*data*

MDPI

*Article*

# DiiS: A Biomedical Data Access Framework for Aiding Data Driven Research Supporting FAIR Principles

**Priya Deshpande [1,*], Alexander Rasin [1], Jacob Furst [1], Daniela Raicu [1] and Sameer Antani [2]**

[1] College of Computing and Digital Media, DePaul University, Chicago, IL 60604, USA; arasin@depaul.edu (A.R.); jfurst@depaul.edu (J.F.); draicu@cdm.depaul.edu (D.R.)
[2] National Library of Medicine, Bethesda, MD 20894, USA; sameer.antani@nih.gov
* Correspondence: pdeshpa1@depaul.edu

Received: 12 March 2019; Accepted: 15 April 2019; Published: 20 April 2019


**Abstract:** Vast amounts of clinical and biomedical research data are produced daily. These data can help enable data driven healthcare through novel biomedical discoveries, improved diagnostics processes, epidemiology, and education. However, finding, and gaining access to these data and relevant metadata that are necessary to achieve these goals remains a challenge. Furthermore, data management and enabling widespread, albeit controlled, use poses a major challenge for data producers. These data sources are often geographically distributed, with diverse characteristics, and are controlled by a host of logistical and legal factors that require appropriate governance and access control guarantees. To overcome these obstacles, a set of guiding principles under the term FAIR has been previously introduced. The primary desirable dataset properties are thus that the data should be Findable, Accessible, Interoperable, and Reusable (FAIR). In this paper, we introduce and describe an abstract framework that models these ideal goals, and could be a step toward supporting data driven research. We also develop a system instantiated on our framework called the *Data integration and indexing System* (DiiS). The system provides an integration model for making healthcare data available on a global scale. Our research work describes the challenges inhibiting data producers, data stewards, and data brokers in achieving FAIR goals for sharing biomedical data. We attempt to address some of the key challenges through the proposed system. We evaluated our framework using the software architecture testing technique and also looked at how different challenges in data integration are addressed by our system. Our evaluation shows that the DiiS framework is a user friendly data integration system that would greatly contribute to biomedical research.

**Keywords:** data integration; data indexing; medical domain datasets; challenges in data integration; electronic health records; medical ontology; FAIR principles

## 1. Introduction

The growing amount of available biomedical data poses new challenges for data management. Data re-usability is a highly desirable goal, both for advancing science as well as replicating or validating results of previous studies. Recognizing this need, publishers as well as funding bodies may require researchers to submit data generated as a result of their work and make it available to the research community. The National Institutes of Health (NIH) is encouraging funded investigators to use cloud computing to conduct research and make their work accessible to larger audiences: "The Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) initiative establishes partnerships with commercial cloud service providers (CSPs) to reduce economic and technological barriers to accessing and computing on large biomedical datasets

to accelerate biomedical advances" [1]. However, in the healthcare domain, datasets are often not shared because of security concerns, lack of integration, or restrictions imposed by governance policy constraints. Several studies have highlighted the need for integration of healthcare data [2–4]. Nevertheless, most of the solutions implemented so far addressed limited aspects of the overall problem such as communication between Electronic Health Records (EHRs) systems across different hospitals. A comprehensive framework that supports global distributed access to healthcare data is needed to address the current limitations on data sharing. The solution needs to be easy to deploy with existing data repositories, cost effective [5], and guarantee requisite governance and access control for both data and system users. This task is made particularly difficult by the inherent heterogeneity of the biomedical data. Data variety and formats create the need for extensive data cleaning before it can be processed. According to a 2016 survey, data scientists across a wide array of fields spend most of their time (about 80 percent) collecting existing datasets and organizing data [4]. A data integration framework would help researchers and practitioners (who provide healthcare services) by making data available, accessible, and support different users by maintaining the quality of data shared from different resources [6]. It would also greatly reduce the need for curation of data sources and data repositories.

Raje et al. [7] discussed the need for integrating multiple data sources that would provide more useful information combined than individual data sources. In fact, repositories such as PubMed [8] and Dryad [9] already provide access to (relatively heterogeneous) medical publications. PubMed is the most prominent repository with collection of biomedical articles and research journals. Dryad is a curated general-purpose repository that makes data (journals) discoverable, freely usable, and citeable. However these repositories still do not fully incorporate datasets (e.g., clinical reports, images), which is desirable in the new realm of multimodal data driven healthcare research. National Cancer Institute data science portal [10] supports data sharing and data access for Genomics data. This system is great inspiration for our proposed system. However, this system is not automated to the level that we proposed in this framework. There are many in between steps that users need to perform to share or access the data.

Our proposed framework, Data integration and indexing System (DiiS), extends beyond merely a communication layer between different data systems. Our goal is to provide an overarching integration model for making healthcare data available on a global scale. DiiS model defines two categories of data donors—those able to share their datasets and those who prefer to provide an index summary of their datasets instead. The latter option would enable data owners to add their data to DiiS search index while retaining full control over access to that data. The DiiS framework proposed in this paper also supports a distributed learning environment across shared data. Distributed learning is a mechanism for training machine learning and artificial intelligence algorithms [11] where training data at different geographical locations is seamlessly used as a resource. Users can be permitted to execute their algorithms directly on the data sites instead of downloading datasets.

The DiiS framework is designed to support distributed clients with authentication and fine-grained access controls to enable data retrieval from heterogeneous and geographically distributed biomedical data sources. Its architecture builds on state-of-the-art research in multimodal, federated, and temporal data integration systems. This is a framework design that has not been implemented yet. Currently, we are focusing on the integration and evaluation of several public biomedical data sources, based on our previous work on IRIS [12]. IRIS is a preliminary work for a future integration system that we proposed in this framework. Our contributions presented in this paper are:

- An overview of significant data integration challenges in the biomedical domain.
- An extensive literature review and gap analysis of existing biomedical data integration systems.
- Design of a biomedical data integration and querying framework that supports FAIR principles.
- Identification and evaluation of diagnostically relevant public datasets for biomedical data integration.

The rest of the paper is organized as follows—in Section 2, we present background and relevant work in the domain of data integration and other research that motivated our approach. We also consider different challenges in data integration. In Section 2.3, we identify challenges in healthcare data integration. In Section 3 we discuss different diagnostically relevant biomedical data sources and users of our proposed system. In Section 4, we discuss the methodology and architecture of the proposed system. In Section 5, we discuss how our framework would facilitate healthcare data integration and how other data intensive research institutes can use our model for data-sharing and indexing. In Section 6, we summarize the steps for data integration and outline both our short- and long-term plans to extend this work.

## 2. Background

In this section, we focus on the research in the field of biomedical data integration and requirements thereof.

### *2.1. Motivation for Data Integration*

We first discuss papers that addressed data integration and studies that focused on providing fine-grained access control to data. Merelli et al. [2] discussed the importance of big data integration in medical bioinformatics. They surveyed the available big data handling techniques (e.g., Hadoop, cloud computing) and mechanisms for developers to implement data access and security by using science gateways and other methods such as cluster computing and GPU computing. However, the survey did not propose a system to implement data integration focusing on a review of relevant technologies. Similarly, Holzinger et al. [13] talked about knowledge discovery and interactive data mining techniques in bio-informatics, the challenges to integrating biomedical data, and open research directions. They have argued that life sciences, biomedicine, and healthcare domains are facing challenges with increased volume and variety of data, and for the increased need for integrative analysis and modeling. Integrative analysis is the simultaneous analysis of multiple datasets, which can be used to extract knowledge from different datasets.

For healthcare records we plan to automate data integration in some aspects such as data cleaning and pre-partition the data based on the type (e.g., structured and unstructured). Providing fine-grained access control over integrated data is also a major challenge. Clinical data often contains sensitive patient information (protected by laws such as HIPAA) so authentication and access control rules are necessary to prevent users accessing data beyond their privilege level. Trifan et al. [14] proposed a system to provide fine-grained clinical data access. They proposed a solution with multiple levels of data visibility, combined with a fine-grained access control over the shared data. Data custodians can provide the authentication to users based on their privileges to access data. The described system included a well defined range of different authentications layers, however, it does not explicitly describe the approach for implementation of these access methods.

Sujansky et al. [15] also describe a method for fine-grained access to personal health records using a relational database system. The proposed system determines with a single SQL query whether a user who accesses patient data from a specific application is authorized to perform the desired operation or not. However, they have not discussed how they support a large number of users combined with many different authentication layers.

### *2.2. Related Work*

In this section, we provide an overview of research groups who took initiative to integrate medical data sources (most have focused on integration of medical articles). McQuilton et al. proposed a system called BioSharing [3], a portal that follows findable and accessible (FAIR) principles. It is a resource for publishers and researchers in the domain of biological sciences that provides integrated access to data sources from biological, environmental, and biomedical sciences.

DataMed [16,17] is an open source system that facilitates discovery and indexing of biomedical datasets. In this system, users can access an integrated data using unified schema (called Data Tag Suite) and search for biomedical articles using a proposed search engine. The DataMed system [18] integrates text data, which may exist in different formats such as XML, CSV, and JSON. However, DataMed does not integrate image datasets and clinical reports, focusing primarily on integration of medical articles and journals.

Wang et al. [19] proposed SciPort, which is a web based collaborative biomedical data-sharing platform to support data sharing across distributed organizations. SciPort provides a mechanism that can integrate different biomedical data sources using a metadata model for data sharing from different organizations.

Trifan et al. [20] describe a web platform for communication between data custodians and biomedical researchers. They provide an integrated system with FAIR principle, so data owners can share their data and users can access the data without performing multiple iterations of security verification. Trifan's framework does not provide implementation details, so it is difficult to evaluate the advantages and limitations of their proposed system. However, we would refer to their system principles and capabilities as a reference benchmark for our proposed framework.

Krumholz [21] discussed the importance of data integration and tools needed for machine learning techniques in healthcare systems. The work focused on how clinical big data can be used by clinicians and researchers and how they can use machine learning techniques to analyze clinical data.

Dey et al. [22] describe the data sharing initiatives and effects of data sharing in cardiology research. They further framed the challenges of data sharing in terms of cost of data integration and who bears these costs. Kansagra et al. [23] presented review of big data in healthcare and use of big data in future radiology informatics. They describe an overview of the big data adoption cycle and how academic radiology departments can use clinical big data for research purposes. We plan to follow systematic data collection methods and standards to maintain transparency in data sharing. Before collecting data we will verify data release information and the associated publications. Deist [11] developed an IT infrastructure euroCAT in five radiation clinics across three countries and provided infrastructure and distributed learning methodology for lung cancer patient data. Angraal et al. [24] discussed data sharing effects in clinical trials and described how the Digitalis Investigation Group trial assessed the effects of data sharing that can be used for further analysis in clinical research. Deshpande et al. developed IRIS [25] Integrated Radiological Image Search (IRIS), which is an integrated radiology teaching file repository that incorporates teaching files from MIRC [26] and MyPacs [27]. IRIS engine is augmented with Radiology Lexicon (RadLex) [28] and Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) [29] ontologies, and interprets several types of natural language queries such as negation statements and term synonyms. RSNA MIRC [26] and MyPacs [27] are well-known public radiology teaching file data sources with huge amount of clinical cases with associated images (discussed in detail in Section 3.1). Deshpande et al. [12,30] describe how integration of medical ontologies helped find new relevant search results within the same dataset. IRIS engine enables a text based, image based, and hybrid query search. IRIS motivates our current proposed methodology and can be considered a preliminary prototype for a biomedical data integration framework.

Open-i® is an experimental Open Access Biomedical Image Search Engine and is a service of the National Library of Medicine that enables the search and retrieval of abstracts and images (including charts, graphs, clinical images, etc.) from the open source literature, and biomedical image collections [31]. Open-i supports text and image based search; however, this system does not support integration of heterogeneous biomedical data sources (e.g., EHRs, teaching files, genome data) and does not support fine-grained access control for users.

FAIR principles are described by Wilkinson et al. and echoed in NIH's Strategic Plan for Data Science [32,33] with the goal of guiding reuse of research data

1. To be *Findable*, data must have unique identifiers, effectively labeling it within searchable resources.
2. To be *Accessible*, data must be easily retrievable via open systems and be subject to effective and secure authentication and authorization procedures.
3. To be *Interoperable*, data should use a formal, accessible, shared, and broadly applicable language for knowledge representation enabling communication between different systems for their meaningful use.
4. To be *Reusable*, data must be adequately described to a new user, have clear information about data-usage licenses, and be associated with their provenance.

We have designed a framework that addresses challenges in biomedical data sharing and supports FAIR principles for fine-grained access to biomedical datasets. Our literature survey shows that although data integration is an important requirement in healthcare domain, very few research groups developed generalized data integration approaches in healthcare domain. To implement an effective integrated data repository, there are many challenges (discussed in Section 2.3) that need to be addressed and a few compromises (e.g., usage policies, cost) that need to be made.

Data integration in healthcare research continues to gain importance as organizations seek to use their data for analytics. Healthcare data is a diverse collection of clinical reports and associated images, EHRs, radiology images, radiology reports, annotations, lab results, and lab images. Usually hospitals own data, however, biomedical research institutes such as National Institutes of Health (NIH) also have large collections of healthcare data, which are shared by different hospitals. However, due to the complexity of integration, security concerns, and governance policy requirements, the data from institutions willing to share it is still not available to researchers.

Health Level Seven International (HL7) [34] provides a framework and standards for the information retrieval between different EHRs and provides an integrated solution across different medical institutions. Further, other data types such as pathology data are even more difficult to convert into structured format compared to relatively structured EHRs. Using data adapters one may combine data from different sources using unified logical schema design, that could be used to gain intelligence and improve analysis process efficiency in clinical research. In practice, it may be necessary to keep structured (e.g., clinical reports) and unstructured (e.g., pathology reports) data separate. We are therefore proposing a hybrid database system with a combination of relational and non-relational storage.

Standards aimed at health data integration and unification have been adopted in many countries. For example, "in the USA the Health Information Technology (HITECH) Act provides incentive payments to health care providers who adopt certified EHR technology and demonstrate a meaningful use of that technology" [35]. Laws such as HIPAA establish the requirements for healthcare data protection; HL7 standard exists to facilitate data communication between different software used by healthcare providers. Data integration has multiple areas of application including patient record management systems, healthcare policies and standards information, EHRs and many more.

O'Dowd [36] discussed specific scenarios where data integration is needed and factors to consider when integrating healthcare data. In business intelligence and data warehouse systems—integrated data is made available to users. One of the important aspects of data integration is data sharing between different medical institutes and hospitals, along with information retrieval techniques. It requires a solution that includes the use of universal identifiers—such as Uniform Resource Identifiers (URIs) and Universally Unique Identifiers (UUIDs)—and mechanisms that would facilitate retrieval of ranked results, domain specific knowledge, and indexing techniques. When transforming data for integration, one also needs to consider data consistency issues (after transformation data should be the same as before transformation), error-free data migration, communication between different data service providers, and different formats. Storage requirements (e.g., space, access readiness) is another challenge that may demand flexible data integration (e.g., migrating data which is historically

important or data which is required for current operations). Based on a literature survey and our study of the biomedical data domain we performed SWOT analysis of our proposed system:

- Strength: Data integration capability would permit users to share and to access biomedical data on a global scale.
- Weakness: Not a fully automated system—some framework features require manual human intervention. Standardized governance policies need to be defined and reconciled across different stakeholders.
- Opportunity: Broad collaboration across different research groups and institutions.
- Threats: Potential misconduct of donors who are sharing data (e.g., sharing data without verifying data correctness or obtaining necessary patient consent).

DiiS would provide data sharing across different research institutions (see Section 4). However, it will depend on data governance policies that need to be defined by relevant authorities (e.g., government health organizations). A practical limitation of our design is that we ask data donors to sign an agreement about accuracy of the data they are sharing. However, we do not validate the quality of shared data. We will also cross check sample datasets de-identification (use pattern matching to detect possible personal data and validate a sample of data items with human input).

### *2.3. Challenges in Data Integration for Healthcare Domain*

In this section we describe some of the major challenges in biomedical data integration process. The NIH strategic plan for data science research work specifically highlights the challenges in data science and data integration [33].

1. The growing costs of managing data would diminish an organization's ability to enable scientists to generate data for understanding biology and improving health.
2. The current data-resource ecosystem tends to be "siloed" and is not optimally integrated or interconnected.
3. Important datasets exist in many different formats and are often not easily shared, findable, or interoperable.
4. Historically, funding approaches designed for research projects have not required release of research data.
5. Funding for tool development and data resources has become entangled, making it difficult to assess the utility of each independently and to optimize outcome value and funding efficiency.
6. There is currently no general system to transform, or harden, innovative algorithms and tools created by academic scientists into enterprise-ready resources that meet industry standards of ease of use and efficiency of operation

Coordination between research groups working on medical data and sites that generate that data is a challenge, where one should update data periodically while also complying with relevant security policies. Another challenge is the lack of protocols that would provide knowledge about accessibility of data for biomedical data sharing. Our survey indicated that there is still a pronounced need of well defined data sharing protocols in the medical community. Data access policies (e.g., only doctors can access patient history details) also require extensive fine grained access controls, where authorization is provided based on user privileges.

Data storage issues are also a significant challenge—storage capacity, access, availability, security, performance, retrieval capabilities, and scalability issues need to be addressed to facilitate data gathering from distributed sites. Management of different collaborators and peer (e.g., National Institute of Health & Family Welfare, India; NIH, USA; MD Anderson, National Healthcare Group USA) relationships is an important aspect of data integration. Data sources could contain images—reports, EHR lab reports (blood/clinical), pharmacy, statistical analysis, articles/documents (PubMed),

and radiology teaching files; to convert this variety of data, custom data adapters are needed. Data overlap, purpose, scope, definition (metadata), types, and sources could all be managed efficiently using the proposed data integration framework.

### *2.4. Data Distribution Across Different Locations*

Healthcare data is distributed across different geographical locations; the list includes different source systems with EHRs—different departments, radiology, pathology labs, or pharmacies. Healthcare data is also stored in different formats (e.g., on-paper, text, numeric, images, videos, other multimedia). Radiology uses images, while old medical records exist in paper format as well. One patient EHR entry can hold hundreds of rows with textual and numerical data. In some cases, the same data exists in different systems and is kept in different formats (e.g., insurance claims data versus clinical data). Collecting data at one place and integrating it into unified format remains a big challenge in health data integration.

### *2.5. Rate of Data Generation*

Another current challenge in data integration is that hospitals are continuously generating data at a much faster rate than their ability to analyze and use that data. Collection and organization of data is thus a necessary feature for a viable data integration solution.

### *2.6. Complex Systems and Lack of Structure in Data*

Each hospital uses their own systems to record patient information, generate reports, and store data in different formats, making data integration inherently more complex. Moreover, hospitals still use paper format to record patient data. It is difficult to transform complex and heterogeneous data into an easy-to-use format that can be used by researchers. Due to the current integration challenges, almost 44% of healthcare organizations are unable to use all of the available data [37].

Kho et al. [38] discussed challenges to integration of genomic data into electronic health records. Genomic datasets contain complex test results, large datasets, and different standards as compared to clinical data. Storage format differences and data volume make these data hard to integrate. From Kho's perspective, genomic data integration can be best achieved through a redesign of existing electronic health record systems.

### *2.7. Process-Oriented Challenges*

The efficacy of data integration is also affected by the integration between teams who handle the data. In practice, organizations follow different work processes and methods for data collection and processing. Thus, the team responsible for data collection is typically different from the team using that data. There is a need to help bridge the gap between different teams through a shared data repository available across the health care domain [37].

### *2.8. Privacy Preservation in Data Integration*

Privacy preservation in data collection and integration is particularly important when working with healthcare data. Christen et al. [39] discuss challenges in achieving scalable privacy preservation and improving accuracy of the data integration process. Clifton et al. [40] discuss privacy preservation challenges of data integration and proposed techniques to address these challenges. For example, creating different privacy views and policies that would restrict users from unauthorized access to the data repository. Based on those policies, views are created (the admin creates views) that will provide access to integrated data. The authors also propose a technique to identify similarity between different datasets without disclosing dataset record identity and anonymization techniques that can be applied to data before it is integrated. DiiS incorporates two privacy preservation modules. The offline module requires data donors to provide de-identified data (e.g., removing patient's personal information) and holds donors responsible for the accuracy of the shared data. In applying de-identifying techniques

for ensuring patient privacy, the accuracy of data matters because the anonymization process could change its meaning (e.g., changing patient visit dates results in incorrect duration between two visits). The online module controls privacy through authentication, providing access to only authorized users based on access policies. Users will be provided access based on their privileges and roles (e.g., a doctor can access patient personal history data – only if that particular doctor is treating that patient).

### *2.9. The Need for Skilled Professionals*

In the medical domain, different databases are maintained within an organization. For example, EHRs will have their own schema and different data structure for patient data and for disease registries.

Thus there is a need for skilled professionals who can understand and interpret data correctly in order to successfully implement an integrated solution for heterogeneous biomedical data. Gomez-Cabrero et al. [41] discuss the challenges in data integration in the era of omics datasets (genomics, proteomics or metabolomics data). Data integration in this domain is a challenge because understanding the structure of data and prior knowledge of data is required while collecting, processing, and integrating these datasets.

### *2.10. Data Governance Policies*

When integrating biomedical data sources, one has to have an understanding of the laws pertaining to data confidentiality. Data privacy concerns are one of the main reasons why few organizations are willing to share their data for research purposes. There is a need to change this culture, so that hospitals and research institutes can take the initiative towards data sharing.

We need to establish a fine-grained security access control over integrated patient data. In the healthcare domain, governance policies and rules vary based on the geographical location of the hospitals and research institutes (rules further vary from country to country). Data integration may require data sharing between geographically distributed institutes with different governance policies guiding data access rules. Government institutions should take the initiative to form global governance policies that will support distributed data integration.

### *2.11. Data Interoperability*

Data interoperability deals with gathering patient data, applying security constraints (e.g., de-identifying patients data) and maintaining quality of data subject to defined standards (e.g., coding standards and term definitions provided by medical ontologies such as Systematized Nomenclature of Medicine—Clinical Terms or Radiology Lexicon). It can be conceptually divided into three levels of complexity [42]:

1. Foundational interoperability is about the communication between different service providers who exchange data across different organizations.
2. Structural interoperability deals with the formats of data sharing used to exchange data between different hospitals. For example, HL7—"Structural interoperability in healthcare, to some extent, is provided by the Health Level 7 (HL7) series of standards, which provide guidelines on how messages should be structured" [34].
3. Semantic interoperability pertains to exchanging data across different hospitals without changing the meaning of the data or the definition of the terms. Different standards such as UMLS [43] code concepts can be used to identify terms and map those terms to the concept ID, that would provide uniform meaning across different institutes or hospitals. For example, radiologists can use same terms (using the concept id) while writing clinical report about particular disease. In ACR index (American College of Radiology - Index for Radiological Diagnosis) – neoplasms and neoplastic-like conditions are represented by 3; primary malignant neoplasms are represented by 32; carcinoma-type neoplasms are represented by 321 [44].

Masseroli et al. [45] discussed the design of data integration solutions. They presented a survey of early community efforts and outlined different challenges of the data integration system. In our approach, we have considered existing and prior systems as a reference for our proposed framework and intend to compare our approach to them as we develop DiiS.

## 3. DiiS Environment

In this section we review the data sources that can be integrated into a biomedical data warehouse and the users of the proposed system. Huesch [46] wrote that 30% of the entire world's stored data is generated by the healthcare domain and described how integration of this data would help the healthcare industry. Currently, we aim to integrate public data sources and in this section we discuss the heterogeneity of these data sources and the difficulties associated with integrating them in a uniform format.

### *3.1. Biomedical Data Sources and Data Integration*

The goal of a data integration system is to make data sources available in a uniform and supported format to users while applying data governance policies, maintaining data standards, and presenting data in a meaningful and user-friendly way. It is also important to understand data properties and requirements to enable seamless integration. We primarily focus on three types of data (a) Radiology teaching files or teaching files used by doctors and radiologists; (b) Electronic health records; (c) Research datasets. We next discuss each domain in detail:

(a) Medical Teaching Files

A radiology teaching files system is a collection of important cases for teaching and clinical follow-up, and references to diagnose a variety of a diseases. Teaching files share a similar overall structure but significant variations exist even within the same data sources. Teaching files can include information such as patient history, findings, diagnosis, differential diagnosis, discussion, comments, references, and images related to clinical reports. There are many public online teaching file data sources available such as Radiology Society of North America Medical Imaging Resource Community (RSNA MIRC) [26], MyPacs [27], EURORAD [47].

- RSNA MIRC

  Radiology Society of North America Medical Imaging Resource Community is a large repository with more than 2500 teaching files and more than 12,000 images. Radiological terms are highlighted and linked to RadLex browser [28] (RadLex is a Radiology Lexicon that provides definitions and synonyms information about radiology terms).

- MyPacs.net [27]

  Publicly available teaching file resource, where radiologists can create, modify and upload teaching files. More than 35,000 cases are available with 200,000 images.

- EURORAD [47]

  European Society of Radiology is a peer-reviewed educational tool based on teaching cases. There are more than 7000 teaching cases.

Despite similar properties, these public data sources are often incompatible, with differences in structure and data representation. A sample teaching file from both MIRC and MyPacs is shown in Figure 1, illustrating the heterogeneity of these data sources.

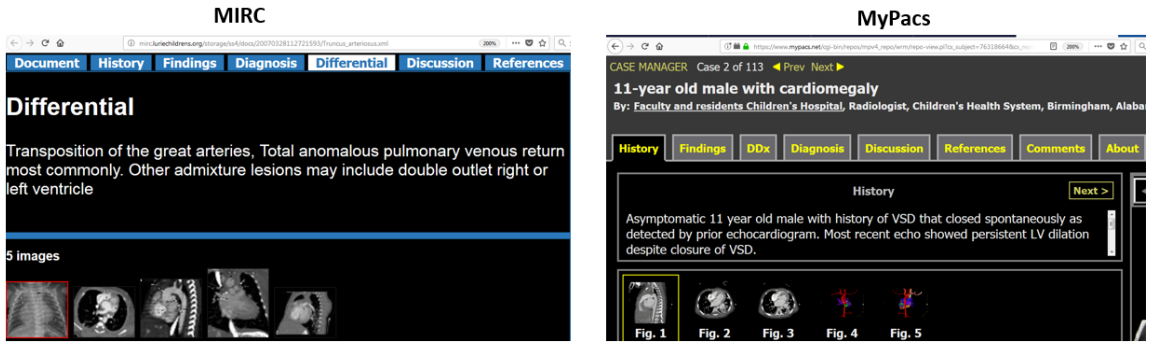


**Figure 1.** Sample teaching file from MIRC [26] and MyPacs [27].

Our project IRIS [12] is an initiative focused specifically on this type of data and developing a multimodal (text, image and hybrid text+image) search engine. IRIS is a prototype that presents preliminary results towards medical data integration. IRIS integrated public teaching file data sources (MIRC and MyPacs) along with two medical ontologies RadLex and SNOMED CT. The ontology integration provides us with a vocabulary which can be used for query expansion to facilitate searches for terms and term synonyms. The IRIS engine is tailored to interpret several aspects of natural language including negation statements, synonym terms, adjectives, and different sources of text. In addition, the IRIS search engine is designed with an integrated module for image-based search to allow finding of visually similar cases. IRIS also facilitates hybrid search query support, in which users can search with both text and image queries. Our hybrid query search technique will provide more accurate and relevant results to the user.

(b) Electronic Health Records (EHRs)

An electronic health record is a digital version of a patient's record. EHRs are maintained at different hospitals and store patient information such as history, medical test results, allergies, immunization details, radiology images, clinical reports, or treating doctor information. EHRs are shared within the hospital as necessary; however, it is unusual for an EHR system to share information across different hospitals. Sharing EHRs while enforcing appropriate data governance policies would enable doctors and researchers to have access to a wider variety of cases, expediting healthcare research.

(c) Research datasets

From our survey of different research institute datasets, we determined that most of the data in healthcare domain are images (CT, X-ray, MRI, and other modalities) [48,49]. Those images are most commonly stored in JPEG, DICOM, or PNG format; most of the time images have associated text data, typically of patient data such as age, date of birth, gender, diagnosis, findings, and case status (normal/abnormal). There are many diagnostically relevant public image datasets that can potentially be integrated in a shared repository. For example, Center for In Vivo Microscopy (CIVM) dataset [50], NIH chest x-ray dataset [48], Neuroimaging data by OpenNeuro [51]. Note that not all datasets have text associated with image data, and not all images have metadata associated with it.

### *3.2. Different Types of Users*

In this section we discuss the different types of biomedical data clients/users. When integrating biomedical data we also consider different types of users who will interact with the system. We primarily consider users who are researchers or doctors, but our system would also support other users interested in biomedical data. We also consider a category of users who want to share their data, referred to as *data donors*. We assume two categories of donors: donors who offer to fully share their

data and donors willing to only index their data (while data will remain at donor's location). (1) Full donors of data: Many institutes would like to share their data for research and development in the healthcare domain. However, seamlessly integrating these data sources into single repository is a challenge as discussed in Section 2.3. In our proposed architecture, data access adapters would support integration of data sources with a variety of data in different files and formats. (2) Users ready to share an index of their data repository: Research institutes, hospitals, or any data sources who do not want to donate their data but are interested in making their data findable will be supported via remote indexing. Our system will index external data based on Uniform Resource Identifiers (URIs) and Universally Unique Identifiers (UUIDs). Our hierarchical data storage system (discussed in Section 4.2) would allow us to efficiently index data across distributed locations. DiiS would also incorporate a facility to permit users to run their analytic code without sharing the actual data with them.

## 4. Methodology

In this section we discuss our research methodology for the proposed integration framework. To develop this architecture, we first performed an extensive literature survey that identified the need for biomedical data integration, existing systems that address integration problems, and limitations of those systems. We also enumerated the challenges of data integration and, using publicly available datasets, we performed a case study analysis that shows the benefits and applications of an integrated database for biomedical datasets. In this section, we show how individual modules of our framework contribute towards our goal of ensuring FAIR principles. We have designed a layered architecture because our system will be built on the core database component [52]. This framework would enable different institutes and stakeholders (e.g., programmers, data governance policy writers, data donors) to work independently, and DiiS layered architecture would facilitate this functionality over the different layers. The framework also relies on human participation in some steps of data integration. For example, we will be asking donors to sign an agreement before sharing data to confirm the accuracy and de-identification of the data.

After users share their data, our integration module will classify that data into different categories. For example, clinical reports might be partitioned into reports consisting primarily of images and reports that have few images and significant amount of text. Such classification would improve DiiS capability to search and retrieve data by adapting the search algorithm to each different category. To improve classification of data into categories we will need human validation; e.g., to validate the accuracy of our classification module, or to set a threshold for similarity between different datasets. There are several aspects of our system that cannot be fully automated such as data classification techniques (where we need domain experts to set categorization parameters), data governance policies, and data authorization rules from different research institutes.

### *4.1. Architecture of Proposed Data Integration Model*

As shown in Figure 2, we designed our system using a layered architecture; we envision a loose coupling between each individual layer of DiiS architecture. Foundation layer, on which the whole system would be built is the operating system layer. We prefer an open source operating systems such as Linux—however, one can also use other operating systems. As we plan to maintain replicas of data in a distributed environment, we prefer to use Hadoop Distributed File System (HDFS) for storage. The next layer facilitates data storage in different files and formats. Here, we are considering two types of data—the in-house data and shared donor's data that also resides on the DiiS servers. Data adapter layer is responsible for converting donor's data and mapping it to correspond to in-house data (e.g., categorize clinical reports and images based on anatomical structure or diseases) and providing data to the next layer. Once we have gathered all data in one shared format, it will be loaded into database layer which will provide results for query searches. We would be maintaining two versions of the data: a global collection of all datasets (a backup copy of the original) and a second copy which is maintained and processed at a database level (categorized based on data source type, e.g.,

EHRs or teaching files). The database would be implemented as a hybrid of a relational and NoSQL database system. The database would maintain Interoperable—FAIR principle, abstracting the internal details of how the data is stored when sharing it with the users. The next system layer is a search engine such as Solr, Elasticsearch, or Lucene where all data retrieval algorithms and Application Programming Interfaces (APIs) will reside. The search engine layer will rank results based on user queries, incorporating addition data repositories (e.g., medical ontologies) into the search algorithm. We intend to support text-based, image-based, and hybrid (text+image) search queries at that layer. The search engine layer would use Findable–FAIR principle and attach unique identifiers to data, making it search-able across the DiiS repository. While incorporating new data from donors, the master node (user authentication module) would apply rules specified by donors of the data (e.g., which user can access which data) and verify user authorization privileges. All inbound and outbound traffic will be overseen by the master node. Authentication layer will use Accessible–FAIR principle by providing data to users in a secure fashion. The subsequent layer is a search interface; an ability to introduce a user friendly and domain-aware interface is one of our design goals in DiiS. We would be sharing dataset information with users (the owner of the dataset or data sources information) which would deliver on Reusable—FAIR principle, so users can understand and reuse datasets.

This architecture would be composed with Service Oriented Architecture (SOA), and provide the necessary abstraction based on user service requirements. For example, if a data source provider wants to provide a paid service (i.e., access to dataset associated with fees), they could use a built-in payment service module.

In order to support data integration and search, DiiS framework will need to incorporate a wide range of features [45]: this includes context-aware query interpretation (using NLP module and ontologies), collaborative user interaction (e.g., incorporate user feedback and annotations), and efficient information retrieval mechanisms (e.g., maintaining a distributed cache of search results). In order to develop these features, we will be focusing on creation of custom data formats, integration of domain specific ontologies, and support for efficient data indexing and adaptive relevance-based ranking algorithms.

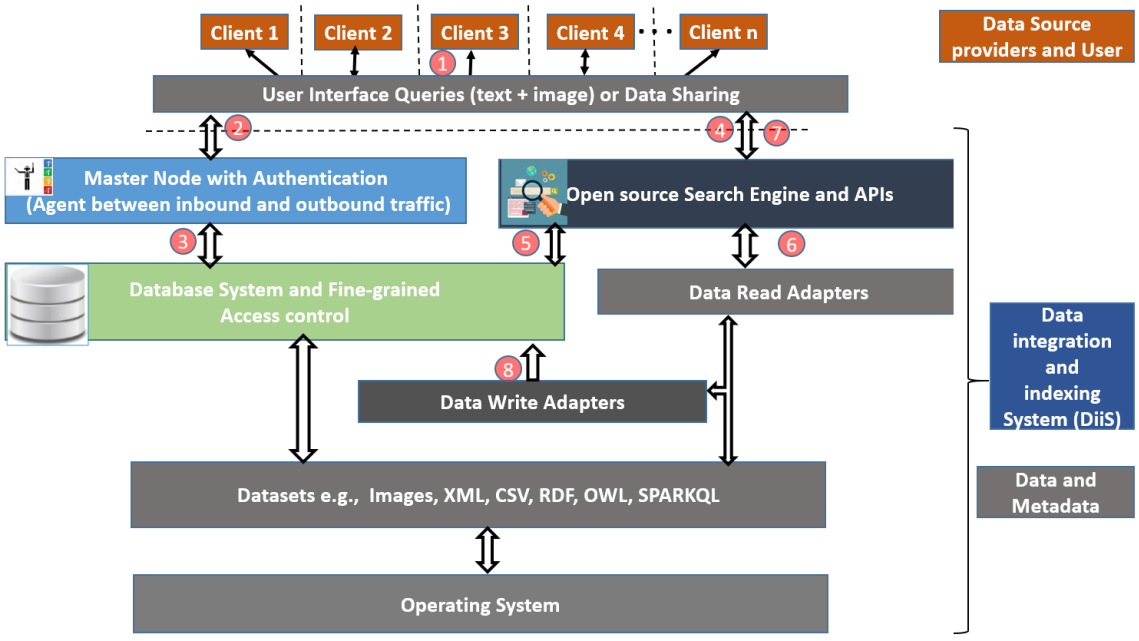


**Figure 2.** Architecture of proposed system.

As shown in Figure 2, we have different clients who can either search for data or are willing to donate and index their data. A user can input their query at the user interface ① and perform a text-based, image-based, or hybrid text+image search. Before interacting with the system, users need

to go through a security check validating their credentials at the master node ② by referencing the database system ③ (where user credentials are stored). At this layer our fine-grained access module would provide user authorization based on their privileges. We anticipate having a large number of geographically distributed users, so our framework will partition user groups based on their role in the system. We will be implementing a layered authentication system, where each layer will handle a different groups of users. For example, although doctors may have access to personal details of the patients—this access will also depend on their current role (i.e., whether doctor is currently treating that patient or not). If the doctor is treating the patient they will get full access to patient data; otherwise they may only access current patient diagnostic information but not other data about the patient. When data donors change access policies, we will need to update internal rules guiding access to those data sources. For this kind of change in requirements, we will be maintaining data source information and policies associated with those data items.

Once a user is authenticated, they do not need to provide credentials for each operation. We will be using Single Sign-On Kerberos algorithm, a user authentication service that permits one to use a single set of login credentials to access multiple applications. Alternative implementations of security layers could be deployed at the master node if desired. Once a user is authenticated, user query is forwarded to the search engine ④, where using information retrieval algorithms our system will identify a set of search results. The search engine will interact with the database system ⑤ and get the current location of the data ⑥. Data adapters are responsible for forwarding data to the search engine; search engine ranks the results based on the relevance to the query and forwards those results to the user ⑦. The same steps ①-③ would be performed when a user wants to share their data (either fully shared or through index-only mechanism, as discussed in Section 3.1). At step ④, a user can share their data using data write adapters ⑧. Data write adapters will be responsible for data cleaning, conversion, storage, and indexing of the stored data.

We did not test the proposed system architecture yet because we anticipate making changes based on the organizational needs (e.g., data governance policies might require introducing platform-as-a-service for computational analysis of data). However, we used the Chemical Abstract Machine Model (CHAM) architectural test plan proposed by Richardson et al. [53] to evaluate our architecture. Based on CHAM evaluation, we confirmed that all data elements, all processing elements, and all proposed internal modules are communicating with each other (as discussed in Section 4). As we will be using HDFS, we can scale processing capacity (up or down) as per current needs of the system. To improve system scalability, our framework would use a dedicated master node for user authentication, while outsourcing processing and storage to geographically distributed worker nodes. Similarly, the system will maintain replica data repositories, which would make our system resilient to hardware or network failure. Currently we are not using intelligent agents because of the lack of training datasets. However, we will consider using intelligent agents in our future work on this project.

### *4.2. Data Indexing*

Figure 3 illustrates the hierarchical structure of the data repository in the proposed system. The root node of DiiS is a centralized unique identifier directory. New data will be routed through the root node; additional copies of the data will be created to maintain reliability and improve query performance. Users would be able to access their target data from the nearest (geographical) node in the data repository. DiiS structure incorporates different types of data sources that include (1) fully shared data sources (copied into the DiiS repository), (2) in-house data sources, and (3) data sources stored at the origin (outside of DiiS repository) and represented by an index over such external data. Data would be categorized and integrated based on the metadata provided by the contributing institutions. DiiS framework would integrate data sources through domain-specific adapters, provide high-availability guarantees, and deliver fine-grained access controls based on defined user privileges.

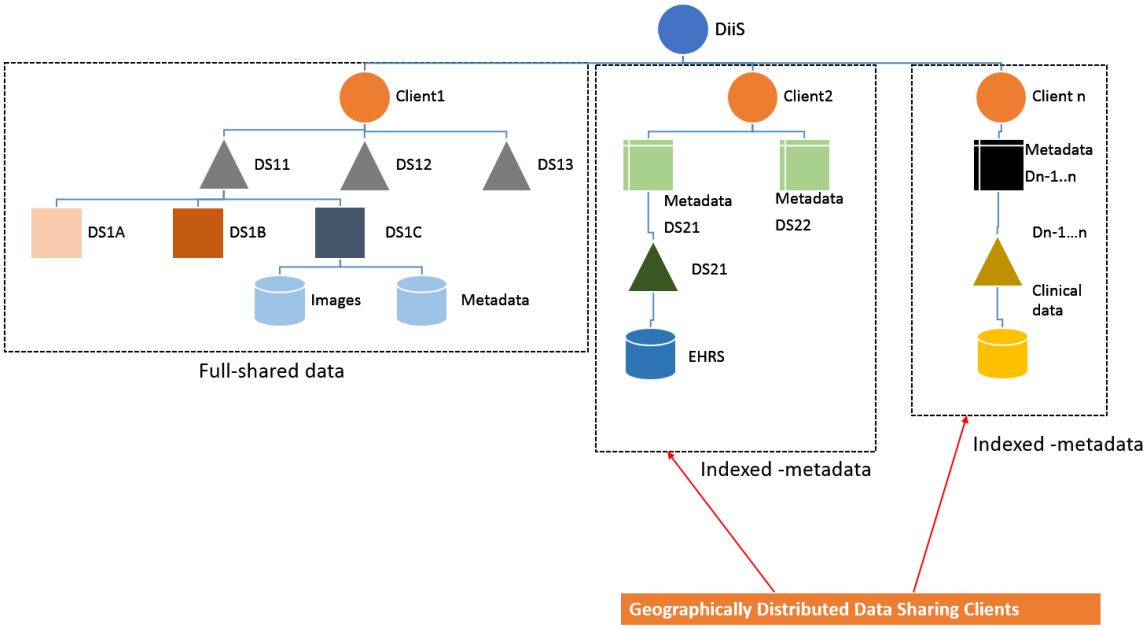


**Figure 3.** Hierarchical structure of Data Sources indexing (DS: Data Source).

Summary of DiiS modules addressing FAIR principles:

1. Findable: Our metadata Uniform Resource Identifiers (URIs) and Universally Unique Identifiers (UUIDs) will make data findable across geographically distributed nodes.
2. Accessible: Our user authentication and search engine modules would make data retrievable and provide access only to authorized users.
3. Interoperable: Our data read and write adapters would simplify the process of data sharing and searching across heterogeneous data sources.
4. Reusable: Our user interface would display all pertinent dataset information (e.g., license, related publications), so users are able to make the decision on whether to reuse these datasets in the future.

## 5. Discussion

We initiated this research project after identifying the need for integrated biomedical database—the need for a repository where users can easily access and share their data. We performed an extensive literature review, went through several case studies, and talked with radiologists, doctors, and medical researchers. From this discussion we realized that there is an urgent need for a data integration system and thus started designing the proposed system. Based on the features we chose to support in DiiS, we realized that some of these features cannot be fully automated and require human participation. To achieve the goals of DiiS, we will need help of research groups, input from different institutes (to form policies), and financial support from government institutions. Our system will be a great contribution towards facilitating the sharing of a variety biomedical data in healthcare research, e.g., in artificial intelligence research, in diagnostic process (doctors can find reference cases in teaching files), and in computer aided diagnosis (using data to create predictive models for diagnostics). As discussed in Section 4, our proposed framework addresses different challenges described in Section 2.3. We believe that DiiS will be instrumental in providing users with data search in the integrated repository. Our system is designed to support data integration transparently; it offers a user friendly interface to data donors who do not need to worry about the internal structure of the data repository.

A limitation of DiiS is the dependency on data donors and organization involvement to form and synchronize different policies for data sharing. To validate the future functionality of our proposed

model we began the integration of different public medical data sources. Our prototype, IRIS [30], successfully integrated two public radiology data sources (MIRC and MyPacs) and two medical ontologies (RadLex and SNOMED CT) into a single repository. We are further extending and adopting this work into DiiS framework; our first step is to integrate new types of data source (e.g., clinical reports). DiiS would provide data interoperability by providing an easy way to import and export datasets from different users. Custom tailored data adapters will play an important role by facilitating access to heterogeneous data sources.

The DiiS framework integrates heterogeneous data sources and enables a centralized data repository with fine-grained access controls. Similar approach can be deployed in a variety of different domains. For example, for the purposes of a NASA repository, we might similarly integrate different datasets such as geological records, air traffic management data, climate records, international space station data, and solar records. DiiS integration for biomedical data is similar to the work done by the Information Integration (I2) Group at NASA [54]. The goal of I2 group is to bring together their data into a single data repository that supports integrated search, access, and comprehensive data analysis. Although NASA data is different from biomedical data that is the current focus of DiiS integration, a taxonomy of data sources (such as the one in Section 3.1) can be readily developed for other source types. Combined with domain-specific ontologies and further domain understanding, we believe that a similar system can be designed for many other fields. Expected benefits of DiiS-like data integration are:

- Simplifying exchange of electronic information.
- Decreasing the cost and complexity of interfacing between different systems.
- Improving data interoperability.
- Facilitating data discovery and supporting reproducibility in research.

## 6. Conclusions

A distributed and integrated data repository is key to advancing biomedical research. The massive increase in data types, data sources and velocity of data collection in healthcare industry further accentuates the need for data integration solutions. Our proposed model, guided by FAIR principles, can help make data available across geographically distributed and varied organizations. We believe that our IRIS system can be used as a prototype for biomedical data integration. DiiS would use the IRIS several data integration approaches (data identification, collection, cleaning, transformation, and loading) that would help education and research institutes in the healthcare domain by making data sources available across a distributed environment—users from heterogeneous locations can share and access data using DiiS.

**Author Contributions:** P.D. conducted R & D into the system design and methods for supporting FAIR principles and fine-grained access control for geographically distributed datasets. She performed an extensive literature study for biomedical datasets integration, learn challenges, and need for proposed system. A.R. guided the design of the DiiS framework and the underlying methodology. J.F., D.R. and S.A. provided feedback on the features of the framework, helping position it with respect to other research. All authors contributed to manuscript writing and editing. Conceptualization, P.D., S.A.; Investigation, P.D., A.R. and S.A.; Methodology, P.D. and A.R.; Supervision, A.R., J.F., D.R. and S.A.; Validation, P.D., A.R. and S.A.; Writing—original draft, P.D.; Writing—review & editing, A.R., J.F., D.R. and S.A.

**Funding:** This research received no external funding.

**Acknowledgments:** This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. NIH. STRIDES Initiative. Available online: https://commonfund.nih.gov/strides/ (accessed on 19 April 2019).
2. Merelli, I.; Pérez-Sánchez, H.; Gesing, S.; D'Agostino, D. Managing, analysing, and integrating big data in medical bioinformatics: Open problems and future perspectives. *BioMed Res. Int.* **2014**, *2014*, 134023. [CrossRef] [PubMed]
3. McQuilton, P.; Gonzalez-Beltran, A.; Rocca-Serra, P.; Thurston, M.; Lister, A.; Maguire, E.; Sansone, S.A. BioSharing: Curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database* **2016**, *2016*. [CrossRef]
4. CrowdFlower 2016. Available online: http://visit.crowdflower.com/ (accessed on 19 April 2019).
5. Raghupathi, W.; Raghupathi, V. Big data analytics in healthcare: Promise and potential. *Health Inf. Sci. Syst.* **2014**, *2*, 3. [CrossRef]
6. Hemler, J.R.; Hall, J.D.; Cholan, R.A.; Crabtree, B.F.; Damschroder, L.J.; Solberg, L.I.; Ono, S.S.; Cohen, D.J. Practice facilitator strategies for addressing electronic health record data challenges for quality improvement: EvidenceNOW. *J. Am. Board Fam. Med.* **2018**, *31*, 398–409. [CrossRef] [PubMed]
7. Raje, S.; Kite, B.; Ramanathan, J.; Payne, P. Real-time Data Fusion Platforms: The Need of Multi-dimensional Data-driven Research in Biomedical Informatics. *Stud. Health Technol. Informat.* **2015**, *216*, 1107.
8. NIH. PubMed. Available online: https://www.ncbi.nlm.nih.gov/pubmed/ (accessed on 19 April 2019).
9. dryad. Available online: https://datadryad.org/ (accessed on 19 April 2019).
10. NCI. NCI data. Available online: https://datascience.cancer.gov/ (accessed on 19 April 2019).
11. Deist, T.M.; Jochems, A.; van Soest, J.; Nalbantov, G.; Oberije, C.; Walsh, S.; Eble, M.; Bulens, P.; Coucke, P.; Dries, W.; et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: EuroCAT. *Clin. Transl. Radiat. Oncol.* **2017**, *4*, 24–31. [CrossRef]
12. Deshpande, P.; Rasin, A.; Brown, E.; Furst, J.; Raicu, D.; Montner, S.; Armato, S., III. An Integrated Database and Smart Search Tool for Medical Knowledge Extraction from Radiology Teaching Files. *Med. Informat. Healthc.* **2017**, *69*, 10–18.
13. Holzinger, A.; Dehmer, M.; Jurisica, I. Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC Bioinform.* **2014**, *15*, I1. [CrossRef]
14. Trifan, A.; Díaz, C.; Oliveira, J. A Methodology for Fine-Grained Access Control in Exposing Biomedical Data. *Stud. Health Technol. Informat.* **2018**, *247*, 561–565.
15. Sujansky, W.V.; Faus, S.A.; Stone, E.; Brennan, P.F. A method to implement fine-grained access control for personal health records through standard relational database queries. *J. Biomed. Informat.* **2010**, *43*, S46–S50. [CrossRef] [PubMed]
16. Chen, X.; Gururaj, A.E.; Ozyurt, B.; Liu, R.; Soysal, E.; Cohen, T.; Tiryaki, F.; Li, Y.; Zong, N.; Jiang, M.; et al. DataMed–an open source discovery index for finding biomedical datasets. *J. Am. Med Informat. Assoc.* **2018**, *25*, 300–308. [CrossRef] [PubMed]
17. Ohno-Machado, L.; Sansone, S.A.; Alter, G.; Fore, I.; Grethe, J.; Xu, H.; Gonzalez-Beltran, A.; Rocca-Serra, P.; Gururaj, A.E.; Bell, E.; et al. Finding useful data across multiple biomedical data repositories using DataMed. *Nat. Genet.* **2017**, *49*, 816. [CrossRef] [PubMed]
18. Ohno-Machado, L.; Sansone, S.A.; Alter, G.; Fore, I.; Grethe, J.; Xu, H.; Gonzalez-Beltran, A.; Rocca-Serra, P.; Soysal, E.; Zong, N.; et al. DataMed: Finding useful data across multiple biomedical data repositories. *bioRxiv* **2016**, 094888. [CrossRef]
19. Wang, F.; Vergara-Niedermayr, C.; Liu, P. Metadata based management and sharing of distributed biomedical data. *Int. J. Metadata Semant. Ontol.* **2014**, *9*, 42–57. [CrossRef] [PubMed]
20. Trifan, A.; Oliveira, J.L. A FAIR marketplace for biomedical data custodians and clinical researchers. In Proceedings of the 2018 IEEE 31st International Symposium on Computer-Based Medical Systems, Karlstad, Sweden, 18–21 June 2018.
21. Krumholz, H.M. Big data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system. *Health Aff.* **2014**, *33*, 1163–1170. [CrossRef] [PubMed]
22. Dey, P.; Ross, J.S.; Ritchie, J.D.; Desai, N.R.; Bhavnani, S.P.; Krumholz, H.M. Data sharing and cardiology: Platforms and possibilities. *J. Am. Coll. Cardiol.* **2017**, *70*, 3018–3025. [CrossRef] [PubMed]

23. Kansagra, A.P.; John-Paul, J.Y.; Chatterjee, A.R.; Lenchik, L.; Chow, D.S.; Prater, A.B.; Yeh, J.; Doshi, A.M.; Hawkins, C.M.; Heilbrun, M.E.; et al. Big data and the future of radiology informatics. *Acad. Radiol.* **2016**, *23*, 30–42. [CrossRef]
24. Angraal, S.; Ross, J.S.; Dhruva, S.S.; Desai, N.R.; Welsh, J.W.; Krumholz, H.M. Merits of Data Sharing. *J. Am. Coll. Cardiol.* **2017**, *70*, 1825–1827. [CrossRef]
25. Deshpande, P.; Rasin, A.; Brown, E.; Furst, J.; Raicu, D.; Montner, S.; Armato, S., III. Big Data Integration Case Study for Radiology Data Sources. *IEEE Life Sci. Conf.* **2018**. [CrossRef]
26. RSNA. RSNA TFS. Available online: http://mirc.rsna.org/query (accessed on 19 April 2019).
27. Weinberger, E; Jakobovits, R; Halsted, M. MyPACS. net: A Web-based teaching file authoring tool. *Am. J. Roentgenol.* **2002**, *3*, 579–582. [CrossRef]
28. RSNA. RadLex Ontology. Available online: http://www.radlex.org/ (accessed on 19 April 2019).
29. SNOMED International. SNOMEDCT Ontology. Available online: http://www.snomed.org/ (accessed on 19 April 2019).
30. Deshpande, P.; Rasin, A.; Brown, E.T.; Furst, J.; Montner, S.M.; Armato, S.G., III; Raicu, D.S. Augmenting Medical Decision Making With Text-Based Search of Teaching File Repositories and Medical Ontologies: Text-Based Search of Radiology Teaching Files. *Int. J. Knowl. Discov. Bioinform.* **2018**, *8*, 18–43. [CrossRef]
31. NIH. Openi. Available online: https://openi.nlm.nih.gov/ (accessed on 19 April 2019).
32. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef]
33. NIH. Data Science at NIH. Available online: https://datascience.nih.gov/ (accessed on 19 April 2019).
34. International, H.L.S. Health Level Seven International. Available online: www.hl7.org (accessed on 19 April 2019).
35. HHS. HITECH. Available online: https://www.hhs.gov/hipaa/for-professionals/special-topics/hitech-act-enforcement-interim-final-rule (accessed on 19 April 2019).
36. O'Dowd, E. Healthcare Data Integration Continues to Challenge Entities. Available online: https://hitinfrastructure.com/news (accessed on 19 April 2019).
37. Shashank, A. Why do Healthcare Organizations Still Struggle with Data Integration. Available online: http://blog.innovaccer.com/healthcare-organizations-still-struggle-data-integration/ (accessed on 19 April 2019).
38. Kho, A.N.; Rasmussen, L.V.; Connolly, J.J.; Peissig, P.L.; Starren, J.; Hakonarson, H.; Hayes, M.G. Practical challenges in integrating genomic data into the electronic health record. *Genet. Med.* **2013**, *15*, 772. [CrossRef]
39. Christen, P.; Vatsalan, D.; Verykios, V.S. Challenges for privacy preservation in data integration. *J. Data Inf. Qual.* **2014**, *5*, 4. [CrossRef]
40. Clifton, C.; Kantarcioğlu, M.; Doan, A.; Schadow, G.; Vaidya, J.; Elmagarmid, A.; Suciu, D. Privacy-preserving data integration and sharing. In Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Paris, France, 13 June 2004; pp. 19–26.
41. Gomez-Cabrero, D.; Abugessaisa, I.; Maier, D.; Teschendorff, A.; Merkenschlager, M.; Gisel, A.; Ballestar, E.; Bongcam-Rudloff, E.; Conesa, A.; Tegnér, J. Data integration in the era of omics: Current and future challenges. *BMC Syst Biol.* **2014**, *8*, I1. [CrossRef]
42. Healthcare Information and Management Systems Society (HIMSS). What is Interoperability? Available online: https://www.himss.org/library/interoperability-standards/what-is-interoperability (accessed on 19 April 2019).
43. UMLS. UMLS. Available online: https://www.nlm.nih.gov/research/umls (accessed on 19 April 2019).
44. Langlotz, C.P. RadLex: A new method for indexing online educational materials. *Radiol. Soc. N. Am.* **2006**, *3*, 1595–1597. [CrossRef]
45. Masseroli, M.; Mons, B.; Bongcam-Rudloff, E.; Ceri, S.; Kel, A.; Rechenmann, F.; Lisacek, F.; Romano, P. Integrated Bio-Search: Challenges and trends for the integration, search and comprehensive processing of biological information. *BMC Bioinform.* **2014**, *15*, S2. [CrossRef]
46. Huesch, M.D. Using It or Losing It? The Case for Data Scientists Inside Health Care. Available online: https://catalyst.nejm.org/case-data-scientists-inside-health-care/ (accessed on 19 April 2019).
47. EURORAD. Available online: http://www.eurorad.org/ (accessed on 19 April 2019).
48. NIH. National Institutes of Health Chest X-ray Dataset. Available online: https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/36938765345 (accessed on 19 April 2019).

49. NIH. NIH Clinical Center Provides One of the Largest Publicly Available Chest X-ray Datasets to Scientific Community. Available online: https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community (accessed on 19 April 2019).
50. CIVM. CENTER for IN VIVO MICROSCOPY (CIVM) dataset. Available online: http://www.civm.duhs.duke.edu/devatlas/index.html (accessed on 19 April 2019).
51. OpenfMRI. Neuroimaging data. Available online: https://openneuro.org/ (accessed on 19 April 2019).
52. Richards, M. *Software Architecture Patterns*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2015.
53. Richardson, D.J.; Wolf, A.L. Software testing at the architectural level. In Proceedings of the Second International Software Architecture Workshop (ISAW-2) and International Workshop on Multiple Perspectives in Software Development (Viewpoints' 96) on SIGSOFT, San Francisco, CA, USA, 16–18 October 1996; Volume 96, pp. 68–71.
54. NASA. Information Integration Overview. Available online: https://ti.arc.nasa.gov/tech/cas/groups/information-integration/ (accessed on 19 April 2019).