DTI data modeling for unlimited query support

Mohammad-Reza Siadat[a,b], Rafat Hammad[a], Anil Shetty[c], Hamid Soltanian-Zadeh[b,d], Ishwar Sethi[a], Ameen Eetemadi[b], and Kost Elisevich[e]

[a] Dept. of Computer Science and Engineering, Oakland University, Rochester, MI 48309, USA
[b] Radiology Image Analysis Lab., Henry Ford Health System, Detroit, MI 48202, USA
[c] Department of Radiology, William Beaumont Hospital, Royal Oak, MI 48073, USA
[d] Electrical and Computer Engineering Dept., University of Tehran, Tehran 14395, Iran
[e] Dept. of Neurosurgery, Henry Ford Health System, Detroit, MI 48202, USA

ABSTRACT

This paper describes Data Modeling for unstructured data of Diffusion Tensor Imaging (DTI). Data Modeling is an essential first step for data preparation in any data management and data mining procedure. Conventional Entity-Relational (E-R) data modeling is lossy, irreproducible, and time-consuming especially when dealing with unstructured image data associated with complex systems like the human brain. We propose a methodological framework for more objective E-R data modeling with unlimited query support by eliminating the structured content-dependent metadata associated with the unstructured data. The proposed method is applied to DTI data and a minimum system is implemented accordingly. Eventually supported with navigation, data fusion, and feature extraction modules, the proposed system provides a content-based support environment (C-BASE). Such an environment facilitates an unlimited query support with a reproducible and efficient database schema. Switching between different modalities of data, while confining the feature extractors within the object(s) of interest, we supply anatomically specific query results. The price of such a scheme is relatively large storage and in some cases high computational cost. The data modeling and its mathematical framework, behind the scene of query executions and the user interface of the system are presented in this paper.

Keywords: Medical Image Databases, Content-Based Image Retrieval (CBIR), Unlimited Query Support, Medical Informatics

1. INTRODUCTION

The entity-relational (E-R) databases are the most common and well-established type of databases with a wide range of desired features. The first step in any data processing and data mining procedure, and prior to that database design and implementation is data modeling [1]. Conventional E-R data modeling for systems as complex as the human brain and their multimodality unstructured data does not lead to a lossless, feasible, and reproducible result [2]. A lossless data model supports all future ad hoc queries and a feasible data model allows doable database implementation and data entry. A reproducible data model promotes inter- and intra-institutional collaborative work. Unstructured data typically comprise about 85% of an organization's data [3], e.g., audio and video clips, body of an email, human brain images, and segmented models of anatomical structures.

Traditionally, modeling unstructured data (e.g., images) leads to certain structured metadata [4,5] (e.g., volume of an anatomical structure in the human brain) as a set of entities and attributes. The lossy nature of such a modeling scheme makes it impossible to answer questions about features that are not part of the database schema [6]. This impedes unrestricted retrieval support from arbitrary aspects of the unstructured data that characterizes a lossy data model.

It is usually very difficult, if not impossible, to enumerate all features that one can extract from an unstructured piece of data. On the other hand, there is a trade-off between the number of attributes in a data model and its feasibility, e.g., more attributes imply data entry at higher cost and almost unreachable agreement between domain experts. Therefore,

Medical Imaging 2009: Advanced PACS-based Imaging Informatics and Therapeutic Applications
edited by Khan M. Siddiqui, Brent J. Liu, Proceedings of SPIE Vol. 7264, 726415 · © 2009 SPIE
CCC code: 1605-7422/09/$18 · doi: 10.1117/12.813748

even if one could list all features, it might still not be feasible to include all those features in the data model. Therefore, feasibility is an important limitation when dealing with unstructured data.

Knowledge engineers usually design data models in consultation with experts of related fields. Since this is a subjective procedure, the data model is not theoretically reproducible. To reduce this limitation, several practical guidelines have been recommended in textbooks for E-R data modeling [7]. However, when modeling very complex and unstructured data like that of the human brain, the above guidelines are not effective [8]. This is due to the complexity of the system and diverse backgrounds of the consulting experts. In short, conventional E-R data modeling for complex systems with unstructured data presents lossyness, infeasibility, and high degrees of intra- and inter-subject variability.

Kirlangic et al [8] have developed a database system for objective therapy planning and evaluation in epilepsy. They have implemented this system for structuring and managing the associated data for different treatment modalities available for epilepsy. The focus of their work is the electroencephalogram (EEG). They use quantitative EEG (QEEG) measures to lessen the subjectivity of the outcome of the EEG reading. The QEEG as well as the electrode position and timing comprise the neuroprofile as a structured set of possible quantitative measures managed in their database system. Barb et al [6], have studied the well established approaches to content management and image retrieval. They have concluded that most of these approaches lack the flexibility of sharing both explicit and tacit knowledge involved in the decision-making processes. They propose a framework using semantic methods to describe visual abnormalities, offering a solution for tacit knowledge elicitation and exchange in the medical domain. To find related functional neuroimaging experiments, Nielsen et al [9] propose a content-based image retrieval technique. Although frameworks and approaches proposed in the above literature and elsewhere [10-11] contribute heavily to the field of decision support systems in medicine and biomedical databases, they do not directly tackle the problem of unstructured image data modeling and content-based data management.

In this paper, we propose a data model for medical imaging data (including DTI) that excludes the content-dependent structured metadata from the process of E-R data modeling. The rationale for this is that the content-dependent structured metadata modeling can only be manifested through a countless number of attributes, e.g., the following features of an anatomical structure in the brain (i.e., hippocampus): volume, surface, curvature, standard deviation of curvature, average intensity, standard deviation of intensities, etc. On the other hand, there is usually a limited number of unstructured metadata pertaining to a piece of unstructured raw data, e.g., limited number of structures in the human brain. The latter is true with content-independent structured metadata as well, e.g., voxel-size of an imaging study. Coupled with a method to navigate through unstructured data and a set of information extraction and fusion procedures, this scheme provides a content-based support environment. Through its query module, this environment engages appropriate feature extraction procedures confined within the brain's structures of interest to retrieve information regarding any arbitrary aspect of the data. This can be indefinitely expanded and, therefore, provide an unlimited query support. Note that with decreasing costs of storage and processing power, the storage of the entire raw data and their real time analyses becomes increasingly feasible. The proposed design can also be considered as plug-in additions to existing PACS systems, after careful time-requirement considerations. The proposed scheme is designed to be as independent as possible on the choice of segmentation, registration and DTI analysis software.

2. METHOD

We have coined the phrase "Content-Based Support Environment (C-BASE)" for systems built upon databases with unparseable and unstructured raw data, which support these features:

1. Automatic navigation through the raw data
2. Capability to manage (store and retrieve) the meaningful segmented objects and episodes
3. Fusion of several modalities of raw data

Switching between different modalities of data while confining the feature extractors within the object(s) of interest potentially yield descriptive, indicative, and distinctive features of the raw data. Such an environment eliminates the need for the modeling of the content-independent structured metadata, which makes the entire procedure of E-R data modeling more objective and robust. In a nutshell, this summarizes our proposed approach to the problem stated in the

previous Section. In the following sections, details of the proposed content-based support environment for DTI data will be described.

2.1. Content-based support environment for OUB_DTI

To make the contents of the image data accessible to future arbitrary queries, we propose a multimodality image database, which manages the raw data, supported by navigational guidance, multimodality data fusion and feature extraction. We assume that the segmentation and registration information as well as DTI data analysis results are available through medical image processing applications and tools (e.g., 3DSlices, DTI-Studio, MedInria, etc.). This makes the proposed system independent of the user's choice of analysis software or package. The flow of information and main modules of the proposed system are depicted in Fig. 1. Fig. 2 emphasizes on the fact that the proposed system is independent of the choice of data analysis software.

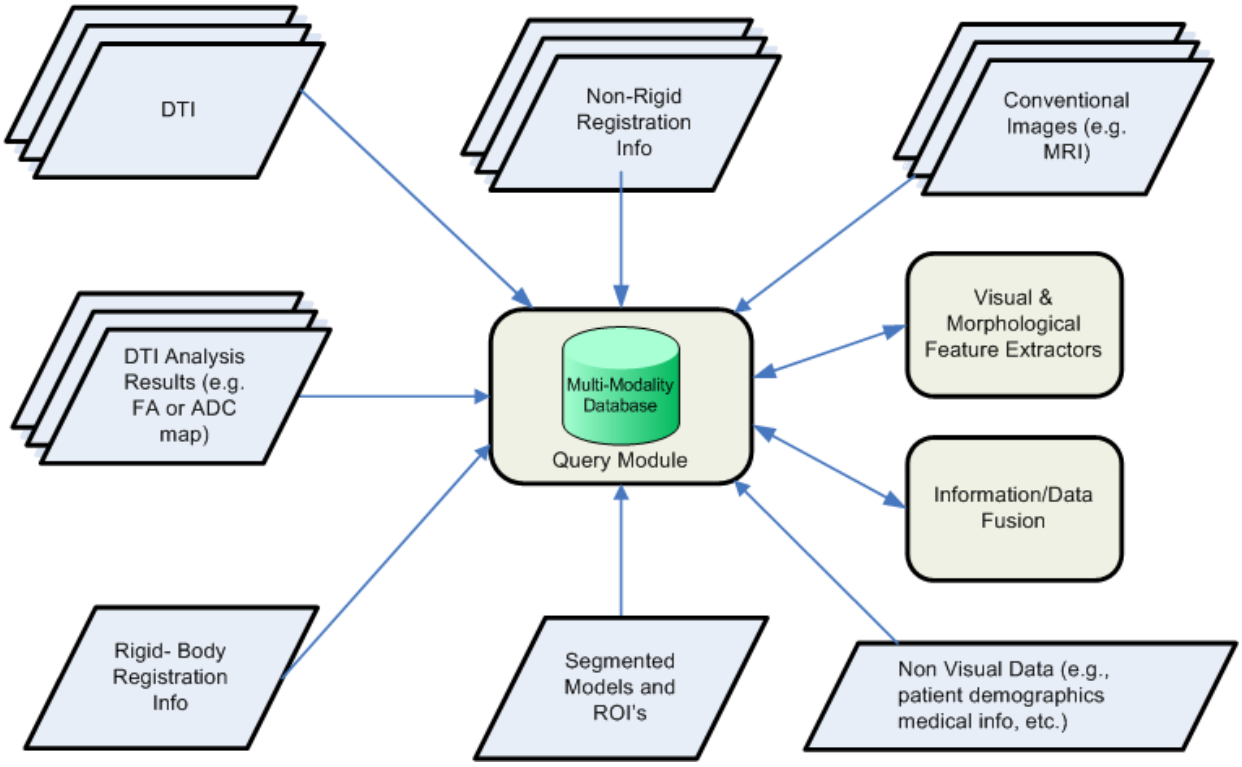

Fig. 1. Flow of information and data in the proposed C-BASE for DTI.

The OUB_DTI is a 3-tier web-based database system, with Oracle Database Management System (DBMS) as the data access tier, Tomcat web server as application tier and a Java-based web browser and related Applets as presentation tier. This design provides all the conveniences supported by web-based applications. On the flip side, this design can suffer when transferring image files over the network, as the medical image data sets are often large. However, ever-increasing speed of the Internet and local networks lessens the negative effect of the above fact. Another major issue of concern when using web-based applications is data security and patients' privacy. We have taken several steps to make the data secure and protect the privacy of the patients. Yet we do not allow any access from outside the hospital. The database is within the secured and firewalled network of William Beaumont Hospital. We are still working towards a secure HIPAA-compliant system to make available through the Internet.

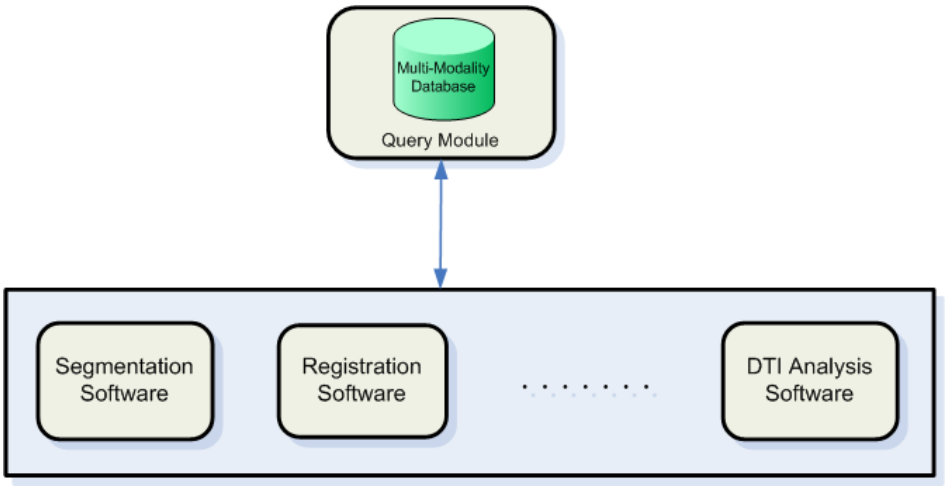

Fig. 2. Image data analysis software-independent design of the proposed DTI database

2.2. E-R data modeling for OUB_DTI

In our application, the goal is to describe the unstructured raw image data so that all aspects of the data, which are of interest, can be queried and mined based on their content in the future. We distinguish between the set of metadata (ε^{C}: content-dependent metadata) that describes the contents of the raw data, and the set of metadata (ε^{NC}: content-independent metadata) that treats the raw data as black boxes and describes them regardless of their contents. Examples of the former and latter cases are the volume of the left hippocampus and the date on which an imaging study has been performed, respectively. We also distinguish between the sets of structured (ε_{SQL}) and unstructured (ε_{NSQL}) metadata. We consider any data that cannot be *directly* used in a basic SQL statement as unstructured, e.g., audio and video clips, body of an email, human brain images, and segmented models of brain structures. We propose to exclude the content-dependent structured metadata in the schema. In other words, the set of metadata that are included in our data model is: $\varepsilon_{SQL}^{NC} \bigcup \varepsilon_{NSQL}^{NC} \bigcup \varepsilon_{NSQL}^{C}$. Table I shows two examples of several levels of data and metadata that the application deals with. The ones that are included in the E-R data modeling are in gray. This table shows that we segment and store anatomical structures (e.g., hippocampus, insula, perisylvian area, corpus callosum, amygdala) in the database, however, the content of the segmented model (e.g., average curvature) will not be included in the database schema as separate entities since this piece of information is content-dependent and structured (ε_{SQL}^{C}). One can simply imagine that such features are almost countless, and therefore, it is not worth including them in the database schema. Similar to the segmented models, the registration information will be part of the database schema since it is in ε_{NSQL}^{C}. Instead, we store all the extracted features (average curvature and intensity, volume, etc.) in an all purpose attribute (APA) as part of a table with a one-to-many relationship to the table that contains the corresponding ε_{NSQL}^{C} item. Fig. 3 shows the E-R data model designed based on the proposed method.

Table I. Examples of data and metadata levels dealt with in the proposed E-R data modeling.

Level	Example I	Example II
Data	Image data, e.g., DTI data	1D signal data, e.g., ECoG signal
Metadata	Segmented structure, e.g., hippocampus model	Electrode location, (x,y,z) coordinates
Meta-metadata	Features of the segmented structure, e.g., average curvature, ...	
…		

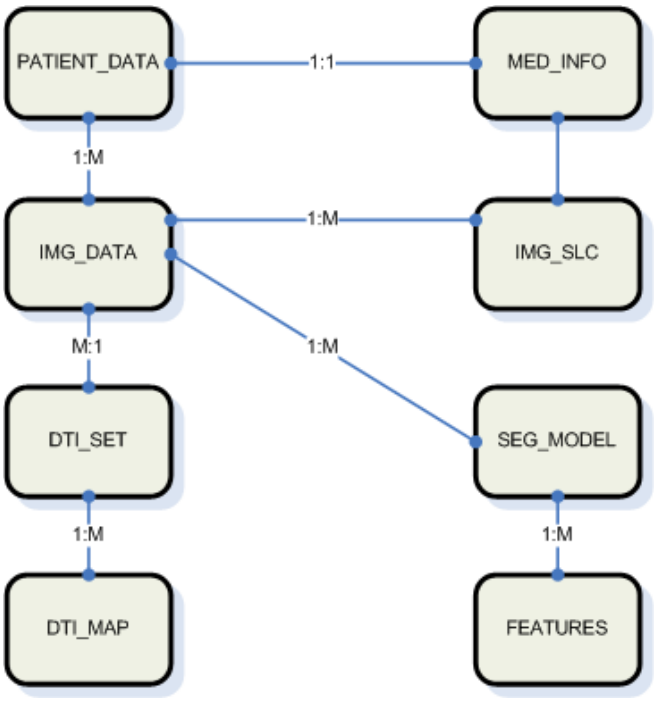

Fig. 3. E-R data model for DTI database.

3. UNLIMITED CONTENT-BASED QUERY SUPPORT FOR OUB-DTI

When querying unparseable raw data stored in a database, one of the following scenarios can be followed:

1. Calculating the quantitative measures of interest using stand-alone software to produce structured data. Adding new tables and attributes to the existing database schema to keep track of the calculated structured data.
2. Integrating the feature extraction routines that calculate the quantitative measures of interest as functions or operators into the query module and make them available within the SQL code.

The advantages of the first scenario are: a) It does not need anything but the requirements of the conventional database management systems; and, b) The calculated quantitative measures will be permanently stored in the database and can be retrieved quickly in the future. The disadvantages of the first scenario are: a) The data model of the database will be constantly changed; b) Managing such variable data model will be difficult, as each user may add new items to query the raw data from their own standpoint, which can produce an endless number of tables and attributes; and, c) The end user of such a system needs to have a high level of expertise in database management systems to add new tables and attributes with correct relationships and within the right tables, respectively. Note that the end users are usually experts in biological and medical fields with limited knowledge of DBMS. In addition, as it has been discussed before in great detail, this is a subjective matter and it is almost impossible to resolve disagreements between different users. Finally and most importantly, this scenario leads to a system with limited query support in practice. On the other hand, the advantages of the second scenario are: a) All features supported by DBMS will continue to be supported by this scheme (e.g., security and privacy at the database level) since the function can encapsulate values that the user does not have the privilege to access or modify; b) All features supported by SQL will be available to the end user; c) It will have the capability to offer a unified integrated interface for query composition module; and, d) There will be no need to change the data model, eliminating the headache of managing a database with variable schema. The disadvantage of the second scenario is that it demands higher expertise in programming during the development phase. Note that only the developers are supposed to meet this requirement and not the end users. The first scenario is more appealing if there are only a few new fields to be added throughout the life expectancy of the database. This may imply that in this situation, the data is intrinsically less unstructured and less unparseable. Therefore, as we move towards non-conventional

applications that use mostly unstructured data, e.g., brain image databases, the advantages of the first scenario fade out and the second scenario will be more appealing.

The functionalities proposed in the second scenario can be implemented through PL/SQL (procedural language/SQL) or some dynamic link libraries (DLL). When using DLL, an all purpose PL/SQL communicates with functions available in DLL files when each function extracts a feature of interest from the raw data. PL/SQL is a procedural extension of the Oracle-SQL that offers language constructs similar to those in imperative programming languages. This scheme allows adding an unlimited number of PL/SQL or DLL functions to extract any arbitrary feature from the raw data. In some cases, the binary images need to be passed to the DLL function as its parameter. Since the PL/SQL cannot pass the binary data, we need a mediator to retrieve the required data from the database and to make it available to the DLL function through the OCI (Oracle Call Interface). The above scheme allows the users to take advantage of all functions available in a DLL file within their SQL code. The flowchart in Fig. 4. shows an example of a generic PL/SQL routine (called MY_AVG) for calculating the average intensities within a given segmented structure in a given image space or DTI analysis map. Here is a query example in which this PL/SQL routine is used:

```
SELECT IDNUMBER, SZ_CLASS FROM MED_INFO
WHERE (MY_AVG(DPSA_ADJ, 'L', FA_MAP, 'AVG_DPSA_ADJ_FA') > 1.2 * MY_AVG(DPSA_ADJ, 'R', FA_MAP, 'AVG_DPSA_ADJ_FA') AND SURGERY_SIDE = 'L') OR (MY_AVG(DPSA_ADJ, 'R', FA_MAP, 'AVG_DPSA_ADJ_FA') > 1.2 * MY_AVG(DPSA_ADJ, 'L', FA_MAP, 'AVG_DPSA_ADJ_FA') AND SURGERY_SIDE = 'R')
```

The PL/SQL routine MY_AVG(SEG_NAME, SIDE, IMAGE/MAP, FEATURE_NAME) calculates the average intensity within the left or right side (SIDE) of the SEG_NAME structure within IMAGE/MAP image space and stores the result in an APA called FEATURE_NAME and finally returns the result to the SQL process. More specifically, MY_AVG(DPSA_ADJ, 'L', FA_MAP, 'AVG_DPSA_ADJ_FA') computes the average intensity of fractional anisotropy (FA_MAP) within the segmented structure called DPSA-ADJ (which is the area adjacent to the deep perisylvian area (DPSA)) on the left side of the brain and the result will be stored in FEATURE_VALUE where the FEATURE_NAME = 'AVG_DPSA_ADJ_FA.' The flowchart in Fig. 4 shows how MY_AVG routine performs its job. This flowchart is quite generic and can be easily adopted for other features (e.g., standard deviation, volume, average curvature, texture, etc.). The MY_AVG starts with checking to see if the value for the feature called 'AVG_DPSA_ADJ_FA' is available in corresponding APA. If it is available, then the value would be retrieved and returned. Otherwise, this routine checks to see if the segmented structure called 'DPSA_ADJ' exists. If it does not, the routine returns -1 which means the value can not be computed. Otherwise, it checks to see if any DTI data has been registered in the native image space of the 'DPSA_ADJ,' if there exists no such DTI data, it means there is no FA_MAP space that can be fused into the native image space and thus it returns -1. If there is such DTI data, then this routine checks to see if there exists a FA_MAP calculated for the DTI data, if no such FA_MAP exists, then it returns -1. Otherwise, all the necessary pieces of data exist and it needs to retrieve segmented model of DPSA_ADJ, registration information that transfers the FA_MAP to native image space of the segmented model and finally it retrieves the required portion of FA_MAP raw data. The next step is to transfer the retrieved FA_MAP to the native image space of DPSA_ADJ model. In practice, we usually transfer the segmented model from its native image space to the image space that the query requires (in the above case, FA_MAP) using the inverse of the registration transformation since this is usually easier. The last step is to calculate the average of FA_MAP pixels within DPSA_ADJ model and to store this value in corresponding APA and to return the value to the ongoing SQL process. Such query can be instrumental in determining whether epileptogenic foci can coexist with higher FA on the gray-white matter interface and in white matter substance adjacent to the foci.

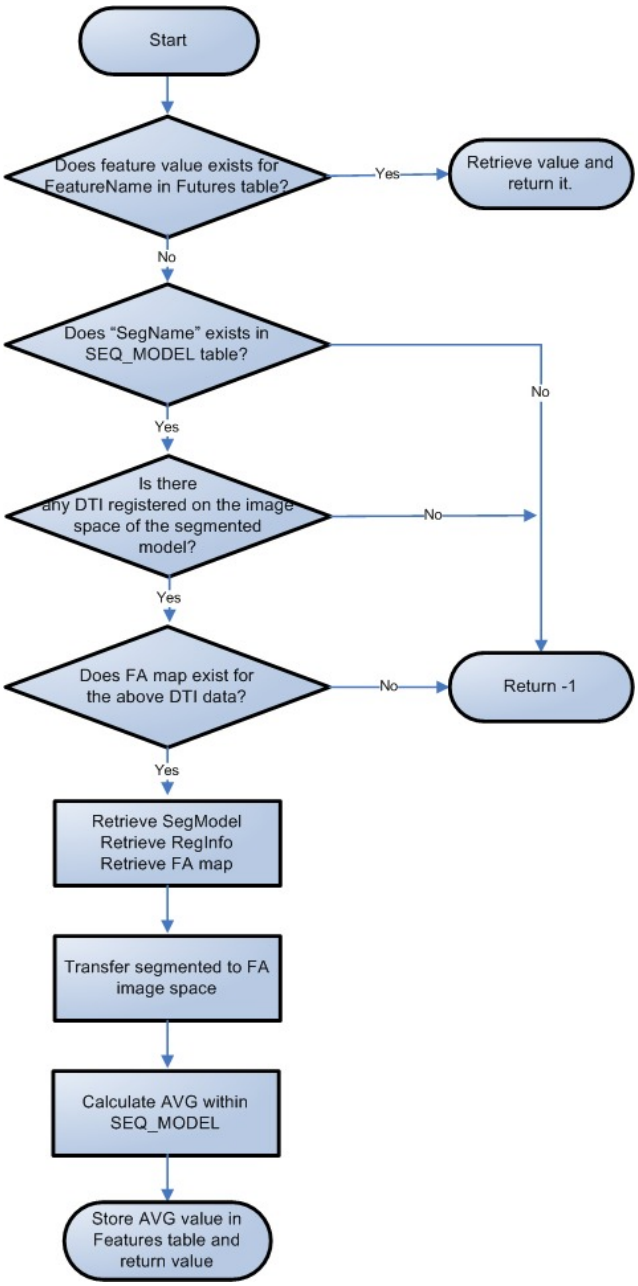

Fig. 4. MY_AVG PL/SQL procedure flowchart.

4. PRELIMINARY RESULTS AND CONCLUSIONS

We have currently developed several forms to allow for storing , browsing/querying and deleting the DTI-related patients data. Fig. 5 shows the available image data of a given patient. This form can be used to insert new dataset through "New Image Data" or "Import Image" buttons. The former provides a manual means of entering the data and the latter requires the user to browse through the file system and select the images that are to be stored to the database and the rest of information will be retrieved from the header of the DICOM files. The "Delete" and "Edit" buttons let the user to delete, and update the existing tuples, respectively. The tabs at the top of the page provide the users with an

arbitrary access to patient medical information, image data, DTI set, segmented models and/or ROIs, and features. In the medical information form, one needs to select a patient to be able to browse through Image Data form. This is done by generating a special session for the selected patient and retrieving only the corresponding data to be displayed on Image Data form. Windows with proper messages will pop up in case correct browsing sequence is not followed or when the transaction involves a referential constraint etc.

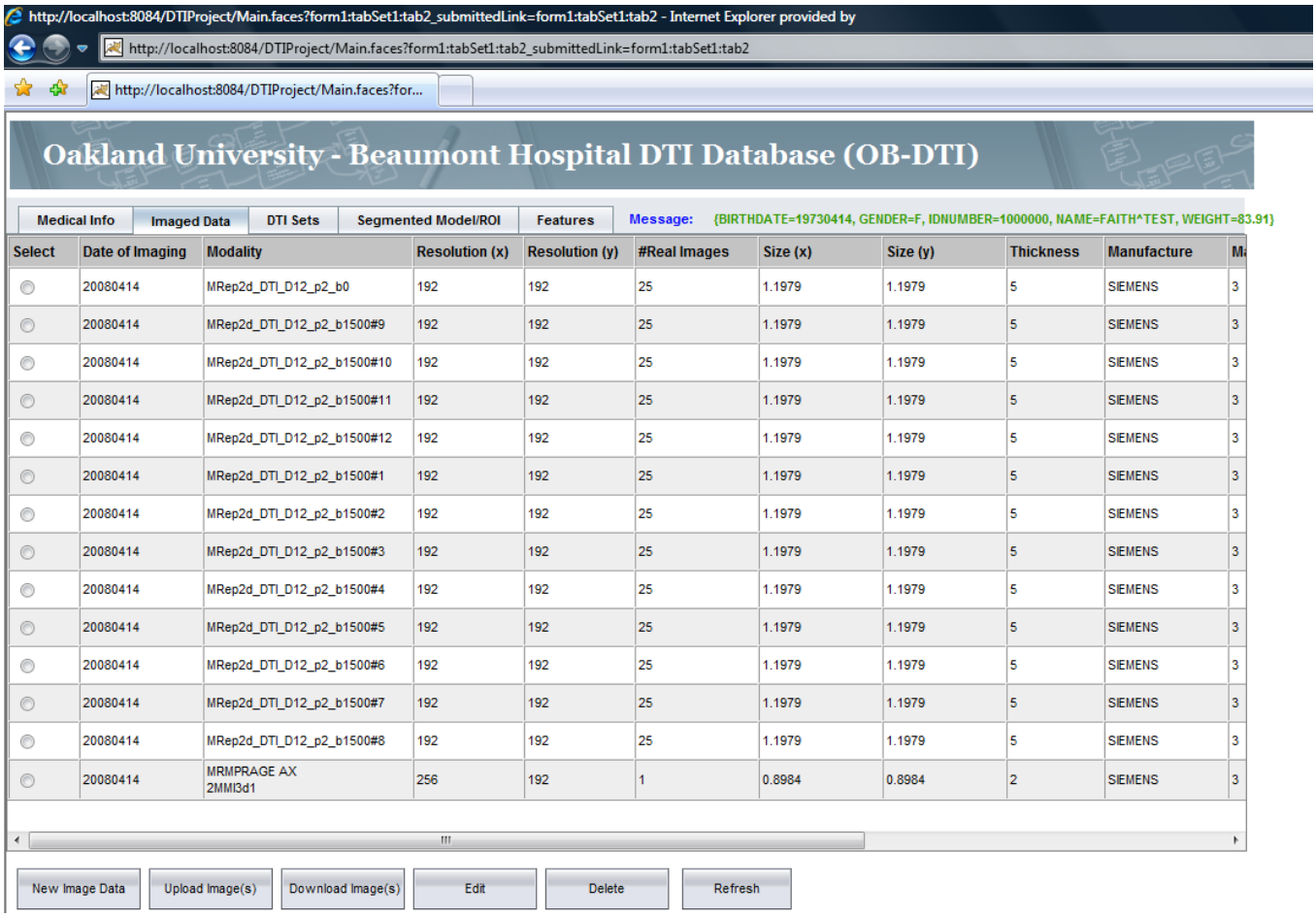

Fig. 5. A form showing "Image Data" with available modalities of data for a given patient, some information about each modality of data is also shown.

Fig. 6 shows the underlying E-R schema at the database level with details of the attributes. In this design there is no content-dependent structured metadata attributes except for FEATUREVALE in FEATURE table. This attribute is an APA that keeps track of all results of the user-defined functions available in the system. One can compute many DTI maps and results from several modalities of DTI data. Therefore, there has to be a many-to-many relationship between DTI_MAP and IMG_DATA tables. The classic way to deal with this situation is to create an new table and have many-to-one relationships between DTI_MAP and IMG_DATA, and the new table. We have named the above table DTI_SET where there is a one-to-many relationship from this table to both DTI_MAP and IMG_DATA tables. The user interface design include a page linked to the "DTI Sets" tab in the main page and the one shown in Fig. 7 that where user can define new DTI sets and corresponding DTI modalities that contribute to calculation of a DTI map. e.g., FA map.

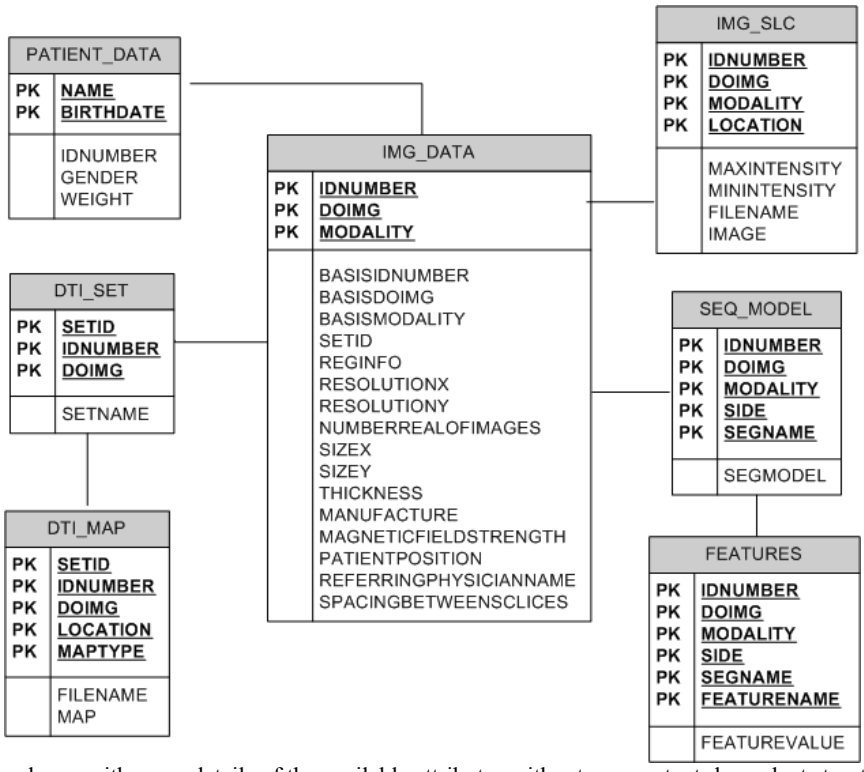

Fig. 6. The E-R schema with some details of the available attributes without any content-dependent structured metadata.

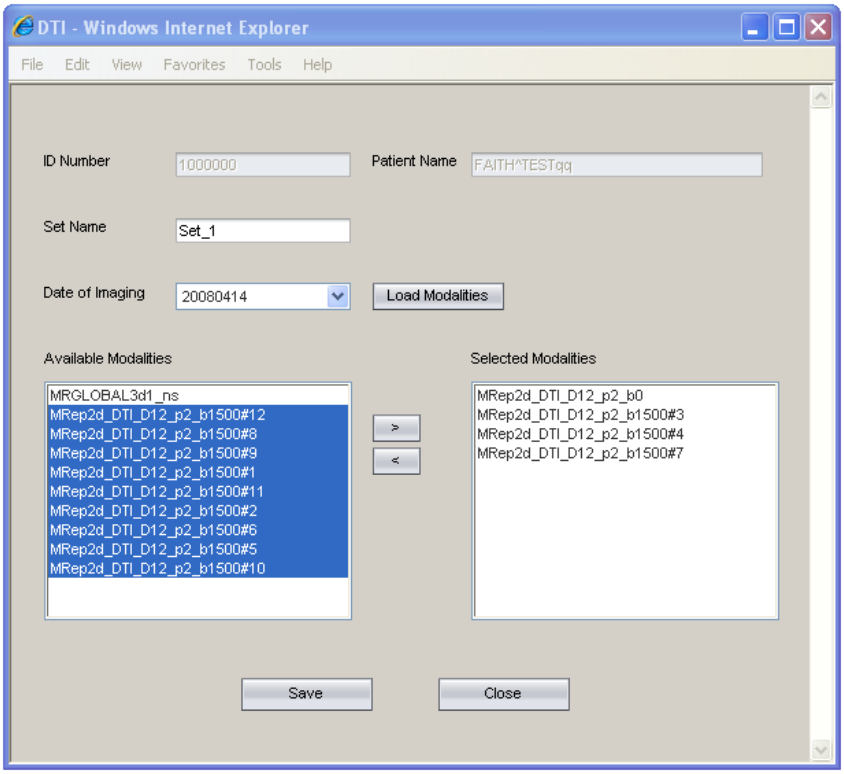

Fig. 7. User interface to create new DTI sets and to define the modalities of DTI that are included in the set.

The proposed system renders efficient analysis of multimodality brain image data including DTI for individual and comparative group studies. The OUB-DTI can be adopted and fine-tuned for several studies, e.g., epilepsy, brain tumor. As a benefit of adding segmentation to the database, the semantic contents of the image data can be queried. The raw image data could be directly retrieved from the hospital's PACS system; however, there are usually limitations in terms of access to those commercial systems since there has to be a guaranteed response time. Systems similar to OUB-DTI that offer content-based unlimited query support are crucial parts of the future developments in evidence-based medicine, decision support systems and medical data mining.

ACKNOWLEDGEMENT

This work has been partly supported through an Oakland University-William Beaumont Hospital multidisciplinary research award. We would like to thank Mr. Sarmad Istephan for editing this article.

REFERENCES

[1] Caverlee, J., Liu, L., "QA-Pagelet: Data Preparation Techniques for Large-Scale Data Analysis of the Deep Web," IEEE Trans. on Knowledge and Data Engineering, vol. 17, no. 9, 2005.
[2] http://en.wikipedia.org/wiki/Data_modeling.
[3] D. Robb, "Getting the Bigger Picture: Dealing with Unstructured Data," in Storage features (http://www.enterpriseitplanet.com/storage/features/article.php/3407161), 2004.
[4] D. Marco, Building and Managing the Meta Data Repository: A Full Lifecycle Guide: Wiley, 2000.
[5] T. Eiter, J. J. Lu, T. Lukasiewicz, and V. S. Subrahmanian, "Probabilistic Object Bases," ACM Transactions on Database Systems, vol. 26, pp. 264-312, 2001.
[6] A. S. Barb, C.-R. Shyu, and Y. Sethi, "Knowledge Representation and Sharing Using Visual Semantic Modeling for Diagnostic Medical Image Databases," IEEE Trans Inf Technol Biomed, vol. 9, pp. 538-553, 2006.
[7] R. Elmasri and S. B. Navathe, Fundamentals of Database Systems, 3rd ed: Addison-Wesley Longman 2000.
[8] M. E. Kirlangic, J. Holetschek, C. Krause, and G. Ivanova, "A database for therapy evaluation in neurological disorders: application in epilepsy," IEEE Trans Inf Technol Biomed, vol. 8, pp. 321-32, 2004.
[9] F. A. Nielsen and L. K. Hansen, "Finding Related Functional Neuroimaging Volumes," Artificial Intelligence in Medicine, vol. 30, pp. 141-151, 2004.
[10] D. S. Obrosky, S. M. Edick, and M. J. Fine, "The Emergency Department Triage of Community Acquired Pneumonia Project Data and Documentation Systems: A Model for Multicenter Clinical Trials," IEEE Trans. on Information Technology in Biomedicine, vol. 10, pp. 377-384, 2006.
[11] A. G. Anagnostakis, M. Tzima, G. C. Sakellaris, D. I. Fotiadis, and A. C. Likas, "Semantics-Based Information Modeling for the Health-Care Administration Sector: The Citation Platform," IEEE Trans. on Information Technology in Biomedicine, vol. 9, pp. 239-247, 2005.