The Author(s). *BMC Genomics* 2016, **17**(Suppl 4):434
DOI 10.1186/s12864-016-2797-9

BMC Genomics

RESEARCH Open Access

The biorepository portal toolkit: an honest brokered, modular service oriented software tool set for biospecimen-driven translational research

Alex S. Felmeister[1,2*], Aaron J. Masino[1], Tyler J. Rivera[1], Adam C. Resnick[1,3] and Jeffrey W. Pennington[1]

From IEEE International Conference on Bioinformatics and Biomedicine 2015
Washington, DC, USA. 9-12 November 2015

Abstract

Background: High throughput molecular sequencing and increased biospecimen variety have introduced significant informatics challenges for research biorepository infrastructures. We applied a modular system integration approach to develop an operational biorepository management system. This method enables aggregation of the clinical, specimen and genomic data collected for biorepository resources.

Methods: We introduce an electronic Honest Broker (eHB) and Biorepository Portal (BRP) open source project that, in tandem, allow for data integration while protecting patient privacy. This modular approach allows data and specimens to be associated with a biorepository subject at any time point asynchronously. This lowers the bar to develop new research projects based on scientific merit without institutional review for a proposal.

Results: By facilitating the automated de-identification of specimen and associated clinical and genomic data we create a future proofed specimen set that can withstand new workflows and be connected to new associated information over time. Thus facilitating collaborative advanced genomic and tissue research.

Conclusions: As of Janurary of 2016 there are 23 unique protocols/patient cohorts being managed in the Biorepository Portal (BRP). There are over 4000 unique subject records in the electronic honest broker (eHB), over 30,000 specimens accessioned and 8 institutions participating in various biobanking activities using this tool kit. We specifically set out to build rich annotation of biospecimens with longitudinal clinical data; BRP/REDCap integration for multi-institutional repositories; EMR integration; further annotated specimens with genomic data specific to a domain; build application hooks for experiments at the specimen level integrated with analytic software; while protecting privacy per the Office of Civil Rights (OCR) and HIPAA.

Keywords: Biorepository research, Translational bioinformatics, Precision medicine, Honest broker, Cancer genomics, Data integration, Data representation, Open source, Patient health information protection, Patient privacy

* Correspondence: felmeistera@email.chop.edu
[1]Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, 3401 Civic Center Blvd, Philadelphia, PA, USA
[2]College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA, USA
Full list of author information is available at the end of the article

Background

Current research is yielding rapid advances in personalized, precision medicine through targeted therapies based on an individual's genome, genomic biomarkers, and cell biology across adult and pediatric translational research [1, 2]. This type of research has become increasingly dependent on the collection of large cohorts of high quality human biospecimens that are paired with clinical annotations [3]. While biospecimen-driven research is widely practiced, it is often limited in scope because it requires time-consuming manual processes such as retrospective annotation, cohort identification and institutional human subjects research oversight [4]. Consequently, many academic medical centers are creating large institutional biospecimen resources that can be leveraged by numerous investigators [5]. There is a trend towards these resources becoming indispensable in academic medical centers [3, 6].

Biorepository data is typically captured in longitudinal, asynchronous workflows that pose software design and data integration challenges [7]. An optimal system must provide de-identified, granular and longitudinal data to researchers while also enabling data collection workflows that require patient identification [8]. The required data often resides in separate systems such as a Laboratory Information Management System (LIMS), Research Data Capture tools, the Electronic Health Record (EHR), genomic data stores and high performance computing clusters [9, 10]. Integrative solutions are necessary at the point of collection and at information and specimen retrieval. The data must be curated to ensure it is persisted in an understandable representation for researchers in a specific medical domain [11, 12]. As biorepository resources include more clinical information and grow in scale, there are more opportunities for protected health information (PHI) to be injected into the process [13, 14]. Therefore, a central component to this toolkit is an informatics-based approach to honest brokering [15]. We build on methods described in Dhir et al. and Boyd et al. that describe specific implementations of software to aid in the honest brokering between various types of clinical data collection and de-identified biorepositories [14, 16]. We take a slightly different approach by creating non-user facing software *service* similar to Boyd, et al. for the honest broker as one of many components of a toolkit of connected operational biorespository informatics resources. We remove the human component completely from the de-identification and re-identification of research records in connected research systems.

In this paper, we address the creation of a robust biorepository management platform that enables association of a physical biospecimens, clinical diagnoses, and patient, genomic and research. This platform utilizes a modular software architecture developed at the Children's Hospital of Philadelphia (CHOP) in partnership with the Children's Brain Tumor Tissue Consortium (CBTTC) [17]. The platform was developed in the specific context of distributed biorepository and biobank studies in biological tissue and genomic research. In this manuscript, we describe the requirements, challenges in architectural design and implementation to create integrated data resources in biospecimen-driven translational research. We designed and utilized an open source, modular software toolkit that supports biorepository operations and de-identified secondary usage. We created an operational and scientific resource that protects subject privacy, allows for variable specimen and data management workflows and flexible resource queries. When new systems and workflows are introduced, the toolkit allows for flexible introduction of new data types, systems and operational workflows spanning specimen, clinical, imaging and genomic data. Our platform allows for extensible software and data resources for biospecimen-driven translational research.

Cancer focus

Cancer is a main focus of current precision medicine initiatives. This is reflected in politics, the media, public funding and medical research community priorities [18]. We are in an age of increased use of web technologies that allow us to reach new levels of productivity and connectivity in business, finance, government and entertainment [19]. The time has come for us to use these same techniques to unravel the complexities of cancer [20]. New breakthroughs are helping us use our own immune systems to target an increasing list of common cancers [21]. Unfortunetely, time is not on the side of children suffering from rare brain tumors. Recent research and government population health programs identify over 120 types of pediatric brain cancers [22]. To make matters more complex, the origins of brain tumors in kids is widely unknown [22–25]. Pediatric cancer patients are treated for cancers with adult-based therapies and there is a lack of investment from pharmaceutical companies in the specific diseases affecting children [26]. It is essential to create biologic- and data-centric resources to find pathways and molecularly describe disease seen in research similar to Bastianos et al. and Parsons et al. where developments, respectively, uncover a molecular pathway in Craniopharyngiomas and a comprehensive molecular description of the common childhood brain tumor, Medulloblastoma [27, 28]. Though molecularly based research has become common with the availability of high throughput technologies, further progress is needed in infrastructure, specifically in cancer research, that allows for complete clinical annotation of specimen and genomic data from consented subjects ([29], p. 549).

Initial research of rare tumors at CHOP brought biorepository data sharing, management, and annotation issues to the forefront. A need for enhanced capabilities was particularly evident in two proposed studies targeted by our initial software system design. These tumor biorepository studies originally used a human honest broker to manage the de-identification and re-identification of records to exclude protected patient information (a/k/a patient health information) from the research LIMS. The process began with manual data intake by a data manager on hardcopy REDCap case report forms (CRFs) [30]. The CRFs contained the patient identifiers: Medical Record Number (MRN), First Name, Last Name and Date of Birth. The CRF was physically delivered to a human honest broker that would create a new electronic REDCap record with a research identifier. The hardcopy CRF was then returned to the data manager with the research identifier attached and the patient identifiers removed. The data manager then abstracted the hand written CRFs to the associated REDCap project record. Each longitudinal data collection event required manual re-identification by the human honest broker. This process became unsustainable as biospecimen and clinical data collection increased and molecular experimentation associated with records began. It was also difficult to complete the CRF in a single patient encounter due to variations in encounter length and frequency. This experience clearly illustrated the need for a scalable solution that would abide by NCI Best Practices for Biospecimen Resources [15]. The Biorepository Portal Toolkit (BRP Toolkit) project was subsequently developed to support biorepository development at institutional scale.

Methods

Modular approach

We took a modular and entity-driven integrated systems approach to facilitate variable specimen acquisition and data collection events. The primary entity is the patient enrolled as a research subject on the study. The subject entity is created in the electronic honest broker (eHB) and assoicated with a master patient index (MPI) and the subject's instutional origin. Each external research record associated with the subject record, in this case the data management tool, REDCap, is not limited to a one-to-one relation of subject-to-REDCap record. The subject entity can be assoicated with many projects, forms and records in a 1-to-many entity relationship [31]. We, in tandem, built a research portal, dubbed the Biorepository Portal (BRP). The BRP can access subject records in the electronic Honest Broker (eHB) and subsequent external research records through token-based authorization from that client system. The BRP reproduces the REDCap electronic Case Report Forms (eCRFs) based on records stored in the eHB with a custom REDCap client utilizing the REDCap application programing interface (API), in real-time. This produces a complete form for that subject at time of access. It displays the subject information and identifiers at the top of the screen at all times during form data entry and while shifting from form to form. A research coordinator or data manager can enter any temporal and longitudinal research data based on their protocol subject list at anytime or in any order (i.e., asynchronously) while maintaining the continuous de-identification and re-identification of research data automatically.

The CHOP Biorepository Core Facility utilizes ThermoFisher Nautilus as its LIMS. As part of our method, we also built a client to this LIMS that allows for association of an arbitrary number of specimen records in the LIMS with the corresponding subject record. In this way specimens can also be collected longitudinally over time. Data and specimen coordinators have the ability to associate sets of specimens with a subject or event and annotate that specimen on the fly in one system. For downstream integration, we use the same eHB software service to perform our Extract Transform Load (ETL) processes that are tailored to each project. The result is a regularly updated non-human subject research database that allows for seamless queries spanning research and clinical data sets. We allow collaborators to access specific sets of data via the data exploration tool, Harvest [32], customized for each project. The phenotypic data associated with specimen records can be integrated with direct clinical data from the EHR with appropriate institutional permissions. The modular approach allows us to integrate genomic visualization tools at the specimen level where applicable. For cancer genomics specifically, we utilize the CBioPortal [33, 34], an open source tool to visualize mutation and gene expression data from The Cancer Genome Atlas (TCGA).

Web service oriented architecture

The integration of tools is accomplished by taking advantage of modern web technologies. Our methods are rooted in web service oriented architecture (SOA). This pattern pervades the current generation of computing and web technology and is rapidly expanding through virtual resources accessed via network resources (i.e., cloud computing) [35]. We created a plug-and-play experience working with multiple tools in a web environment. We employ REST (Representational State Transfer) API architecture over HTTP protocols providing uniform channels for applications integrated into the tool kit [36]. In this section, we describe each specific technology in the stack, each with its own set of RESTful end-points that allow us to guide users through multiple tools as they interact with biorepository resources. Though SOAs are widely used in the field of biomedical informatics to build complex application tool chains, they are not transparent as to their usage and sometimes have very poor adoption because of their complexity [37]. To address this, we mask technical complexity with end-user tools that are familiar

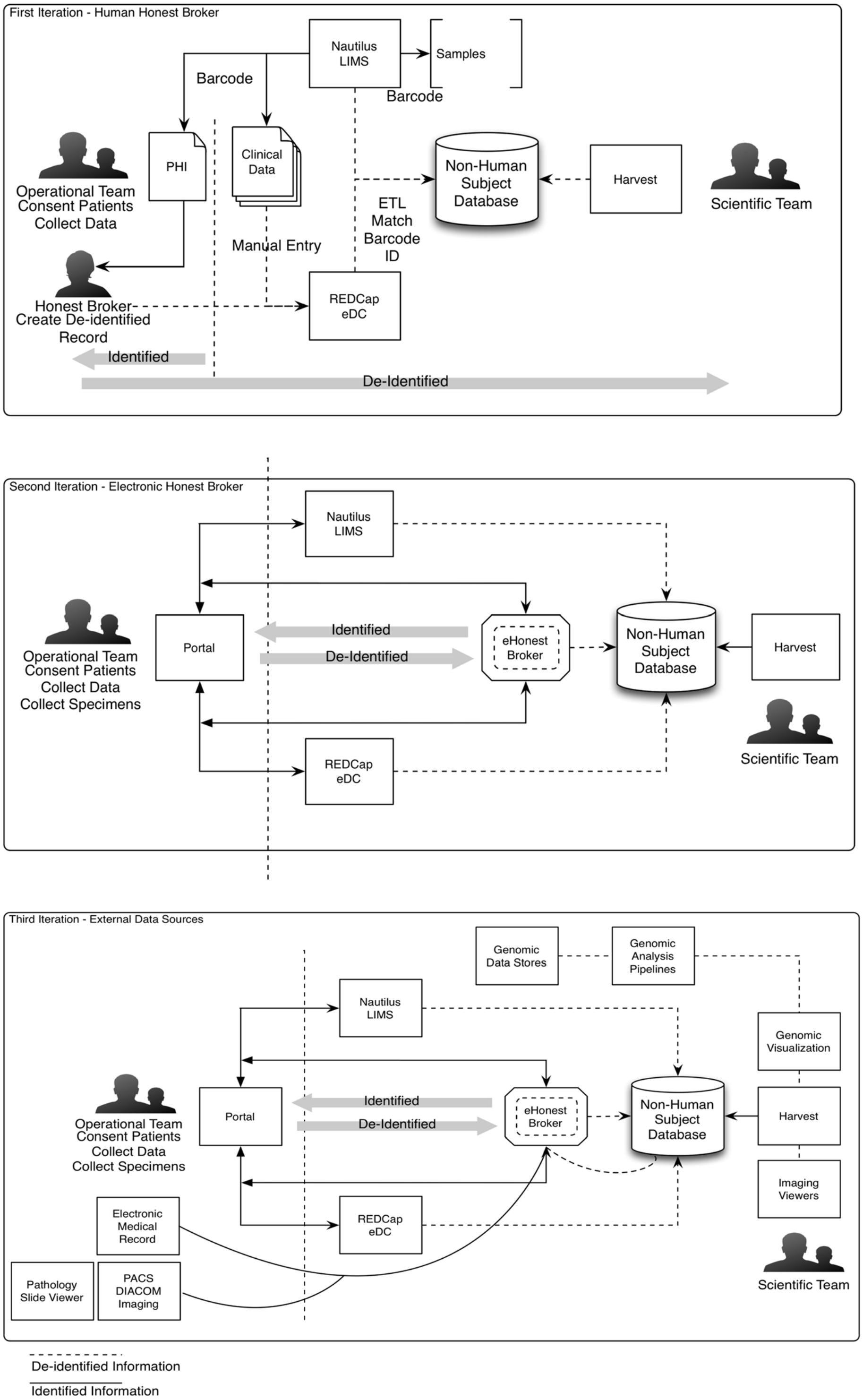

First Iteration - Human Honest Broker
Nautilus LIMS
Samples
Barcode
Barcode
PHI
Clinical Data
Non-Human Subject Database
Harvest
Operational Team Consent Patients Collect Data
Scientific Team
ETL Match Barcode ID
Manual Entry
Honest Broker Create De-identified Record
REDCap eDC
Identified
De-Identified
Second Iteration - Electronic Honest Broker
Nautilus LIMS
Portal
Identified
De-Identified
eHonest Broker
Non-Human Subject Database
Harvest
Operational Team Consent Patients Collect Data Collect Specimens
REDCap eDC
Scientific Team
Third Iteration - External Data Sources
Genomic Data Stores
Genomic Analysis Pipelines
Nautilus LIMS
Genomic Visualization
Portal
Identified
De-Identified
eHonest Broker
Non-Human Subject Database
Harvest
Operational Team Consent Patients Collect Data Collect Specimens
Electronic Medical Record
REDCap eDC
Imaging Viewers
Pathology Slide Viewer
PACS DIACOM Imaging
Scientific Team
De-identified Information
Identified Information

Fig. 1 Three workflow iterations. Iterations of the introduction of the modular tool kit from human intervention, electronic intervention to fully realized modular approach

to research coordinators, specimen coordinators and data managers. There is wide recognition that there is no compliant Health Level Seven (HL7)-type interoperability standards when using SOA's [9]. The operational applications and scientific applications we implemented utilize an honest broker software service, a well-known method of dealing with de-identification in biorepository research that maintain the HIPAA compliance in downstream systems [14, 38–41]. Figure 1 illustrates a high-level architecture description along with external integration points for key scientific and clinically relevant data points for data entry and reuse. This figure is split into three parts, the last of which describes the fully realized modular approach integrating both clinical, genomic and research specimen resources.

Electronic Honest Broker (eHB) software service

The concept of an honest broker has been implemented in other academic medical center environments to protect privacy when integrating research data ([14, 42], pp. 56–107). Central to our solution is the eHB, a web-based software service with end-to-end encryption that maintains an index of subjects linked to their associated research records. The initial studies/projects targeted with this solution began as a manual paper-based process of considerable complexity, and are now an "informatics tool" [15]. The index in the eHB uniquely identifies each patient through a combination of organizational association (e.g., The Children's Hospital of Philadelphia) and a unique organization provided identifier (e.g., Medical Record Number). The eHB associates each subject record with trusted external systems and known record identifiers in those systems. For annotated biorepository studies, the eHB maintains associations to de-identified records in REDCap and the LIMS. Following our SOA approach, the eHB makes data and system functionality available via a REST web service. This allows the addition of new data management application clients in a way that is system- and programming language-agnostic. To control access, the eHB uses token-based authorization, and encrypts its data both at rest and in transit, relying on client-side keys to decrypt the payload received from the API. For known applications, the eHB provides subject data with few restrictions. Client applications determine the context of what information is appropriate to display to a user, thereby enabling flexibility to meet different workflow and protocol requirements.

The eHB has a limited web-facing user interface that allows for the administration of access tokens and users. It can be managed through a comprehensive set of create, read, update, delete (CRUD) operations exposed by the REST API. Client applications, described in subsequent sections, determine the context of the request. The client application requests resources of the eHB service via a URL endpoint secured using transport layer security (TLS) and, with appropriate keys and credentials, can read and write data to and from the eHB

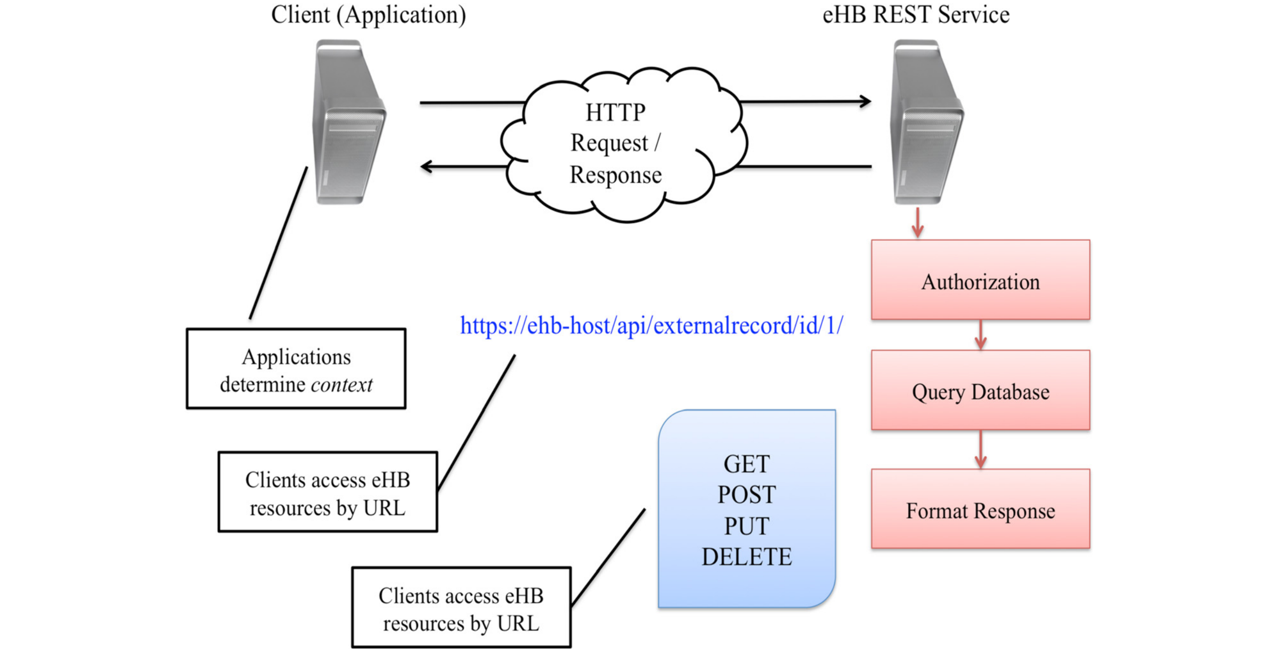


Fig. 2 The interaction between the eHB and client applications. This takes place over HTTP through CRUD (GET, POST, PUT, DELETE) operations. Client applications (i.e., the BRP) determine the context of the information sent and received to the eHB service. Within the eHB, data is encrypted at rest, and in transit. A query of data can only take place through the eHB software service

service. The eHB REST service handles authorization of the application, encrypts data and formats a response. The actual database behind the eHB service stores only encrypted information and would be unreadable if accessed. This type of encryption decouples the identifiers and encrypts any and all information going into the eHB database and can be considered "privacy-by-design" by selectively sending and granting access to information based on context and only storing the minimum set of information needed to stitch together a record for data management or data query [16, 43]. The architectural design of the eHB, illustrated in Fig. 2, utilizes web request type architecture to be a completely independent component of the tool-kit. The eHB model is similar to prior research in clinical informatics and the EHR. Architecturally, the health record must have the element of being future-proof. There is an assurance of openness and portability through standards, flexibility and scalability, semantic interoperability and acceptance from the domain experts. In Blobel et al., the authors discuss the fundamentals of future-proof health systems describing an "atomic component" [44]. We apply this notion by creating an "atomic component" that must be guaranteed utilizing the eHB to associate the subject on the study (as the atomic component) with their biospecimens, phenotypic and genomic data.

The Biorepository Portal (BRP)

The eHB service discussed in the previous section is exposed solely to external client systems through APIs and lacks a user interface. The user-facing BRP component provides an integrated view of multiple data management software applications. We implement a web-based application that uses HTTP protocols to communicate with the eHB and external systems. The BRP allows research staff to work with subject identifiers, subject research data and associate specimen records. The BRP presents the honest brokered data by integrating custom clients using external system APIs to integrate in real-time, the research data and patient identifiers and records. The BRP allows access via authentication utilizing institutional identity management systems that comply with all network access guidelines for hospital clinical systems then limits users to specific cohorts, institutions and data resources accessed through the portal. Figure 3 describes the layers of user access to protocols and data sources representing applications connected to the tool in a sudo-ER diagram.

The BRP provides context to operational tools through clients to tools used in the context of biospecimen management. These clients are configured in the BRP as *data sources*. Display options for each client can be specified by a configuration encoded in JSON [45]. JSON formatted configuration files enable the portal to manage the display and access of REDCap forms and events in the data management processes based on workflows for capturing longitudinal data. The BRP not only relates an honest brokered subject record to external research systems, but also relates records to each other across systems. The following describes two implementations of external data sources; in this case a laboratory system and a data management system. In this case, records in these systems are linked in the BRP to provide further context around the connection of external research system (i.e., specimen collection record to case report form). Figure 4 is an example of the record listing in the BRP.

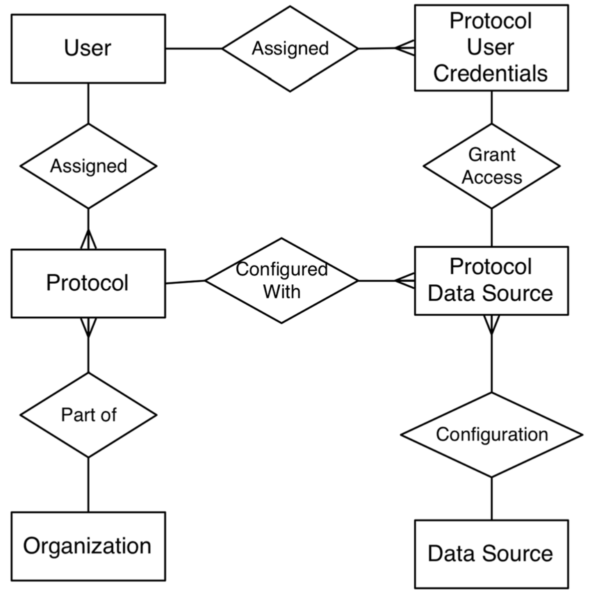
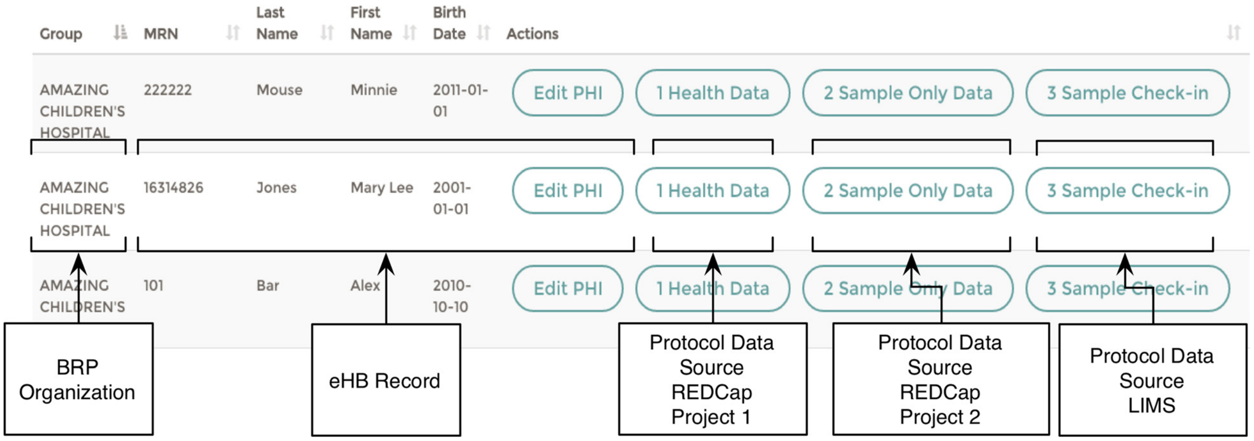


Fig. 3 Sudo ER diagram of complex relationships between users and data sources. Users are granted access to subject records based on protocols. Each protocol groups a list of subjects that only authorized users can manage and are part are assigned to. Within the walls of a protocol, there are groups of subjects that allow for users to differentiate between institutional or site-based cohorts (i.e., organizations). Each unique subject enrollment creates a new record for that subject in the eHB. Within a protocol there are specific data entry points set up as protocol data sources (e.g., a REDCap case report form project or LIMS specimen project), and each user has credentials to each data source. Data sources are set up as clients to each external system represented in the BRP. In our case, we have created data sources for two systems, REDCap and ThermoFisher Nautilus LIMS

Data sources

REDCap client

The REDCap client in the BRP makes a request for the metadata and data from a subject record stored in the eHB and recreates the form requested utilizing the REDCap API. The REDCap form client shows the patient Medical Record Number (MRN), last name, first name



Group	MRN	Last Name	First Name	Birth Date	Actions			
AMAZING CHILDREN'S HOSPITAL	222222	Mouse	Minnie	2011-01-01	Edit PHI	1 Health Data	2 Sample Only Data	3 Sample Check-in
AMAZING CHILDREN'S HOSPITAL	16314826	Jones	Mary Lee	2001-01-01	Edit PHI	1 Health Data	2 Sample Only Data	3 Sample Check-in
AMAZING CHILDREN'S	101	Bar	Alex	2010-10-10	Edit PHI	1 Health Data	2 Sample Only Data	3 Sample Check-in

BRP Organization
eHB Record
Protocol Data Source REDCap Project 1
Protocol Data Source REDCap Project 2
Protocol Data Source LIMS

Fig. 4 Screenshot example. A screenshot from the BRP displaying identifiable data entry along the de-identified entry points for client systems' data sources

and date of birth along with each customized research data capture form. After the form record is saved, the BRP utilizes the eHB software service to either create a new record or modify an existing record in the REDCap project. The REDCap project record identifier is hashed and randomly generated without use of derived patient information. eHB identifiers are generated utilizing the application client key, in combination with a salted hash value which is guaranteed to be unique [46]. Creating a research identifier not derived from a direct patient identifier is required when using patient data for research [13]. Research identifiers are created by the connected research system randomly and are not derived from any patient information.

Adding a layer that removes the subject entity from the REDCap projects associated with a study allows for REDCap to facilitate user access to projects, form building, data logging, and managing a study data dictionary [30]. The ability to supplement an entire REDCap project(s) or form(s) as specimen annotations is accomplished by associating another REDCap project with a portal project and, in turn, a subject. Our approach includes the ability to have variable numbers of projects and nested project records per patient. There are many variations of studies that use a variable number of REDCap projects/forms and project records depending on the need of the domain. For example, a BRP protocol can capture demographics one time in one REDCap project while collecting many diagnosis-type forms with longitudinal events in another project that allows for multiple records per subject. The eHB mediates and stores the links between the subject entity and their project records. Conversely, we allow for the tools to maintain separate cohorts of identified subjects where the data are stored in the same REDCap project. This is particularly important for studies in which multiple institutions are participating in sending data and specimens to one data and specimen-coordinating center (DCC/SCC). The link to REDCap records depends on the domain and temporal requirements of a biorepository study.

LIMS client

A key requirement in our choice of a LIMS was that it implements an API that exposed the ability to create new specimen records, print labels and update specimen records with tags from external systems. The LIMS assigns a unique identifier to a specimen collection event, and this identifier is associated with the subject entity in the eHB by user input via the BRP. The BRP has a custom client that allows specimen coordinators with the proper credentials to associate pre-labeled specimen accession kits with the subject entity. Specimen collection kits with proper collection tubes and labels are created prior to subject enrollment in the CHOP Biorepository Core Facility.

The specimen coordinator then scans the barcoded label on the kit through a LIMS client in the BRP to associate the kit with the patient. Any downstream laboratory work, for example; receiving, processing, analysis and storage are performed directly in the LIMS. The laboratory technicians processing and receiving specimen kits do not see patient identifiers, only the LIMS assigned identifier. This facilitates the longitudinal capture of multiple specimen collection events associated with one subject.

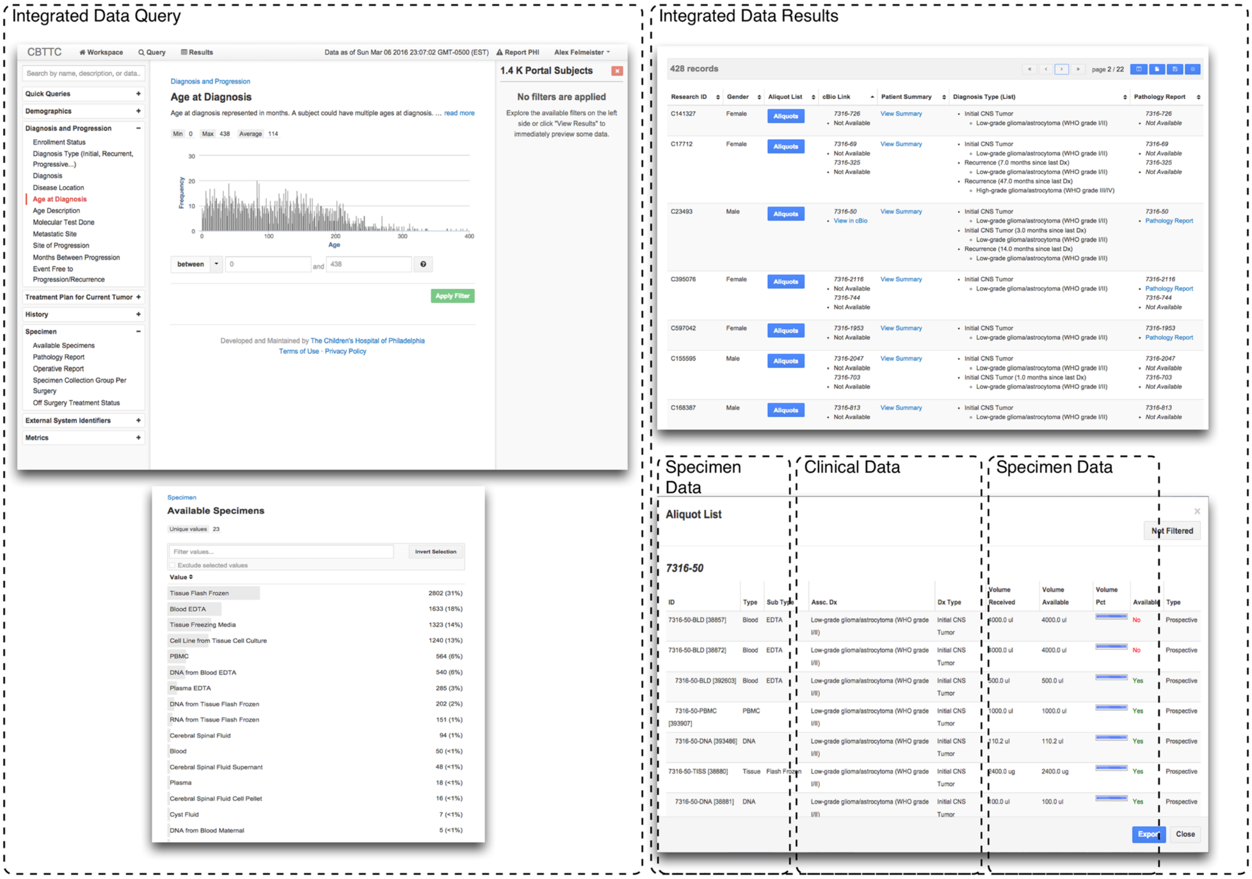
Access

Extract transform load

As Goble, et al., describe service oriented technology mechanisms "…[o]nce plumbed, the data have to be massaged and cleaned to make them fit together or conform to new schema” ([36], p. 689). We meet this requirement with ETL scripts written to integrate the disparate and de-identified data together for scientific use by researchers. This ETL process acts as another application with client access to the eHB. The first part of the ETL script uses application key-driven access to obtain a list of subject entities on specific protocols linked their respective research identifiers in data management applications. This linked list is used throughout the ETL process to join together and integrate data from disparate research systems and perform further de-identification where required by a study protocol. The ETL process produces data in a relational domain model suitable for researcher query. The ETL process is also where we integrate systems and data that are not part of the data entry in the BRP. If allowed by the protocol, the ETL process can query the eHB for patient records and pull clinical data from the health record and move it to the non-human subject biorepository database.

Researcher query tool and non-human subject research data resource

Integrating data sets through an ETL process is the starting point to fully realize the research potential of an integrated biobank. Researchers need to be able to discover available data and formulate queries for case definition and cohort creation. To enable this, we created a query tool that allows researchers to get quick answers to questions using available data without involving the Institutional Review Board (IRB) because all data is de-identified in the query tool [11]. This tool is implemented using the open-source Harvest framework [32]. Harvest gives the informatics team the ability to customize the application where necessary, but also have an out-of-box query tool for domains of relevant data. An example of a Harvest instance where users can search the multi-institutional biospecimen and annotation data of the CBTTC is shown in Fig. 5. This figure shows point-and-click access to multi-dimensional and disparate data in one place.



Data as of Sun Mar 06 2016 23:07:02 GMT-0500 (EST)

Diagnosis and Progression

Age at Diagnosis

Age at diagnosis represented in months. A subject could have multiple ages at diagnosis. ... read more

Min 0 Max 438 Average 114

Developed and Maintained by The Children's Hospital of Philadelphia
Terms of Use · Privacy Policy

1.4 K Portal Subjects

No filters are applied

Explore the available filters on the left side or click "View Results" to immediately preview some data.

Specimen

Available Specimens

Unique values 23

Value	
Tissue Flash Frozen	2802 (31%)
Blood EDTA	1633 (18%)
Tissue Freezing Media	1323 (14%)
Cell Line from Tissue Cell Culture	1240 (13%)
PBMC	564 (6%)
DNA from Blood EDTA	540 (6%)
Plasma EDTA	285 (3%)
DNA from Tissue Flash Frozen	202 (2%)
RNA from Tissue Flash Frozen	151 (1%)
Cerebral Spinal Fluid	94 (1%)
Blood	50 (<1%)
Cerebral Spinal Fluid Supernant	48 (<1%)
Plasma	18 (<1%)
Cerebral Spinal Fluid Cell Pellet	16 (<1%)
Cyst Fluid	7 (<1%)
DNA from Blood Maternal	5 (<1%)

428 records

Research ID	Gender	Aliquot List	cBio Link	Patient Summary	Diagnosis Type (List)	Pathology Report
C141327	Female	Aliquots	*7316-726* • Not Available	View Summary	• Initial CNS Tumor ∘ Low-grade glioma/astrocytoma (WHO grade I/II)	*7316-726* • *Not Available*
C17712	Female	Aliquots	*7316-69* • Not Available *7316-325* • Not Available	View Summary	• Initial CNS Tumor ∘ Low-grade glioma/astrocytoma (WHO grade I/II) • Recurrence (7.0 months since last Dx) ∘ Low-grade glioma/astrocytoma (WHO grade I/II) • Recurrence (47.0 months since last Dx) ∘ High-grade glioma/astrocytoma (WHO grade III/IV)	*7316-69* • *Not Available* *7316-325* • *Not Available*
C23493	Male	Aliquots	*7316-50* • View in cBio	View Summary	• Initial CNS Tumor ∘ Low-grade glioma/astrocytoma (WHO grade I/II) • Initial CNS Tumor (3.0 months since last Dx) ∘ Low-grade glioma/astrocytoma (WHO grade I/II) • Recurrence (14.0 months since last Dx) ∘ Low-grade glioma/astrocytoma (WHO grade I/II)	*7316-50* • Pathology Report
C395076	Female	Aliquots	*7316-2116* • Not Available *7316-744* • Not Available	View Summary	• Initial CNS Tumor ∘ Low-grade glioma/astrocytoma (WHO grade I/II)	*7316-2116* • Pathology Report *7316-744* • *Not Available*
C597042	Female	Aliquots	*7316-1953* • Not Available	View Summary	• Initial CNS Tumor ∘ Low-grade glioma/astrocytoma (WHO grade I/II)	*7316-1953* • Pathology Report
C155595	Male	Aliquots	*7316-2047* • Not Available *7316-703* • Not Available	View Summary	• Initial CNS Tumor ∘ Low-grade glioma/astrocytoma (WHO grade I/II) • Initial CNS Tumor (1.0 months since last Dx) ∘ Low-grade glioma/astrocytoma (WHO grade I/II)	*7316-2047* • *Not Available* *7316-703* • *Not Available*
C168387	Male	Aliquots	*7316-813* • Not Available	View Summary	• Initial CNS Tumor ∘ Low-grade glioma/astrocytoma (WHO grade I/II)	*7316-813* • *Not Available*

Aliquot List

7316-50

ID	Type	Sub Type	Assc. Dx	Dx Type	Volume Received	Volume Available	Volume Pct	Available	Type
7316-50-BLD [38857]	Blood	EDTA	Low-grade glioma/astrocytoma (WHO grade I/II)	Initial CNS Tumor	4000.0 ul	4000.0 ul	▬	No	Prospective
7316-50-BLD [38872]	Blood	EDTA	Low-grade glioma/astrocytoma (WHO grade I/II)	Initial CNS Tumor	4000.0 ul	4000.0 ul	▬	No	Prospective
7316-50-BLD [392603]	Blood	EDTA	Low-grade glioma/astrocytoma (WHO grade I/II)	Initial CNS Tumor	500.0 ul	500.0 ul	▬	Yes	Prospective
7316-50-PBMC [393907]	PBMC		Low-grade glioma/astrocytoma (WHO grade I/II)	Initial CNS Tumor	1000.0 ul	1000.0 ul	▬	Yes	Prospective
7316-50-DNA [393486]	DNA		Low-grade glioma/astrocytoma (WHO grade I/II)	Initial CNS Tumor	110.2 ul	110.2 ul	▬	Yes	Prospective
7316-50-TISS [38880]	Tissue	Flash Frozen	Low-grade glioma/astrocytoma (WHO grade I/II)	Initial CNS Tumor	2400.0 ug	2400.0 ug	▬	Yes	Prospective
7316-50-DNA [38881]	DNA		Low-grade glioma/astrocytoma (WHO grade I/II)	Initial CNS Tumor	100.0 ul	100.0 ul	▬	Yes	Prospective

Fig. 5 Integrated researcher view. Integrated non-human subject data and specimen query tool built on the Harvest platform. This interface allows for query across multiple systems in one place. The platform can be customized to allow for only certain data elements to be utilized that depend on domain and researcher requirements. Exposed in this example are elements captures in CRFs and the LIMS. There are also links out to genomics analysis tools

Results and discussions

Usage results

The toolkit supports multiple biospecimen-driven research studies. In these studies, the accessioning of specimens and related data has grown and completely changed in scale and volume. Figure 6 is a graph showing specimen accessioning 2011 to January, 2016 for the CBTTC. In 2012 the toolkit was adopted and the graph illustrates the change in accession events happening after the toolkit adoption with a varied pace of specimen accession. This means data is available to researchers in near real-time as accessioning happens compared to pre 2012 where data was only available as the operations center could sort and enter data. Table 1 is a list of projects that utilize the toolkit with summary counts of subjects, specimens and scientific data points as of the end of 2015.

Operational and scientific complexity

Our methods allow for a significant amount of complexity in scientific data management. With a modern, web-service oriented architecture we abstract the subject entity from multiple related project data in supporting research systems, providing increased flexibility and adaptability over comparable monolithic systems. The tools facilitate the longitudinal collection of clinical phenotypic data over an arbitrary time period. We also promote asynchronous and variable collection of specimens.

Protection of subject privacy

The architecture of the eHB helps to protect patient privacy by limiting exposure of patient identifiers to research staff. Identifiers are only available to study operations staff responsible for associating data and specimens with subjects in the context of specific IRB protocols. The toolkit shares implementation of security and access controls with the connected downstream research system (e.g., access and logging features of REDCap). Similarly, authorization to connect LIMS specimen records to a patient record are configured with LIMS named users, and the toolkit relies on the security protocols exposed through the LIMS API. Researchers only access secondary data resources that result from

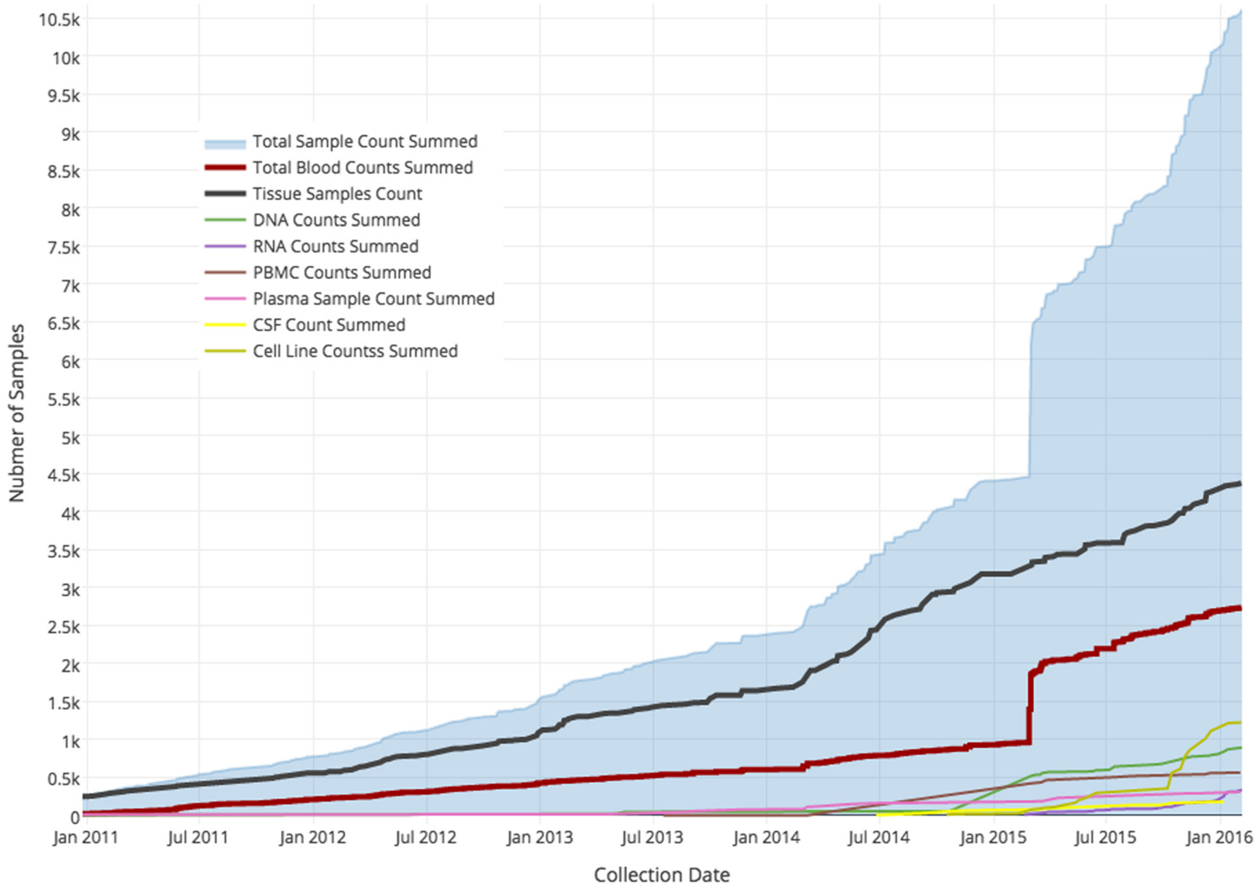


Fig. 6 Usage. A graph of specimen accessioning for the CBTTC which has accessioned over 10,000 specimens to date (end of January 2016). The blue shaded area represents the sum of all specimens collected by the project

Table 1 Project list this table is a list of select high-volume projects utilizing the modular tool-kit architecture described in this manuscript that particularly integrate multiple research resources through the toolkit

Project	Subjects (rounded)	Specimens accessioned (rounded)	Data integration points
The Children's Brain Tumor Tissue Consortium	1400	9200	-Case Report Forms -LIMS -Cancer Genomics
University of Pennsylvania-CHOP Neurosurgery Tissue Collaborative	1500	14,000	-Case Report Forms -LIMS
IBD Center Biorepository Studies	177	1100	-Case Report Forms -LIMS -Electronic Medical Record
Center for Childhood Cancer Research	470	2200	-Case Report Forms -LIMS -Electronic Medical Record
PennCHOP Microbiome Center	60	1500	-Case Report Forms -LIMS -Electronic Medical Record -Genomic Analysis Pipelines

The table includes the project, a rounded number of individual subjects consented to the study, specimens and the integrated data points for the project. These numbers are rounded and are from the end of year, 2015

downstream ETL and are never directly connected to the honest broker component of the tool kit. The ability to query the eHB allows the team to build domain specific data repositories and a web query tool limited risk of exposing patient information.

Evolving data types and workflows

Downstream data management systems and workflows can change project to project or as the science changes within a project. We allow all downstream applications do what they were designed to do without any impact on the underlying architecture of the toolkit. The toolkit is resilient to change, and the service-oriented architecture provides hooks to incorporate additional systems, reflecting the certainty of changing requirements, data and workflows.

Discussion

Annotation of biospecimens with longitudinal clinical data was the initial impetus of creating this toolkit, but it has gone much further. The BRP and REDCap integration have enabled multiple multi-institutional biorepository studies, both international and national. As in Harris et al., we enable concurrent multi-institutional projects to be controlled by staff through a single common interface similar to other new research informatics initiatives [30]. Exploration of HL7 for Oncology and honest brokering is ongoing, but we have not been able to implement these medical informatics type interchange languages for this purpose [47]. Additionally, web services were not built to be HIPAA compliant [9], but we find that designing modular tools that have privacy built in by their specific usage to be sufficient in research. Further, we find the modern web architecture to allow for the integration of the favored web tools scientists are using for their research to be a novel approach.

The Children's Brain Tumor Tissue Consortium

In Fig. 5, the data and specimen query tool of the Children's Brain and Tumor Tissue Consortium, is some of the results of ambitious biorepository collaboration between six children's hospitals. Informatics staff utilize this set of integrated tools to support the data and specimen-coordinating To date, the toolkit has been central to a number of grant-funded projects focused on next generation sequencing of biospecimens, and the identification of causative mutations in tumors and data integration heavy microbiologic research (See Table 1). A researcher can quickly determine study feasibility by reviewing available specimens and data in childhood brain cancer research and formulate new studies focused on finding pathways in rare cancer similar to Brastianos et al. and Parsons et al. [27, 28]. Current efforts are focused on integration of highly dimensional genomic data sets generated by such studies.

EMR integration

Integrating or retrieving data from the Electronic Medical Record (EMR) is a growing need for users of the toolkit. Projects originating at CHOP have been able to take advantage of the eHB to re-identify patients for the purpose of extracting clinical data from the CHOP EMR. We are able to, with appropriate IRB permission, obtain and de-identify clinical data directly from the EMR and incorporate these data as annotations on specimens and genomic data. This has proven useful in studies that require observational data such as medication at time of specimen collection, particularly in high frequency specimen collection in microbiologic research.

Genomic data integration

Continued success of biorepository-driven research relies on tight integration of other scientific data points. A primary example of this is the phenotypic and genomic information relation. In the case of the cancer tissue biorepository, the specific cancer genomic alterations are specifically important to annotate specimens. We found that allowing a researcher to query specific somatic mutations from comparative analysis of sequencing the germ-line and the tumor and find the physical specimens with that mutation to be a powerful feature. We accomplish the query of specific tumor mutations by integrating the CBioPortal into our toolkit. The CBioPortal an open source tool developed to view cancer-based genomic analysis data originally developed at Memorial Sloan Kettering Cancer Center (MSKCC) is an open access resource for exploration of multi-dimensional cancer genomics data sets [33]. Figure 7 shows how the integration of the CBioPortal with the clinical and specimen based query tool with a mix of traditional ETL and web based endpoints. Figure 8 is a set of screen shots of this cancer genomic data integration. This integration point opens up a capability unavailable in previous tissue focused biorepositories. Users can search for and view mutations called on specimens in the local biorepository data repository to pull a physical specimen for biological research. Conversely, the user can search for pathways of interest across published studies and discover if those pathways are shown in local specimens. The user can be moved back to the specimen request tool with the context of specimens from a genomic mutation-based query. Similarly as we continue to add scientific platforms to the scientific side of the toolkit pertinent to domains we are integrating with multiple commercial data management service providers. We will extend this concept by pushing the platform further to incorporate workflows dedicated to the classification of pathology diagnoses and collaborative discussion about pathology calls derived from high-resolution pathology slide scans. All of this would not be possible without a modular approach.

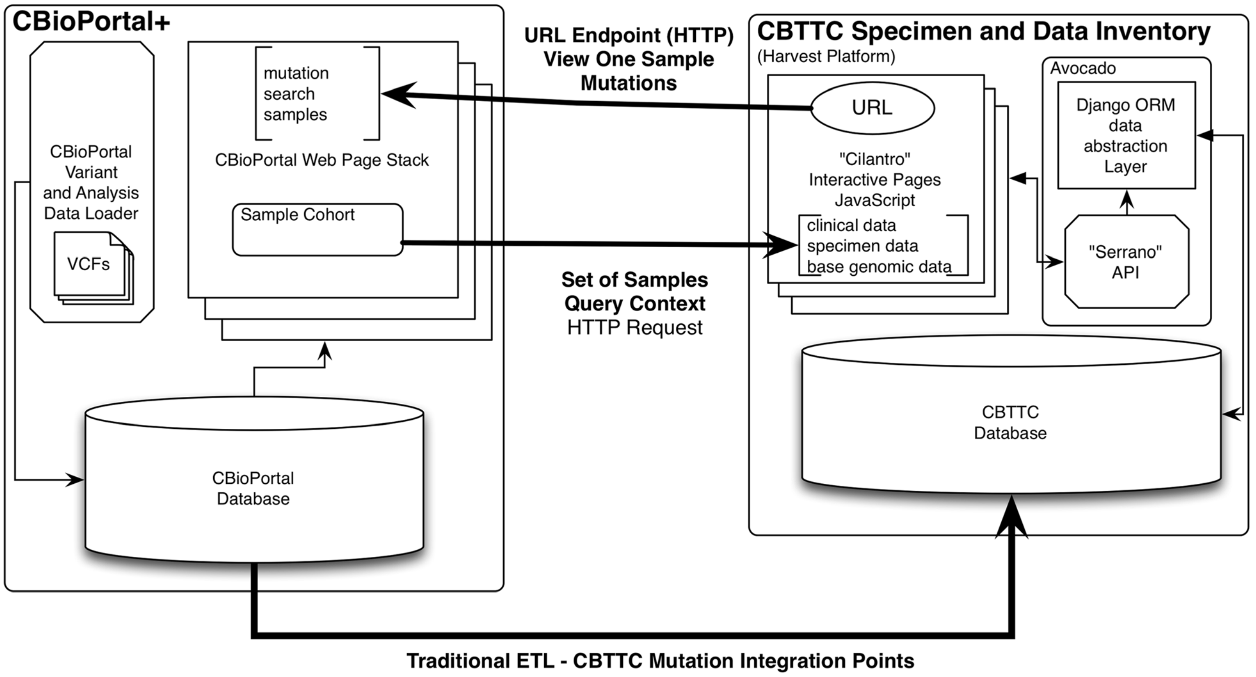
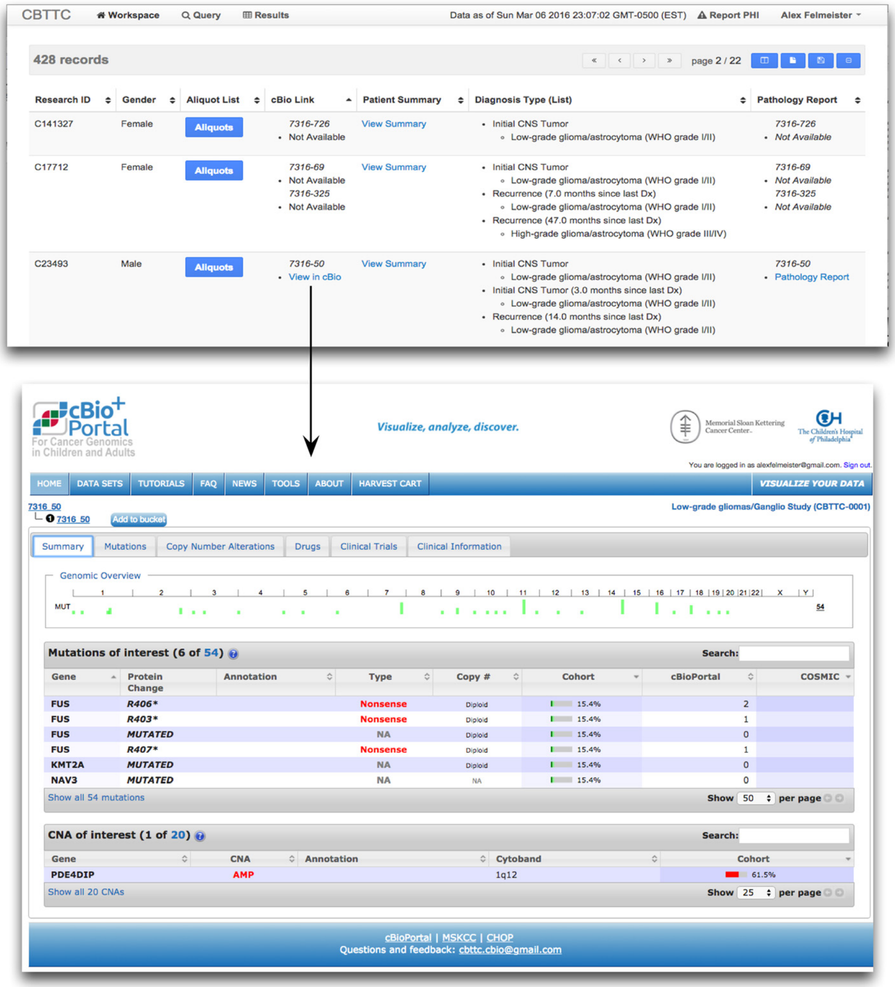

CBioPortal+
CBioPortal Variant and Analysis Data Loader
VCFs
mutation search samples
CBioPortal Web Page Stack
Sample Cohort
CBioPortal Database
URL Endpoint (HTTP) View One Sample Mutations
Set of Samples Query Context HTTP Request
CBTTC Specimen and Data Inventory
(Harvest Platform)
URL
"Cilantro" Interactive Pages JavaScript
clinical data specimen data base genomic data
Avocado
Django ORM data abstraction Layer
"Serrano" API
CBTTC Database
Traditional ETL - CBTTC Mutation Integration Points

Fig. 7 CbioPortal – harvest integration architecture. The integrated query tool allows for a back and forth search between genes of interest and visualization in CBioPortal, and the tool for phenotype and specimen requests. We perform this integration in a similar fashion to the other tools in the tool-kit. We utilize a combination of constructing web endpoints and traditional ETL. The cancer genomics integration starts with a scripted pull of mutation data via a secure database connection utilizing elements of CBioPortal's relational data structure to store this large set of data. Specimens known to the repository are loaded into the CBioPortal by the bioinformatics team with known specimen identifiers from the LIMS. This creates a natural link between any granular genomic data, the specimen and ultimately the subject. URLs are constructed in the query platform that allow for researchers to move from clinical and specimen driven queries directly to CBioPortal to visualize mutation data of interest



Data as of Sun Mar 06 2016 23:07:02 GMT-0500 (EST)

428 records page 2 / 22

Research ID	Gender	Aliquot List	cBio Link	Patient Summary	Diagnosis Type (List)	Pathology Report
C141327	Female	Aliquots	*7316-726* • Not Available	View Summary	• Initial CNS Tumor ∘ Low-grade glioma/astrocytoma (WHO grade I/II)	*7316-726* • *Not Available*
C17712	Female	Aliquots	*7316-69* • Not Available *7316-325* • Not Available	View Summary	• Initial CNS Tumor ∘ Low-grade glioma/astrocytoma (WHO grade I/II) • Recurrence (7.0 months since last Dx) ∘ Low-grade glioma/astrocytoma (WHO grade I/II) • Recurrence (47.0 months since last Dx) ∘ High-grade glioma/astrocytoma (WHO grade III/IV)	*7316-69* • *Not Available* *7316-325* • *Not Available*
C23493	Male	Aliquots	*7316-50* • View in cBio	View Summary	• Initial CNS Tumor ∘ Low-grade glioma/astrocytoma (WHO grade I/II) • Initial CNS Tumor (3.0 months since last Dx) ∘ Low-grade glioma/astrocytoma (WHO grade I/II) • Recurrence (14.0 months since last Dx) ∘ Low-grade glioma/astrocytoma (WHO grade I/II)	*7316-50* • Pathology Report

You are logged in as alexfelmeister@gmail.com. Sign out.

7316_50
└ ❶ 7316_50

Low-grade gliomas/Ganglio Study (CBTTC-0001)

Mutations of interest (6 of 54) Search:

Gene	Protein Change	Annotation	Type	Copy #	Cohort	cBioPortal	COSMIC
FUS	*R406**		Nonsense	Diploid	15.4%	2	
FUS	*R403**		Nonsense	Diploid	15.4%	1	
FUS	*MUTATED*		NA	Diploid	15.4%	0	
FUS	*R407**		Nonsense	Diploid	15.4%	1	
KMT2A	*MUTATED*		NA	Diploid	15.4%	0	
NAV3	*MUTATED*		NA	NA	15.4%	0	

Show all 54 mutations Show 50 per page

CNA of interest (1 of 20) Search:

Gene	CNA	Annotation	Cytoband	Cohort
PDE4DIP	AMP		1q12	61.5%

Show all 20 CNAs Show 25 per page

cBioPortal | MSKCC | CHOP
Questions and feedback: cbttc.cbio@gmail.com

Fig. 8 Specimen to cancer genomics. Screen shot of integration of the Harvest-based data and specimen query tool with a the CBioPortal for cancer genomic visualization of a specific case

Conclusions

As of January of 2016 there are 23 unique protocols/patient cohorts being managed in the Biorepository Portal (BRP). There are over 4000 unique subject records in the electronic honest broker (eHB), over 30,000 specimens accessioned and 8 institutions participating in various biobanking activities using this tool kit. We specifically set out to build rich annotation of biospecimens with longitudinal clinical data; BRP/REDCap integration for multi-institutional repositories; EMR integration; further annotated specimens with genomic data specific to a domain; build application hooks for experiments at the specimen level integrated with analytic software; while protecting privacy per the Office of Civil Rights (OCR) and HIPAA. To meet this challenge, we created an open source modular software toolkit that automates many manual components of biorepository data workflows while also protecting patient privacy. Conversely the modular solutions allow for novel integration points for translational research spanning clinical and genomic data. We believe this work advances the state of the art within the biomedical domain by moving towards modern technologies and architectures to provide translational research resources.

Availability of software

Project Name: The Biorepository Portal Toolkit
Project home page: http://www.brptoolkit.com
Code Repositories:

- Electronic Honest Broker Service: https://github.com/chop-dbhi/ehb-service
- Electronic Honest Broker Client: https://github.com/chop-dbhi/ehb-client
- Biorepository Portal: https://github.com/chop-dbhi/biorepo-portal
- Data Sources: https://github.com/chop-dbhi/ehb-datasources

Operating Systems: Linux

Other Requirements: Docker (optional but recommended)

License: We believe in open source software and open-source our work. The Biorepository Toolkit and all integration work with non-proprietary systems is licensed under BSD 2-clause License.

Restrictions to use by non-academics: None.

Ethics approval: No ethics approval was required for this work.

A demonstration of this software is available at http://www.brptoolkit.com. This website also contains documentation, webinars, descriptions and pointers to code repositories in context. Software discussed in this paper is available on the CHOP Department of Biomedical and Health Informatics github repository at https://github.com/chop-dbhi. A specific implementation of the toolkit is available through the Children's Brain Tumor Tissue Consortium at http://www.cbttc.org.

Declarations

The Children's Hospital of Philadelphia Clinical and Translational Science Award (CTSA) through the Clinical and Translational Research Center (CTRC), U54 KL2RR024132, funded this project and publication. The project is also made possible by the generous philanthropic effort of the Children's Brain Tumor Foundation (CBTF).

This article has been published as part of *BMC Genomics* Vol 17 Suppl 4 2016: Selected articles from the IEEE International Conference on Bioinformatics and Biomedicine 2015: genomics. The full contents of the supplement are available online at http://bmcgenomics.biomedcentral.com/articles/supplements/volume-17-supplement-4.

Authors' contributions

ASF was the main author of this manuscript and envisioned this research method along with AJM who is also a main contibutor to this manuscript, envisioned and developed large components of the software. TJR is the current lead developer for the project and a contributor to the manuscript. ACR is the scientific domain contributor, and JWP is a main contributor to design of the project and author of this manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics, consent to participate and consent to publish

Not applicable for this research.

Author details

[1]Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, 3401 Civic Center Blvd, Philadelphia, PA, USA. [2]College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA, USA. [3]Department of Neurosurgery, Perelman School of Medicine at the University of Pennsylvania, 3400 Civic Center Boulevard, Building 421, Philadelphia, PA, USA.

Published: 18 August 2016

References

1. Brisson a R, Matsui D, Rieder MJ, Fraser DD. Translational research in pediatrics: tissue sampling and biobanking. Pediatrics. 2012;129(1):153–62.
2. Colman E, Golden J, Roberts M, Egan A, Weaver J, Pharm D, Rosebraugh C. The path to personalized medicine. N Engl J Med. 2010;363(4):2012–4.
3. Hirtzlin I, Dubreuil C, Préaubert N, Duchier J, Jansen B, Simon J, Lobato De Faria P, Perez-Lezaun A, Visser B, Williams GD , Cambon-Thomsen A. An empirical survey on biobanking of human genetic material and data in six EU countries. Eur J Hum Genet. 2003;11(6):475–88.
4. Altekruse SF, Rosenfeld GE, Carrick DM, Pressman EJ, Schully SD, Mechanic LE, Cronin KA, Hernandez BY, Lynch CF, Cozen W, Khoury MJ, Penberthy LT. SEER cancer registry biospecimen research: yesterday and tomorrow. Cancer Epidemiol Biomarkers Prev. 2014;23(12):2681–7.
5. Compton C. Getting to personalized cancer medicine: taking out the garbage. Cancer. 2007;110(8):1641–3.
6. Végvári A, Welinder C, Lindberg H, Fehniger TE, Marko-Varga G. Biobank resources for future patient care: developments, principles and concepts. J Clin Bioinf. 2011;1(1):24.
7. Vaught J, Rogers J, Carolin T, Compton C. Biobankonomics: developing a sustainable business model approach for the formation of a human tissue biobank. J Natl Cancer Inst Monogr. 2011;2011(42):24–31.
8. Kamm L, Bogdanov D, Laur S, Vilo J. A new way to protect privacy in large-scale genome-wide association studies. Bioinformatics. 2013;29(7):886–93.
9. Wang X, Liu L, Fackenthal J, Cummings S, Olopade OI, Hope K, Silverstein JC, Olopade OI. Translational integrity and continuity: personalized biomedical data integration. J Biomed Inform. 2009;42(1):100–12.
10. Wang X, Olopade O, Foster I. Personalized biomedical data integration. Pers Biomed Data Integr Biomed Eng Trends Electron Commun Software Mr Anthony Laskovski. 2011;1:100–12.
11. Dhir R, Patel AA, Winters S, Bisceglia M, Swanson D, Aamodt R, et al. A multidisciplinary approach to honest broker services for tissue banks and clinical data: A pragmatic and practical model. Cancer. 2008;113:1705–15.
12. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) P.L. No. 104-191, 110 Stat. 1938 (1996).
13. The Office for Civil Rights (OCR), B. Malin. Guidance regarding methods for de-identification of protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Heal Inf Priv. 2012;1–32. http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.
14. Dhir R, Patel A a, Winters S, Bisceglia M, Swanson D, Aamodt R, Becich MJ. A multidisciplinary approach to honest broker services for tissue banks and clinical data: a pragmatic and practical model. Cancer. 2008;113(7):1705–15.
15. National Cancer Institute. NCI best practices for biospecimen resources. 2010.
16. Boyd AD, Saxman PR, Hunscher DA, Smith KA, Morris TD, Kaston M, Bayoff F, Rogers B, Hayes P, Rajeev N, Kline-Rogers E, Eagle K, Clauw D, Greden JF, Green LA, Athey BD. The University of Michigan Honest Broker: a Web-based service for clinical and translational research and practice. J Am Med Inform Assoc. 2009;16(6):784–91. doi:10.1197/jamia.M2985.
17. The Children's Brain Tumor Tissue Consortium. [Online]. Available: www.cbttc.org. Accessed 25 June 2015.
18. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372(9):2012–4.
19. Rainie L, Fox S, Duggan M. The Web at 25 in the U.S. Pew Res Cent. 2014;1:1–5.
20. S. Mukherjee. The emperor of all maladies: a biography of cancer. New York, NY: Scribner. 2011.
21. June CH, Maus MV, Plesa G, Johnson L a, Zhao Y, Levine BL, Grupp A, Porter DL. Engineered T cells for cancer therapy. Cancer Immunol Immunother. 2014;63(9):969–75.
22. Ostrom QT, Gittleman H, Chen Y, Wolinsky Y, Barnholtz-Sloan J. CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2005–2009. Neuro Oncol. 2012;14 suppl 5:1–21.
23. American Brain Tumor Association. Brain tumor information. 2014. [Online]. Available: http://www.abta.org/brain-tumor-information/types-of-tumors/. Accessed 25 June 2015.
24. Cancer in Children & Adolescents. 2014.
25. Baldwin RT, Preston-Martin S. Epidemiology of brain tumors in childhood - a review. Toxicol Appl Pharmacol. 2004;199(2):118–31.
26. Boklan JJ. Little patients, losing patience: pediatric cancer drug development. Mol Cancer Ther. 2006;5(8):1905–8.
27. Brastianos PK, Taylor-Weiner A, Manley PE, Jones RT, Dias-Santagata D, Thorner AR, Lawrence MS, Rodriguez FJ, Bernardo LA, Schubert L, Sunkavalli A, Shillingford N, Calicchio ML, Lidov W, Taha H, Martinez-Lage M, Santi M, Storm PB, Lee JYK, Palmer JN, Adappa ND, Scott RM, Dunn IF, Laws ER, Stewart C, Ligon KL, Hoang MP , Van Hummelen P, Hahn WC, Louis DN, Resnick AC, Kieran MW, Getz G, Santagata S. Exome sequencing identifies BRAF mutations in papillary craniopharyngiomas. Nat Genet. 2014;46(2):161–5.
28. Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin JC-H, Boca SM, Carter H, Samayoa J, Bettegowda C, Gallia GL, Jallo GI, Binder ZA, Nikolsky Y, Hartigan J, Smith DR, Gerhard DS, Fults DW, VandenBerg S, Berger MS,

Marie SKN, Shinjo SMO, Clara C, Phillips PC, Minturn JE, Biegel JA, Judkins AR, Resnick AC, Storm PB, Curran T, He Y, Rasheed BA, Friedman HS, Keir ST, McLendon R, Northcott PA, Taylor MD, Burger PC, Riggins GJ, Karchin R, Parmigiani G, Bigner DD, Yan H, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE. The genetic landscape of the childhood cancer medulloblastoma. Science. 2011;331:435–9.

29. Chin L, Hahn WC, Getz G, Meyerson M. Making sense of cancer genomic data. Genes & Development. 2011;25:534–555.
30. Harris P a, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42(2):377–81.
31. Chen PP-S. The entity-relationship model—toward a unified view of data. ACM Trans Database Syst. 1976;1(1):9–36.
32. Pennington JW, Ruth B, Italia MJ, Miller J, Wrazien S, Loutrel JG, Crenshaw EB, White PS. Harvest: an open platform for developing web-based biomedical data discovery and reporting applications. J Am Med Inform Assoc. 2013;21(2):379–83.
33. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4.
34. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Poznań, Poland). 2015;19(1A):A68–77.
35. Alkhatib H, Faraboschi P, Frachtenberg E, Kasahara H, Lange D, Laplante P, Merchant A, Burgess A. IEEE CS 2022 Report. 2014.
36. Goble C, Stevens R. State of the nation in data integration for bioinformatics. J Biomed Inform. 2008;41(5):687–93.
37. Goble C, Stevens R, Hull D, Wolstencroft K, Lopez R. Data curation + process curation=data integration + science. Brief Bioinform. 2008;9(6):506.
38. Boyd AD, Hosner C, Hunscher D a, Athey BD, Clauw DJ, Green LA. An 'Honest Broker' mechanism to maintain privacy for patient care and academic medical research. Int J Med Inform. 2006;76(5–6):407–11.
39. McConnell P, Dash RC, Chilukuri R, Pietrobon R, Johnson K, Annechiarico R, Cuticchia AJ. The cancer translational research informatics platform. BMC Med Inform Decis Mak. 2008;8:60.
40. Kho AN, Cashy JP, Jackson KL, Pah AR, Goel S, Rn Boehnke J, Humphries JE, Kominers SD, Hota BN, Sims SA, Malin BA, French DD, Walunas TL, Meltzer DO, Kaleba EO, Jones RC, Galanter WL. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. J Am Med Inform Assoc. 2015;22(5):1072–1080. doi:10.1093/jamia/ocv038.
41. Manion FJ, Robbins RJ, Weems W a, Crowley RS. Security and privacy requirements for a multi-institutional cancer research data grid: an interview-based study. BMC Med Inform Decis Mak. 2009;9:31.
42. Hutton JJ. Pediatric biomedical informatics, vol. 2. Dordrecht: Springer Netherlands; 2012.
43. Boyd AD, Hunscher DA, Kramer AJ, Hosner C, Saxman P, Athey BD, Greden JF, Clauw DC. The 'Honest Broker' method of integrating interdisciplinary research data. AMIA Annu Symp Proc. 2005;902.
44. Blobel B, Nordberg R, Davis JM, Pharow P. Modelling privilege management and access control. Int J Med Inform. 2006;75:597–623.
45. Ecma International. ECMA-404: The JSON data interchange format. 1st ed. 2013.
46. Morris R, Thompson K. Password security: a case history. Commun ACM. 1979;22(11):594–7.
47. Warner JL, Maddux SE, Hughes KS, Krauss JC, Yu PP, Shulman LN, Mayer DK, Hogarth M, Shafarman M, Fiscalini AS, Esserman L, Alschuler L, Koromia GA, Gonzaga Z, Ambinder EP. Development, implementation, and initial evaluation of a foundational open interoperability standard for oncology treatment planning and summarization. J Am Med Informatics Assoc. 2015; 22(3):577–86.