Bio-Image Warehouse System: Concept and Implementation of a Diagnosis-Based Data Warehouse for Advanced Imaging Modalities in Neuroradiology

L. Minati,[1,2] F. Ghielmetti,[2] V. Ciobanu,[3] L. D'Incerti,[2] C. Maccagnano,[2] A. Bizzi,[2] and M. G. Bruzzone[2]

Advanced neuroimaging techniques, such as functional magnetic resonance imaging (fMRI), chemical shift spectroscopy imaging (CSI), diffusion tensor imaging (DTI), and perfusion-weighted imaging (PWI) create novel challenges in terms of data storage and management: huge amounts of raw data are generated, the results of analysis may depend on the software and settings that have been used, and most often intermediate files are inherently not compliant with the current DICOM (digital imaging and communication in medicine) standard, as they contain multidimensional complex and tensor arrays and various other types of data structures. A software architecture, referred to as Bio-Image Warehouse System (BIWS), which can be used alongside a radiology information system/picture archiving and communication system (RIS/PACS) system to store neuroimaging data for research purposes, is presented. The system architecture is conceived with the purpose of enabling to query by diagnosis according to a predefined two-layered classification taxonomy. The operational impact of the system and the time needed to get acquainted with the web-based interface and with the taxonomy are found to be limited. The development of modules enabling automated creation of statistical templates is proposed.

KEY WORDS: PACS, image repository, data warehouse, web-based, neuroradiology

INTRODUCTION

C linical neuroradiology is witnessing a profound change from being based mainly on morphological analysis toward a novel approach in which several types of functional data have central importance. For example, functional magnetic resonance imaging (fMRI), chemical shift spectroscopy imaging (CSI), diffusion tensor imaging (DTI), and perfusion-weighted imaging (PWI) are set to become routine use techniques delivering information on the localization of eloquent areas, on the concentration of the major neurometabolites, on the layout of the major fiber tracts, and on perfusion. Their growing clinical importance and the ability to acquire these data without any hardware intervention during MR sessions will fuel their rapid diffusion in the forthcoming years.[1,2]

Alongside with new opportunities, these techniques generate novel challenges. In contrast with morphological MR imaging, processing the raw data that these techniques generate is not straightforward, and the final output obtained may considerably depend on the algorithms and parameters that have been used. Moreover, advanced MR techniques tend to generate huge amounts of raw data, on the order of 1 GB for a typical

DTI–fMRI examination. The situation is further complicated by the fact that many of the intermediate files involved in the processing of fMRI, CSI, DTI, and PWI data are not compliant with the current DICOM (digital imaging and communication in medicine) standard, and they cannot be made such given that they contain multidimensional complex arrays, tensor values, and various types of data structures. Therefore at present, they often cannot be stored on a picture archiving and communication system (PACS).[3]

The task of creating radiological teaching files has traditionally been accomplished by downloading images from the PACS system, saving them in a widely accepted format such as JPEG (Joint Photographic Experts Group) or TIFF (tagged image file format), and inserting them into a structure, for example, a slideshow or a folder. Such an operation can be performed manually, or with the support of systems such as the one described by Rosset et al.[4] Although handling images in this manner has excellently served teaching purposes for many years, the stored images are usually not suitable for postprocessing.

A unitary storage suitable for both DICOM images and postprocessed data is called for. The rapidly growing amount of research underway on advanced imaging techniques fuels a need for storage solutions enabling rapid group retrieval by radiological and pathological classification, for educational and statistical purposes; several image databases dedicated to educational purposes have been proposed.[4–6]

Most notably, the MyPacs.net project, based on Web Interfacing Repository Manager (WIRM) technology initially developed in the context of the UW Human Brain Project, has extended the concept of web-based repositories to a freely accessible Internet portal.[7] At the time of writing, the repository contains over 10,000 cases, spanning over the majority of anatomical districts of radiological interest.[8] Although the MyPacs.net project is primarily devoted to the storage of morphological images, in 2001 a multicentric National Institutes of Health (NIH)-led effort known as the Biomedical Informatics Research Network (BIRN) was started, with the purpose of sharing functional and morphological data sets, study protocols, and processing capabilities among U.S. organizations.[9]

We report a preliminary experience with the development of a web-based repository, akin to the MyPacs.net project in its structure, designed for storage of raw and postprocessed data from advanced neuroimaging techniques, meant to form the basis for a future multicentric cooperation infrastructure analogous to the BIRN.

ARCHITECTURE

Two computed tomography (CT) scanners and two magnetic resonance (MR) scanners are in operation at our institution; one of the scanners, a 0.5-T system, is used only for conventional imaging, whereas the other, a 1.5-T system, performs an increasing number of fMRI, CSI, DTI, and PWI examinations in the clinical routine and for research purposes, currently accounting to about 20% of its load.

From a management viewpoint, our choice was to centralize storage for clinical purposes on the PACS system, and to centralize storage for research purposes on a system to be developed in-house, referred to as Bio-Image Warehouse System (BIWS), whose key feature had to be its ability to store the non-DICOM file types in use (e.g., tensor field data files used in fiber tracking), alongside with DICOM images.

There are several major differences between the PACS and BIWS systems. Although the PACS system is intended for permanent legal storage of all examinations, only studies deemed to be of research relevance are to be sent to the BIWS system. Although the PACS system can only store bitmap image data, it was decided that all intermediate files resulting from the postprocessing of fMRI, DTI, PWI, and spectroscopy examinations deemed of research relevance should strictly be stored in the BIWS system. In terms of retrieval process, although the radiology information system (RIS)/PACS system is mainly targeted at retrieval by patient name and exam date, the BIWS system was designed mainly to enable group querying by neuroradiological, neuropathological, or clinical diagnosis. Finally, although data are always permanently stored on the PACS system, and retrieval may sometimes take several minutes (due to jukebox queuing), the BIWS system was designed to enable deletion of cases under predefined policies (by a restricted set

of users, eg, following invalidation of a previous diagnosis), and was designed accordingly in order to have all of its storage on-line.

As found with other radiology applications, our current prototype implementation of the BIWS system is fully web-based.[6–13] We chose as server a Pentium 4 machine, at the present time having 1600 GB of HDD storage, networked through a 1000BASE-T link. The software architecture for implementation of the BIWS system was chosen by the software engineer (V.C.). Because of its performance and security characteristics, the OpenBSD 3.8 system was chosen. Most of the BIWS code has been written in PHP, with some DICOM file parsing and database access modules having been developed in C++ following performance considerations. Because of their popularity and security, the mySQL 5 database engine and the Apache web server provided with the operating system have been employed. Such a set-up has already proven effective for similar applications.[6]

A custom-designed DICOM storage node, implemented in C++, is used to receive images from CT and MR consoles, and from other DICOM work stations. Images received through this DICOM node are put in a temporary buffer, and their headers immediately read; to prevent accumulation of unnecessary files, images remain available within this temporary buffer only for 48 h, waiting for user action to transfer them to permanent storage. Moreover, a DICOM client library has been developed, in the C++ language, which is used to retrieve images from the PACS system.

The BIWS system can store images and other files in a directory hierarchy, reflecting patient exam series levels. The main conceptual difference with DICOM nodes is that object entities may contain DICOM series, or sets of non-DICOM files. To enable backup to be performed on DVD-R disks, the system limits the size of each directory to 4.5 GB. BIWS objects are stored as TAR files. A cron-tab script compresses TAR files into BZIP2 format between 00:00 a.m. and 06:00 a.m., thus optimizing usage of storage space, without loss of information. A set of PHP scripts and C++ programs is periodically invoked by the cron-tab, performing maintenance operations on the SQL database, as well as system health checks (number of processes running, number of open and idle sessions, etc.). A small set of Java applets is used on the web client side,

for image and data visualization purposes. At present, the BIWS system is only able to provide web-based preview of DICOM-compliant images, whereas all other file types need to be downloaded and visualized off-line.

The BIWS system will soon be extended in order to be able to import data from our RIS and EMR (Electronic Medical Record) databases, through dedicated modules written in C++.[14]

Access security is implemented through individual login/password identification, strict IP filtering policy allowing accesses only from a predetermined list of computers, and full logging of all database operations.

The resulting system architecture is depicted in Figure 1, with dashed lines representing modules under development.

After definition of the system architecture, the operational protocol, data fields, and diagnosis taxonomy were discussed in roundtable form with radiologists, consultants, and technologists.

PROCEDURES, DATA FIELDS, AND CLASSIFICATION TAXONOMY

As a result of roundtable discussions, the following criteria were chosen.

When it is known in advance that an examination will be of research interest, technologists are instructed from the beginning to send data during the examination to the DICOM node in the BIWS server. If an examination is found to be of research interest during the session, it is sent to the BIWS server at completion. If an examination is found to be of research interest afterwards, it can be sent to the BIWS server from any DICOM work station, or raw DICOM images can be retrieved from the PACS system through the web-based interface.

On completion of reporting, the radiologist is to access the BIWS web interface from any computer.

A database architecture based on three entity types, as represented in Figure 2, was chosen as the most appropriate. Registration of a new examination is accomplished in three steps, corresponding to three types of entities: retrieval of an existing patient or addition of a new patient, addition of a new examination, and addition of new objects. The patient entity contains standard fields (such as patient name, date of birth, sex, and referring doctor) whose values can be picked from
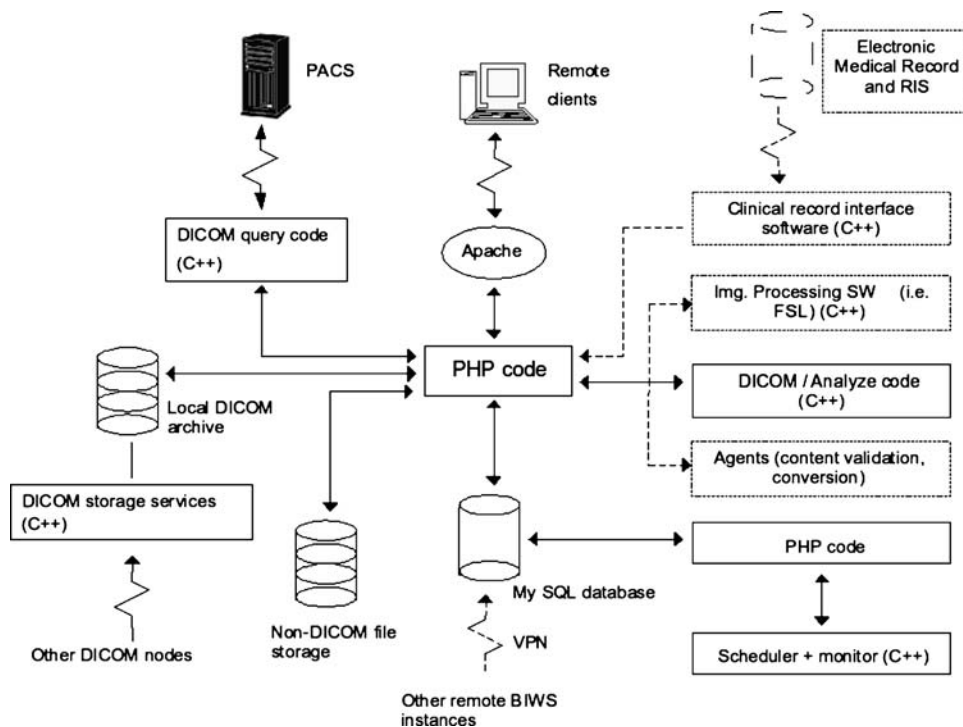
Fig 1. System architecture.

DICOM images of the patient. In the future, a link will enable a foreign key to be established to the electronic medical record (EMR) of the patient, if available, in a way conceptually similar to some PACS systems described elsewhere.[14] Such a link will enable the user to directly access the medical record of the patient by requesting the relevant web page from the web server of the EMR system, without duplication of information, and without introducing complex interactions between the BIWS and EMR databases. This is akin to a star architecture.

Alongside standard fields (such as performing physician, exam date, modality, scanner type), the examination entity contains fields that are central to the BIWS concept of enabling diagnosis-based retrieval. To prevent confounding effects from variability in naming conventions, each examination is classified according to a taxonomy that has been commonly agreed among the neuroradiologists at our institution.

To be usable, such a taxonomy must eliminate or minimize classification ambiguities, and meet a trade-off in being sufficiently analytical while keeping the number of classes and levels down at a practical number. Such a need has been recently acknowledged through the establishment by Radiological Society of North America (RSNA) of the RadLex project, aimed at the development of a unified radiological lexicon.[15] At the time of writing, a RadLax lexicon exists for neuroradiological anatomy, but not for diagnoses of neuroradiological relevance. As we deemed classification according to type of diagnosis central to the BIWS concept, a custom taxonomy had to be developed.

A dedicated panel comprising eight neuroradiologists from our institution agreed on a classification structured on two layers, as a compromise between analytical detail and ease of use.

The entries in the upper layer are divided in two separate sections, that is brain and spine, which are further subdivided. As established in a number of neuroradiology textbooks, it was found to be more natural to classify some conditions by the affected anatomical area, rather than by type of pathology.[16] The brain section in the upper layer therefore comprises entries corresponding to pathology groups, such as stroke and trauma, and entries corresponding to anatomical areas, such as orbit and auditory canal. Table 1 reports the structure of the upper layer of the classification taxonomy. Each branch of the lower layer con-
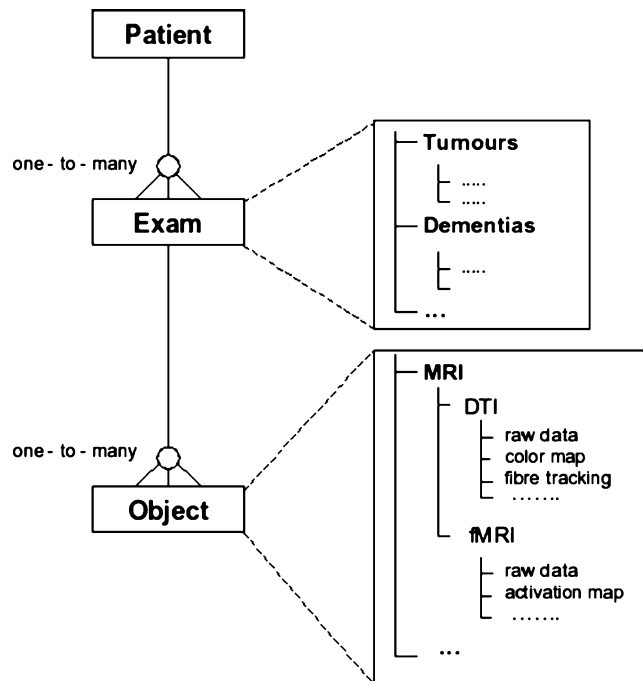
Fig 2. Database entity diagram.

tains an average of 10 entries corresponding to specific pathologies.

The usage of a classification taxonomy is necessary because, even though diagnostic data are available in the RIS system, they are stored in free-text form—which is not ideal for performing mass queries, because of the very large amount of text to be searched, and because of the potentially confounding effects of variability in naming conventions. Although the classification taxonomy reported in Table 1 is specific for neuroradiology, it could be naturally extended to other pathologies of interest, if necessary.

A "Location" field was included, containing a list of combinations of the form "Hemisphere(s), Lobe(s)," to be used only for focal pathologies. Furthermore, a field representing the type and confidence of diagnosis was included, possible choices being: "Confirmed," "Neuroradiological diagnosis, likely," "Neuroradiological diagnosis, possible," "Clinical diagnosis, likely," "Clinical diagnosis, possible." With the exception of follow-up examinations, in the context of neurological pathologies diagnosis is usually not certain at the time of imaging. As a consequence, radiologists entering new cases need to take care of updating the fields in the examination entity, changing or confirm-

ing the diagnosis once it is finalized. A free-text field captures in natural language a short description of why the case was deemed to be of research interest.

Each object entity can be linked either to a DICOM series, extracted from the local DICOM storage or from the PACS system through a query, or to a TAR archive containing any type of files, uploaded with the file transfer (FTP) or hypertext transfer protocol (HTTP).

A two-layered taxonomy was used also to classify the content of objects. The upper layer contains entries corresponding to imaging techniques, that is, "CT," "MRI—0.5 Tesla," "MRI—1.5 Tesla," "Functional MRI," "MR Spectroscopy/Chemical Shift Imaging," "Diffusion Tensor Imaging," "Perfusion Weighted Imaging." Furthermore, a "Nonimage data" entry was added, whose lower layer branch contains the following entries: "Neuroradiological Report," "Clinical Report," "Histological Report," "Surgical/Cortical Mapping Record," "Relevant Article," "Notes." The lower layer branch for each imaging technique contains entries corresponding to all relevant file types in use.

Figure 3 reports the appearance of some pages from the current web interface. The main page, which includes links for addition and retrieval of

Table 1. Entries in the upper layer of the classication taxonomy

Brain	Spine
B1. Anatomical variants	S1. Anatomical variants
B2. Cranial congenital abnormalities	S2. Postop changes
B3. Hydrocephalus	S3. Congenital malformations
B4. Chiari malformations	S4. Spinal trauma
B5. Cerebellar malformations	S5. Infectious and inflammatory disease
B6. Neurocutaneous syndromes	S6. Spine tumours
B7. Malformations of cortical development	S7. Spinal vascular disorders
B8. Paediatric neurodegenerative diseases	S8. Degenerative diseases
B9. Preterm infants	
B10. Hypoxic–ischemic encephalopathy	
B11. Paediatric metabolic disorders	
B12. Congenital–neonatal infections	
B13. Intracranial haematomas	
B14. Trauma	
B15. Intracranial vascular malformations and aneurysms	
B16. Primary nonneoplastic cysts	
B17. Glial tumours	
B18. Mixed neuronal–glial tumours	
B19. Nonglial tumours	
B20. Pineal tumours	
B21. Meningeal tumours	
B22. Germ cell tumours	
B23. Pituitary	
B24. Stroke	
B25. Intracranial infections	
B26. Neurodegenerative disorders	
B27. Demyelinating diseases	
B28. Adult toxic and metabolic disorders	
B29. Radiation therapy sequelae	
B30. Chemotheraphy sequelae	
B31. Epilepsy	
B32. Intracranial hypotension	
B33. Dural abnormalities	
B34. Bone abnormalities	
B35. Temporal bone	
B36. Internal auditory canal	
B37. Orbit	

patients, examinations, and objects, is depicted in Figure 3a. The object query page, including search criteria at the patient, examinations, and object levels, is shown in Figure 3b. The exam view page, including the exams fields and the links to objects, is shown in Figure 3c.

Four permission levels have been defined: "Guest," "Radiologist," "Editor," and "Administrator." Guests can execute queries, but they cannot perform any addition or modification of existing data. A radiologist can add new patients and exams, but modify or add objects only to exams she/he added. An editor can perform additions, deletions, and modifications without restrictions. In addition to editing rights, administrators can manage users.

It was not possible to define an objective criterion to determine whether an examination should be stored on the BIWS system. As a consequence, at the present time such a decision is left to the arbitrary choice of the neuro-radiologists.

INITIAL EXPERIENCE

The time needed to enter patient, examination, and objects through the web-based system ranged between 3 and 6 min, thereby having a negligible operational impact, compatible with the clinical routine of our department.
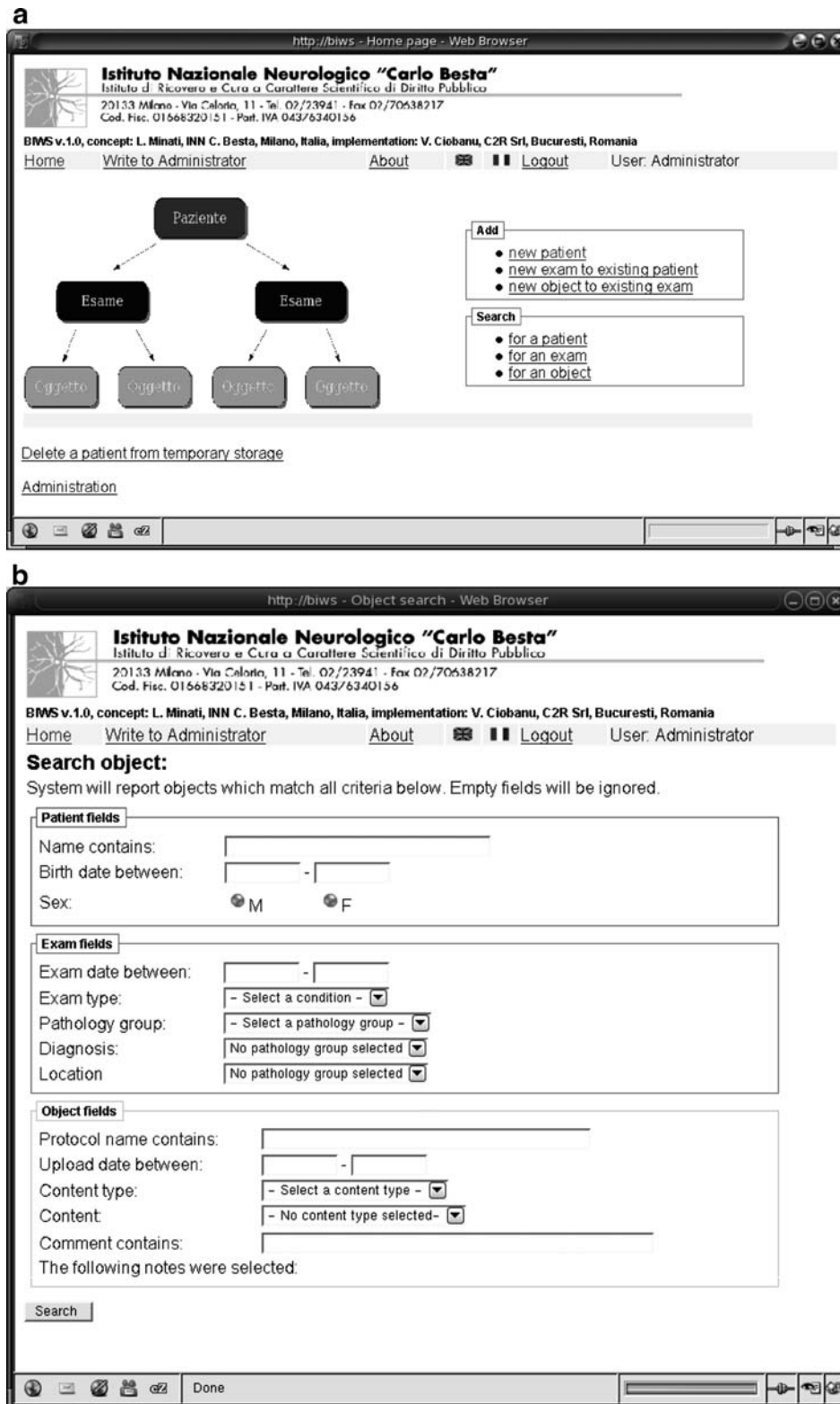
Fig 3. Web interface of the system: (a) main page, (b) object query page, and (c) exam view page.

Fig 3. Continued

About six neuroradiologists from our institution and five residents/guests used the system. They were able to get acquainted with the classification scheme, to perform queries, and to learn how to classify new cases in an average of about 3 h. Such a finding confirmed that the system held a promise of deindividualizing classification schemes, and of enabling a large database of cases to be shared among neuroradiologists from different institutions.

At the time of writing, about 200 cases (patients) have been stored on the BIWS system, corresponding to about approximately 230 examinations and 1,150 objects. Although a formal survey has not yet been conducted, major usability obstacles were not reported. The majority of users found that getting acquainted with the taxonomy was the most time-consuming part of the initial learning process, while selecting the sequences to save was the most time-consuming task in routine use.

In the forthcoming months the BIWS system may begin to be used as a core learning tool during resident discussion sessions dedicated to specific pathologies.

FUTURE DEVELOPMENTS

A usability survey will be performed in the forthcoming months, and revisions to the web-based interface will be considered as a consequence; as explained in the previous section, selection of image sequences to save as objects

needs optimization: alternative models of interaction will be considered, in particular the use of drag-and-drop applets.

A major priority in the short term is to develop integration with the departments carrying out neuropathological and genetic analysis, in order to associate neuroradiological images with microscopic images and genetic data. This is important because neuropathology and genetics are the diagnostic gold standards for a wide spectrum of pathologies: the former in the differential diagnosis and staging of tumors, the latter in all genetically determined pathologies, such as Huntington's disease.

At an initial stage, such a target may be achieved by uploading files through the web interface and attaching them to examinations. In the long term, more sophisticated means can be considered, such as standardized structures to represent genetic data, which would imply defining specific fields in the BIWS database. As an alternative, in an analogy to the link to the EMR system, a foreign link may be established, enabling users to directly access the relevant web page on a dedicated web server, which will be devoted to the storage of genetic or pathological data. Although having the advantage of avoiding duplication of information, such a solution would have the disadvantage of not enabling to use entries in those databases to query the BIWS database.

Two major aspects for mid-long term future development have been envisaged: automated group processing of data for statistical analysis, and the usage of intelligent agents for data mining purposes and to exchange cases with other warehouses.

For example, research on voxel-based morphometry techniques and on diffusion tensor imaging is increasingly relying on brain templates computed from data of controls in various age groups. An introduction to their role in imaging research can be found in the report of Ashburner and Friston.[17] The creation of templates is relatively simple from a conceptual viewpoint, but often very time consuming (several man-hours for a typical template), as it involves a large number of steps to be performed on very large amounts of data. Such a process can be implemented in a fully automated way on a system such as BIWS. A set of dedicated HTML pages, generated with PHP, might in the future enable the user to select the objects to include for creation of the template, using a drag-and-drop or a selection list approach. A C++ module could then be used to manage the data processing steps, which can be implemented with established open-source toolboxes, such as the FSL package.[18] Most of these toolboxes have a command-line interface, and are therefore rather easy to integrate. To avoid degradation of access performance, data processing can be performed remote on a separate server, the storage tree being shared through the NFS protocol on a dedicated peer-to-peer Ethernet link.

In the case of voxel-based morphometry, processing would involve four main steps. A coherency check would have to be performed to ensure that the volumetric (e.g., MPRAGE) images are homogeneous in terms of acquisition parameters, that is, for example, TR, TE, TI, field-of-view, and number of slices. After conversion to the Analyze format, multiple instances of the FLIRT program (from the FSL package) can launch, performing coregistration of the volumes to an institutional reference, for example, to the MNI reference brain. The FAST segmentation tool can then be run for segmentation of white matter and gray matter, and the resulting Analyze volumes then be fused into a template through user-written code. In the case of diffusion tensor imaging, a similar sequence of steps might be used, with the segmentation step being substituted by tensor fitting and determination of diffusion-related parameters, such as mean diffusivity and fractional anisotropy, which could be implemented using the FDT tool from the FSL package. The templates could then be made available in a dedicated folder, and the intermediate files automatically associated to their exams as new objects.

The scientific relevance of warehousing neuroimaging data can be greatly enhanced by sharing warehouses among institutions to which patients with different diseases are referred to leading to different research interests.

Imposing a universally agreed taxonomy, although conceptually highly desirable, might in some cases turn out to be impractical, due to the differences that are seen in classification schemes among institutions: intelligent agents might provide a solution to such a problem, automatizing classification conversion among taxonomies according to predefined schemes, and automatizing the process of periodically importing cases that match specified criteria of interest from remote warehouses.

Privacy issues are of central importance when multicentric sharing of patient data is considered.

Although anonimization of images is not mandatory if circulation is strictly limited to a single department, it becomes absolutely necessary if images are to be exchanged among centers, unless explicit written consent is given by the patient. In the future perspective of multicentric data sharing, privacy can be assured by blinding all database fields that are related to the patient's identity, and by embedding into the software an anonimization process meant to remove all sensitive fields from DICOM image headers. Although anonimization is technically straightforward for DICOM files, it may be remarkably difficult for other file types, as it implies parsing the contents of all fields. When multicentric data sharing will be implemented, initially only sharing of DICOM files will be permitted, whereas, if necessary, anonimizing scripts for other file types will have to be developed on a case-by-case basis.

CONCLUSIONS

Clinical and research neuroradiological practice are characterized by different requirements in terms of data storage. Integrating a preexisting RIS/PACS system with a storage system dedicated to research may be advisable. Web- and SQL-based systems may provide a convenient way to minimize learning time and to enable exchange of cases among geographically distributed centers. A two-layered taxonomy might serve as a common reference for the establishment of large-scale imaging database networks, increasingly perceived as a necessity for the future.[19]

REFERENCES

1. Le Bars E, Gondry-Jouet C, Deramond H, et al: MR diffusion and perfusion imaging in clinical practice. J Neuroradiol 27:39–51, 2000

2. Sabbah P, Leveque C, Dutertre G, et al: Clinical application of functional MRI: a strategic tool for neurosurgery. J Neuroradiol 27:226–232, 2000

3. Bidgood WD, Horii SC, Prior FW, Van Syckle DE: Understanding and using DICOM, the data interchange standard for biomedical imaging. J Am Med Assoc 4:199–212, 1997

4. Rosset A, Ratib O, Geissbuhler A, Vallee JP: Integration of a multimedia teaching and reference database in a PACS environment. Radiographics 22:1567–1577, 2002

5. Stacy MC, Augustine KE, Robb RA: Image BOSS: an image database system designed for research. J Digit Imaging 10:56–59, 1997

6. Lee WJ, Yang CY, Liu KL, et al: Establishing a web-based DICOM teaching file authoring tool using open-source public software. J Digit Imaging 18:169–175, 2005

7. Weinberger E, Jakobovits R, Halsted M: MyPACS.net: a web-based teaching file authoring tool. Am J Roentgenol 179(3):579–582, 2002

8. http://www.my-pacs.net, accessed on Jun 12, 2006

9. Grethe JS, Baru C, Gupta A, et al: Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease. Stud Health Technol Inform 112:100–109, 2005

10. Fernandez-Bayo J, Barbero O, Rubies C, et al: Distributing medical images with internet technologies: a DICOM Web server and a DICOM Java viewer. RadioGraphics 20: 581–590, 2000

11. Dai W, Liang Y, Zhou ZH: Web portal to an image database for high-resolution three-dimensional reconstruction. J Struct Biol 144:238–245, 2003

12. Pinciroli F, Masseroli M, Acerbo LA, et al: BIRD: Bio-Image Referral Database. Design and implementation of a new web based and patient multimedia data focused system for effective medical diagnosis and therapy. Stud Health Technol Inform 98:301–303, 2004

13. Cao P, Hashiba M, Akazawa K, Yamakawa T, Matsuto T: An integrated medical image database and retrieval system using a web application server. Int J Med Inform 71:51–55, 2003

14. Ratib O, Swiernik M, Mc Coy JM: From PACS to integrated EMR. Comput Med Imaging Graph 27:207–215, 2003

15. http://www.rsna.org/RadLex, accessed on Jun 12, 2006

16. Osborn A, Blaser S, Salzman K. Diagnostic Imaging: Brain. USA: W.B. Saunders, 2004

17. Ashburner J, Friston KJ: Voxel-based morphometry—the methods. Neuroimage 11:805–821, 2000

18. Smith SM, Jenkinson M, Woolrich MW, et al: Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23:208–219, 2004

19. Yousem DM, Bryan RN, Beauchamp NJ, Arnold AM: A national neuroimaging database: a call to action. AJNR Am J Neuroradiol 25:908–909, 2004