J Digit Imaging (2013) 26:1045–1057
DOI 10.1007/s10278-013-9622-7

The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository

Kenneth Clark · Bruce Vendt · Kirk Smith · John Freymann · Justin Kirby · Paul Koppel · Stephen Moore · Stanley Phillips · David Maffitt · Michael Pringle · Lawrence Tarbox · Fred Prior

Published online: 25 July 2013

Abstract The National Institutes of Health have placed significant emphasis on sharing of research data to support secondary research. Investigators have been encouraged to publish their clinical and imaging data as part of fulfilling their grant obligations. Realizing it was not sufficient to merely ask investigators to publish their collection of imaging and clinical data, the National Cancer Institute (NCI) created the open source National Biomedical Image Archive software package as a mechanism for centralized hosting of cancer related imaging. NCI has contracted with Washington University in Saint Louis to create The Cancer Imaging Archive (TCIA)—an open-source, open-access information resource to support research, development, and educational initiatives utilizing advanced medical imaging of cancer. In its first year of operation, TCIA accumulated 23 collections (3.3 million images). Operating and maintaining a high-availability image archive is a complex challenge involving varied archive-specific resources and driven by the needs of both image submitters and image consumers. Quality archives of any type (traditional library, PubMed, refereed journals) require management and customer service. This paper describes the management tasks and user support model for TCIA.

Keywords TCIA · NBIA · Cancer imaging · Image archive · Biomedical image analysis · Cancer detection

K. Clark (✉) · B. Vendt · K. Smith · P. Koppel · S. Moore · S. Phillips · D. Maffitt · M. Pringle · L. Tarbox · F. Prior
Mallinckrodt Institute of Radiology, Washington University School of Medicine, ERL 510 South Kingshighway Boulevard, St. Louis, MO 63110, USA
e-mail: clarkk@mir.wustl.edu

J. Freymann · J. Kirby
Clinical Research Directorate/CMRP, SAIC–Frederick, Inc., Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA

Background

The Human Genome Project pioneered the creation of large sharable databases [1, 2] in a successful effort to accelerate our understanding of the genetic material of which we are made and usher in the era of big data in biomedical research [3]. More recently, the Human Connectome Project (HCP) [4, 5] is accumulating vast amounts of image data in order to accelerate our understanding of brain function. This and predecessor programs such as the Bioinformatics Research Network (BIRN) [6, 7] have established medical imaging firmly in the realm of big-data-based science.

The Cancer Genome Atlas (TCGA) researchers are collecting tissue samples (brain, breast, gastrointestinal, head and neck, hematologic, skin, thoracic, urologic) and are mapping the genetic changes in 20 cancers [8]. The TCGA Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA. Current National Institutes of Health (NIH) research funding favors both collaborative efforts and sharable data in hopes of decreasing the time to achieve new levels of understanding and therapies. This, in turn, has stoked demands for collaborative initiatives to produce large and sharable data repositories, along with tools and resources to manage and analyze these data.

In 2005, and driven by projects that required standardized imaging data sets to support cancer research, the National Cancer Institute (NCI) initiated the development of a software environment that would support research-centric archiving of cancer imaging data. The software was to be open-source, vendor neutral, and to support the submission, curation, and public distribution of cancer image data. Through collaboration with the RSNA Medical Imaging Resource Community (MIRC), later the Clinical Trial Processor or CTP [9] effort, NCI-managed software teams developed the open source National Cancer Image Archive application (NCIA), around

which was built the public NCI-hosted National Cancer Imaging Archive [10] web site. Since the application evolved beyond cancer images, its name was changed to the National Biomedical Imaging Archive (NBIA) [11], and the hosting web site name was likewise changed for consistency (throughout this paper, the acronym "NBIA" refers to the software application). NBIA software, available through the NCI Center for Bioinformatics [12], has also been adopted for non-cancer repositories, most notably the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) Osteoarthritis Initiative (OAI) [13, 14].

As NBIA matured, it became evident that robust software combined with a basic help-desk alone was not sufficient to provide the support demanded by the varied needs of the cancer community. In 2010, NCI released a Request for Proposal for the development and management of a cancer imaging archive service that would provide the cancer research community the critical image-data-sharing resource that it required. In December, 2010, Washington University's (Saint Louis, Missouri) Electronic Radiology Laboratory (hereafter "WUSTL") was awarded the sub-contract to build and manage a full-featured cancer imaging archive service that would support NCI-funded research activities and the cancer research community at large.

WUSTL has a significant history of managing research imaging repositories and creating open-source software to support image transport and de-identification, most notably an acquisition-node software application for clinical trials [15] and regulatory compliance requirements for open-source image-trial management [16]. WUSTL has served as an imaging core in multi-center clinical trials, e.g., the Silent Cerebral Infarct Transfusion Trial (SITT) (~1,000 patients; 1,552 examinations; 850,000 images) [17] and the CT Image Library for the Lung Screening Study of the National Lung Screening Trial (NLST) (17,309 patients with serial CT screens; 48,723 CT examinations; 12 million images) [18].

With news of the cancer imaging archive award, a team of experienced experts (in network management, software design and implementation, systems management, operations management, DICOM standard, systems security, and image quality-control) quickly assembled, designed network and systems configurations, ordered equipment, and began mapping standard operating procedures (SOPs). In May 2011, the new project, The Cancer Imaging Archive (TCIA) launched. Described here are the management tasks and user support model for TCIA. Though briefly described, the hardware/network/software architecture is not the subject of this paper.

Methods

Washington University School of Medicine IRB Protocol 201108194: Image Archive Hosting allows Washington University in Saint Louis (WUSTL) to receive image data from submitting sites that may contain Protected Health Information (PHI) in DICOM private tags or a small set of identified text fields. All images must be submitted following a standard de-identification pass compliant with the DICOM Standard. This is accomplished using CTP and a script provided by the WUSTL TCIA staff. All data transfer employs encryption in transit. Data are received on an isolated quarantine system (Intake) where they are analyzed for residual PHI, which is then removed using a second CTP de-identification script. Prior to transfer to the public TCIA, all images are reviewed.

System Architecture TCIA hardware is housed in two independent data centers on the WUSTL School of Medicine campus. Operational software is based on multiple instances of NBIA deployed on XEN virtual machines (VMs) configured in a high-availability VM cluster fed by a load-balancing network switching (Coyote Point Systems, San Jose, CA, USA) infrastructure. The primary hardware cluster consists of two identical Dell (Round Rock, TX, USA) 510 servers ("intake" and "public") operating as VM hosts. NBIA, with its associated web presence, is deployed in a redundant set of VMs, providing high-level performance and fault tolerance. Using MySQL clustering, all VM-instance NBIA databases on each server contain the same information. The physical machines are located in separate buildings. A BlueArc shared storage system (BlueArc Corporation, San Jose, CA, USA) is used as a high-reliability principal data storage device for the VM cluster. Its redundant storage and snap-shot features assure that data are not lost. Mirrored direct attached disk arrays on the physical servers act as high-speed caches to maintain peak performance. Each server's operating system is CentOS 5.8. Additional details of the architecture are available elsewhere [19].

Operations Overview

TCIA operations include secure transmission of DICOM images and metadata (hereafter "images") from image submitters to the TCIA intake server, the curation of these images to remove any personally identifiable information, and the arrangement of images into cancer-specific and/or research-group "collections" that may be downloaded by anyone with access credentials. TCIA requires credentials to protect against spam and to track usage. Individuals are not tracked, but TCIA follows usage: number of users and countries, what kinds and numbers of images/collections they download, etc. The overall process is depicted in Fig. 1.

Operational Details

New Collection Preparation NCI must approve all submission proposals. NCI Cancer Imaging Program (CIP) staff

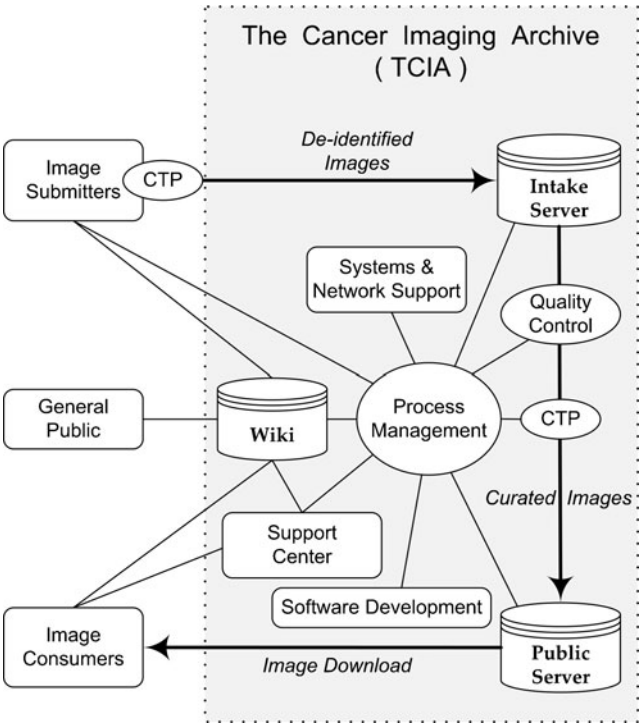


Fig. 1 TCIA operations overview. TCIA project managers negotiate new collection details with each image submitter, then supply submitter with a de-identification/re-identification script and Clinical Trial Processor (CTP) software that image submitter uses to transmit de-identified images to the TCIA intake server. The TCIA management team reviews images to make sure that counts match those of the submitter, no images have been quarantined, and re-identified IDs are as they should be. Management then directs quality control (QC) processing involving visual inspection and a thorough analysis of DICOM headers. Armed with QC results and a new CTP script to cleanse images of unwanted DICOM tag values, management moves the images to the TCIA public server from which properly authenticated image consumers may download images. A Support Center assists image consumers with gaining such access. A TCIA wiki hosts collection-specific details, a FAQ for typical TCIA questions and answers, and user guide for submitting images. TCIA managers use a private portion of the wiki to shepherd collection accrual and documentation thereof. Systems and network personnel facilitate a streamlined operation; software developers improve operations mechanisms and reporting

may be approached by a prospective image submitter wanting to contribute images to TCIA, or CIP may reach out to a group known to be active in cancer imaging and invite them to contribute images to TCIA. Either way, once the image submitter and CIP have agreed to proceed, TCIA staff gather preliminary information and post it to the TCIA wiki (Table 1).

Information gathering complete, TCIA staff begin a detailed process ending with images stored on the TCIA public server. The process is guided by the TCIA New Submissions SOP. Accomplished steps throughout the process are noted on the TCIA wiki, documenting progress for TCIA management; Table 2 lists specific steps, and details follow.

New Collection Submission The sequence of events surrounding the image submission process are as follows:

- The TCIA staff set permissions for the submitter to transmit images to the intake server.
- The TCIA staff prepare scripts based on information gleaned from the TCIA wiki and discussions with submitter. Such information includes imaging modalities, body part(s) imaged, details of a prior de-identification, knowledge of PHI stored in screen-save or overlay objects, and DICOM private tag retention requirements.
- Scripts are delivered to and installed by submitter.
- The submission process is tested by executing the scripts, using a test image.
- The collection is submitted over the internet.
- TCIA staff and the submitter communicate regularly to verify images have been transmitted and received.

Setting Permissions First, the submitter registers for a TCIA account. The submitter creates a logon name and password, then adds contact information. TCIA staff then create the necessary permissions that will allow the submitter to transmit images to the intake server as well as to view and download images from the public server. Permissions are also set for curators, support help-desk, and system administrators and managers. If the collection to be submitted is limited-access, then these permissions allow only the submitter and persons designated by the submitter to view and download these images once the images are moved to the public server. Permissions are granted via the public server's Common Security Module User Provisioning Tool (UPT) [20].

Script and File Preparation The script and files required by Clinical Trial Processor (CTP) [9] include a config.xml file, an ID-mapper file template that, when filled, translates submitter patient IDs to TCIA IDs, a burned-in pixel filter, and a de-identification script.

- The config.xml file is a series of CTP pipelines that tell the CTP executable how images will get into the pipeline (from PACS or file-folder storage); the pointers to the ID mapper file, the burned-in-pixel filter, and the de-identification script; the uniform resource locator (URL) destination to which the images will be sent; and the order in which pipeline stages are executed. The file also specifies quarantine-storage areas for problematic images.
- The ID-mapper file is a template showing how the submitter would map the submitter IDs to TCIA IDs. Because the submitter's IDs may constitute PHI, the submitter must complete the table. As CTP sees each new image, it simply replaces the submitter's patient ID with the TCIA ID in the header of each file before transmitting the image to TCIA.

Table 1 Preliminary information gathered from image submitter

Question	Information sought
Data Owner/Primary Investigator:	(name, email, phone)
Point of Contact via whom we will get the images:	(name, email, phone)
Are there any usage restrictions on this data?	(e.g., cannot be used for commercial purposes, patents, copyrights, etc.)
Expected availability or relevant deadlines:	Any gaps in availability to work with us?Any deadlines related to this submission?
Is there a batch schedule for the submission? If so, what is it?	
Data transfer mechanism (how you will get the data to CTP for Internet transmission to us):	PACS or stored files on CTP PC
Verify no prior de-identification tools used:	(check to ensure no tools used, and list any tools they say they have to use to get images to us)
Collection access:	Public or Limited?
Modalities:	e.g., MRI, CT, etc. (may be multiple)
Re-identified Patient ID Format:	e.g., TCGA-14-xxxx
Body parts examined:	e.g., brain, lung, etc.
Number of patients:	(number of patients in collection)
Studies per patient:	(number of time points per patient)
Approximate date range of image studies:	(general idea of when images were collected so it is clear during QC that dates were modified)
Series per study:	(number of scans or reconstructions per time point)
Is any potential PHI stored in the image pixels—i.e., "SCREEN SAVE" or "OVERLAY" objects?	(yes/no—if yes, please provide as much info as possible about the usage of this in your images)
Private tag requirements:	(Are there any you need to keep for scientific analysis? If so, please provide any info you have about which ones to keep)
Is there any accompanying metadata (xml, pdfs, xls, etc.)?	(yes/no—if yes, what kind?)
Would you like a wiki page for your collection? See: https://wiki.cancerimagingarchive.net/x/mgAe	(useful for telling people the scientific value of your data, if yes, please provide summary for us to use on the page)
Who should provide attribution once the data are posted?	(assumes a wiki page exists for the collection)
What are the details of DICOM Series Descriptions?	(useful for image consumers deciding which images to download)

- The burned-in pixel filter checks specified DICOM tags for specific values suggesting the possibility of PHI on an image that, if found, would prevent the image from being transmitted to TCIA.
- De-identification. All images must be submitted following a standard de-identification process specified by DICOM PS 3.15, Appendix E: Attribute Confidentiality Profiles [21]. This is accomplished using CTP and a script provided by TCIA staff. The de-identification script, at minimum, directs CTP to remove or blank certain standard tags that are known to contain, or possibly contain, PHI. In addition, the script uniformly offsets each DICOM-header date by a day-interval assigned by TCIA (that varies by collection and site), assigns the value to be written into the DICOM tag Body Part Examined (0018,0015), makes a hash of each submitter-side unique identifier (UID) that includes a TCIA-side embedded root ID to help further avoid the possibility of collisions between institutions, and assigns DICOM header provenance information to be authenticated against TCIA's intake server. What the script does not do is to assess the private tags (specific to the manufacturer of the scanner) for PHI because the meaning of the tags varies from vendor to vendor and even among scanner models from the same vendor, making the design of a single de-identification script virtually impossible.

Script Delivery and Installation Once these files and script are prepared, they are bundled with the current CTP package obtained from the RNSA–CTP-sponsored web site, along with TCIA contact information and a pointer to an Image Submitter Site User's Guide [22], and deposited in the Washington University DropBox. The DropBox notifies the submitter that the CTP bundle is available and provides directions for retrieving. The submitter retrieves the bundle and installs CTP, following specific steps in the Image Submitter Site User's Guide, which includes directions for monitoring the submission process.

Submitting Test Image Submitter installs CTP with associated files and attempts to transmit a test image. If successful,

Table 2 Image submission task list for TCIA management

Task	Date	Note	Initials
Contacted Image Submitter for information gathering			
Received required information from Image Submitter			
Submitter's intake NBIA account and collection//site UPT elements created			
CTP scripts/config files created			
CTP and config files sent to Image Submitter (in comments, type CTP Date from CTP Launcher's Version Tab)			
Image transmission test complete			
Image Submitter begins collection transmission			
Collection transmission complete			
Begin collection QC on Intake server			
Run DICOM tag analyzer TagSniffer on Intake server			
Review Element Inventory and Values reports			
Curator visually inspects images, changes series from "Not Reviewed" to "Visible"			
Collection QC on Intake completed			
Image Submitter asked to review and approve DICOM changes			
Collection signoff by Image Submitter (provide name)			
Prepare final script for upload to Public server			
Use CTP to upload to Public server			
Run DICOM tag analyzer TagSniffer on Public server			
Review Element Inventory and Values reports			
Curator visually inspects images, changes series from "Not Reviewed" to "Not Visible"			
Project manager visually inspects images, changes series from "Not Visible" to "Visible"			
Download Manager tested			
Collection QC on Public completed			
Collection completion status updated on Wiki			

TCIA staff inspect image quality, check for pixel-embedded PHI, and verify that the de-identification script has properly updated the image header. If unsuccessful, TCIA staff assist the submitter to troubleshoot the problem until the issue is resolved and a test image successfully transmitted.

Full Collection Submission TCIA staff direct the submitter to commence transmission of the available collection. The Image Submitter's User Guide details instructions for monitoring the submitter-side transmission (CTP status, quarantines, and logging). TCIA staff monitor the receiving-side intake server (CTP status, quarantines, and logging) and report back to the submitter if images are arriving but being quarantined. All data are transmitted encrypted (https). The submitter can determine when the transmission is complete by checking the status page accessible as a link from the CTP client. The submitter may verify that the images have arrived by invoking CTP's Database Verifier that checks the TCIA intake server's database and reports back to the submitter's CTP client the number of images successfully received. If the number of images sent and received do not match, TCIA staff work with the submitter until issues are resolved and all images have been successfully received.

Problematic Images On occasion, images arrive but are quarantined (set aside from accepted images) for a variety of reasons. Chief among these are images that were previously transmitted with a different Patient ID. A less frequent issue is an improper value representation in a DICOM tag (too many characters, string rather than number, negative number when nonnegative expected, etc.). A duplicate image (identical DICOM Service Object Pair Instance UID and DICOM Patient ID) simply replaces the prior image. NBIA logs and dialog with the image submitter are typically required to resolve issues, most often resulting in the image submitter transmitting corrected images.

Intake Server Quality Control Arriving images and their file headers undergo extensive quality-control checks, both semi-automated and visual. Figure 2 provides an overview of the quality control process; details follow.

Image-Header Inspection A TCIA-developed program TagSniffer [23] searches through image DICOM headers and reports all unique string values and their tag numbers from among the standard tags. These are then visually inspected for PHI. TagSniffer also reports all dates found among the standard tags; these are visually inspected to make sure they have all been uniformly decremented. TagSniffer also reports each scanner manufacturer, scanner model, and all private tags and values found for each model. Next, the manufacturer's DICOM conformance statement for each model is checked to make sure private tags observed in the images are specified in the conformance statement (if not, unknown tags will be removed) and whether conformance tags specified as likely to contain Protected Health Information (PHI) do appear among the images (if so, these tags will be removed or blanked when the images are moved to the public server).

Image Visual Inspection Experienced quality-control reviewers or "curators" inspect images using the Quality Control (QC) Tool of the NBIA application. Slice data are viewed in cine mode. Curators examine each image for any pixel-burned-in PHI as well making sure each image is visible and uncorrupted. DICOM-header tags for each image are displayed next to the image, and curators will reference these tags in conjunction with TagSniffer reports to help

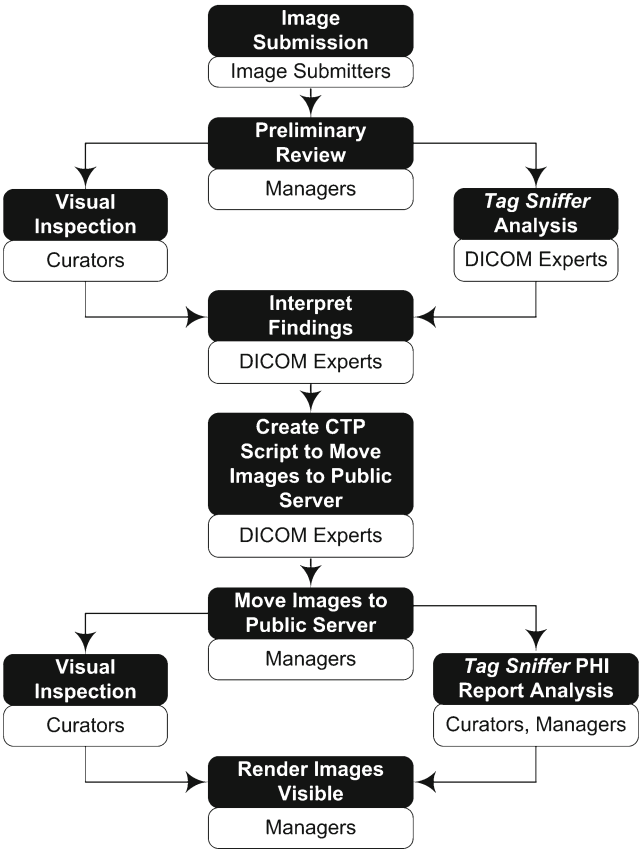


Fig. 2 Image quality control process. *Darkened boxes* are step processes; *light boxes* specify roles required to complete processes. Images submitted to the TCIA intake server are reviewed by managers to make sure DICOM patient IDs have been properly assigned, no images have been quarantined, and image counts match submitter's counts. Managers notify curators to begin visual inspection tasks and request TCIA DICOM experts to perform a TagSniffer analysis. A DICOM expert uses the analysis findings are used to create a CTP script that a manager will use to move the images to the TCIA public server. Curators perform another visual inspection of the public images and both curators and managers review a new TagSniffer report designed to expose any lingering PHI. Public-server images are then rendered "Visible"

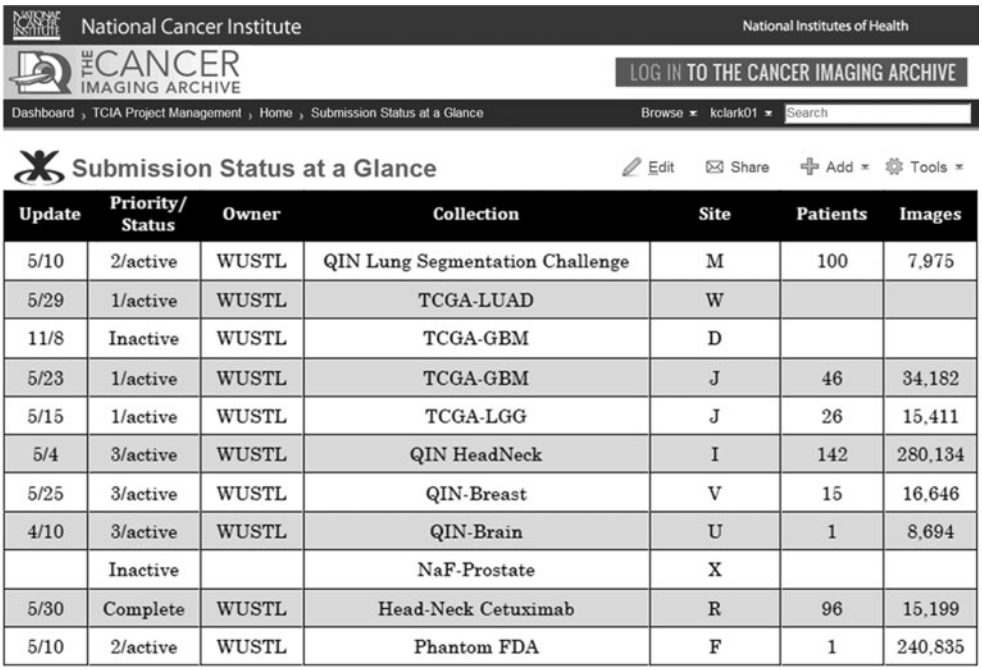
identify any PHI in standard or private tags. Suspect PHI is reported to the image submitter, and the image submitter must respond whether or not it is indeed PHI. Any confirmed PHI is expunged when the collection is moved to the public server.

Moving Images to Public Server Once a collection has been curated and all issues documented, the submitter is offered a chance to inspect images and provide approval before images are moved to the public server. The move is not unlike the image submitter transmitting images to the TCIA intake server. The chief difference is that the de-identification script includes only adjustments based on information learned from TagSniffer analysis and curation-visual-inspection. The modification of DICOM tags takes place when the images are moved from the intake server to the public server as specified in the CTP de-identification script effecting that move. Once the images have been moved to the public server, another TagSniffer report is run and reviewed to make sure the de-identification script functioned as intended. The curators also spot-check each image series to be assured that the images arrived and are viewable. Queries are made against both intake-server and public-server databases to verify matching counts for patients, studies, series, and images. The counts are recorded on the TCIA wiki page Submission Status at a Glance (examples in Fig. 3), and the images are rendered visible. Those TCIA account holders (users) with permissions to view and download images may do so. Users may access images by logging into the public server's NBIA application. Here, they may search for different kinds of images (by collection, modality, scanner vendor, etc.). Selected images are saved to a Data Basket. After filling the Data Basket with desired images, the user invokes NBIA's Download Manager that transmits the Data Basket images to a storage location specified by the user.

Support of Operations

Systems Administration A TCIA systems administrator installed and tested the NBIA software the TCIA intake and public operational servers as well as on a half dozen virtual machines (VMs) used for testing NBIA improvements and CTP upgrades. The systems administrator also arranges the proper mount points of BlueArc storage to TCIA servers. The administrator monitors the health and heartbeat of all servers and their VMs with Nagios (Nagios Enterprises, Minneapolis, MN, USA). He and the network administrator coordinate system hardware and software upgrades during monthly scheduled maintenance windows. In addition, the system administrator assists all WUSTL–TCIA personnel with issues regarding, and upgrades to, their TCIA-related computers and software.

Network Administration A TCIA network administrator set up TCIA's VMs and the virtual local area network (VLAN) on which they reside, and he configured the TCIA servers. The network administrator manages the VLANs and operational security status for TCIA, using advanced routing features, including health checks to route traffic to the servers that can best handle the load through the Coyote Point Load Balancer. He is also responsible for any Coyote upgrades to be applied during a monthly scheduled TCIA outage for system and software updates and upgrades. The systems and network administrators are responsible for keeping the TCIA infrastructure (hardware, NBIA application, NBIA database) available for public access with a 99 % uptime rate.



Update	Priority/ Status	Owner	Collection	Site	Patients	Images
5/10	2/active	WUSTL	QIN Lung Segmentation Challenge	M	100	7,975
5/29	1/active	WUSTL	TCGA-LUAD	W		
11/8	Inactive	WUSTL	TCGA-GBM	D		
5/23	1/active	WUSTL	TCGA-GBM	J	46	34,182
5/15	1/active	WUSTL	TCGA-LGG	J	26	15,411
5/4	3/active	WUSTL	QIN HeadNeck	I	142	280,134
5/25	3/active	WUSTL	QIN-Breast	V	15	16,646
4/10	3/active	WUSTL	QIN-Brain	U	1	8,694
	Inactive		NaF-Prostate	X		
5/30	Complete	WUSTL	Head-Neck Cetuximab	R	96	15,199
5/10	2/active	WUSTL	Phantom FDA	F	1	240,835

Fig. 3 Submission status at a glance. By way of example, several collections of varying priority and status are shown. Status "active" means images are being received, being curated, or expected shortly; "inactive" implies submitter has agreed to send images but is not quite ready to do so; "complete" means all images for the collection have arrived and are available on the public server. Priorities are assigned by TCIA management and are based on needs of identified research groups

Web Page Support The TCIA webmaster is responsible for updating the TCIA home page [24], the public face to the outside world (Fig. 4). Besides a pointer to the TCIA logon page, the home page provides links in three general categories: About Us, For Researchers, and Image Submissions. The webmaster is responsible for updating these links and their content.

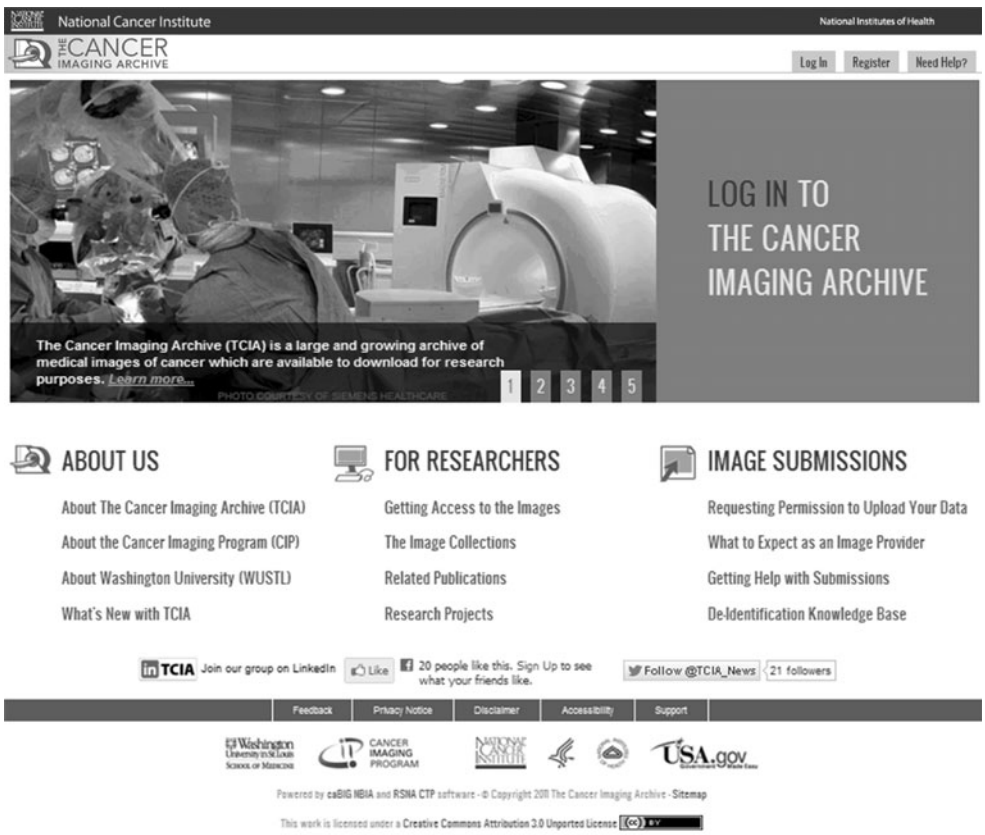


Fig. 4 TCIA home page (https://www.cancerimagingarchive.net)

Software Development Several software development efforts have aided in the access to stored images, an improved inspection of DICOM private tags, a more reliable Download Manager, and a dashboard for current TCIA status.

- Access to stored images. Images arriving to the TCIA intake server are not stored by collection but simply into a general storage area. Images arriving simultaneously from different collection sites are put into adjacent storage under control of CTP, with the location saved in the MySQL database of the NBIA application. The same occurs when images are moved from intake to public. At times, it makes sense to point all images of a specific collection and site without having to query the database for each image. Preparing a move of intake images to the public server is a good example. A WUSTL-developed program, Extraction Tool, allows one to specify collection, site, date range, intake or public server, and destination; the Extraction Tool then sweeps the database for storage locations and creates a folder of pointers in the specified destination.
- Improved analysis of DICOM private tags. When TCIA first launched, TCIA staff provided image submitters with an early version of a TagSniffer tool that would generate reports regarding the DICOM private tags of the images about to be transferred to TCIA. The thinking was that the reports would document for the submitter those images containing PHI; the submitting site would then report to TCIA which tags contained PHI (but not share tag values). TCIA staff would then provide the site with a CTP de-identification script that would delete those tags. It soon became apparent that the process took far too much time and placed an undue burden on the submitter who might, or might not, have the DICOM expertise to efficiently evaluate TagSniffer output, and TCIA shifted to a new model. Under the WUSTL–IRB arrangement now in effect, sites may securely transmit their images without deleting private tags; instead, TagSniffer is applied after the images arrive at TCIA. A TagSniffer report is generated for each scanner model and software version encountered. From each report, TCIA staff consult the appropriate manufacturer's DICOM conformance statement and construct a new de-identification script that is applied as images are moved from the intake server to the public server. The new script removes tags containing or likely to contain PHI and removes private tags not found in the conformance statement. With each new scanner-model-software-version encountered, the de-identification script is saved and details recorded in a spreadsheet; the spreadsheet is checked when new collections arrive; if the new scanner-model-software-version of the new collection can be found in the spreadsheet, the TagSniffer work for the new collection is greatly reduced. The TagSniffer software and related documents are open-source and available via the MIR-GForge [25].
- More reliable Download Manager. When TCIA launched, TCIA management had access to NBIA developers to provide understanding of the NBIA software, fixing bugs, and providing workarounds. Soon after the launch, the NBIA developers disbanded, and TCIA was left to fend for itself. Users reported issues with the Download Manager, specifically, failing to complete the download process. TCIA software developers were able to modify the NBIA code, and the problem disappeared; an NBIA–JIRA report was filed. Later, other users reported download problems when large numbers of image series were requested. Users may diminish the problem by increasing the Java-heap size on their local computers, but the problem persists, and users are cautioned to limit downloads to no more than 3,500 series at a time until the problem can be addressed in full.
- Dashboard reporting. TCIA management requires frequent updates on the status of the public server's collections. An evolving dashboard currently reports weekly download activity, weekly number of TCIA accounts, and number series, by collection, available on the public server. The download activity may be filtered by collection and/or date range. The dashboard also shows a Nagios digest of current server uptime and a Google-Analytics report of accesses to the TCIA home page.
- MIR-GForge. TCIA software developers use a GForge (GForge, LLC, West Des Moines, IA, USA) to track software revisions and store final product. TCIA operations managers also use the MIR-GForge to archive CTP configuration files and de-identification scripts related to the various TCIA collections.

DICOM Expertise The TCIA team is staffed with DICOM experts who have been involved with the DICOM standard since its early years. They have shared their knowledge with other TCIA team members and supervised the construction of an evolving publicly available TCIA-wiki-based De-Identification Knowledge Base [26] stemming from the TagSniffer work.

Security Expertise TCIA staff administrate an open-source implementation of the Lightweight Directory Access Protocol (LDAP) for TCIA logon and password authentication. With LDAP, a user registers once; the user account allows access to the NBIA application and the public portion of the TCIA wiki. All registered users have access to all of TCIA's public collections and may download images from those collections via the NBIA Download Manager.

Support Center The TCIA Support Center operates a normal-working-hours Help Desk that services users via

email and telephone. The Support Center sets UPT permissions for users creating new accounts and provides end-user support. All user issues are documented and tracked using a trouble-ticket program, Request Tracker (Best Practical Solutions LLC, Somerville, MA, USA). Trouble tickets are automatically created for email requests and manually entered for users who report problems with TCIA logons or downloading images with NBIA's Download Manager. The Support Center staff handle basic trouble tickets (Level 1) such as helping users create accounts, changing passwords, using the NBIA application, or pointing them to the right web sites. More difficult issues such as the inability to logon with seemingly correct credentials, problems downloading images, technical questions about the NBIA application, or specific questions about certain TCIA collections are routed to Level-2 TCIA operations managers. Crucial issues such as unscheduled system and network outages are routed to Level-3 systems and network administrators. Level-2 and Level-3 notifications are by email sent automatically by the Request Tracker system with phone calls made by the Support Center directly to Level-2 or Level-3 staff on critical issues.

TCIA Wiki The TCIA wiki has both a public space for dissemination of information and a private space for program management. The public space has details for every collection whose submitter expresses the desire to have a collection-specific wiki page. In the case of multi-site collections, there are links to the project in which the submitters are participating. As users enquire about certain kinds of images, the answers are compiled on a public-faced wiki page or a list of Frequently Asked Questions (FAQ). The wiki gives data submitters a platform to describe the scope and intent of their image collection and to provide metadata and/or ways for users to contact them. The wiki supports research groups by summarizing their research objectives and posting conference abstracts and publications. The public space also provides access to user guides. The private space of the wiki is used for TCIA internal management. The "Submission Status at a Glance" page (Fig. 3) gives an overall view of collections in progress and those soon-to-be started. Each collection is linked to that collection's specific details page. The private wiki also includes internal conference-call agendas and SOPS.

TCIA SOPs, User Guides, Checklists, and FAQ Because of the many details involved in processing each collection and the need to simultaneously process multiple new collections, TCIA has developed a number of SOPs to help guide the process. For example, the New Collections SOP outlines the steps required to start up a new collection all the way through images being rendered visible on the public server. Separate checklists are used for image-visual-review and moving images from intake to public. The User Guides assist image submitters and TCIA–NBIA users. A FAQ, on the TCIA wiki, attempts to answer the questions posed by image submitters and users. Each new collection can reveal new issues that contribute to the submission knowledge base. With each new experience, the SOPs, User Guides, checklists, and FAQ (Table 3) are amended and tuned to accommodate new information.

Results

In the first year of operation, May 2011–May 2012, TCIA grew to 23 collections from 31 sites (3,268,644 images, 1.3 terabytes). Imaging modalities represented are computed radiography (CR), computed tomography (CT), digital radiography (DX), magnetic resonance (MR), mammography (MG), nuclear medicine (NM), and positron emission tomography (PT). Anatomical sites include brain, breast, chest, colon, head and neck, kidney, and lung. A significant number of phantom images are also archived. In addition, a few collections include non-image DICOM objects such as presentation state; structured report; and radiotherapy dose, plan, and structure. TCIA attracted more than 1,000 new

Table 3 Standard Operating Procedures, User Guides, Checklists, and Frequently Asked Questions

Standard Operating Procedures (SOPs)	
Name	Description or Purpose
New Collections SOP	Procedures for adding new collections
User Account Management SOP	Creating and maintaining user accounts
User Support SOP	Support Center assistance for TCIA users
Server Administration SOP	Scheduling and performing scheduled hardware and software maintenance
User Guides	
Name	Description or Purpose
Image Submitter Site User's Guide	Detailed assistance for image submitter
TCIA User's Guide	Detailed assistance for TCIA–NBIA application
Checklists	
Name	Description or Purpose
New Collection Submission Checklist	Review of new collection details gathered from image submitter by TCIA management
Submission Review Checklist	TCIA management procedures for new collections
Frequently Asked Questions (FAQ)	
Name	Description or Purpose
TCIA FAQ	Common questions by, and answers for, image submitters, image consumers, and the general public

users who collectively downloaded more than 228,000 image series (4.5 terabytes).

The first collections contributed to TCIA were the Reference Image Database to Evaluate Therapy Response (RIDER), Lung Image Database Consortium (LIDC), CT Colonography, and Federal Drug Administration (FDA) Phantom collections transferred from NCI's Cancer Imaging Program (CIP). The first new image-contributor group to TCIA was The Cancer Genome Atlas (TCGA) [27] Glioblastoma Multiforme (GBM) initiative (five contributing sites). Additional GBM sites as well as other TCGA initiatives [Brain Lower Grade Glioma (LGG), Breast Carcinoma (BRCA), Renal Clear Cell Carcinoma (KIRC), and Lung Adenocarcinoma (LUAD)] have and/or are scheduled to contribute images. In addition, the Quantitative Imaging Network (QIN) [28, 29] group has contributed or plans to contribute brain, breast, head–neck, phantom, and prostate images. Individual institutions have contributed breast, prostate, radiation treatment planning, and a variety of phantom images. Table 4 lists the various general- and limited-access collections and the numbers of images, and Fig. 5 shows monthly accumulations of images and contributing collection sites. Figure 6 shows the collection percentage by imaging modality, differentiating percentages by study (Fig. 6a), series (Fig. 6b), and image (Fig. 6c). Figure 7 shows, for all collections, image percentages by anatomy. Figure 8 shows the number of new

Table 4 TCIA collections with access types and image counts

Collection	Access	Images
Breast Diagnosis	General	105,050
CT Colonography	General	941,771
Head–Neck Cetuximab	Limited	15,199
LIDC–IDRI	General	244,527
NaF Prostate	Limited	41,404
Phantom FDA	General	634,256
Prostate Diagnosis	General	18,584
Prostate MRI	Limited	22,036
QIBA CT-1C	General	69,258
QIN Breast	Limited	16,646
QIN Lung	Limited	1,168
QIN Phantom	Limited	466
QIN Prostate	Limited	25,981
REMBRANDT	General	110,020
Renal Training	Limited	3,905
RIDER Breast MRI	General	2,400
RIDER Lung CT	General	15,716
RIDER Lung PET–CT	General	269,511
RIDER Neuro MRI	General	70,220
RIDER Phantom MRI	General	7,061
RIDER Phantom PET–CT	General	2,231
TCGA-BRCA	General	47,829
TCGA-GBM	General	603,425
Total images		3,268,664

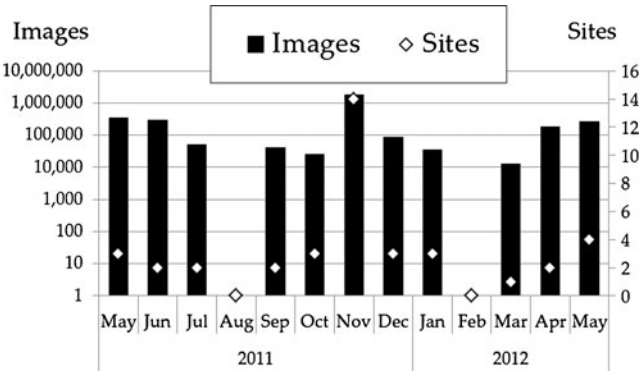


Fig. 5 TCIA public-server number of images (*vertical bars*; left vertical axis log-scale) and number of contributing sites (right vertical axis) by month

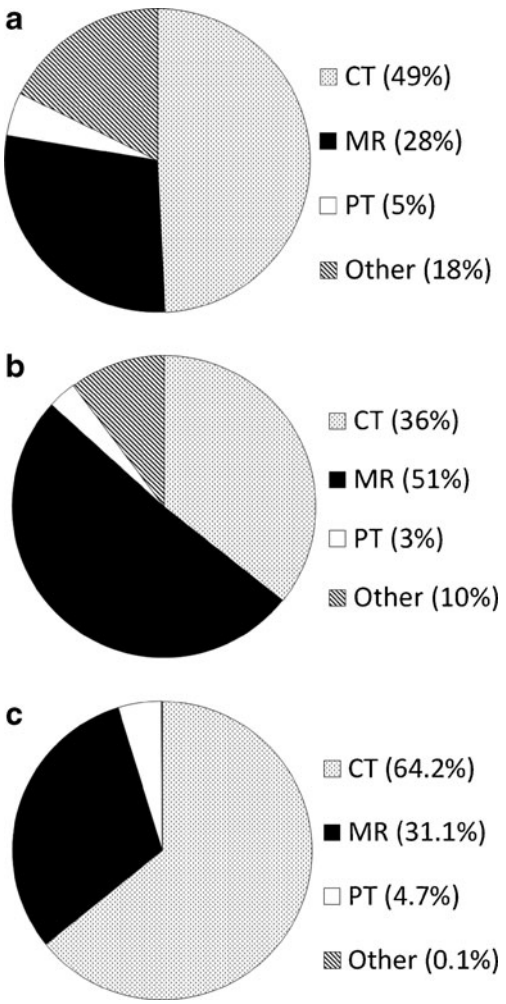


Fig. 6 **a** Study percentages by modality. **b** Series percentages by modality. **c** Image percentages by modality

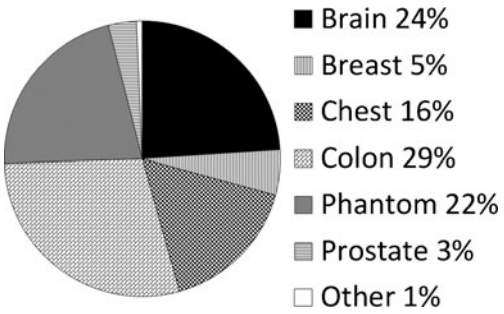


Fig. 7 Image percentages by anatomy. "Other" includes head–neck, kidney, lung, and non-image objects

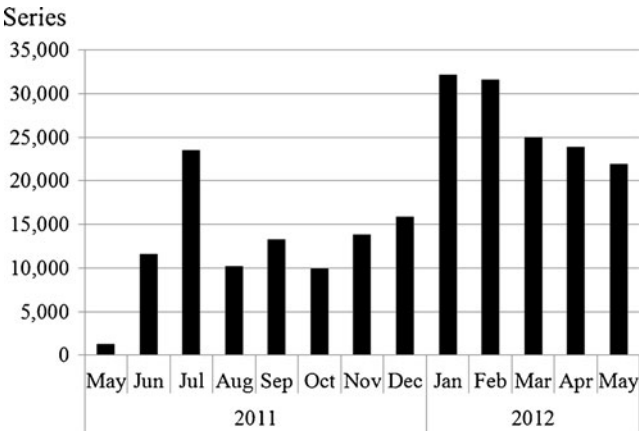


Fig. 9 TCIA image-series downloads by month

and cumulative user accounts by month and clearly demonstrates the steady growth of the user community. Figure 9 shows image-series downloads by month as a measure of system usage. Figure 10 shows the Help Desk system issue-tickets by month. User Accounts tickets (account creation, re-setting password, etc.) dominate (90 %). General tickets (7 %) are often queries regarding specific collections. All Other tickets (3 %) are tracked as Image (images seemingly unavailable or missing), New Collection (enquiries by potential image submitters), and Critical (server and web-site outages). Most ticket activity is via email; the Support Center received fewer than 75 telephone calls among a total 1,439 tickets.

Discussion

Before the advent of image repositories such as TCIA, it was difficult, if not impossible, for investigators to share or find research-relevant clinical image data collections. The NCI vision of a managed resource to supply a quality set of data with customer support actively facilitates reuse of existing imaging and clinical data. This TCIA resource supports the researcher who wants to test new algorithms or to validate the procedures in published studies.

Though NCI must approve all submission proposals, NCI encourages users and groups of users to contribute their images and participate in this public-sharing resource. A continuous infusion of both new and/or unique images will keep TCIA viable and attractive to clinical researchers and disease-detection/analysis software developers.

TCIA also facilitates image sharing among investigators from different institutions who are, perhaps, collaborating in common multicenter research programs. The variety of groups (TCGA and QIN) and individual institutions that have contributed images, as well as the variety of cancers types represented, bodes well for the success of the TCIA as a diversified public archive of cancer images.

An important part of the image-submission process includes working with image submitters to properly de-identify the submitted images while retaining scientifically valuable metadata. The de-identification process is managed by knowledgeable staff and provides a uniform mechanism that has been reviewed and approved by the Institutional Review Board of Washington University. Once images are transferred to Washington University, the image-curation process normalizes collection names and assures image quality and data integrity. In addition, TCIA staff work with image submitters to understand whether provided image DICOM Series Descriptions

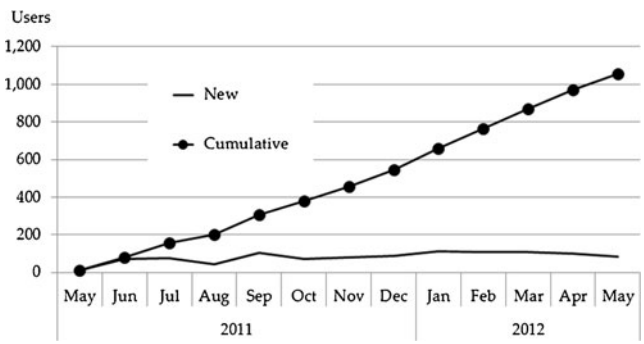


Fig. 8 TCIA registered user accounts by month

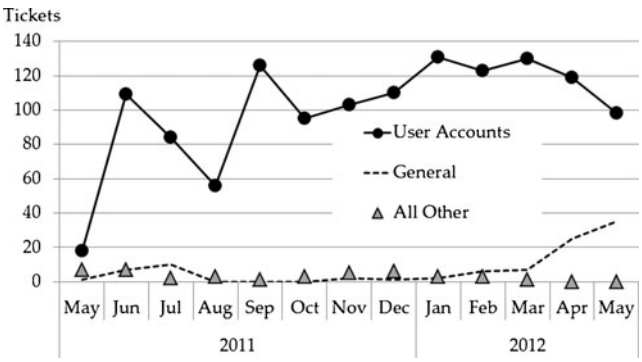


Fig. 10 Request Tracker (RT) tickets by month

adequately define the image series to which they belong and, if inadequate, to determine how to re-cast them to be meaningful for image consumers who will download them at some future date.

While TCIA personnel had experience with other imaging archives prior to TCIA, the variety of TCIA collections has taught that nearly every collection is unique and offers its own set of challenges: submission process (single shots, batches), single versus mixed modalities, unique versus multiple scanner vendors/models/software levels, no prior de-identification versus prior de-identification (known details versus not), and the imaging and DICOM experience levels (novice to professional) of image submitters. These variations sometimes challenge the speed at which collections are submitted and become available for public access; at the same time, TCIA staff gain additional experience and add to the knowledge base of managing large image archives. Any measurement of level of effort is difficult to come by because of the variety and magnitude of collections as well as issues encountered as they are processed. For example, a collection with 1,000 images from multiple modalities and even more scanners might take longer than a collection 100 times the size from a single modality and single scanner.

Bottlenecks are not unusual for a number of reasons. The typical steady state is 10–20 collections in process at various stages. Should multiple image submitters transmit images simultaneously, the intake server handles the load, but ongoing dialog between submitters and TCIA management is sometimes delayed by the multiplicities, especially if one or more submitters report submissions problems and/or some of their images are quarantined at their sites or on the intake server. While multiple curators are available to visually inspect new-collection images, they often experience database contention problems when they are trying to curate at the same time, whether or not they working on the same collection. Collections with images from multiple scanners complicate the TagSniffer analyses; this issue becomes even more acute when images are from a scanner model not previously encountered so that its DICOM conformance statement needs to be analyzed to understand which vendor private tags could contain PHI.

The gridlock caused by bottlenecks is attenuated with several strategies. First, there are multiple personnel in every nearly every role level; this permits flexible involvement when collections processing becomes intense with multiple collections and deadlines, and it allows for illness and vacation coverage. Second, managers are gradually being trained to assist with the reading of conformance statements, the analyses of TagSniffer reports, and the casting of new CTP de-identification scripts. Third, the visual-inspection contention issue seems to have disappeared with a clustering scheme whereby the NBIA application exists on a ring arrangement of multiple VMs (on both intake and public servers), with each curator assigned to a specific VM. The database, replicated on the multiple VMs, is updated instantaneously.

The number and variety of collections available to the general public, together with a broad base of more than 1,000 users, demand a high-availability repository that TCIA has provided during its first year of operation. The outlook promises additional collections from more sites and an increasing demand for images, particularly from large research groups such as TCGA and QIN.

TCIA is built upon Open-Source resources. The NBIA application (NCI), the TCIA wiki (Confluence), the ticket-tracking system (RT Tracker), and the software-revision-control system (GForge) are external resources. TCIA has contributed TagSniffer and the Extraction Tool and facilitated enhancements to both NBIA and CTP (RSNA). The regular application of the TCIA TagSniffer process contributes to the knowledge base (maintained as a public resource on the TCIA wiki) of understanding the mining of significant data from DICOM private tags while at the same time exposing the dangers of private tag PHI.

Described here are the details of operation and management of TCIA. Additional information, including system/network/software architecture, is available elsewhere [19, 30, 31]. Replication of TCIA, or some variant thereof, though achieved with Open-Source resources, would be a non-trivial task without the right team of expertise. In the spirit of Open Source, the TCIA team is willing to dialog with those entertaining such thoughts.

Conclusion

TCIA has upgraded the NBIA software program and added specialized tools to create an easily accessible archive of cancer images. In 1 year of operation, the archive has grown to a resource of more than 3 million images available to more than 1,000 worldwide users.

The NCI has funded a managed resource to support a centralized collection of cancer imaging data. NCI's Cancer Imaging Program and WUSTL have collaborated to build the infrastructure using NBIA open-source software and to put in place processes and staff members who actively manage the archive and provide industrial-level support to both image submitters and end users. This resource helps those investigators who need to publish their collections as well as those who are looking for high-quality data sets to further their research.

Acknowledgments This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. The Washington University component is funded by SUBCONTRACT 10XS220: SAIC-F (PI: Prior) Image Archive Hosting.

We thank John Perry, TCIA consultant, who provided expert CTP and de-identification guidance throughout. We also thank our image curators Joan Moulton, Mary Wolfsberger, and Tracy Nolan for their detailed image-quality-assurance efforts.

References

1. Birney E, et al: Mining the draft human genome. Nature 409(6822):827–828, 2001
2. Benson DA, et al: GenBank. Nucleic Acids Res 25(1):1–6, 1997
3. Howe D, et al: Big data: the future of biocuration. Nature 455(7209):47–50, 2008
4. Van Essen D, et al: The Human Connectome Project: a data acquisition perspective. Neuroimage 62:2222–2231, 2012
5. Marcus D, et al: Informatics and data mining tools and strategies for the human connectome project. Front Neuroinform 5:4, 2011
6. Grethe JS, et al: Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease. Stud Health Technol Inform 112:100–110, 2005
7. Keator DB, et al: A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN). IEEE Trans Inf Technol Biomed 12(2):162–172, 2008
8. Hampton T: Cancer Genome Atlas. JAMA 296(16):1958–1958, 2006
9. RSNA. The RSNA Clinical Trial Processor. 2012 [cited 2013 4/19/2013]; Available from: http://mircwiki.rsna.org/index.php?title=CTP-The_RSNA_Clinical_Trial_Processor
10. NCI. National Cancer Imaging Archive (NCIA). 2008 [cited 2013 04/19/2013]; Available from: http://cabig.cancer.gov/objects/pdfs/NCIA_toolsheet_060608_508.pdf
11. NCICB. National Biomedical Imaging Archive. 2011 [cited 2013 4/19/2013]; Available from: https://imaging.nci.nih.gov/ncia/login.jsf
12. NCICB. NCI Center for Bioinformatics Downloads [cited 2013 4/19/2013]; Available from: http://ncicb.nci.nih.gov/download/#NTools
13. NAIMS. Osteoarthritis Initiative (OAI). 2010 [cited 2013 4/19/2013]; Available from: http://www.niams.nih.gov/funding/funded_research/osteoarthritis_initiative/oai_study_update.asp
14. NAIMS. Osteoarthritis Initiative Image Archive. 2011 [cited 2013 4/19/2013]; Available from: https://niams-imaging.nci.nih.gov/ncia/login.jsf
15. Moore SM, et al: Workstation acquisition node for multicenter imaging studies. Proc SPIE 4323, Medical Imaging 2001: PACS and Integrated Medical Information Systems: Design and Evaluation, 271 (August 7, 2001); doi:10.1117/12.435489, 2001
16. Rhodes C, et al: Regulatory Compliance Requirements for an Open Source Electronic Image Trial Management System. In IEEE EMBS Conf 2010. 2010. IEEE Xplore
17. Vendt B, et al: Silent Cerebral Infarct Transfusion (SIT) trial imaging core: application of novel imaging information technology for rapid and central review of MRI of the brain. J Digit Imaging 22(3):326–343, 2009
18. Clark K, et al: Collecting 48,000 CT exams for the Lung Screening Study of the National Lung Screening Trial. J Digit Imaging 22(6):667–680, 2009
19. Tarbox L, et al: The Cancer Imaging Archive (TCIA): creating a large public image collection. [Presentation] 2012 June 7, 2012 [cited 2013 4/19/2013]; Available from: http://erl.wustl.edu/documents/posters/SIIM2012-TCIAtechnical.ppt
20. CBIIT. Common Security Module (CSM) [cited 2013 4/19/2013]; Available from: https://wiki.nci.nih.gov/display/caCORE/Common+Security+Module+%28CSM%29
21. NEMA. Digital Imaging and Communications in Medicine (DICOM) [cited 2013 4/19/2013]; Available from: http://medical.nema.org
22. TCIA. Image Submitter Site User's Guide. 2012 [cited 2013 4/19/2013]; Available from: https://wiki.cancerimagingarchive.net/display/Public/Image+Submitter+Site+User%27s+Guide
23. Moore S: TagSniffer. 2012 [cited 2013 4/19/2013]; Available from: https://mirgforge.wustl.edu/gf/project/dicomtagsniffer/docman/?subdir=56
24. Prior FW, et al: The Cancer Imaging Archive. 2012 [cited 2013 4/19/2013]; Available from: http://www.cancerimagingarchive.net/
25. MIR. Mallinckrodt Institute of Radiology (MIR) gForge. 2011 [cited 2013 4/19/2013]; Available from: https://mirgforge.wustl.edu/gf
26. Moore S, et al: De-identification Knowledge Base. 2012 [cited 2013 4/19/2013]; Available from: https://wiki.cancerimagingarchive.net/display/Public/De-identification+Knowledge+Base
27. CIP. The Cancer Genome Atlas (TCGA). 2012 [cited 2013 4/19/2013]; Available from: http://cancergenome.nih.gov
28. Clarke LP, et al: Quantitative imaging for evaluation of response to cancer therapy. Transl Oncol 2(4):195–197, 2009
29. CIP. The Quantitative Imaging Network (QIN). 2012 [cited 2013 4/19/2013]; Available from: http://imaging.cancer.gov/programsandresources/specializedinitiatives/qin
30. Vendt B, et al: The Cancer Imaging Archive (TCIA). [Poster] 2013 March 26, 2013 [cited 2013 4/19/2013]; Available from: http://erl.wustl.edu/documents/posters/TCIA-WashU-March-2013-v5.pdf
31. Clark K, et al: The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. [Poster] 2012 June 7, 2012 [cited 2013 4/19/2013]; Available from: http://erl.wustl.edu/documents/posters/siim2012.pdf