An Open-Source Digital Archiving System for Medical and Scientific Research

Julien Jomier, Adrien Bailly, Mikael Le Gall and Ricardo Avila
Kitware Inc, 28 Corporate Drive, Clifton Park, NY, USA.

ABSTRACT

In this paper, we present MIDAS, an open-source web-based digital archiving system that handles large collections of scientific data. We created a web-based digital archiving repository based on open standards. The MIDAS repository is specifically tuned for medical and scientific datasets and provides a flexible data management facility, a search engine, and an online image viewer. MIDAS allows researchers to store, manage and share scientific datasets, from the convenience of a web browser or through a generic programming interface, thereby facilitating the dissemination of valuable imaging datasets to research collaborators. The system is currently deployed at several research laboratories worldwide and has demonstrated its ability to streamline the full scientific processing workflow from data acquisition to analysis and reports.

INTRODUCTION

Scientific research, and especially in the medical imaging field, often involves interdisciplinary teams, each performing a separate task from acquiring datasets to analyzing the processing results. Moreover, the number and size of the datasets continue to increase every year due to advancements in magnetic resonance (MR) and computer tomography (CT) technologies. In order to streamline the management of images coming from clinical scanners, hospitals have been relying on picture archiving and communication systems (PACS). However, research teams can rarely access PACS located in hospitals due to security restriction and confidentiality agreements. Furthermore, PACS have become increasingly complex and often do not fit in the scientific research pipeline. Similar to PACS, the MIDAS system targets research and development units and aims at managing digital contents. The MIDAS system provides most of the functionalities of PACS but is not limited to Digital Imaging and Communications in Medicine (DICOM) images and stores any digital media such as clinical notes and 3-dimensional models.

The system stores and archives digital media and make them easily accessible to other research units via several methods of transfer: DICOM, REST application programming interface (API), programmable API (C++) and direct file sharing (SCP/FTP). The stored digital content can also be searched, visualized and downloaded directly from the MIDAS website. Next, we present the main components of the MIDAS system as well as some current applications of our system.

DIGITAL ARCHIVING FOR SCIENTIC RESEARCH

In this section, we review the different technologies related to MIDAS. We also present the core functionalities of the system.

Background

In the past few years, the medical research community has been investigating the use of digital management systems. Historically, hard-drives and permanent storage (tapes, DVD, etc.) were the technology of choice to manage datasets. Nowadays, researchers must search, retrieve and share datasets on a daily basis and, more importantly, they expect a tight integration of these capabilities directly into end-user applications to streamline the processing workflow.

For instance, the Extensible Neuroimaging Archive Toolkit (XNAT)[1] has been recently developed has a data management tool for the Neuroimaging research community. XNAT is written in Java and exclusively supports the DICOM standard. The system allows researchers to store and share research datasets along with the metadata information.

In parallel, the scholar publishing community has been actively developing digital archiving tools. Successful examples include DSpace[2] and ePrints[3]. DSpace has been initiated in 2002 by Hewlett Packard and MIT, with funding from the Library of Congress, to provide a central repository for digital archive. DSpace is intended as a platform for digital preservation activities and was primarily designed to archive text documents. It is most commonly used by research libraries as an institutional repository; however, many organizations currently use the software to manage digital data for a project, subject repository, web archive, and dataset repository. In addition to DSpace, the Handle system[4] provides a unique identifier to repositories worldwide and acts as a digital object identifier (DOI). A handle is comparable to a URL in the internet world and allows moving transparently the digital repository to a different location. Moreover, systems such as DSpace implement the Open Archive Initiatives Protocol for Metadata Harvesting (OAI-PMH). This protocol allows search engines, such as Google and Yahoo!, to query DSpace for specific digital documents. As a result most of the DSpace repositories worldwide are referenced on Google Scholar[5].

MIDAS Core functionalities

Initially based on DSpace, MIDAS extends its capability to support medical and scientific metadata. MIDAS is written in PHP[6] and supports different open-source database backend such as MySQL and PostGreSQL. MIDAS is an archiving platform meant to be versatile and therefore can be easily extended via plug-ins. MIDAS also supports any type of digital media, integrates with the Handle system for object identifier and implements the OAI-PHM.

MIDAS supports scientific and medical research in several ways. First, an upload workflow has been designed with the capability of running external filters at each stage of the upload process. For instance, an anonymization tool is automatically applied on clinical datasets to help ensure the data is anonymized before storage and thus compliant with the Health Insurance Portability and Accountability Act (HIPAA) regulations. We have also developed other automated filters. The DICOM metadata filter extracts relevant patient and scanner information and stores the metadata for later retrieval.

Second, MIDAS provides a wide range of data transfer mechanisms. We have integrated a DICOM server into MIDAS so that the system can act as a DICOM node and receive data from clinical scanners as well as export clinical data stored in MIDAS to PACS systems. For algorithm developers, MIDAS provides a file-system access as well as interfaces in common programming languages (C/C++). MIDAS also implements a generic web API based on the representational state transfer (REST) [7] architecture. The REST API provides access flexibility for other programming tools to store, manage, search and retrieve datasets.

Finally, we have been extending MIDAS in two main directions. The first direction is the ability to perform online visualization of massive datasets stored in the system. With the recent advances in technology (microscope imaging for instance), the amount of data is massive and transferring the data remotely for processing for processing often takes more time than the processing itself. For these reasons, researchers have been investigating a data-centric approach that either divides a dataset into smaller pieces or brings the processing close to the data storage.

The second direction provides a generic framework for distributed computing from MIDAS using Condor[8]. The goal is to provide a generic interface to grid-computing through the MIDAS platform, allowing researchers to run complex processing with data and associated metadata stored in the system.

USE CASES

In this section we present how the MIDAS system is currently used at three major institutions: The Optical Society of America, The University of North Carolina and Kitware Inc.

The Optical Society of America

In October 2008, the Optical Society of America (OSA) launched the first interactive scientific publishing (ISP) system[9]. This initiative allows scientists to expand upon traditional research results by providing software for interactively viewing underlying source data. The ISP system enhances the standard scientific publishing by adding interactive visualization. Using ISP, authors have the ability to create 3-dimensional visualization of their datasets, add 3-dimensional annotations and measurements and make the datasets available to reviewers and readers. The system is composed of two main components: the archiving system and the visualization software. A customized version of MIDAS provides the data storage, delivers low-resolution datasets for pre-visualization, and in the background serves the full-resolution dataset. MIDAS has also been extended to support MeSH, the U.S. National Library of Medicine (NLM) controlled vocabulary used for indexing scientific publications. The second component, the ISP visualization software, interacts directly with the MIDAS system in order to retrieve stored datasets. Readers of an ISP-enabled manuscript can automatically launch the ISP software by clicking on a web link directly in the PDF. Within ten seconds, a low-resolution dataset is loaded from MIDAS and can be interactively manipulated in 3D via the ISP software.

The OSA and the NLM have collaborated with Kitware Inc. to provide this service. OSA's MIDAS repository[10] currently hosts more than 30 ISP manuscripts with over 20GB of associated data. These datasets are made freely available to the scientific community.

The University of North Carolina

The Neuro-Image Research and Analysis Laboratory (NIRAL)[11] at the University of North Carolina has been using the MIDAS system for internal image management and processing for several years. The NIRAL distributed computing environment, currently composed of 56 dedicated cores, uses the different MIDAS interfaces (DICOM, Web and C++ API) to collect and retrieve original datasets and store processed data. The NIRAL area of research includes shape analysis of the brain and diffusion tensor imaging analysis in the context of clinical research. The NIRAL has been developing computational modules for the analysis of brain structures related to autism and schizophrenia studies. These modules are distributed as part of Slicer[12]: an open-source visualization and processing platform funded by NIH.

The MIDAS installation at the NIRAL shows how a research laboratory can benefit from a centralized data repository. Clinical collaborators, located worldwide, upload clinical datasets to the MIDAS system. Algorithm developers can then retrieve and process the selected datasets and upload the results. Finally, statisticians can access the processed datasets and upload analysis results.

Kitware Inc.

Kitware Inc. is the original developer of the MIDAS system and has been running a public digital repository[13] for more than five years. Kitware delivers scientific content through two main websites. The first website is a standard MIDAS instance, which hosts more than 100 healthy brain magnetic resonance (MR) datasets provided by Dr. Bullitt at the University of North Carolina. These datasets are freely available for everyone to use. This MIDAS instance also hosts clinical datasets from the Retrospective Image Registration Evaluation Project[14] and the Slicer project. The second website, based on the MIDAS framework, hosts an open-science journal initiated by Kitware and the National Library of Medicine: the Insight-Journal[15]. The Insight-Journal is an open access online publication covering the domain of medical image processing and visualization. The unique characteristics of the Insight Journal include: (a) open-access to articles, data, source code, and reviews, (b) open peer-review that invites discussion between reviewers and authors, (c) emphasis on reproducible science via automated code compilation and testing and (d) support for continuous revision of articles, code, and reviews. Scientists and researchers can submit freely source code and datasets to the Insight Journal. Currently, the Insight-Journal and a companion journal, the Midas Journal, are counting more than 1,400 subscribers and more than 370 submitted open-science articles. Kitware, as well as many other institutions, also run several internal instances of MIDAS to manage its own digital documents and image processing pipeline.

CONCLUSION

We have presented a novel digital archiving and retrieval system for scientific research. By providing a platform for data management, sharing and dissemination, the scientific research pipeline can be streamlined across the different organizations. MIDAS also facilitates open-access across institutions and allows researchers to disseminate their work in an efficient manner. The system is under continuous development by the open source community and is available under an unrestrictive open-source license. For more information visit http://www.kitware.com/products/midas.html.

REFERENCES

[1] The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data, Marcus DS, Olsen TR, Ramaratnam M, Buckner RL, Neuroinformatics. 2007 Spring;5(1):11-34

[2] DSpace, and open-source solution for accessing, managing and preserving scholarly works: http://www.dspace.org

[3] EPrints: http://www.eprints.org

[4] Handle System Overview, S Sun, L Lannom, B Boesch - 2006 - Digital Library

[5] Google Scholar: http://scholar.google.com

[6] PHP: Hypertext Preprocessor: http://www.php.net

[7] Representational State Transfer (REST). Roy T. Fielding. Architectural Styles and the Design of Network-based Software Architectures. Chapter 5.

[8] Litzkow M, Livny M, and Mutka M, "Condor - A Hunter of Idle Workstations", Proceedings of the 8th International Conference of Distributed Computing Systems, pages 104-111, June, 1988.

[9] OSA's Interactive Scientific Publication system: http://www.opticsinfobase.org/isp.cfm

[10] OSA's MIDAS installation: http://midas.osa.org

[11] Neuro Image Research and Analysis Laboratory at UNC: http://www.niral.unc.edu

[12] 3D Slicer: http://www.slicer.org

[13] MIDAS at Kitware Inc: http://www.insight-journal.org/midas

[14] The Retrospective Image Registration Evaluation Project: http://www.insight-journal.org/rire

[15] The Midas Journal: http://www.midas-journal.org