Medici: A Scalable Multimedia Environment for Research

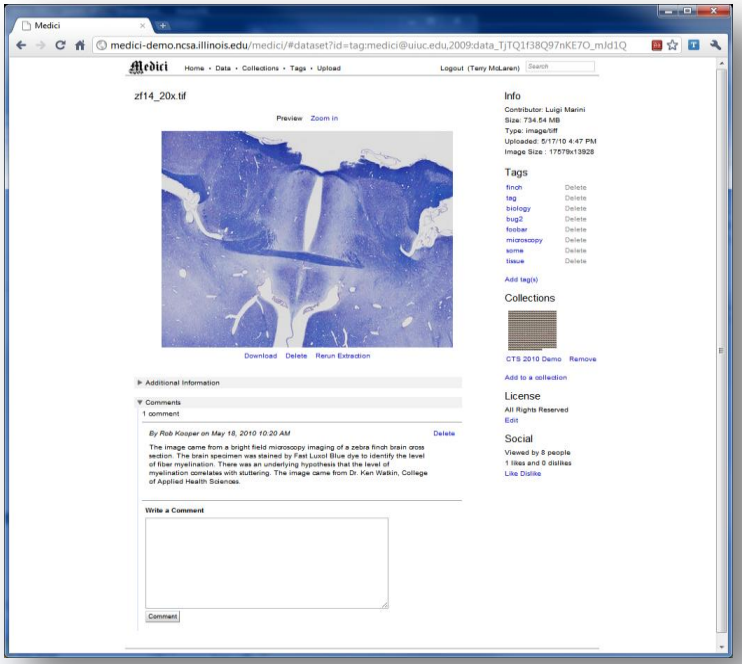
Luigi Marini

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

OVERVIEW

Medici in a nutshell

- A multimedia content management system based on
  - Web 2.0 interfaces
  - Semantic web technologies (RDF)
  - Cloud-based processing and preprocessing



Community Drivers

- Large-scale community collections of heterogeneous data (documents, images, video, etc.) are a critical resource in many areas of research and education
  - physical sciences, biology, medicine, humanities, arts, and social sciences
- Medici does not place any requirements on the nature of the data
  - Preprocessing is applied to the data based on the available extensions in a particular instance

Community Drivers

- Rapid growth from
  - High throughput instruments
  - Digitization efforts
  - More sources (e.g. everyone's cameras)
- Medici provides previewers to stream the data to the client
  - Large files don't have to be downloaded before being viewed and acted upon
- Medici provides different ways of ingesting the data
  - RESTful web service for batch uploading

Community Drivers

- Researchers face coupled problems in
  - managing large amounts of data
  - organizing metadata and provenance information
  - sharing data
  - curating and preserving data
- Medici stores data and metadata as semantic RDF triples and blobs of binary data
  - Open and portable data

Project Goals

- Provide a (customizable) out of the box solution to store, organize, analyze, view, share, and preserve research content
- Ability to handle heterogeneous files
- Portable and open data and metadata standards
- Support for extensible analytics

Would Flickr/YouTube work?

- Maybe
  - Web-accessible tools are relatively generic
  - Users like not having to manage storage
  - Metadata, tagging, linking, etc. are effective means of organizing information (i.e., no need for “folders”)
- No
  - Limits on type, volume, throughput, resolution
  - No community ownership / control of resources
  - Inadequate privacy (e.g., for unpublished work)
  - No domain-specific processing
  - No provenance (everything is a stream of “posts”)

Different Communities and Drivers

- Development and use cases supported by
  - US Office of Naval Research (ONR)
  - US National Archives and Records Administration (NARA)
  - US National Endowment for the Humanities (NEH)
  - US National Institute of Health (NIH)
  - US National Science Foundation (NSF)
  - EU LinkSCEEM-2

FEATURES

Upload

- Drag and Drop from web and desktop clients
- Write custom code against a RESTful web service
- Single files or whole directories
  - Directories become collections when using the Java Applet uploader

Data Preview

- Preview datasets without having to download the full resolution data. For example zoom and pan over large images in the browser. Preview video and audio files
  - zoomable images (seadragon)
  - playable movies (jwplayer)
  - 3D objects (webgl and HTML5)
- Display metadata embedded in files
  - Geospatial location embedded in TIFF files

Retrieval

- Search the repository using text based search
- Browse tags and collections on specific topics
- Download original datasets based on access control

Scalability

- Support for large datasets (1GB+) and millions of metadata records
- Asynchronous extraction of metadata and creation of preview to scale to large volumes of traffic and data
- Write custom extractors to run specific analytics

Sharing

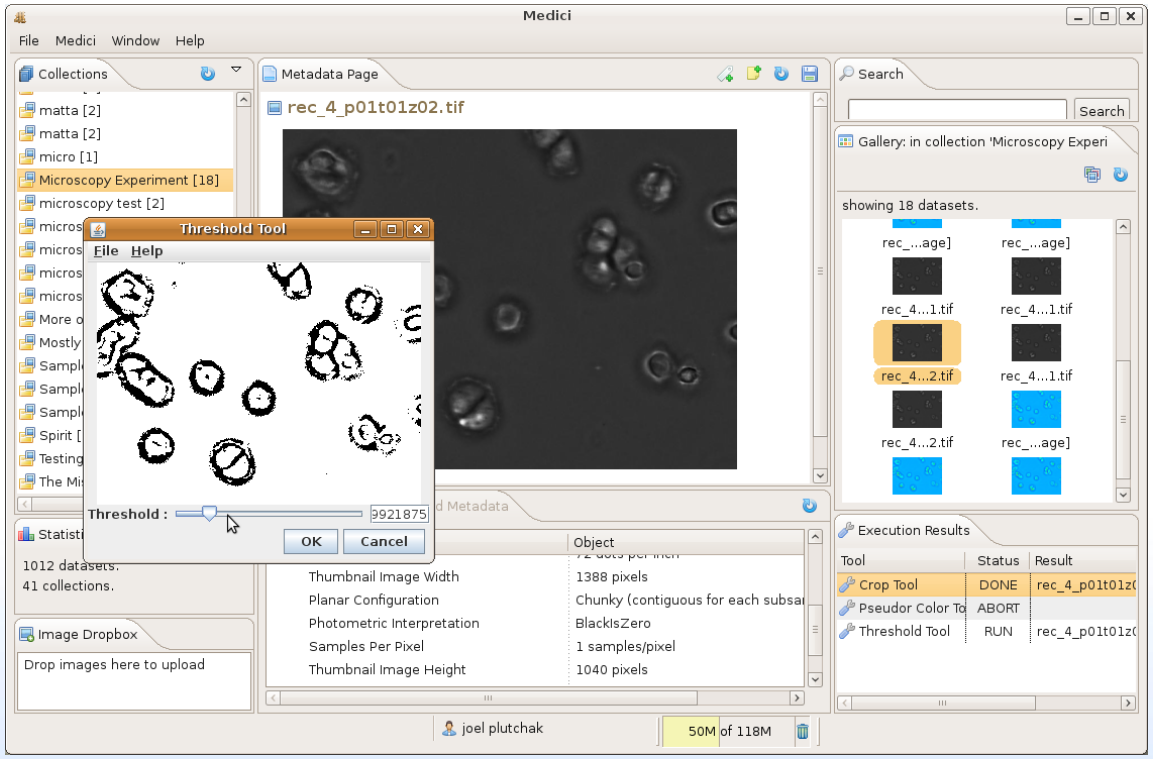
- Send citable persistent URLs to colleagues to share data.
- Set copyright and attribution information to protect the rights to use, copy, and distribute the data
- Track the provenance of derived data

Social Annotation

- Tag datasets and/or create collections to organize datasets
- Leave notes and comments on specific datasets or collections
- Fill out additional fields based on community specific ontologies
- Establish typed relationship between datasets

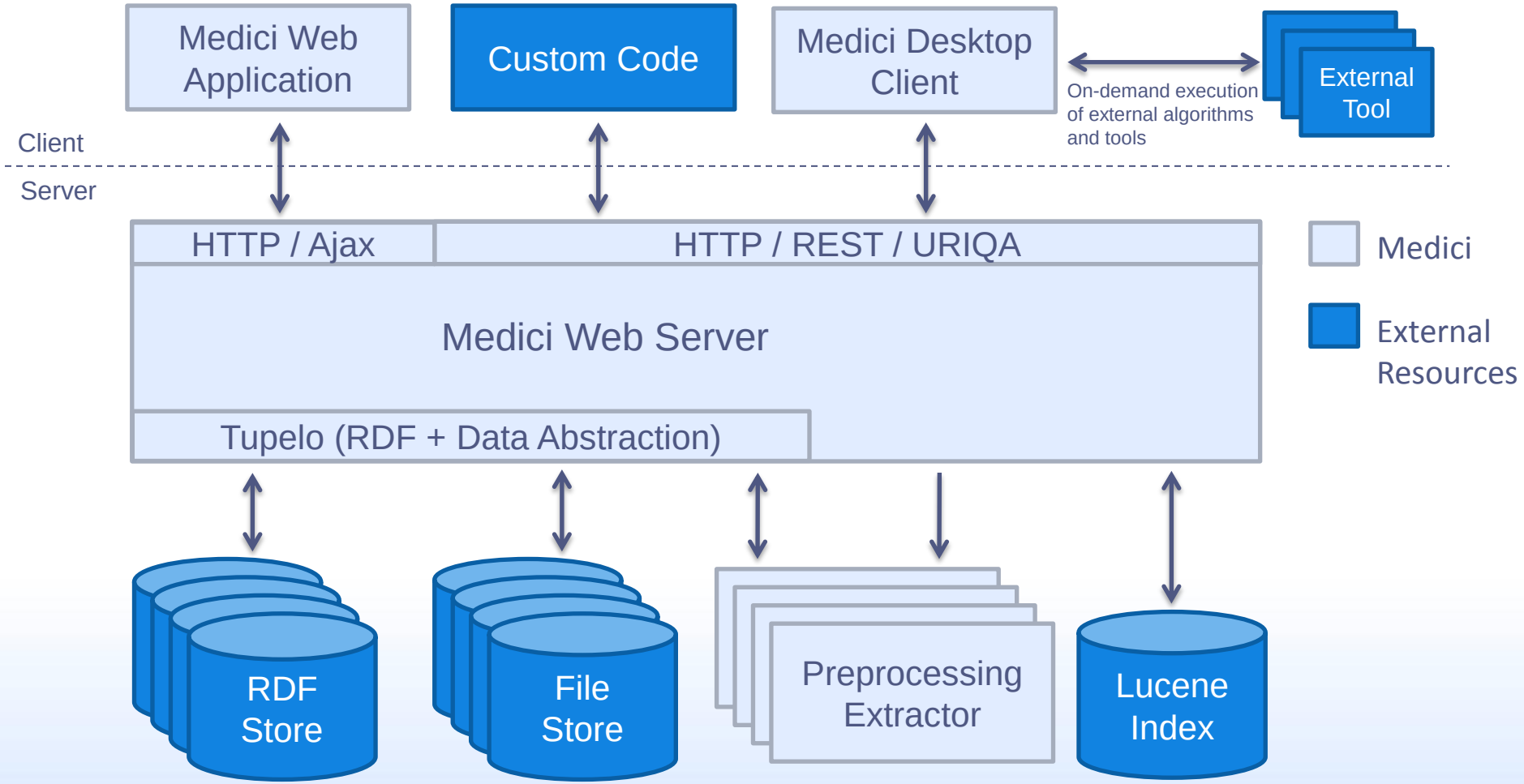
Processing

- Extensible preprocessing using plug-in based extractors
- On demand data processing using community shared tools



TECHNOLOGY

Medici Architecture



Technologies

- Web application
  - Google Web Toolkit
  - Java Servlets
  - Plain Javascript
  - Viewers: Flash, Java Applet, HTML, etc.
  - Apache Lucene
  - Mysql
- Extraction Service
  - Eclipse RCP (Java)
  - Large collection of external applications
- Desktop Client
  - Eclipse RCP (Java)
  - Cyberintegrator Workflow Management System

Repository Middleware - Tupelo

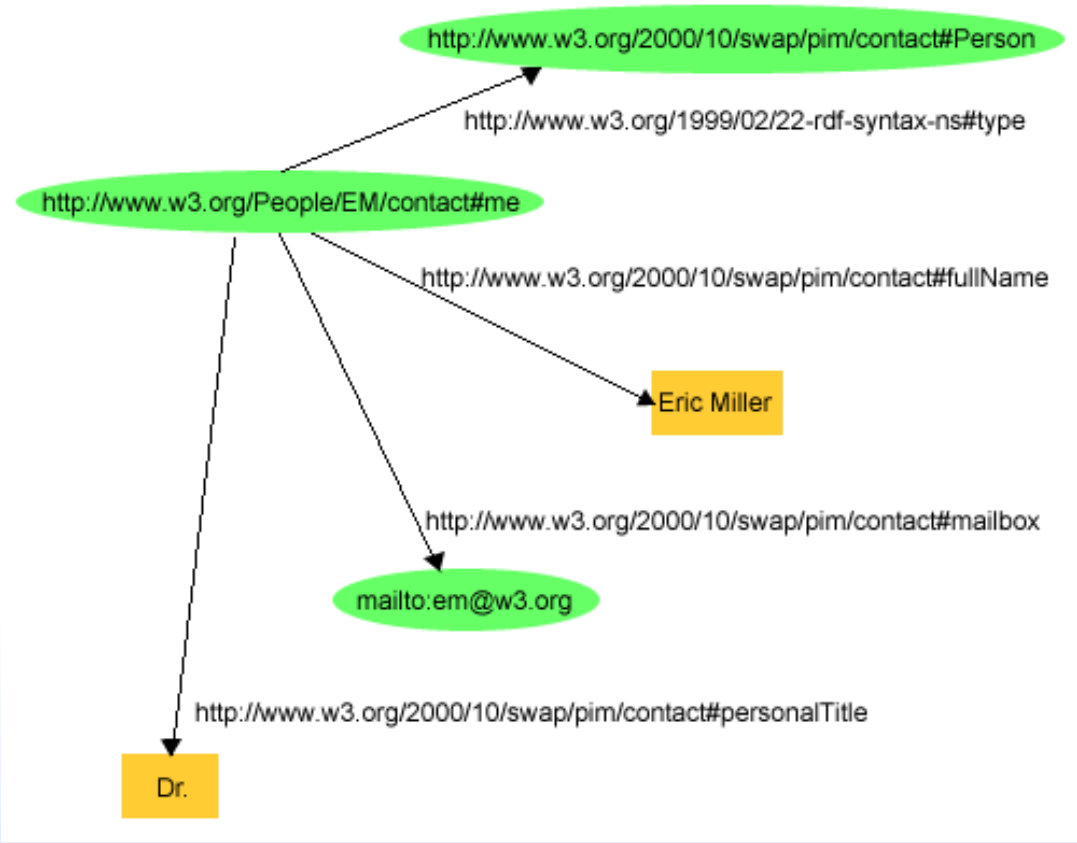
- *Lightweight* middleware “fabric”
- Minimal model of binary content (URI-addressed “blobs”) with no format restrictions
- Standard, portable model of metadata (linked data – style RDF) allowing arbitrary extensibility
- “Context” allows generic data / metadata operations to be interpreted on the fly
- Documentation available at
  - http://tupeloproject.ncsa.uiuc.edu

Sesame

WebDAV

RDF Primer

- http://www.w3.org/TR/2004/REC-rdf-primer-20040210/



Extraction service

- Multiple, extensible pre-processing pipelines
- Asynchronous, distributed, triggered by ingest
  - Processing selected on basis of content type
  - Recursive (products can trigger additional extractions)
- Used to produce web-viewable previews
  - Image pyramids, audio/video previews, thumbnails, pdf to plain text
- Used for domain-specific pre-processing, e.g.,
  - Metadata extraction (e.g., FITS headers, geolocation)
  - Feature detection
  - Specialized OCR for non-standard textual types (e.g., 18th-century manuscripts)

For more information please visit

http://medici.ncsa.illinois.edu

email us at

medici@ncsa.illinois.edu

join the discussion at

medici-users@ncsa.illinois.edu