Safe Storage and Multi-modal Search for Medical Images

Jukka KOMMERI [a], Marko NIINIMÄKI [a] and Henning MÜLLER [b,c]
[a] *Helsinki Institute of Physics, CERN, Switzerland*
[b] *University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland*
[c] *University Hospitals and University of Geneva, Switzerland*

Abstract. Modern hospitals produce enormous amounts of data in all departments, from images, to lab results, medication use, and release letters. Since several years these data are most often produced in digital form, making them accessible for researchers to optimize the outcome of care process and analyze all available data across patients. The Geneva University Hospitals (HUG) are no exception with its daily radiology department's output of over 140'000 images in 2010, with a majority of them being tomographic slices.
In this paper we introduce tools for uploading and accessing DICOM images and associated metadata in a secure Grid storage. These data are made available for authorized persons using a Grid security framework, as security can be a main problem in secondary use of image data, where images are to be stored outside of the clinical image archive. Our prototype tool combines the security and metadata access of a Grid middleware with the visual search that uses GIFT.

Keywords. grid networks, multi–modal information search, security

Introduction

Images are getting increasingly important in modern diagnosis and treatment planning. Through the large variety in radiology protocols and modalities, a detailed image interpretation is not always simple. By producing extremely large volumes of imaging data, tomographic modalities such as CT (Computed Tomography), MRI (Magnetic Resonance Imaging) but also combined PET/CT (Positron Emission Tomography) and PET/MRI can lead to an information overload and create a need for new tools to help interpreting images. Content–based image retrieval has been proposed as one of the potential tools to aid diagnosis and use the large amount of visual data available [1]. The amount of data to be analyzed could benefit from a parallel/distributed data processing solution, such as a computation Grid.

So far, Grid technologies have been successfully employed inside hospitals to speed up image analysis by distributing the visual feature extraction of images to a computer cluster [2], and by integrating image analysis, for example with the ProVision PACS of a hospital [3]. Such an in–house solution has the advantage that images do not need to leave the hospital network for analysis or treatment. The image analysis and retrieval software used in our case is the GNU Image Finding Tool (GIFT). GIFT is used for content–based visual retrieval, whereas so called multi–modal systems combine visual and textual information in retrieval. This was demonstrated to often give better results than either textual or visual information alone [4]. Often,

security constraints are not taken into account when discussing an inclusion of image retrieval into the PACS (Picture Archival and Communication System) or RIS (Radiology Information System) [5] and images are stored unencrypted.

Steps towards integrating analysis with a secure storage of medical images were taken in the Medical Data Manager (MDM) software [6]. MDM uses technologies of the EGEE (Enabling Grids for E–science in Europe) project's gLite middleware [7]. The medical images themselves are stored in an encrypted format in a Disk Pool Manager (DPM) Grid storage [8]. Their meta data are stored in AMGA (gLite Grid MetaData Catalog) [9,10]. The symmetric encryption key is split into a number of pieces and stored in the distributed Hydra storage [11,12] according to the well–known Shamir's Secret Sharing Scheme (SSSS). Even if one node of the Hydra storage is compromised, one piece of the key is not enough to reconstruct the actual symmetric encryption key to de-crypt the data in question. This enhanced security measure is a feature requested by the EGEE BioMedical user group. The gLite security system was audited by the Centre National d'Etudes Spatiales (CNES) and was validated.

In this paper, we describe an integration between GIFT and an MDM–like system to implement analysis of images. This integration combines the content-based image retrieval capabilities of GIFT with a secure image and meta data storage based on Grid technologies. The evaluation of this new technical concept is thus the goal of the text. Two usage scenarios, metadata and multimodal search, are described in Section 1. The components needed to enable these scenarios are described in Section 2. Section 3 contains a discussion and directions for future work.

1. Functionality

The main goal of our system is to enable two functionalities:

- *Meta data search*: The user is identified by a digital certificate (can be carried on a physical device like a SmartCard or a USB disk), and authenticated using a VOMS (Virtual Organization Membership Service) server [13]. The user can then start the metadata search. Access is granted based on the VOMS role and the results of the search are returned. The keys to decrypt the images are obtained and applied. The resulting images are returned.
- *Multi–modal search*: In the PACS system, the user selects an image. This image has a DICOM header containing structured information. Similar images are searched by the visual image content using GIFT and by the metadata using a textual search. Only data matching the user's privacy (authorization) level are returned and shown on screen.

This requires complete system integration as follows:

1. The image and its meta data are extracted from the PACS system.
2. A request containing the image data is sent to GIFT.
3. A request containing the meta data is sent to AMGA.
4. The results are combined and shown to the user based on the role.

The combined system then allows for a safe access to the distributed image data based on the privacy levels (authorizations) of the users.

For adding images and metadata to a secure storage system, the following steps are taken: (i) The user is authenticated, (ii) the metadata of images are loaded to AMGA, (iii) GIFT carries out a feature extraction of the image and the features are stored by

GIFT, (iv) the image is encrypted and encryption keys are stored by Hydra, (v) the encrypted image is stored in DPM. As the system tests were carried out outside of the hospital network, a test database of DICOM files and files from the medical literature used in the ImageCLEF[1] benchmark were used for testing. In Figure 1, the methods for data access are described. The user authentication is performed via digital certificates with the VOMS server that defines the organization to which the user belongs and the role he has in the organization. Queries of the metadata can be performed with AMGA, and visual retrieval with GIFT. In practice this means that using our integrated system, a clinician, for example for a research project, can search for a patient's images by the patient data, or search similar images visually. The data that clinician is able to see can be filtered based on the role in the organization so only images with access rights can be accessed.

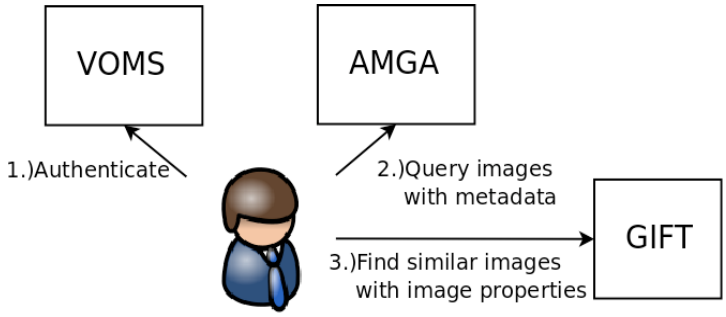

Figure 1. Work flow of multi–modal search for images and associated metadata.

2. Components

The system uses components based on existing Grid tools from the official gLite[2] repositories. Their structure is shown in Figure 2. The software components were installed on several virtual machines so that each virtual machine contained a logical collection of the software. On one virtual machine we installed AMGA, GIFT, the gLite user interface and one Hydra server. Then, we installed two separate Hydra servers on two separate virtual machines. We used an existing VOMS server of the Swiss Multi Science Computing Grid (SMSCG). The components are described below.

The AMGA Metadata Catalog is an EGEE gLite service allowing metadata handling on the Grid. AMGA is a front end file metadata service, providing means of describing and discovering data files required by users. It can also be used as a Grid–enabled database for applications that require structured data; AMGA provides a database–like service that supports Grid security features, including the VOMS authentication and authorization system. It allows access to existing databases from a Grid environment, so that Grid security features can be applied to existing databases.

Even if the access to the data is made secure by AMGA, it is still possible that the data itself can be stolen – for instance in the case that someone can physically remove a hard disk on which the medical images and their metadata are stored. Hydra provides

[1] http://www.imageclef.org/
[2] http://glite.cern.ch/

means of preventing this. Hydra, part of the European Middleware Initiative (EMI), encrypts data using a distributed key storage system. The passkey is generated and then split into components, which are shared across multiple key stores on different servers and, if possible, in different countries. This is more secure than a central key storage system, which requires only one security breach to be compromised. To obtain the passkey generated by Hydra, a coordinated attack on multiple servers is required.

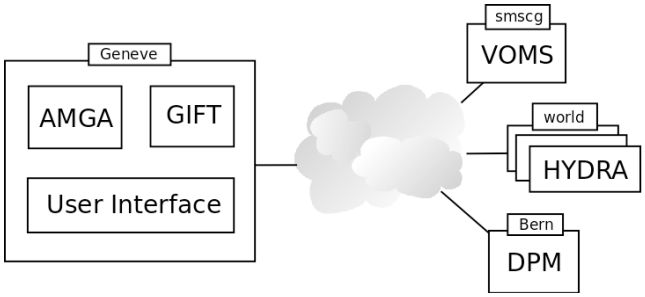

Figure 2. System component network topology for the setup of safe image storage and access.

The GNU Image Finding Tool, GIFT[3], is a content–based image indexing and retrieval package developed at the University of Geneva in the late 1990's. GIFT uses techniques common for textual information retrieval and creates a large set of mainly binary features (global and local color and texture features) [14]. The retrieval quality in terms of precision and recall is evaluated in [14] and also for medical images in the ImageCLEF competition. GIFT extracts the visual features and stores them in an inverted file. In a typical desktop PC, the speed of this feature extraction is about 1 or 2 images per second. An inverted file is created after the feature extraction enabling quick retrieval. Through the Multimedia Retrieval Markup Language (MRML) the system can easily be integrated with other applications.

3. Discussion and Future Work

This article describes a safe storage and access system for medical images and associated metadata using methods based on standard Grid tools. The tools allow for an easy integration of safe storage of all data in encrypted form and an access to the data via meta-data search and content-based image retrieval. Initially the use of the MDM (Medical Data Management) system [6] was planned but the software turned out to not be maintained anymore and the security framework was unfortunately outdated. Thus we decided to change the architecture for the meta data search.

The system uses a role-based access via VOMS servers to potentially confidential medical data. Our test system uses several storage servers of the Universities of Geneva and Bern, and a standard VOMS server of the Swiss Grid community. A similar structure can also be implemented inside hospitals, with the encryption keys being distributed on several machines. To limit security risks, all data (images, thumbnails) are always stored in encrypted format and access to meta data is protected by digital certificates. Encryption keys are stored in a distributed fashion, so a single security breach does not give access to the encryption keys. Such an architecture also allows

[3] http://www.gnu.org/software/gift/

data for research projects to be extracted from the PACS and stored in safe format, quicker than accessing via the often overloaded PACS system. All data are accessed via the role definition of a user, according to which his access rights are defined. The usage of the test system is still limited, and therefore future studies are needed to measure the usability of the system and the potential advantages in terms of security vs. maintenance time.

Acknowledgements

This work was partly supported by the SWITCH AAA project MedLTPC and the European Union in the context of the Khresmoi project (grant agreement no 257528).

References

[1] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, A review of content-based image retrieval systems in medicine–clinical benefits and future directions, *International Journal of Medical Informatics*, **73**, pp. 1–23, 2004.

[2] M. Niinimäki, X. Zhou, A. Depeursinge, A. Geissbuhler, and H. Müller, Building a community grid for medical image analysis inside a hospital, a case study, *Medical imaging on grids: achievements and perspectives (Grid Workshop at MICCAI 2008)*, New York, USA, pp. 3–12, 2008.

[3] M. Niinimaki, X. Zhou, E. de la Vega, M. Cabrer, and Müller, A web service for enabling medical image retrieval integrated into a social medical image sharing platform, in *MEDINFO 2010*, *Studies in Health Technology and Informatics*, **160**, pp. 1273–1276, IOS press, 2010.

[4] I. Eggel and H. Müller, Indexing the medical open access literature for textual and content–based visual retrieval, in *MEDINFO 2010, Studies in Health Technology and Informatics*, **160**, pp. 1277–1281, IOS press, 2010.

[5] P. Welter, T. M. Deserno, B. Fischer, B. B. Wein, B. Ott, and R. W. Günther, Integration of CBIR in radiological routine in accordance with IHE, in *SPIE Medical Imaging 2009: Advanced PACS–based Imaging Informatics and Therapeutic Applications*, **7264**, 2009.

[6] J. Montagnat, A. Frohner, D. Jouvenot, C. Pera, P. Kunszt, B. Koblitz, N. Santos, C. Loomis, R. Texier, D. Lingrand, P. Guio, R. Brito Da Rocha, A. Sobreira de Almeida, and Z. Farkas, A secure grid medical data manager interfaced to the glite middleware, *Journal of Grid Computing*, **6**, pp. 45–59, 2008.

[7] E. Laure, S. M. Fisher, A. Frohner, C. Grandi, P. Z. Kunszt, A. Krenek, O. Mulmo, F. Pacini, F. Prelz, J. White, M. Barroso, P. Buncic, F. Hemmer, A. Di Meglio, and A. Edlund, *Programming the grid using glite, Computational Methds in Science and Technology*, **12**(1), pp. 33–45, 2006.

[8] G. A. Stewart, D. Cameron, G. A. Cowan, and G. McCance, Storage and data management in EGEE, in *Proceedings of the fifth Australasian symposium on ACSW frontiers*, Darlinghurst, Australia, Australia, pp. 69–77, 2007.

[9] N. Santos and B. Koblitz, Metadata services on the grid, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **559**(1), pp. 53–56, 2006.

[10] B. Koblitz, N. Santos, and V. Pose, The AMGA Metadata Service, *Journal of Grid Computing*, **6**, pp. 61–76, 2008.

[11] L. Abadie, P. Badino, J.-P. Baud, J. Casey, A. Frohner, G. Grosdidier, S. Lemaitre, G. Mccance, R. Mollon, K. Nienartowicz, D. Smith, and P. Tedesco, Grid–enabled standards-based data management, *Mass Storage Systems and Technologies*, pp. 60–71, 2007.

[12] A. Frohner, J.-P. Baud, R. M. G. Rioja, G. Grosdidier, R. Mollon, D. Smith, and P. Tedesco, Data management in EGEE, *Journal of Physics: Conference Series*, **219**(6), 2010.

[13] R. Alfieri, R. Cecchini, V. Ciaschini, L. Dell'agnello, Frohner, A. Gianoli, K. Lõrentey, and F. Spataro, VOMS, an Authorization System for Virtual Organizations, *Grid Computing*, pp. 33–40, 2004.

[14] D. M. Squire, W. Müller, H. Müller, and T. Pun, Content–based query of image databases: inspirations from text retrieval, *Pattern Recognition Letters*, **21**(13–14), pp. 1193–1198, 2000.