Medici: A Scalable Multimedia Environment for Research

Luigi Marini, Rob Kooper, Joe Futrelle, Joel Plutchak, Alan Craig, Terry McLaren and James Myers

National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

Large-scale community collections of images, videos, and other media are a critical resource in many areas of research and education including the physical sciences, biology, medicine, humanities, arts, and social sciences. Researchers face coupled problems in managing large amounts of data, analysis and visualization over such collections, and managing descriptive metadata and provenance information. NCSA is involved in a wide range of projects targeting collections that involve terabytes to petabytes of data, complex image processing pipelines, and rich provenance linking. Based on this experience, we have developed Medici – a general multimedia management environment based on Web 2.0 interfaces, semantic content management, and service/cloud-based workflow capabilities that can support a broad range of high-throughput research techniques and community data management. Medici provides scalable storage and media processing capabilities, simple desktop and web 2.0 user interfaces, social annotations, preprocessing and preview capabilities, dynamically extensible metadata, provenance support, and citable persistent data references. This talk will provide an overview of Medici's capabilities and use cases in the humanities and microscopy as well as describe core research and development challenges in creating usable systems incorporating rich semantic context derived from distributed automated and manual sources.

Social media services such as Flickr, Picasa, and YouTube provide ways of storing and sharing images and videos that are highly intuitive and consequently have become widely used. However, they have a number of limitations that reduce their value in research. Compared to social media services, research collaborations involve heterogeneous data collections, larger amounts of data, ingestion and previewing of large datasets, project-specific privacy concerns, distributed data, richer and higher dimensional metadata, coupled with domain-specific multistage analysis.

The heterogeneous nature of research data requires systems of this kind to be open to any data format. Medici does not impose any restrictions on the type, size or nature of the data uploaded. Researchers can store and refine not just images and videos, but also documents, 3D models, audio, and any of their multimedia files. Medici is built on the Tupelo semantic content management system and thus gains the benefits inherent in RDF including global

identification and ability to mix generic and domain-specific ontologies. By building on Tupelo, there are no constraints imposed on the file types. Every file uploaded receives a unique URI to which the binary content of the files are attached. Moving the data and metadata from private repositories to public ones also becomes a built-in feature. Information initially available on a local machine that is accessible only by a small research group can be moved to a larger public repository without having to worry about broken links, identifier conflicts, or metadata crosswalks. Tupelo also enables us to scale from lab machine to distributed file system, by abstracting the underlying repository implementation for both data and metadata. This keeps the basic requirements for the system low and allows it to scale up to larger repositories. Researchers have two options of installing and managing such a system on a lab machine as well as on high throughput resources. In one instance we use the Lustre high performance parallel file system to store the binary data on the Indiana University's Data Capacitor and the metadata about the files on a local machine at NCSA. This context instance provides members of the scientific community access to a multi-terabyte data store via a 10Gb/sec network connection.

Users upload files without having to specify much more than the file itself by using the standard file browse interface, or by clicking and dragging folders onto a drop box. Single files are added to the default file collection where folders become a new collection of files based on the folder name. Additionally, a RESTful service interface is available for batch processing, mashups and script development. Once a file is added to a Medici context, the system extracts basic file information such as the filename that becomes the title by default, it's assigned a unique URI, and extraction or analysis services are executed on the file based on its MIME type. Researchers can also add, edit or delete the metadata associated with a specific file or collection; for example, change the file name to something more meaningful. The user can make any statement about the data, such as adding a tag, leaving a comment, or adding a specific metadata field specific to a particular ontology, such as Dublin Core, or adding arbitrary metadata without having to understand the underlying bits that make up the data. By treating data independent of its format, researchers can make connections and analyze heterogeneous datasets. This allows them to link a paper to observational data stored in a CSV file, or to a movie taken the day the data was collected. Medici also provides privacy controls for the data using license controlled access which can restrict data sets from being downloaded.

Medici is designed to support any data format and multiple research domains and contains three major extension points: preprocessing, processing and previewing. Every time new data is added to the system, whether it is via the web application, the desktop client or through the RESTful services, preprocessing is off-loaded to extraction services in charge of extracting appropriate data and metadata. The extraction services attempt to extract information and run preprocessing steps based on the type of the data. For example, in the case of images, a preprocessing step takes care of creating the previews of the image, but also of extracting EXIF and GPS metadata from the image. This raw metadata is presented to the user in both the Medici web and desktop clients. For example, if GPS information is available, the web client maps the location of the dataset on a Google map, while the desktop client shows the latitude and longitude values as textual fields in the additional information view. By making the clients and preprocessing steps independent from each other, and using RDF as a common but domain-neutral representation, the system can grow and adapt to different user communities and research domains.

A Medici desktop client is provided to demonstrate how the processing step can be supported when focusing on data management. The primary concern with processing and data management is to keep track what happens to the data. This history of the data, also known as data provenance, is fundamental for experimental reproducibility. Medici enables users to launch external tools such that Medici itself can capture the inputs and outputs of the processes and store this information as part of the metadata. The web client displays this information on the page so that researchers can easily identify derived data and its source. This is not a requirement of the system but an example of how one can support data provenance as an integral part of the research cycle.

Finally, when browsing the data, researchers need to be able to gain a sense of the results not just from the metadata perspective (tags, comments, metadata fields), but also by previewing information about the bits making up the data itself. Medici provides an environment where a researcher can see the end-to-end results from data source, through the types of processing to the derived data products and the previewing components are key to accessing data that is potentially very large without moving this data around which can be incredibly costly. For example, microscopy images can be as large as a gigabyte each. It's not feasible to download such large datasets on demand, however, it is feasible to provide the user with a preview of the large image for preliminary investigation. We demonstrate how we use Microsoft Seadragon to provide the ability to preview such large image. The Medici extraction services

creates the image pyramids of such large image files as part of the preprocessing step, and Microsoft Seadragon allows us to smoothly visualize the resulting image pyramid within the preview frame of Medici.

Diverse use cases involving high-throughput microscopy and sharing of cultural collections (quilts, historical maps) have driven the design of the system. With a flexible metadata model, the ability to store any data format, and extensible preprocessing, processing and previewing steps, different research communities can customize the system to meet their needs and make the data and metadata portable and open.

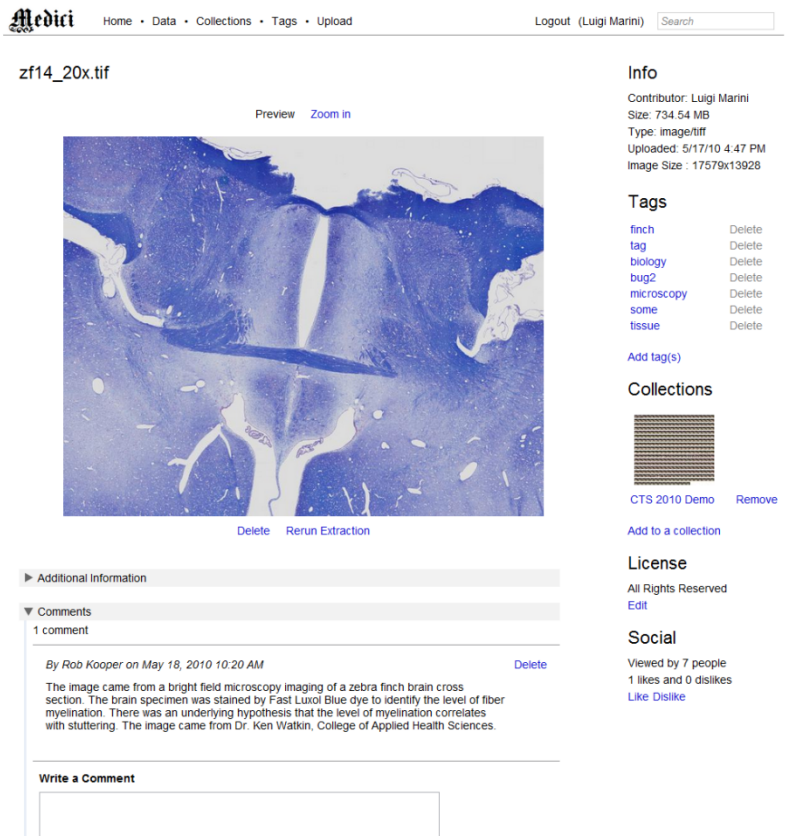

Figure 1. Medici web application

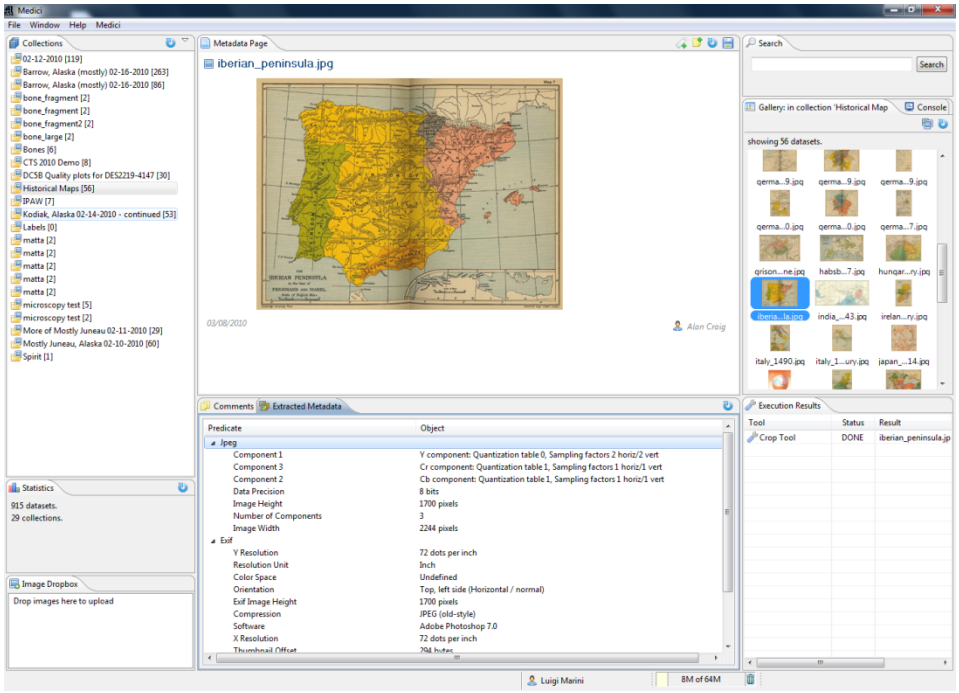

Figure 2. Medici desktop application