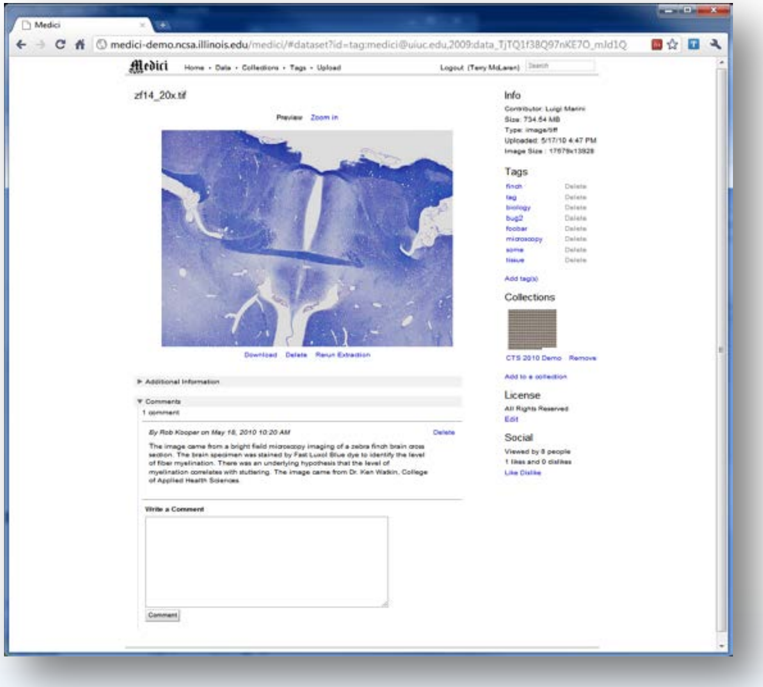
Medici in a nutshell

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

Medici in a nutshell

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



zf14_20x.tif

Info

Contributor: Luigi Marini
Size: 734.54 MB
Type: image/tiff
Uploaded: 5/17/10 4:47 PM
Image Size : 17679x13928

Tags

finch
tag
biology
bug2
foobar
microscopy
some
tissue

Collections

CTS 2010 Demo

License

All Rights Reserved

Social

Viewed by 8 people
1 likes and 0 dislikes

Comments

1 comment

By Rob Kooper on May 18, 2010 10:20 AM

The image came from a bright field microscopy imaging of a zebra finch brain cross section. The brain specimen was stained by Fast Luxol Blue dye to identify the level of fiber myelination. There was an underlying hypothesis that the level of myelination correlates with stuttering. The image came from Dr. Ken Watkin, College of Applied Health Sciences.

Motivations

- Address research and education data collection and analytic needs
  - Manage large collections of heterogeneous data
  - Organize data with metadata and provenance information
  - Facilitate collaborations and data sharing
  - Enable curation and data preservation
- Support community collections of heterogeneous data (documents, images, video, sensor, modeling, etc.)
- Enable automated data extraction, analytic and preprocessing services on local and remote systems
- Provide data preview capabilities specific to different data types

Why Not Flickr or YouTube

- 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”)
- Maybe not?
  - No individual or community ownership
  - No control of resources
  - Inadequate privacy (e.g., for unpublished work)
  - Limits on format, volume, throughput, resolution
  - No domain-specific processing
  - No provenance (everything is a stream of “posts”)

Community Drivers

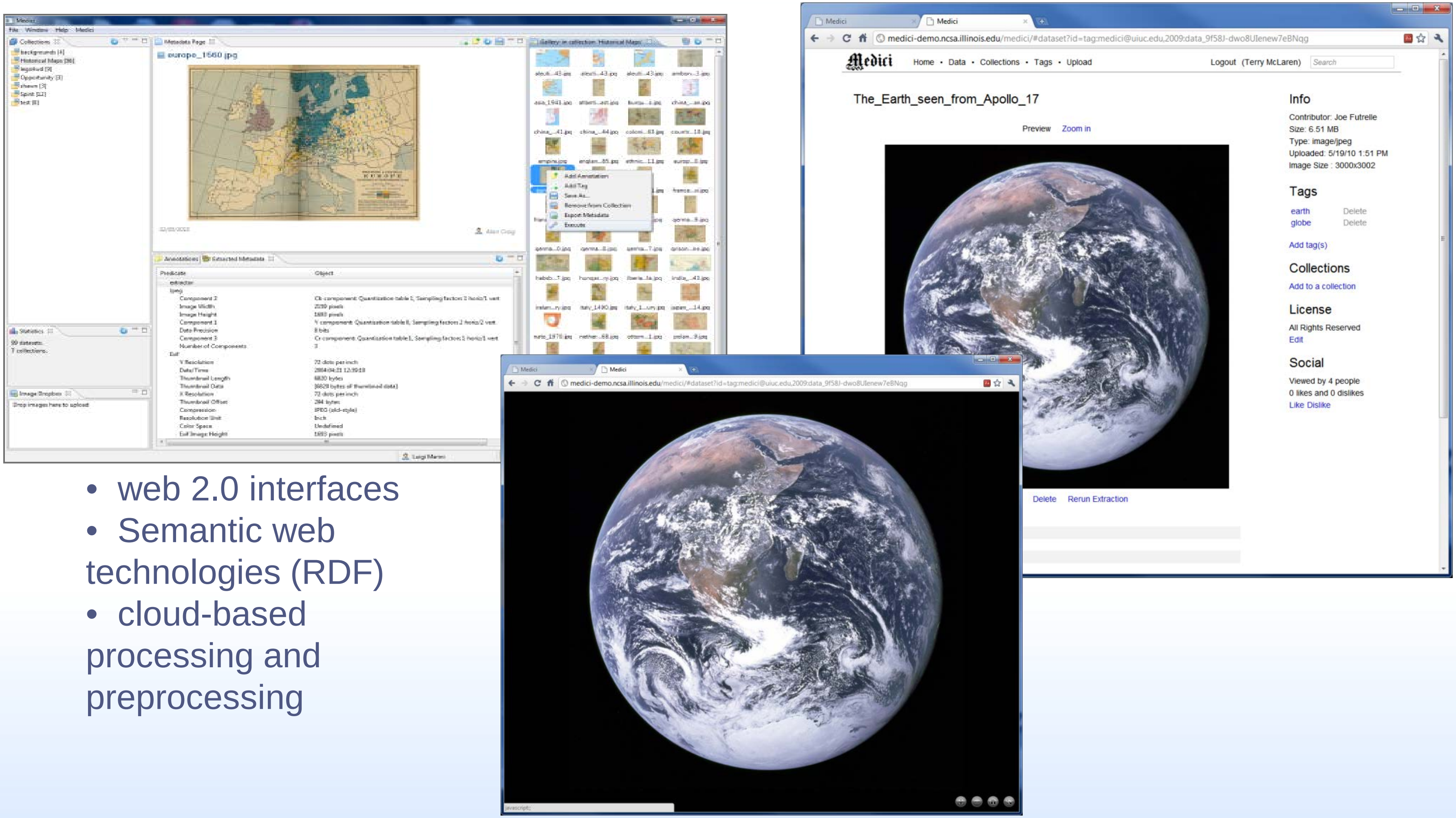
- Use cases and requirements driven by
    - US Office of Naval Research (ONR)
    - US National Archives and Records Administration (NARA)
    - US National Endowment for the Humanities (NEH)
    - US National Institute of Health (NIH)
    - US National Science Foundation (NSF)
    - Institute for Advanced Computing Applications and Technologies (IACAT)
    - Seagrant/EPA
    - EU LinkSCEEM-2 (Cyprus Institute)

Why Medici?

- Provides a customizable turn key solution to store, organize, analyze, view, share, and preserve research content
- Supports heterogeneous files
  - Single file or directory upload via click-n-drag
  - RESTful web service for batch or script based uploading
  - Owner defined copyright and download permissions
- Standards based (RDF) semantic content model
- Conforms to open data and metadata standards
  - Supports OPM (Open Provenance Model)
  - Tags, comments, ratings
- Supports customizable automated extraction services
  - E.g.: Image pyramid creation, OCR for scanned text, movie frame extraction (.mpeg), file transformation, etc…
- Leverages proven technologies
  - Lucene indexing, MySQL, any command-line tool for extraction/analytic services
- Open source, public APIs

Medici – Semantic Data Repository

- Web and Desktop access to a semantic content repository.



59 datasets.
7 collections.

Drop images here to upload

Predicate	Object
extractor	
jpeg	
Component 2	Cb component: Quantization table 1, Sampling factors 1 horiz/1 vert
Image Width	2239 pixels
Image Height	1683 pixels
Component 1	Y component: Quantization table 0, Sampling factors 2 horiz/2 vert
Data Precision	8 bits
Component 3	Cr component: Quantization table 1, Sampling factors 1 horiz/1 vert
Number of Components	3
Exif	
Y Resolution	72 dots per inch
Date/Time	2004:04:21 12:39:18
Thumbnail Length	6820 bytes
Thumbnail Data	[6820 bytes of thumbnail data]
X Resolution	72 dots per inch
Thumbnail Offset	284 bytes
Compression	JPEG (old-style)
Resolution Unit	Inch
Color Space	Undefined
Exif Image Height	1683 pixels

Luigi Marini

Medici
medici-demo.ncsa.illinois.edu/medici/#dataset?id=tag:medici@uiuc.edu,2009:data_9f58J-dwo8Ulenew7eBNqg
Medici
Home • Data • Collections • Tags • Upload
Logout (Terry McLaren)
Search

The_Earth_seen_from_Apollo_17

Info

Contributor: Joe Futrelle
Size: 6.51 MB
Type: image/jpeg
Uploaded: 5/19/10 1:51 PM
Image Size : 3000x3002

Tags

earth Delete
globe Delete

Add tag(s)

Collections

Add a collection

License

All Rights Reserved
Edit

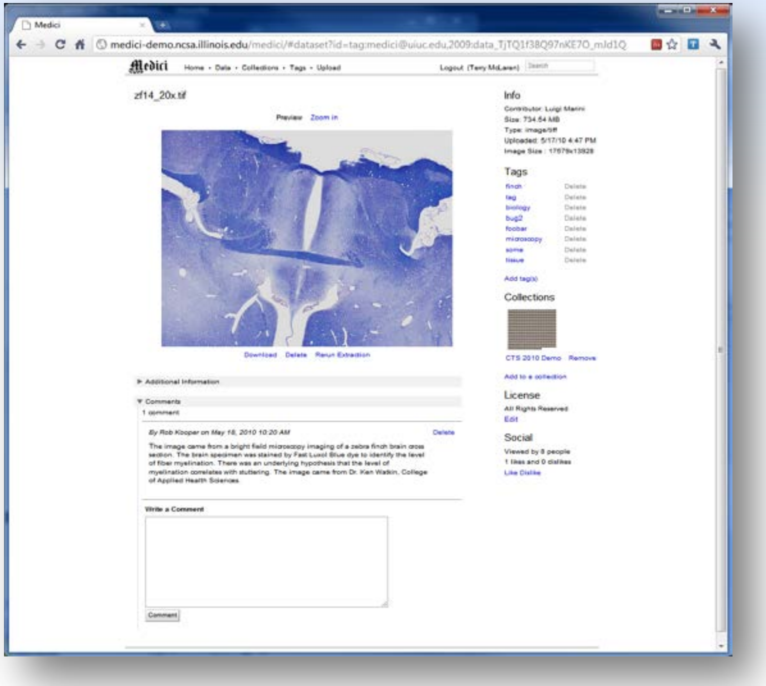
Social

Viewed by 4 people
0 likes and 0 dislikes
Like Dislike

- web 2.0 interfaces
- Semantic web technologies (RDF)
- cloud-based processing and preprocessing

Client features

- Upload / download
- Search / browse
- Tag / comment
- Create collections
- Geo-locate data (map view)
- Content-type-specific previewing
  - e.g., zoomable images (Seadragon), playable movies (jwplayer), rotatable 3D objects (HTML5)
- Define a specific taxonomy
- Access statistics, provenance
- Citable persistent URLs
- Set copyright and license attributes
  - View only, prevent download
- Define dataset relationships



Medici

medici-demo.ncsa.illinois.edu/medici/#dataset?id=tag:medici@uiuc.edu,2009:data_TjTQ1f38Q97nKE7O_mJd1Q

zf14_20x.tif

1 comment

By Rob Kooper on May 18, 2010 10:20 AM Delete

The image came from a bright field microscopy imaging of a zebra finch brain cross section. The brain specimen was stained by Fast Luxol Blue dye to identify the level of fiber myelination. There was an underlying hypothesis that the level of myelination correlates with stuttering. The image came from Dr. Ken Watkin, College of Applied Health Sciences.

Write a Comment

Info

Contributor: Luigi Marini
Size: 734.54 MB
Type: image/tiff
Uploaded: 5/17/10 4:47 PM
Image Size : 17679x13928

Tags

finch	Delete
tag	Delete
biology	Delete
bug2	Delete
foobar	Delete
microscopy	Delete
some	Delete
tissue	Delete

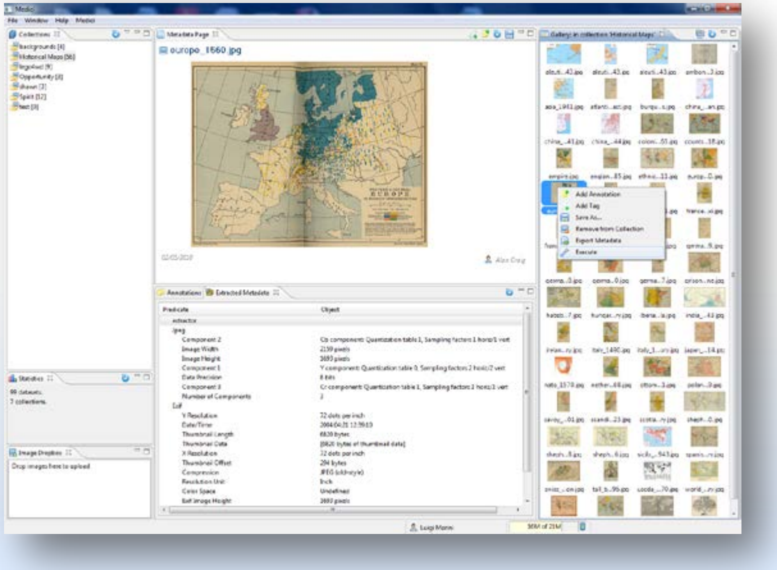
Collections

License

All Rights Reserved

Social

Viewed by 8 people
1 likes and 0 dislikes



backgrounds [4]
Historical Maps [56]
lego4ad [9]
Opportunity [3]
dhawn [3]
Spirit [12]
test [3]

europe_1660.jpg

02/05/2010 Alan Craig

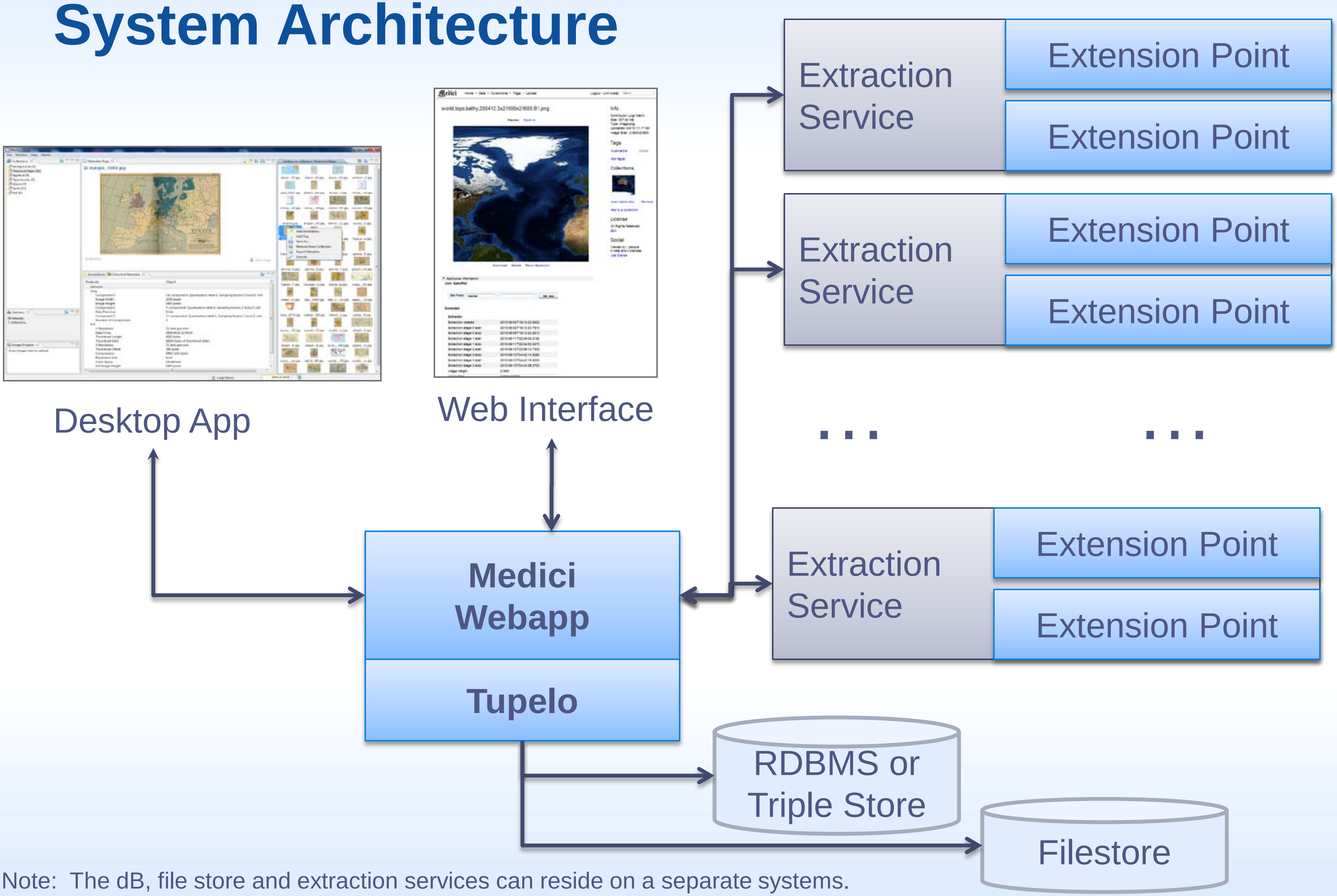
98 datasets
7 collections

Drop images here to upload

Predicate	Object
extractor	
jpeg	
Component 2	Cb component: Quantization table 1, Sampling factors 1 horiz/1 vert
Image Width	2139 pixels
Image Height	1693 pixels
Component 1	Y component: Quantization table 0, Sampling factors 2 horiz/2 vert
Data Precision	8 bits
Component 3	Cr component: Quantization table 1, Sampling factors 1 horiz/1 vert
Number of Components	3
Exif	
Y Resolution	72 dots per inch
Date/Time	2004:04:21 12:39:13
Thumbnail Length	6820 bytes
Thumbnail Data	[6820 bytes of thumbnail data]
X Resolution	72 dots per inch
Thumbnail Offset	294 bytes
Compression	JPEG (old-style)
Resolution Unit	Inch
Color Space	Undefined
Exif Image Height	1693 pixels

Luigi Marini 36M of 21M

System Architecture



Extraction Service

Extension Point

Extension Point

Extraction Service

Extension Point

Extension Point

. . .

. . .

Extraction Service

Extension Point

Extension Point

Desktop App

Web Interface

Medici Webapp

Tupelo

RDBMS or Triple Store

Filestore

Note: The dB, file store and extraction services can reside on a separate systems.

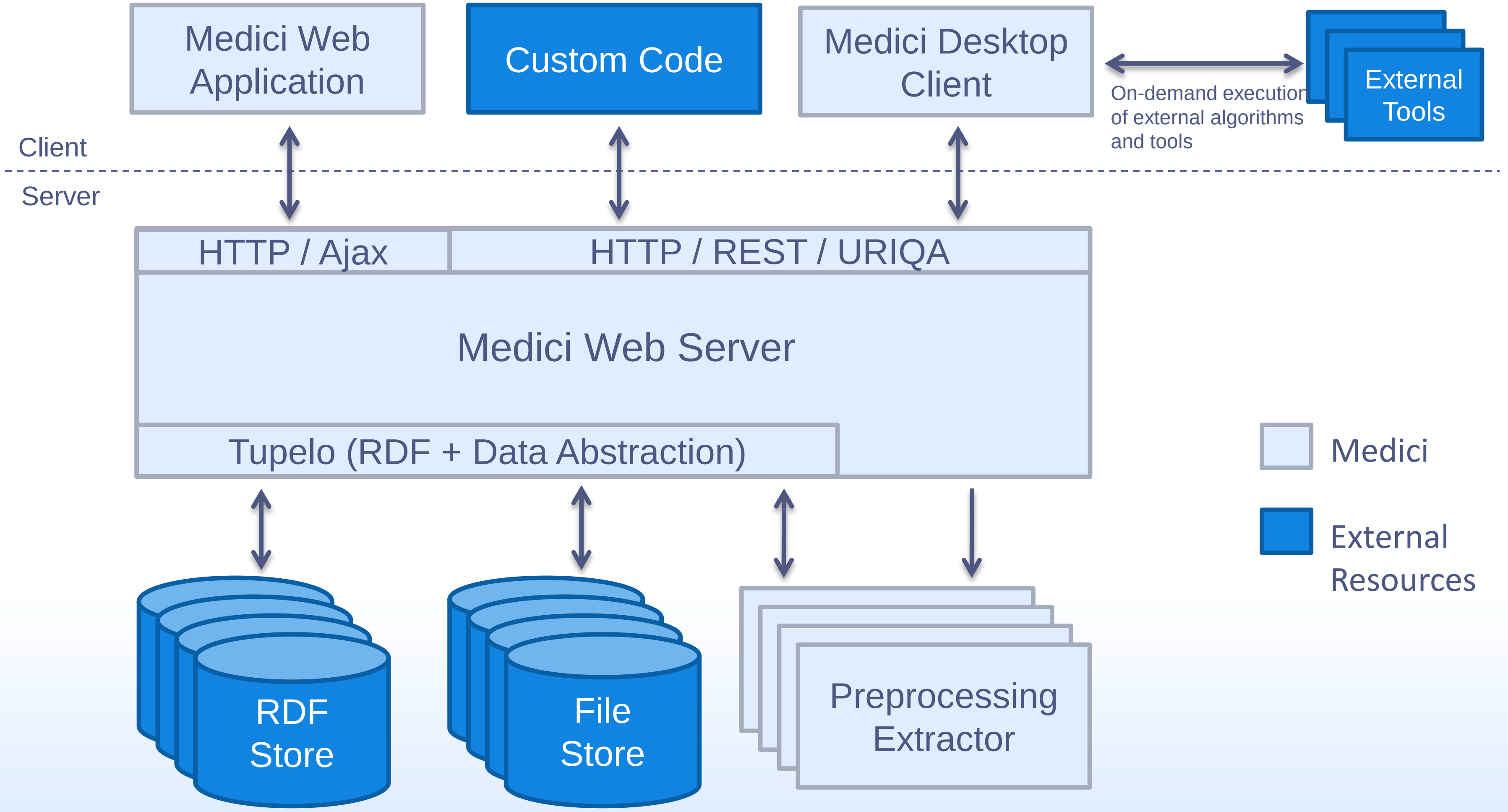
Extraction services

- Multiple, extensible pre-processing pipelines
- Asynchronous, distributed, triggered by upload
  - Processing selected on basis of file content type (MIME 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)

Medici 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

Software Architecture



Medici Communities

- Cyprus Institute (Digital Cultural Heritage)
  - 3D object archive for artifacts
- Datanet: sead.ncsa.illinois.edu
- Medici-demo.ncsa.illinois.edu
  - An open public server, upload requires account
- InvertNet.org
  - Digitization of Biological Collections
- Digging into Data
  - University of Sheffield, MATRIX Center
  - Given a set of images of historical artefacts, discover what salient characteristics make an artist different from others using computational image analysis
  - Enable statistical learning about individual and collective authorship.
- Walker Institute – Rule of Law
  - Repository of reports, video, satellite images for the Rule-of-Law in different locations around the world
- 18Connect – OCR of 18th century Manuscripts
  - Institute for Computing in Humanities, Arts, and Social Science
  - Extraction service to OCR manuscript images using Gamera OCR toolkit

Acknowledgements

- Institute for Advanced Computing Applications and Technologies (IACAT)
  - UIUC Campus collaboration
- NIH - (Image repository)
- iChass - (Digging into Data, 18Connect)
- NSF - (InVertnet, Datanet:SEAD)
- EPA / Seagrant
- Cyprus Institute Collaboration

ONR
Office of Naval Research

National Archives and
Records Administration

This work is supported by the Office of Naval Research (ONR), grant N00014-04-1-0437, and by the National Archives and Records Administration (NARA) as a supplement to NSF PACI cooperative agreement CA #SCI-9619019.

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