Biomedical Information Integration Middleware for Clinical Genomics

Simona Rabinovici-Cohen

IBM Haifa Research Lab
Haifa University, Mount Carmel, Haifa 31905, Israel
simona@il.ibm.com

Abstract. Clinical genomics, the marriage of clinical information and knowledge about the human or pathogen genome, holds enormous promise for the healthcare and life sciences domain. Based on a more in-depth understanding of the human and pathogen molecular interaction, clinical genomics can be used to discover new targeted drugs and provide personalized therapies with fewer side effects, at reduced costs, and with higher efficacy. A key enabler of clinical genomics is a sound standards-based biomedical information integration middleware. This middleware must be able to de-identify, integrate and correlate clinical, clinical trials, genomic and images metadata from the various systems. We describe MedII, a novel biomedical information integration research technology that some of its components were integrated in IBM Clinical Genomics solution. We also introduce the need for biomedical information preservation to assist in ensuring that the integrated biomedical information can be read and interpreted decades from now.

Keywords: information integration, biomedical systems, de-identification, long term digital preservation.

1 Introduction

On June 26, 2000, the first draft of the human genome was announced, significantly boosting research on correlations between phenotypic and genotypic data. Today, some genomic data is produced by DNA microarrays that generate high throughput molecular data and by several hundred non-expensive genetic tests that probe small portions of the genome. While sequencing a full human genome is still very costly, taking approximately six months and $10 million to $50 million to complete, new cheaper and faster sequencing machines are being developed (such as the Helicos BioSciences machines). It is assumed that within a few years, a human genome could be completely sequenced in one day at the cost of $1000, creating new healthcare and life sciences practices in which sequencing an individual's genome becomes a commodity (see the NIH request for application number RFA-HG-08-008 titled "Revolutionary Genome Sequencing Technologies – The $1000 Genome").

Consequently, a new field known as Clinical Genomics is emerging in the healthcare and life sciences domain that may contribute to a revolution in the health of humankind.

Y. Feldman, D. Kraft, and T. Kuflik (Eds.): NGITS 2009, LNCS 5831, pp. 13–25, 2009.

Clinical genomics is the correlation of clinical information—such as patient records including environmental data, family histories, medications, and lab tests—with knowledge about the human or pathogen genome. This correlated information will impact decisions about diagnosis, prognosis**,** treatment, and epidemiology. By understanding illnesses on the molecular level, including gene variations linked to disease or drug response, doctors may be able to make more precise diagnoses and tailored treatment decisions. Pharmaceuticals researchers will discover new drugs and develop targeted treatments that have better safety and higher efficacy. Clinical genomics will also improve healthcare guidelines and protocols, leading the way towards truly personalized healthcare and information-based medicine.

However, realizing clinical genomics has difficulties. Much of the data is still in propriety formats in various silos, and privacy issues associated with sharing that data still exist. A biomedical information integration technology is needed that can standardize, de-identify, integrate, and correlate clinical trials with genomic and image metadata described in diverse formats and vocabularies and scattered in disparate islands of data. The data needs also be integrated with public data sources such as PubMed and GenBank.

We describe a novel research technology, named MedII (previously known as Shaman-IMR), for biomedical information integration, which some of its components were incorporated in IBM Clinical Genomics solution [1]. MedII includes a set of services to standardize, de-identify, integrate and correlate the various biomedical data. The services utilize XML technology [2], Model Driven Development (MDD), and emerging healthcare and life science standards. XML technology is well-situated to serve as the glue for integrating the various data sources because XML includes both data and metadata in a universal format, namely text, and it is agnostic to the platform or application used to create or consume the data. However, XML lacks semantic at its core, so standards are needed to provide the semantics, and indeed vast standardization efforts by various organizations are witnessed. MDD aims to separate domain knowledge from the underlying technology specifics, enabling to build tools that ease the generation and update of transformation services and data models.

MedII utilizes an integrated target schema that is a union of sub-schemas; one for each standard in the domain. The records of the various standards are correlated via a global privacy id which is the same for all documents of the same individual, as well as via standardized controlled codes and vocabularies. This MedII approach reduces cost and complexity as a new data source can be added or removed without changes in the other data sources or target schema. Moreover, data de-identification is an integral part of the system and the generated privacy id is used to replace all the identifying information as well as to correlate among the records of the same individual.

The rest of this paper is organized as follows. The next section provides related work and our summarized contribution. Then, we describe an overview of MedII various layers followed by additional three sections that dive deeper into those layers. The integration layer section describes the services, which generally run in the source premises, to standardize and enrich the data so that it could be integrated with data from other sources. The de-identification section describes a service which is required by government legislation in the lack of user consent, and thus has special interest. The index layer section describes the services to manage and index the data in order to

be efficiently queried and searched by various applications. Finally, as much of the integrated data should be preserved for decades, we describe the need for biomedical information preservation and propose future work in that space.

2 Related Work

The direct effect of healthcare and life sciences on each individual, the ever growing volumes of biomedical data and the advances in genomics encouraged research in biomedical information integration for several years [3]. One approach, also termed "integration at the glass", performs integration in the eyes of the human user by putting data from different sources side by side on the same screen. PharmGKB [4] which integrates literature, diseases, phenotypes, drugs and genomic information is an example of such approach. This approach is sometimes combined with data grids such as in the BIRN [5] project funded by NIH, which is a data grid that targets shared access to neuro-imaging studies. While this approach is effective for human users, it is not adequate for machines which need to process and mine the data.

Other approaches offer integration which either merges the query result sets (e.g. OPM [6], DiscoveryLink [7]), or merges the sources data (e.g. GenMapper [8], BioMediator [9]) and thus are adequate for human and machine users. OPM offers a propriety query language to join remote databases via wrappers. DiscoveryLink which is now embedded in IBM DB2 is a federation technology in which integration is done in query time using wrappers to the source databases. GenMapper integrated data from more than forty data sources into a 4-table propriety schema. BioMediator is similar but uses a more elaborated target schema defined using the Protégé knowledgebase system. All these projects are schema focused and tend to employ propriety schemas, and require learning propriety and complicated languages. Creating the semantic mappings from the sources to the target schema is tedious and is generally done manually, but this may be relaxed by using automatic schema mapping methods [10]. Also, these projects do not deal with privacy issues associated with the data.

Aladin [11] is a system which combines the schema focused approach with a data focused approach and utilizes automatic mechanisms for finding links among source objects. Aladin is mostly for biological data and it assumes the data is semi structured and text centric rather than a structured data centric database. Adding or removing data sources as well as updates to the current data sources are expensive in the Aladin system because all links must be recomputed even if only a small fraction of the data of a data source changed.

In MedII, each data source is cleansed, augmented with terminology codes and transformed to the standard for that data type, in case the source data wasn't in the required standard format initially. Then, the standardized data is assigned with a privacy id, de-identified and stored in a data model based on that standard. Thus, the integrated schema is a union of sub-schemas; one for each standard. The records of the various standards are correlated via the global privacy id which is the same for all documents of the same individual, and via the standardized controlled codes and vocabularies.

MedII offers integration which is adequate for a human or machine user. It employs standard-based schemas and does not require learning any propriety language. Using MDD based tools, data transformation and generation of data models is simplified. Additionally, adding, removing, or updating a data source has a contained impact and there is no effect on the other data sources or the target schema. Finally, in MedII, data de-identification is an integral part of the system and the generated privacy id is used to replace all the identifying information as well as to correlate among the records of the same individual.

3 MedII Conceptual Layers Overview

Clinical and genomic data are scattered among different archives (computerized and paper-based) in various locations. Data is generally stored where it was created, and is not always available to researchers. Furthermore, this information is often expressed using different vocabularies, terminologies, formats, and languages, and is retrieved using different access methods and delivery vehicles. Some of the data resides in external open databases, which include rich data but are not easy to navigate.

MedII goal is to integrate those diverse data sources with a low cost and while considering privacy limitations. The cost remains minimal even when adding, removing or updating a data source. Update in one source does not cause updates to the integration of the other data sources. Similarly, adding or removing a data source does not impact the other parts of the system.

MedII uses standardized formats and workflows which facilitates the transition from information in silos to cross-institutional information integration. It includes three conceptual layers as shown in Figure 1 below. The first is an **integration layer**, which includes tools and services for transforming data from propriety formats to standard XML-based representation. This layer also includes data cleansing and enrichment services, such as adding codes from controlled vocabularies or adding annotations from public data sources. Finally, it includes a de-identification service for de-identifying protected data prior to leaving its source premises, as required by government legislation. The **index layer** generates efficient standardized data models and indexes for structured, semi-structured and unstructured data so that applications on top can query, perform free text search and retrieve the data in an efficient and powerful way. The **EHR layer** (Electronic Health Record layer) is designed for applications in the healthcare domain, and creates longitudinal and cross-institutional individual-centric objects that are compiled from the documents indexed by the previous layer. As this layer is not an integral part for the integration, it is not discussed further in a separate section. Mining and query tools, predictive engines or other applications can reside on top of either the second layer or the third layer [12]. MedII facilitates complex queries such as: “What protocols were used for tumors that produced similar staining sections and were from patients aged 40-60 with the same ‘Yakamura’ polymorphism in their genes?” The various services of MedII layers can be exploited in a Service Oriented Architecture and attached to an enterprise service bus.

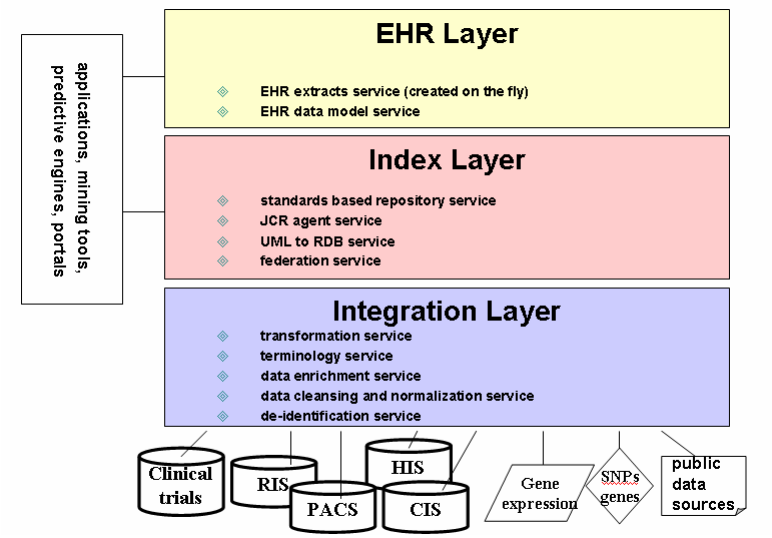

Fig. 1. MedII layers and services

4 Integration Layer

The integration layer includes several services, only some of which may be used and the order in which those services are called is flexible. The services employ MDD and include a transformation service, a terminology service, a data enrichment service, a data cleansing and normalization service, and a de-identification service, required by law in the lack of user consent. The de-identification service is described in more detail separately in the next section. Usually, it is preferable to apply the transformation service first because then you can employ the other services on the obtained standardized formats, thereby simplifying their employment.

The transformation service transforms the various proprietary biomedical data formats to standardized formats, or transforms one standard to another, thereby enabling the generic correlation of data from cross-institutional sources while reusing the domain knowledge exploited by the standard organizations. This knowledge is obtained over many years, from various contributors coming from diverse organizations. In some cases, only some of the data to be shared is transformed to a standardized format along with a link to the full original data. These cases are usually known as metadata extraction and occur primarily when some of the data is in an uninterpreted format, such as binary format. For example, in DICOM images, only some of the DICOM tags are extracted and transformed to a standard along with a link to the full DICOM object in the original Picture Archive Communication Systems (PACS) system.

Biomedical data includes various types of information, including clinical data from Clinical Information Systems (CIS) in hospitals and HMOs, demographic data from Hospital Information Systems (HIS), diagnostic and imaging data from Radiology

Information System (RIS) and PACS, clinical trials data, environmental and life style data, and genomic data that partially reside in Laboratory Information System (LIMS). Currently, the standards for representing biomedical data are still emerging and some of them have not yet been widely adopted. Some data types do not have standards, while other data types have several standards. Most of the new standards are XML-based and still evolving. As a result, one of the challenges is to decide what standard to use for what data.

The standards for clinical data include the widely used HL7 V2 and its XML version HL7 V2.XML. HL7 is also developing the new V3 family using MDD methodology, including among others the Clinical Document Architecture (CDA) and Clinical-Genomics message, but this new family is still in the early adoption phase. Digital imaging data is dominated by the DICOM standard. The new DICOM Structured Reports (SR) for imaging reports is still in early adoption phases.

The standards for genomic data include OMG MGED MAGE for gene expression and the related Minimum Information About a Microarray Experiment (MIAME), NIH HapMap for haplotype data, BSML for bioinformatic sequence, SBML for Systems Biology, and NCI caBIO for cancer bioinformatics infrastructure objects. All these standards are developed with MDD methodology and are still in the early adoption phase.

The standards for clinical trials data are CDISC ODM for operational data that is moved from a collection system to the central database of the sponsor of the clinical trial. CDISC SDTM is used to submit data to regulatory authorities such as the FDA. CDISC Define.xml is an extension of ODM that includes metadata on the domains and variables used in SDTM. All these standards are still in early adoption phases.

In many standards organizations and large consortiums, domain knowledge is increasingly captured by domain experts using UML conceptual models, which include data elements without methods. In the MDD approach, the source and target schemas are represented in UMLs, and UML profiles are introduced to extend and constrain UMLs for a particular use. Accordingly, the transformation service, which utilizes the MDD approach, includes a transformation tool that imports the source and target UML models and then configures the UML profile and mappings from the source to the target schema. By using MDD, the service becomes generic and agnostic to changes in the models that occur during the standardization process.

The terminology service augments the data with codes from controlled vocabularies, thus facilitating more accurate search and mining of the integrated data. The primary controlled vocabulary is SNOMED, a very extensive vocabulary adapted by the US government and available free for distribution. Additional controlled vocabularies include LOINC for laboratory results, test orders and document types; ICD-9 and ICD-10 for diseases; RxNORM, which is adapted by FDA for medications; CPT for procedures and genomic tests; and HGNC for nomenclature of gene symbols and names. The terminology service can be built on top of open existing terminology services such as HL7 Clinical Terminology Service (CTS) or NCI Enterprise Vocabulary Services (caEVS).

The data enrichment service adds to the data annotations and related data from public data sources. PubMed is a prominent public source for clinical data that includes medical publications. Additionally, there are rich public data sources for genomic data, such as EBI ArrayExpress for microarray data, GenBank for genetic sequence, dbSNP

for simple genetic polymorphisms, UniProt and PDB for proteins, and OMIM for correlations of human genes and genetic disorders. The relevant content of those public data sources are sometimes embedded within the standardized data to be shared, while in other cases it is federated in the index layer. The first case is used when it is important to document the evidence that brought to a specific conclusion, e.g., for compliance purposes. The latter case is preferred when it is important to enrich the data in real time with the most recent information from public sources.

In MedII, we implemented transformations to HL7 V3, specifically CDA, as well as to MAGE, ODM, SDTM, BSML, and performed metadata extraction from DICOM. We have used the SNOMED, LOINC and ICD9 terminologies and performed a federation with PubMed public data source. However, there are no limitations to adding additional standards, terminologies or public data sources to the system.

5 De-identification

De-identification issues usually arise when enterprises need to share data that is related to individuals or includes sensitive business data. For example, in clinical genomics, de-identification is often mandated when cross-institutional data needs to be integrated and correlated. Medical information is naturally associated with a specific individual, and when this data leaves the source premises, it must be altered so that it cannot be re-associated with that individual by the recipient. Even in this case, the process should preserve the ability to correlate the de-identified document with other de-identified documents or records from the same individual.

Government guidelines and legislations in various countries define Protected Health Information (PHI) data, which if lacking user consent must be de-identified before leaving the data source premises. This includes the Health Insurance Portability and Accountability Act of 1996 (HIPAA) in the United States, the Freedom of Information and Protection of Privacy Act (FIPPA) in Canada, and the Personal Information Protection and Electronic Document Act (PIPEDA) in the European Union, Australia, and New Zealand. For genomic data, there is also the NIH-DOE Joint Working Group on Ethical, Legal, and Social Implications of Human Genome Research (ELSI).

HIPAA, the least stringent of the legislation listed above, defines seventeen identifiers that must be de-identified. Some of the identifiers that must be removed includes the name, certificate/license numbers, diagnostic device ID and serial number, biometric identifier (e.g., voice, finger print, iris, retina), full face photo or comparable image, SSN, fax numbers, electronic mail address, URL, IP address, medical record number, health plan number, account numbers, vehicle ID, serial number, and license plate number. The other identifiers are less restrictive. In address identifiers, the city/town, state, and first three digits of zip code can be kept if the population in the city is greater than 20,000. Similarly, in date identifiers (e.g., DoB, ADT, DoD), the year can be kept, and in telephone numbers identifiers, the area code and prefix can be kept if geographical information is missing. Additionally, the eighteenth item defined by HIPAA is called “Other”, and it refers to information that exists in free text sections of patient records and can be used to identify the patient. Examples of phrases that must be de-identified according to the “Other” identifier are “the British Queen” and “The

patient suffered from serious leopard bite acquired at the De Moines Zoo fire incident". Note that according to HIPAA, gender, race, ethnic origin, and marital status can be kept.

In MedII, the UDiP technology is used within the integration layer to provide de-identification service, as depicted in Figure 2 below. UDiP includes a highly configurable and flexible engine, generally located on the source premises, that receives various types of data, including clinical, imaging, and genomic data, and various formats, including comma-separated-values (CSV) files, DICOM images, and XML data. UDiP de-identifies the PHI while maintaining the correlation between the various records belonging to the same individual, and then transfers it (pass the red line) to the shared data area. By transferring de-identified data to the shared area, we prevent access to the data in the source premises, thus avoiding the risk of a direct privacy attack on the original data. Maintaining the correlation between the various records is obtained by either PHI encryption or by using an Anonymous Global Privacy Identifier (AGPI) server. Although PHI encryption is a simpler method, it is not always preferable. According to some interpretation of HIPAA, PHI encryption violates the "Other" field by keeping the PHI (although encrypted) together with the medical data in the shared area.

With the AGPI server method, the PHI is de-identified (removed) and replaced by an AGPI opaque value assigned by an AGPI server. That AGPI server resides in a secured area with strict access control. Only authorized people can access it with an AGPI value for re-identification. Note that in this case, there is no focal point to include the PHI and the actual medical data – they are disjoint. The AGPI is used to maintain the correlation between documents belonging to the same individual, without identifying the individual. The AGPI server is required to fulfil two criteria. The server gives the same AGPI to the same individual (regardless of the source of the document, the spelling of the patient's name, incompatible fields, etc.), and it never gives the same AGPI to two different individuals.

Since privacy restrictions change with time and are distinctive in different locations, UDiP is highly configurable and extensible. A model-driven tool that can be used by domain experts configures the data types to de-identify, the locations of the PHI, the action to apply on each location, and the method for generating and storing the AGPI.

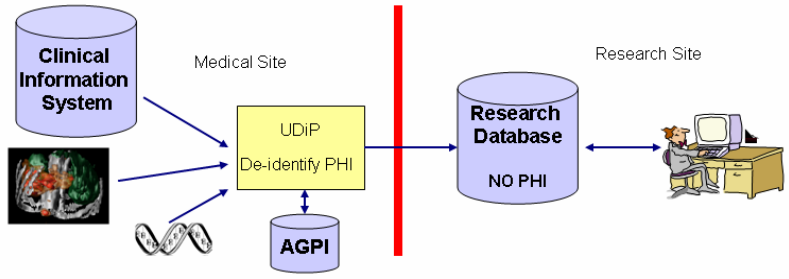

Fig. 2. De-identification with AGPI server

6 Index Layer

The index layer is responsible to manage and index all the normalized standardized documents obtained from the integration layer, so that applications on top can efficiently perform powerful queries, searches and data retrieval. It mainly contains information management services such as standard based repositories services and services that employ MDD to generate those repositories. The data models used in the index layer are standard based and include a separate schema for each standard. The records of the various standards are correlated via the AGPI which is the same for all documents of the same individual and via the standardized controlled codes and vocabularies. This design allows the flexible update of a data model for a specific standard that evolved without affecting the other data models of the other standards.

There are several options to manage and index the data that came from numerous sources. One option is the aggregation model in which the data is physically stored outside the sources premises in a shared repository and the indexation and queries are performed on that repository. In some cases, the repository can include some metadata of some other documents such as in the case of imaging where only the image metadata is kept in the repository with a link to the full image in the PACS system. Another option is the federation model in which no data is stored outside the sources premises. Instead, mappings are defined from the shared system to the sources. In query time, the query is divided to sub-queries which are posed against the various data sources. Then, the correspondent sub-results are jointed to one result to form the query result. In the federation model there is no data stored outside of the sources premises, but on the other hand it requires that the sources are on-line all the time and the performance of the queries is decreased with respect to the aggregation model.

A third option, which is used in MedII, is a hybrid model of aggregation and federation. Data from operational systems are aggregated to prevent them from interfering during operational time. Data from public sources that are not under institutional control are generally federated. Accessing the external reference data via federation ensures data currency, but at the cost of performance degradation and at the risk of security compromise.

The biomedical data includes structured as well as unstructured data. A case report form, for example, includes patient name, gender, age, and so forth, which is structured data, along with a description of the medical history, which is unstructured data. To query and search the data, there is a need for a combined relational database index and an information retrieval (IR) index. Today, the data is either treated as structured data and stored in a relational database, or treated as unstructured data and stored in a content management system. In the first case, the data is shredded (decomposed) in a relational database but then the free text search capabilities are limited. To this end, an IR technology is added to convert unstructured data to semi-structured content using some annotators, and then the associated meta-data can be joined with the relational repositories.

In the latter case where the documents are stored in a content management system, the queries for these documents are generally less powerful. For example, it's difficult to find patients that have two medications, where each medication appears in a different document (a.k.a. the join operation in the relational database). The new Java Content Repository (JCR) specification holds the promise to bridge the gap between the structured and unstructured worlds, and provide an interface to combine query and free

text search. The JCR has a hierarchical data model and allows to store various structured and unstructured typed properties (leafs of the hierarchy) such as strings, binaries, dates. JCR query manager facilitates structured SQL queries as well as XPATH queries with free text search capabilities.

As stated earlier, the domain knowledge is increasingly represented in conceptual UML models. Standard organizations and government sponsored initiatives build public UML models for biomedical data representation as well as models for disease specific data. Examples of such models are HL7 RIM, OMG MAGE, NCI caBIO, and HUPO PSI. Moreover, it is assumed that pharmaceutical companies and academic medical research centers will also build conceptual UML models for their specific use cases by either tailoring those publicly available models or by building their own models.

Thus, the MDD approach is vastly used in the index layer as well. It includes tools and runtimes to semi-automatically generate a relational data model or a JCR data model for a given UML. The services such as the JCR agent service introduce UML profiles to extend and constrain the UMLs to describe primary keys, JCR nodes references, etc. According to the UML profile configuration, the JCR agent builds a JCR data model and the runtime receives the XML documents that are compliant with the defined UML model as input and stores them in the JCR-compliant repository. The UML to RDB service provides similar functionality but for the relational database case. Another example of an MDD based service is a service that builds a domain text annotator for a given model.

MedII has implemented the JCR agent and the UML to RDB services which generate standard based repositories. Additionally, it uses DB2 underneath, so it inherits the federation service of DB2.

7 Future Work – Biomedical Information Preservation

The amount of biomedical information is constantly growing and while some of the data includes small text objects like CDA documents, the data generated by medical devices include large (gigabytes) born-digital binary objects such as large images and gene expressions. To assure a lifetime EHR as is proposed nowadays by a new US bill, and also to support compliance legislations, this data should be preserved for the individual whole life time and even beyond that for research purposes and treatments of descendents.

This poses a new challenge to technologists – the challenge of biomedical information preservation, namely how to ensure that the biomedical data can be read and interpreted many years (tens or hundred years) from now when current technologies for computer hardware, operating systems, data management products and applications may no longer exist. As the cost of biomedical information integration is high, it is important to add to it digital preservation capabilities from the beginning, so that the integrated data will be interpretable for many years to come and the high cost is shared with the future over a long period.

A core standard for digital preservation systems is the Open Archival Information System (ISO 14721:2003 OAIS) [13], which targets the preservation of knowledge rather than the preservation of bits, and provides a set of concepts and reference model to preserve digital assets for a designated community. OAIS defines the preservation object that is the basic unit to be stored for preservation. It is the Archival Information

Package (AIP) which consists of the preserved information, called the content information, accompanied by a complete set of metadata.

Figure 3 below depicts an example AIP for a CDA document. The content data object is the raw data intended for preservation namely the CDA document itself. The representation information consists of the metadata that is required to render and interpret the object intelligible to its designated community. This might include information regarding the hardware and software environment needed to render the CDA or the specification of the CDA data format. The other AIP metadata, called the Preservation Description Information (PDI) is broken down by OAIS into four well-defined sections:

- **Reference information** - a unique and persistent identifier of the content information both within and outside the OAIS such as the CDA ClinicalDocument.id object.
- **Provenance information** - the history and origin of the archived object such as a description of the organization in which the CDA was created (ClinicalDocument.Custodian object).
- **Context information** - the relationship to other objects such as related X-rays, related lab tests, previous encounters for the same theme, consent document, encompassing encounter, etc.
- **Fixity information** - a demonstration of authenticity, such as checksums and cryptographic hashes, digital signatures and watermarks.

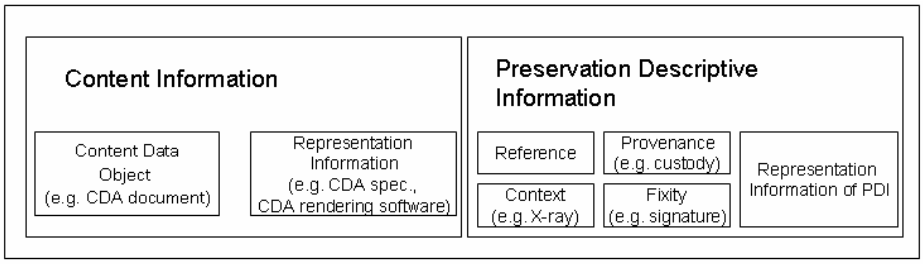

Fig. 3. Example AIP for CDA

At the heart of any solution to the preservation problem, there is a storage component, termed archival storage entity in OAIS. We argue that biomedical information integration systems will be more robust and have less probability for data corruption or loss if their storage component is a preservation aware storage, namely if their storage has built-in support for preservation. As more and more storage systems offload advanced functionality and structure-awareness to the storage layer, we propose to offload OAIS-based preservation functionality to the storage system. To that end we are developing Preservation DataStores (PDS) which realizes a new storage paradigm based on OAIS [14].

For future work, we propose building preservation objects for biomedical information and storing those preservation objects in an OAIS-based preservation aware storage such as PDS.

8 Conclusions

The recently new field of clinical genomics holds the promise to a more in-depth understanding of the human and pathogen molecular interaction. This will lead to a personalized medicine with more precise diagnoses and tailored treatment decisions. A key enabler of these new field is a biomedical information integration technology that can standardize, integrate and correlate the disperse data sources. Various clinical and genomic data sources are extremely valuable independently, but may contain even more valuable information when properly combined.

In this paper we described MedII, a biomedical information integration research technology that utilizes XML technology, model driven development, and emerging healthcare and life sciences standards. The use of standards and XML technology enable semantic integration in a universal format while using MDD makes the technology agnostic to standards evolvement and adaptive to each specific solution extension. MedII features flexible integration of new data sources or update of existing ones without affecting the other parts of the system. Furthermore, the various MedII services, that some of them are incorporated in IBM Clinical Genomics solutions, can be utilized within SOA architecture.

New technology advances and legislations will require preserving the individual integrated biomedical information for his whole lifetime and even beyond that for research purposes and treatments of descendents. We propose for future work to build preservation objects for the integrated biomedical information and utilize an OAIS-based preservation-aware storage for long-lived storage and access to those objects.

Acknowledgements. We would like to thank Barry Robson and Pnina Vortman who were visionary and initiated the Shaman-MedII project. Additional thanks to Houtan Aghili, OK Baek and Haim Nelken who managed IBM Clinical Genomics solution. Finally, thanks to the great team that helped creating MedII various components including Flora Gilboa, Alex Melament, Yossi Mesika, Yardena Peres, Roni Ram and Amnon Shabo.

References

1. IBM Clinical Genomics solution, `http://publib.boulder.ibm.com/infocenter/eserver/v1r2/index.jsp?topic=/ddqb/eicavcg.htm`
2. Shabo, A., Rabinovici-Cohen, S., Vortman, P.: Revolutionary impact of XML on biomedical information interoperability. IBM Systems Journal 45(2), 361–373 (2006)
3. Altman, R.B., Klein, T.E.: Challenges for biomedical informatics and pharmacogenomics. Annual Review of Pharmacology and Toxicology 42, 113–133 (2002)
4. Hewett, M., Oliver, D.E., Rubin, D.L., Easton, K.L., Stuart, J.M., Altman, R.B., Klein, T.E.: PharmGKB: the Pharmacogenetics Knowledge Base. Nucleic Acids Research (NAR) 30(1), 163–165 (2002)
5. Astakhov, V., Gupta, A., Santini, S., Grethe, J.S.: Data Integration in the Biomedical Informatics Research Network (BIRN). In: Ludäscher, B., Raschid, L. (eds.) DILS 2005. LNCS (LNBI), vol. 3615, pp. 317–320. Springer, Heidelberg (2005)

6. Chen, I.A., Kosky, A.S., Markowitz, V.M., Szeto, E., Topaloglou, T.: Advanced Query Mechanisms for Biological Databases. In: 6th Int. Conf. on Intelligent Systems for Molecular Biology (1998)
7. Haas, L.M., Schwarz, P.M., Kodali, P., Kotlar, E., Rice, J., Swope, W.C.: DiscoveryLink: A System for Integrated Access to Life Sciences Data Sources. IBM Systems Journal 40(2), 489–511 (2001)
8. Do, H.H., Rahm, E.: Flexible Integration of Molecular-biological Annotation Data: The GenMapper Approach. In: Bertino, E., Christodoulakis, S., Plexousakis, D., Christophides, V., Koubarakis, M., Böhm, K., Ferrari, E. (eds.) EDBT 2004. LNCS, vol. 2992, pp. 811–822. Springer, Heidelberg (2004)
9. Shaker, R., Mork, P., Brockenbrough, J.S., Donel-son, L., Tarczy-Hornoch, P.: The BioMediator System as a Tool for Integrating Biologic Databases on the Web. In: Workshop on Information Integration on the Web, IIWeb 2004 (2004)
10. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The VLDB Journal 10(4), 334–350 (2001)
11. Leser, U., Naumann, F.: (Almost) Hands-Off Information Integration for the Life Sciences. In: Proceedings of the Conference on Innovative Data Systems Research, CIDR (2005)
12. Mullins, I., Siadaty, M., Lyman, J., Scully, K., Garrett, C., Miller, W., Muller, R., Robson, B., Apte, C., Weiss, S., Rigoustsos, I., Platt, D., Cohen, S., Knaus, W.: Data Mining and Clinical Data Repositories: Insights from a 667,000 Patient Data Set. Journal of Computers in Biology and Medicine (2005)
13. ISO 14721:2003, Blue Book. Issue 1. CCSDS 650.0-B-1: Reference Model for an Open Archival Information System, OAIS (2002)
14. Factor, M., Naor, D., Rabinovici-Cohen, S., Ramati, L., Reshef, P., Satran, J., Giaretta, D.L.: Preservation DataStores: Architecture for Preservation Aware Storage. In: 24th IEEE Conference on Mass Storage Systems and Technologies (MSST), San Diego, pp. 3–15 (2007)