 1 is an overview of the system architecture. I4DW relies on the DICOM protocol. This server, using the Spring Framework[1], provides a HTTP REST API (Application Programming Interface) on the CDW side and a Dicom standard interface on the other side, using the PixelMed Java library[2], to connect to one or more PACS.

To avoid unnecessary solicitations of the IT resources, a DataBase cache is dedicated to store image collections (one or more cohorts) loaded from the PACS.

I4DW is based on a "data extraction/enrichment" plugin system that facilitates the addition of new functionalities without increasing the application burden and without any impact on the API. The ETL process uses a transitional "pivot" XML file. Each document to be integrated in the CDW goes through this format before its integration into the target data model. For more flexibility, we chose to create a dedicated data extraction plugin that produces this pivot format as a result. The main parameters of the REST API call performed by the ETL are the list of accession numbers to retrieve, the list of features needed (i.e the list of plugins to call) and a flag indicating the need to store the images in the cache.

Periodically, the ETL component calls the I4DW, passing as parameter a list of accession numbers obtained via the already integrated radiology reports data stream (coming from the RIS). This method allows retrieving the link between the targeted Dicom image data and the patient ID. The DICOM protocol allows to make queries on a limited set of fields indexed by the PACS (CFIND query) or queries requiring the retrieval of at least one image to obtain the whole header (CMOVE). As metadata extraction via CFIND is limited, we opted to routinely query imaging data with the CMOVE method.

Integration in the CDW data model: eHOP is a CDW technology [9] developed by our team and currently used by 17 academic hospitals in France. eHOP data model is similar to that of the currently most recognized solutions (I2B2, OMOP). However, it introduces the notion of "document" entity that groups a set of atomic data elements in a specific context (e.g. a laboratory test report assembles all the measurements made, a drug prescription report lists all the drug prescriptions and administrations during a stay). To integrate imaging data in this model, we considered the Dicom series as a document entity. During integration, each series from a Dicom study becomes a document. To keep track of the original study, all these documents and the radiology report document are linked by an accession number. This logical view of a Dicom study is computed at the application level and available in the dedicated view on the eHOP interface (Figure 3B).

eHOP allows querying all the patient's data (age, sex, etc.), hospital stays (dates, medical unit, etc.), and documents. Documents have a text field for text search and are linked to structured elements stored in a dedicated table. Structured elements are atomic values of different types (number, text, code, date) associated with terminologies. For instance, a document representing a surgery report can be associated with a structured element with the code 'JGFC001' (i.e prostatectomy) from the terminology of the French medical classification for clinical procedures (CCAM).

The Dicom standard describes images with attributes and can organize them hierarchically with "sequence attributes" grouping subsets of other attributes (e.g. extensively used in Dicom structured reports). When a Dicom series is integrated into eHOP as a document, structured elements linked to this document are created and rely on the "DCMEHOP" terminology, based on the Dicom attributes and their position in the hierarchy of sequences. This terminology is built dynamically in eHOP as Dicom data are progressively integrated, and thus users can search through a terminology organised like the data is organized in Dicom. The simple "Attribute Tag" top node of the terminology contains the attributes used in modalities other than Dicom SR. For each newly integrated type of SR modality, a new top node is created that contains the attributes organized according to the report template. The querybuilder allows setting constraints depending on the structured element type. Using the Value Representation (VR) of Dicom tags, each attribute type can be identified. eHOP also keeps track of all possible values for a

[1] The Spring framework https://spring.io/projects/spring-framework

[2] PixelMed™ Java Dicom Toolkit https://www.pixelmed.com/Dicomtoolkit.html

given structured element, e.g., when defining a constraint on modality the system suggests a set of possible values, “MR”, “CT”, etc. Figure 2 shows how this terminology is displayed in the query builder GUI.

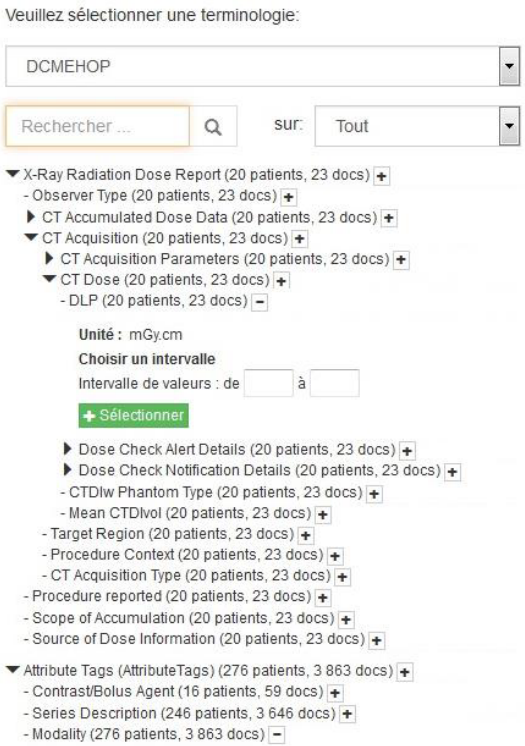

Figure 2 - Generated Dicom terminology displayed in the eHOP query builder

Moreover, some elements in the Dicom files are based on external terminologies such as LOINC and SNOMED. For instance, the attribute "Anatomic Region Sequence" could contain the attribute "coding scheme designator" with the value "SRT" (i.e SNOMED) and the attribute "code value" with the value "T-D3000" to designate the chest. As eHOP can do mapping between terminologies, when integrating these elements, the mapping must be maintained between the generated terminology and the SNOMED Clinical Terms or LOINC elements already in use in eHOP. eHOP documents created from the integration of imaging tests are also textually indexed based on a subset of Dicom attributes.

Before being cached or sent to the viewer, data passing through the I4DW are deidentified. Dicom headers are anonymized in a configurable way, by attribute, defaulting to the Dicom “Basic ApplicationLevel Confidentiality Profile” and retaining temporal and device identity information.

GUI functionalities: Dicom unstructured and structured reports (“SR” modality) are displayed as documents, like any other eHOP document type. We added a set of specific functionalities to help users to easily navigate and visualize the imaging document:

- possibility to browse other documents belonging to the same Dicom study;
- possibility to view and handle images (Dicom series) through an integrated Dicom viewer (Papaya viewer library [3]);
- possibility to open a "Dicom Study View" that presents side by side the report and a viewer with a selector to display any study series (see Figure 3B).

Evaluation: We evaluated the performance of our system for selecting a cohort and generating a dataset that was compared to an existing cohort of 271 patients with prostate cancer used as “gold standard”. The objective was to generate, from the data of 1.4 million patients available in our CDW, a datamart containing all clinical data including imaging reports and tests for this cohort. Inclusion criteria were (i) patients who underwent prostatectomy between 01/01/2014 and 31/12/2019, (ii) and patients with at least one pre-op MRI. Standard metrics of information retrieval (Precision, Recall, and F-measure) were computed to assess whether our system could retrieve all patients of the existing prostate cancer cohort and their imaging data.

We evaluated I4DW data transfer performance both in “routine” mode (i.e query a single image from the PACS for each Dicom series to obtain all metadata) and in “caching” mode (i.e all pixel data and metadata were retrieved from the PACS and images were cached in I4DW for future use).

[3] Papaya, JavaScript medical research image viewer https://github.com/rii-mango/Papaya

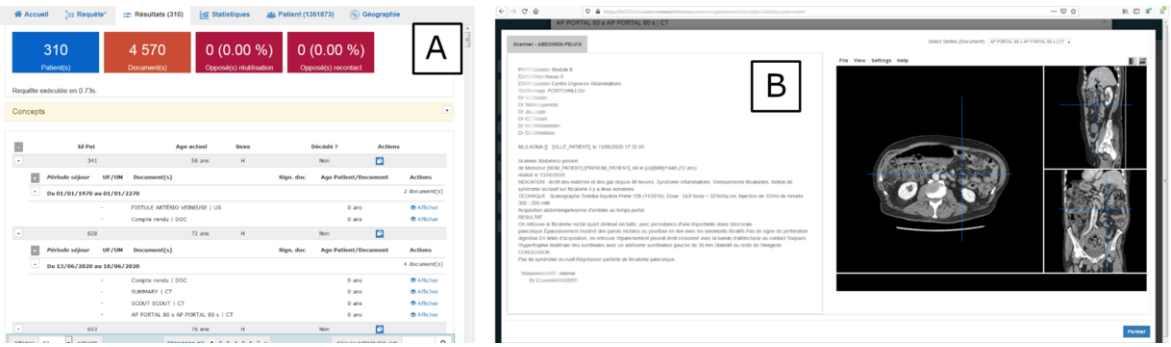

*Figure 3 - Screenshots of the eHop interface. **A** Result panel in eHOP. **B** The "Dicom Study" view.*

Results

Data integration performance: The ETL sends queries to the I4DW server through a list of three accession numbers on each call, specifying whether to store the images in the cache (a CMOVE for all images is performed) or not (a CMOVE is performed for a single image per series). This limit of three accession numbers has been chosen because the I4DW response that contains the list of eHOP pivot files must not exceed a specific size limit. In function of the implied Dicom modalities (e.g. MRI, CT, ultrasonography), each HTTP call will include different numbers of series.

Table 1 shows the mean durations (in seconds) of the image retrieval tasks for the whole series and for the first image of a series. For the first retrieval task, on a set of 1200 Dicom series, all images were retrieved and cached in the I4DW component. In routine mode, on a set of 300 Dicom series, only the first image of each DICOM series was retrieved to get all the metadata.

Table 1 – I4DW query performance

Tasks	Whole series	First image of each series
Total http request time per series (mean)	7.74 s	2.06 s
CMOVE Query time per series (mean)	5.94 s	0.9 s

Retrieval performance:

We recreated the cohort in our CDW based on the inclusion criteria used for the "gold standard" prostate cancer cohort. Table 2 presents the precision, recall and F1-measure for this retrieval task.

Table 2 - Retrieving an imaging cohort

	Patients and their MRI
Precision	0.95
Recall	1
F1-measure	0.975

All patients in the cohort were found and some patients who were not included in the cohort were found in addition in eHOP. Once this cohort is in our data warehouse, we need to classify the patients according to the magnetic field strength of the MRI. This is possible thanks to the integration and indexing of the attribute "Magnetic Field Strength" which comes from the dicom metadata.

Query builder and data visualization: Users can query the integrated imaging data with all other data available in eHOP. The structured searching mode allows querying any attribute from the Dicom metadata. Figure 2 shows the query builder with the hierarchy of data elements organized according to the generated Dicom terminology. Each node is associated with the number of patients and documents available in the CDW or in the current datamart. According to the data element type, users can query using numerical or textual criteria. Relevant Dicom attributes ("Study Description", "Series Description" and "Body Part Examined") are indexed as text during the integration phase, and this allows the eHOP text searching mode to easily retrieve imaging tests. After query completion, results are sorted and displayed by patient and hospital stay (Figure 3A).

Users can browse, open and visualize clinical images to check the image quality or to ensure that the images in the cohort match his expectations, by viewing a document as they would view any other document in eHOP. It is also possible to directly check the consistency between the report and the Dicom study series thanks to the "Dicom study" view.

Discussion

In this work, we presented the prototype we developed to integrate imaging data as a new information source for our CDW. The implemented system addresses the problem of exhaustive and semantic integration of imaging data. We tested the performance of the prototype by creating an imaging cohort and we demonstrated that this approach is feasible. The integration of fine-grained data allowed the advanced query of imaging data with all the other data gathered in eHOP and leverages coded Dicom attributes (using LOINC or SNOMED). This metadata extraction method goes beyond in terms of data indexing than the CFIND based approach implemented in the study by Kaspar et al. [8]. We plan to keep the plugin approach to implement a set of services such as classification in the RadLex ontology [10].

As a limitation, our prototype was evaluated using a specific prostate cancer cohort. However, the system was designed independently of a specific clinical domain and can be used for many use cases.

The fine-grained integration of metadata generates a very rich collection of technical and clinical data elements. However, some are not useful to clinicians (e.g., Manufacturer), and some can be confusing (e.g. elements in depth in the hierarchy of a structured report). An improvement would be to customize the DCMEHOP terminology from a technical or clinical point of view.

Currently, metadata integration only covers the Dicom public attributes. The management of Dicom private attributes (vendor-defined attributes) is an important problem, as mentioned by Langer [5] and by Doran et al [6]. For some specific cases where these attributes are necessary, we could consider developing an I4DW plugin focused on the conciliation of private attributes from different vendors. This plugin could map attributes with the same meaning on a unique new DCMEHOP code.

Like in the study by Kaspar et al [8], our PACS does not implement the WADO and QIDO protocols [11], which seems still underused by vendors [12]. Theoretically, WADO might improve the ETL process performance because it does not require images to extract metadata. Although WADO is not designed to retrieve data in a bulk mode, it would be interesting to develop and evaluate a WADO-based retrieval approach.

We now plan to technically improve the ID4W component. Specifically, we want to optimize the ETL process by managing an asynchronous task system that will enable pooling more queries to the PACS and optimize the bulk metadata retrieval. We also want to improve the pixel data privacy using a ML-OCR method as recommended in good practice guidelines [13].

Acknowledgments

We would like to thank the French National Research Agency (ANR), for funding this work in the framework of the LabCom LITIS project (grant no. ANR-17-LCV1-0004) and the Canceropole GO for funding the Oncoshare project.

References

[1] Huang, SC., Pareek, A., Seyyedi, S. *et al.* Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digit. Med.* **3,** 136 (2020). https://doi.org/10.1038/s41746-020-00341-z

[2] Murphy P, Koh DM. Imaging in clinical trials. *Cancer Imaging*. 2010;10 Spec no A(1A):S74-S82. Published 2010 Oct 4. doi:10.1102/1470-7330.2010.9027

[3] Marcus DS, Olsen TR, Ramaratnam M, Buckner RL. The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics*. 2007;5(1):11-34. doi:10.1385/ni:5:1:11

[4] E. Micard, D. Husson, and J. Felblinger, ArchiMed: A Data Management System for Clinical Research in Imaging, *Frontiers in ICT*. **3** (2016) 31. doi:10.3389/fict.2016.00031.

[5] Langer SG. Dicom Data Warehouse: Part 2. *J Digit Imaging*. 2016;29(3):309-313. doi:10.1007/s10278-015-9830-4

[6] Doran SJ, d'Arcy J, Collins DJ, et al. Informatics in radiology: development of a research PACS for analysis of functional imaging data in clinical research and clinical trials. *Radiographics*. 2012;32(7):2135-2150. doi:10.1148/rg.327115138

[7] Rajala T, Savio S, Penttinen J, et al. Development of a research dedicated archival system (TARAS) in a university hospital. *J Digit Imaging*. 2011;24(5):864-873. doi:10.1007/s10278-010-9350-1

[8] Kaspar M, Liman L, Ertl M, et al. Unlocking the PACS Dicom Domain for its Use in Clinical Research Data Warehouses. *J Digit Imaging*. 2020;33(4):1016-1025. doi:10.1007/s10278-020-00334-0

[9] Madec J, Bouzillé G, Riou C, et al. eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. *Stud Health Technol Inform*. 2019;264:1536-1537. doi:10.3233/SHTI190522

[10] Lemordant P, Gibaud B, Garde C, et al. Ontology-based classification of radiological procedures for consistent sharing in Clinical Data Warehouses. 11th International Conference on Biomedical Ontologies (ICBO). Published in CEUR-WS Proceedings

[11] DICOMweb™ : https://www.dicomstandard.org/dicomweb (last accessed: 22 April 2021)

[12] Connectathon Results Browsing : https://connectathon-results.ihe.net/advanced.php (last accessed: 22 April 2021)

[13] Elngar, Ahmed, Ambika Pawar, and Prathamesh Churi, eds. *Data Protection and Privacy in Healthcare: Research and Innovations*. CRC Press, 2021.

Address for correspondence

Corresponding Author, Pierre Lemordant, Univ Rennes 1, Inserm, LTSI UMR 1099, Rennes, France; E-mail: pierre.lemordant@univ-rennes1.fr