UCSF

UC San Francisco Previously Published Works

Title

Creation and Curation of the Society of Imaging Informatics in Medicine Hackathon Dataset

Permalink

Journal

ISSN

Authors

Kohli, Marc
Morrison, James J
Wawira, Judy
et al.

Publication Date

DOI

CrossMark

Creation and Curation of the Society of Imaging Informatics in Medicine Hackathon Dataset

Marc Kohli [1] · James J. Morrison [2] · Judy Wawira [3] · Matthew B. Morgan [4] ·
Jason Hostetter [5] · Brad Genereaux [6] · Mohannad Hussain [6] · Steve G. Langer [7]

Abstract In order to support innovation, the Society of Imaging Informatics in Medicine (SIIM) elected to create a collaborative computing experience called a "hackathon." The SIIM Hackathon has always consisted of two components, the event itself and the infrastructure and resources provided to the participants. In 2014, SIIM provided a collection of servers to participants during the annual meeting. After initial server setup, it was clear that clinical and imaging "test" data were also needed in order to create useful applications. We outline the goals, thought process, and execution behind the creation and maintenance of the clinical and imaging data used to create DICOM and FHIR Hackathon resources.

Introduction

Radiology has benefitted tremendously from broad vendor adoption and implementation of standards such as DICOM, HL7, as well as the integration profiles defined by IHE. As radiology transitions to the enterprise, our community hopes to expand on the success of DICOM and HL7 with Fast Healthcare Interoperability Resources (FHIR) [1] and DICOMweb [2]. Adoption of these new integration standards will benefit imaging informatics professionals who are commonly requested to provide integration solutions between various imaging and health information technology (HIT) platforms. Recognizing the need for new skills and experience The Society of Imaging Informatics in Medicine (SIIM), led by Don Dennison, and Chris Meenan created the hackathon.

The SIIM Hackathon Committee oversees the event, which has three parts: (1) the persistent cloud-based instantiations of reference implementations, (2) the hackathon dataset, and (3) the hackathon events during the annual meeting. This article will focus on the creation of the hackathon dataset.

During the first hackathon in 2014, it was quickly realized that building novel applications using the FHIR and DICOMweb specifications were difficult without test data. For example, the 2014 winning project [3] required real radiology report data in order to provide a meaningful demonstration. As a result, some of the coder's time was spent crafting test data (FHIR DiagnosticReport elements), which may have been better spent developing the solution. To avoid this in the future, the committee decided to streamline the process by providing not only server resources but also the test clinical data. The committee set out to explore whether existing

✉ Marc Kohli
marc.kohli@ucsf.edu

1   UCSF, Level P1 Room AC09H, 400 Parnassus Ave, Irving St, San Francisco, CA 94143, USA

2   Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, 3181 SW Sam Jackson Park Rd, Portland, OR 97239, USA

3   Department of Radiology and Imaging Sciences, Indiana University, 550 N. University Blvd. Room 0641, Indianapolis, IN 46202-5149, USA

4   Department of Radiology and Imaging Sciences, University of Utah, 30 North 1900 East #1A071, Salt Lake City, UT 84132-2140, USA

5   Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland, 22 S. Greene St, Baltimore, MD 21201, USA

6   Agfa HealthCare, 375 Hagey Blvd, Waterloo, ON N2L 6R5, Canada

7   Imaging Physics and Imaging Informatics, Mayo Clinical Radiology, Rochester, MN 55904, USA

datasets were available or whether a new resource should be created.

Goals of the Hackathon Dataset

In order to provide a realistic sandbox for development of integrated solutions, the hackathon committee looked for medical data that included three components: clinical data (allergies and current medications), imaging data sets, and associated radiology reports. Through the Internet searches and queries to medical informatics experts across the country, the committee identified a list of existing datasets and evaluated each based on the availability of the three data components. A summary of the search results is presented in Table 1.

Since none of the publically available datasets met the requirements, the hackathon committee set out to create a set of standardized data to populate the reference implementations with the following goals:

1. Create a robust set of clinically coherent data, including imaging and clinical elements
2. Make it easily updatable, upgradable
3. Allow for contributions from multiple sources

Since the creation of the hackathon dataset in 2015, the availability of public FHIR servers with a broad base of resource types has increased substantially; however, at the time of publication, we are unaware of other FHIR servers that also include matching image data.

Process

The Cancer Imaging Archive (TCIA) [4] was identified as an ideal source of multimodality imaging data. Established in 2005, the TCIA provides an online repository for image data associated with tissue samples collected as part of The Cancer Genome Atlas (TCGA) project. TCIA has quickly grown into a massive repository of multimodality image data. TCIA

Table 1 Summary of known public medical and imaging data sets

Dataset	Description	Patient records	Clinical data	Radiology reports	DICOM image data
MIMIC[a]	Critical care—demographics, vital signs, lab values	40,000	Yes	No	No
SynPUFs[b]	Synthetic medicare claims data	6,873,274 benefit summaries	No	No	No
MTSAMPLES[c]	Transcription samples	4999	Yes	Yes	No
TREC 2011[d]	Text reports over several topic areas	101,712	Yes	Yes	No
TREC 2012[e]	Text reports over several topic areas	Unknown	Yes	No	No
Public FHIR servers[f]	Mixed, constantly changing	Unknown	Yes	Yes	No
TCIA[g]	Imaging of oncology patients	34,483	Yes	No	Yes
NLM/IU Chest Radiographs[h]	Chest radiographs	4000	No	Yes	Yes
JSRT Chest Radiographs[i]	Chest radiographs	247	No	No	Yes
ADNI[j]	Brain MRI and PET scans	1150	Yes	Yes	Yes
NBIA[k]	Various	Unknown	No	Yes	Yes

[a] https://mimic.physionet.org/

[b] https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-files/SynPUFs/index.html

[c] http://www.mtsamples.com

[d] http://trec.nist.gov/data/medical/11

[e] http://trec.nist.gov/data/medical/12

[f] http://wiki.hl7.org/index.php?title=Publicly_Available_FHIR_Servers_for_testing

[g] http://www.cancerimagingarchive.net/

[h] https://openi.nlm.nih.gov/gridquery.php?q=&it=xg&coll=cxr

[i] http://www.jsrt.or.jp/jsrt-db/eng.php

[j] http://adni.loni.usc.edu/about/

[k] https://imaging.nci.nih.gov/ncia/

collections are licensed via a Creative Commons Attribution 3.0 license. Using the web-based search interface, patients were chosen from different sections within TCIA in order to include the following modalities: computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), PET/CT, mammography, and planar nuclear medicine. A summary of the selected patients is presented in Table 2.

The images were downloaded from the TCIA web page via their Java applet and stored locally. The committee decided to name each patient with a pseudonym consisting of a unique first name (Andy, Joe, Ravi, Neela, and Sally), and SIIM as the surname. The committee felt that it was important to retain linkage and attribution in our dataset to the TCIA so the PatientID dicom header element was left intact. Accession numbers were added to allow linkage between the imaging data and FHIR resources. Accession numbers were populated with random values using the maximum length specified by DICOM and entered into the ID of the corresponding FHIR report resources. The reidentified images were originally stored on Amazon S3, which has been recently replaced with a github sub-repository. Images are loaded into the hackathon DICOM server nightly.

As mentioned earlier, one of our goals was clinically coherent data, and by this, we aim to replicate the data that would be found in a typical picture archive and communication system (PACS), electronic medical record (EMR), and radiology information system. Thus, we required a report for each of our image sets. We also created a few other artifacts that patients would have in a typical clinical setting, including medication orders and allergy information. With the clinical data defined, we set out to create corresponding FHIR resources. Each FHIR resource is a single text file written in JavaScript Object Notation (JSON). The resources are arranged in a directory structure, organized by patient, and then resource type.

FHIR resources are split into six sections: clinical, identification, workflow, financial, conformance, and infrastructure. The current hackathon dataset includes examples of the following patient-level resource types: Patient, DiagnosticReport, ImagingStudy, AllergyIntolerance, Condition, MedicationOrder, and Specimen. All of the imaging studies in the dataset have corresponding DiagnosticReport and ImagingStudy resources. The DiagnosticReport resources were created by radiologist and resident volunteers who interpreted

the imaging studies. ImagingStudy resources were created through scripting by one of the committee members.

In addition, a few of the patients have additional resources outside the imaging domain. For instance, patient Sally SIIM has an allergy to IV contrast that is documented with an AllergyIntolerance resource and has a MedicationOrder for thyroid hormone replacement.

Several FHIR resources such as organization, practitioner, and medication exist at a system level rather than the patient level and are referenced by patient-level resources. For example, a medication resource was created to fulfill the MedicationOrder for Sally SIIM. An organization resource using location information for SIIM headquarters was created, and a practitioner resource was created as the author of some of the DiagnosticReport resources.

Dataset Tools

Several tools were created in order to assist in working with the dataset. First, the Readme.md document in the github repository lists helpful information about how to obtain the image data from github, conventions that are used throughout the dataset, and instructions to use the provided ruby scripts to upload to a FHIR server. This is intended to make the dataset useful for FHIR servers outside the hackathon. The Readme.md also lists a human-readable summary of the clinical information available for each patient.

There are two helper ruby scripts included in the repository. The first (upload.rb) allows command like uploading of the resources into a FHIR server. The second script generates FHIR resources to support Integrating the Healthcare Enterprise (IHE) Mobile access to Health Documents (MHD). MHD provides a RESTful discovery and retrieval service in addition to a RESTful facade to Cross-Enterprise Document Sharing (XDS). MHD requires the creation of DocumentReference resources for each of the DiagnosticReport and ImagingStudy objects. As these are derived, rather than being primary sources, a ruby script was created to dynamically generate DocumentReference objects. MHD further requires that DocumentReference objects be grouped together into DocumentManifest resources, so the same script also creates DocumentManifest resources.

Table 2 Table listing the patients contained within the SIIM Hackathon dataset

SIIM name	TCIA name	TCIA collection	Primary diagnosis
Andy	TCGA-50-5072	TCGA-LUAD	Lung adenocarcinoma
Joe	TCGA-17-Z058	TCGA-LIHC	Hepatocelluar carcinoma
Neela	TCGA-BA-4077	TCGA-HNSC	Head and neck squamous cell carcinoma
Ravi	LIDC-IDRI-0132	LIDC-IDRI	Lung cancer
Sally	Breastdx-01-0003	Breast diagnosis	Ductal carcinoma in situ

Challenges

One challenge the committee has faced is the mainte-
nance of the dataset to match release upgrades in
FHIR. When the project was originally started, FHIR
was a draft specification at version 0.1 DSTU (Draft
Standard for Trial Use). Currently, FHIR is at version
3.0.1-STU (Standard for Trial Use) [1]. There have been
many updates and changes to the resource specification
along the way, and other changes are anticipated. In or-
der to assist with keeping the dataset current, the github
repository includes a set of tests for each resource type.
The tests are built using a ruby testing framework
(Rspec) and include creation, deletion, and retrieval of
each resource from a FHIR server. The retrieval re-
sponses are compared with the original file system re-
source ensuring appropriate function. The DICOM data
has remained constant and has not required updates.

Collaboration

As one of the core values of the hackathon dataset was the
ability to upgrade and update as well as the ability to accept
updates from multiple users, the committee decided to host the
dataset on github. If one examines the github statistics, there
have been 180 updates to the github repository by five con-
tributors. However, this is an under-representation of the col-
laborative effort, because several people have augmented the
dataset through e-mailing changes to github contributors.
Actual collaboration included eight individuals from nine dis-
tinct organizations.

The repository has been forked by 13 github users.
Additionally, our repository has eight stars, which is a
fair number for such a niche project. Github has certainly
facilitated discussion and work among committee mem-
bers, but has not yet fulfilled the vision of expansion of
the dataset.

Conclusions

The hackathon dataset has reached the goals outlined above:

1. A clinically coherent dataset including a wide range of
   imaging modalities and supporting FHIR resources

   a. Five patients
   b. 12 FHIR resource types
2. Infrastructure that allows for updating as the FHIR stan-
   dard matures

   a. Rspec testing
   b. Ruby FHIR upload scripts
3. Mechanisms to allow collaborative editing from multiple
   sources

   a. Public github repository
   b. Eight individual dataset contributors, representing
      nine distinct organizations

In order to remain relevant, the dataset must continue to
evolve and increase in FHIR resource coverage, adding more
patients with longer and more complex imaging histories, as
well as adding more coded values to the existing resources
(LOINC-RadLex procedure codes, SNOMED-CT codes for
diagnosis).

References

1. FHIR v1.0.2. https://www.hl7.org/fhir/. Accessed 12 Jul 2016
2. DICOMweb: http://dicomweb.hcintegrations.ca/#/home. Accessed 12 Jul 2016
3. Morrison JJ, Hostetter JM, Aggarwal A, Filice RW: Constructing a Computer-Aided Differential Diagnosis Engine from Open-Source APIs. J Digit Imaging:1–4, 2016. doi: 10.1007/s10278-016-9874-0
4. Clark K, Vendt B, Smith K et al.: The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging 26:1045–1057, 2013. doi: 10.1007/s10278-013-9622-7