Chris A. Mattmann, Ph.D.
Jet Propulsion Laboratory, California Institute of Technology /
University of Southern California / Apache Software Foundation

Content Detection and Analysis for Big Data

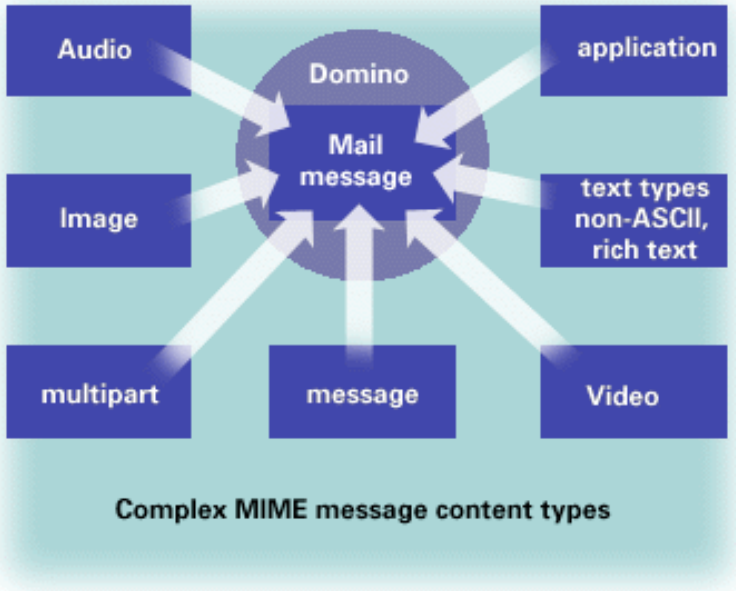
Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

Outline

- The Information Landscape
- Importance of Content Detection
- Challenges
- Approaches
- Introduction to Apache Tika
- Using Apache Tika
- Wrap-up

Goals

- Identify and classify file types
  - MIME detection
    - Glob pattern
      - *.txt
      - *.pdf
    - URL
      - http://…pdf
      - ftp://myfile.txt
    - Magic bytes
    - Combination of the above means



- Classification means reaction can be targeted

Goals

- Parsing
  - Based on MIME type in an automated fashion
  - Extraction of Text and Metadata

- Text content can be fed into
  - Search engine
  - Machine learning/Statistical analysis
  - Used to subset data from a formatted document
- Metadata can be used for field/faceted search

Many custom applications and tools

- You need this to read thi:

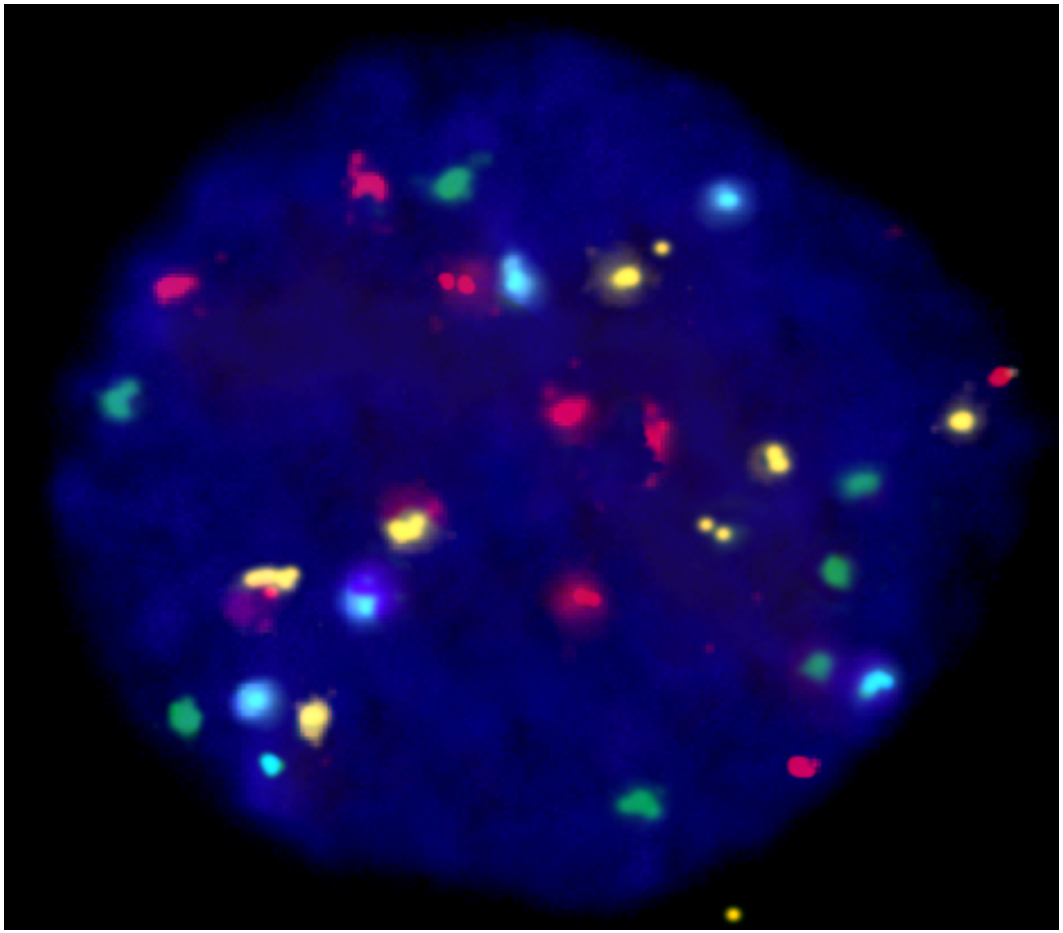
Third-party parsing libraries

- Most of the custom applications come with software libraries and tools to read/write these files
  - Rather than re-invent the wheel, figure out a way to take advantage of them
- Parsing text and structure is a difficult problem
  - Not all libraries parse text in equivalent manners
  - Some are faster than others
  - Some are more reliable than others

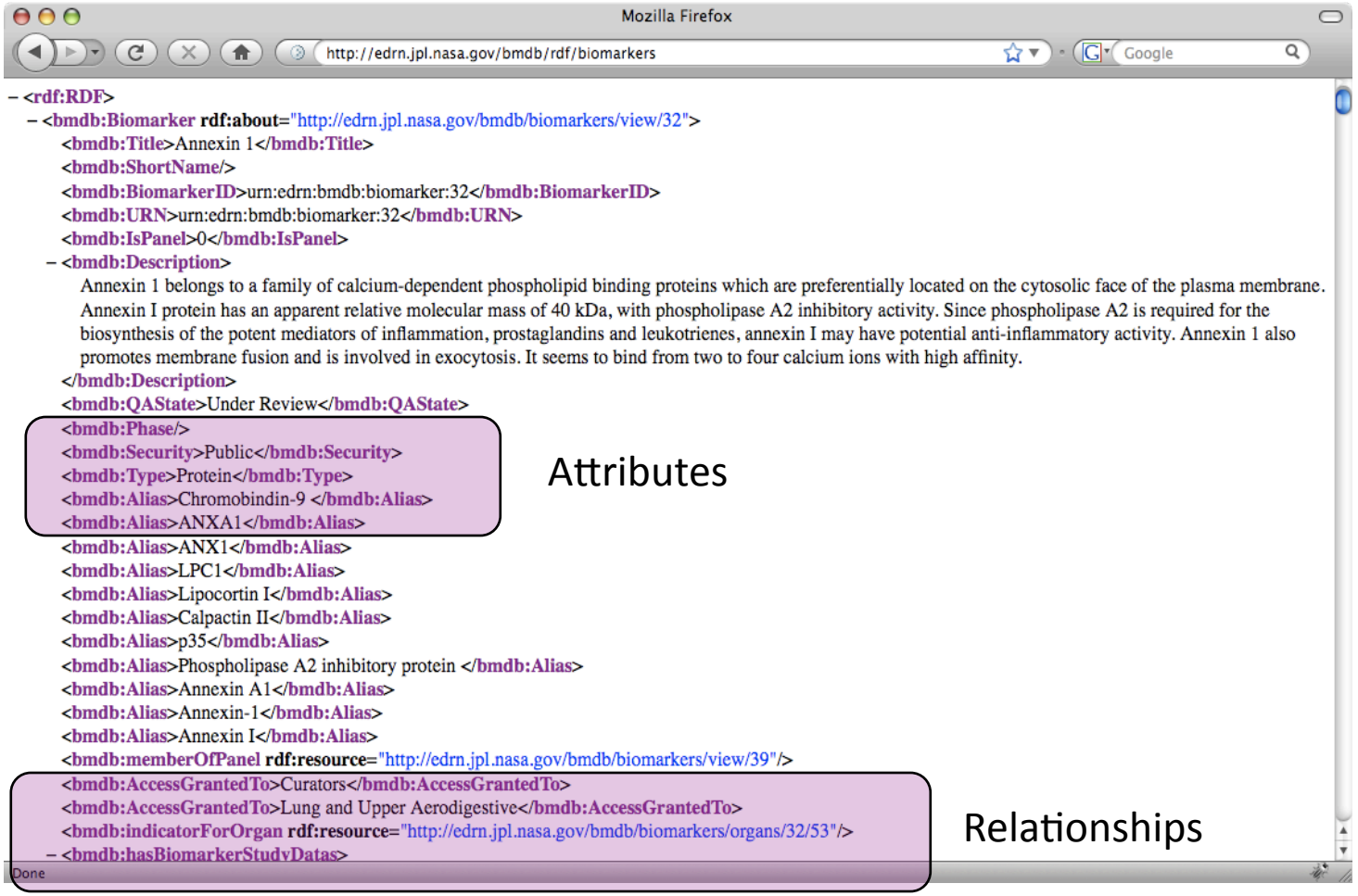
Extraction of Metadata

- Important to follow common Metadata models
  - Dublin Core
  - Word Metadata
  - XMP
  - EXIF
- Lots of standards and models out there
  - The use and extraction of common models allows for content intercomparison
  - All standardizes mechanisms for searching
  - You always know for X file type that field Y is there and of type String or Int or Date

Cancer Research Example



Cancer Research Example




```
– <rdf:RDF>
  – <bmdb:Biomarker rdf:about="http://edrn.jpl.nasa.gov/bmdb/biomarkers/view/32">
      <bmdb:Title>Annexin 1</bmdb:Title>
      <bmdb:ShortName/>
      <bmdb:BiomarkerID>urn:edrn:bmdb:biomarker:32</bmdb:BiomarkerID>
      <bmdb:URN>urn:edrn:bmdb:biomarker:32</bmdb:URN>
      <bmdb:IsPanel>0</bmdb:IsPanel>
    – <bmdb:Description>
        Annexin 1 belongs to a family of calcium-dependent phospholipid binding proteins which are preferentially located on the cytosolic face of the plasma membrane.
        Annexin I protein has an apparent relative molecular mass of 40 kDa, with phospholipase A2 inhibitory activity. Since phospholipase A2 is required for the
        biosynthesis of the potent mediators of inflammation, prostaglandins and leukotrienes, annexin I may have potential anti-inflammatory activity. Annexin 1 also
        promotes membrane fusion and is involved in exocytosis. It seems to bind from two to four calcium ions with high affinity.
      </bmdb:Description>
      <bmdb:QAState>Under Review</bmdb:QAState>
      <bmdb:Phase/>
      <bmdb:Security>Public</bmdb:Security>
      <bmdb:Type>Protein</bmdb:Type>
      <bmdb:Alias>Chromobindin-9 </bmdb:Alias>
      <bmdb:Alias>ANXA1</bmdb:Alias>
      <bmdb:Alias>ANX1</bmdb:Alias>
      <bmdb:Alias>LPC1</bmdb:Alias>
      <bmdb:Alias>Lipocortin I</bmdb:Alias>
      <bmdb:Alias>Calpactin II</bmdb:Alias>
      <bmdb:Alias>p35</bmdb:Alias>
      <bmdb:Alias>Phospholipase A2 inhibitory protein </bmdb:Alias>
      <bmdb:Alias>Annexin A1</bmdb:Alias>
      <bmdb:Alias>Annexin-1</bmdb:Alias>
      <bmdb:Alias>Annexin I</bmdb:Alias>
      <bmdb:memberOfPanel rdf:resource="http://edrn.jpl.nasa.gov/bmdb/biomarkers/view/39"/>
      <bmdb:AccessGrantedTo>Curators</bmdb:AccessGrantedTo>
      <bmdb:AccessGrantedTo>Lung and Upper Aerodigestive</bmdb:AccessGrantedTo>
      <bmdb:indicatorForOrgan rdf:resource="http://edrn.jpl.nasa.gov/bmdb/biomarkers/organs/32/53"/>
    – <bmdb:hasBiomarkerStudyDatas>
```

Language Identification

- Hard to parse out text and metadata from different languages
  - French document: J'aime la classe de CS 572!
    - Metadata:
      - Publisher: L'Universitaire de Californie en Etas-Unis de Sud
  - English document: I love the CS 572 class!
    - Metadata:
      - Publisher: University of Southern California
- How to compare these 2 extracted texts and sets of metadata when they are in different languages?

Methods for language identification

- N-grams
  - Method of detecting next character or set of characters in a sequence
  - Useful in determine whether small snippets of text come from a particular language, or character set
- Non-computational approaches
  - Tagging
  - Looking for common words or characters

Machine Translation

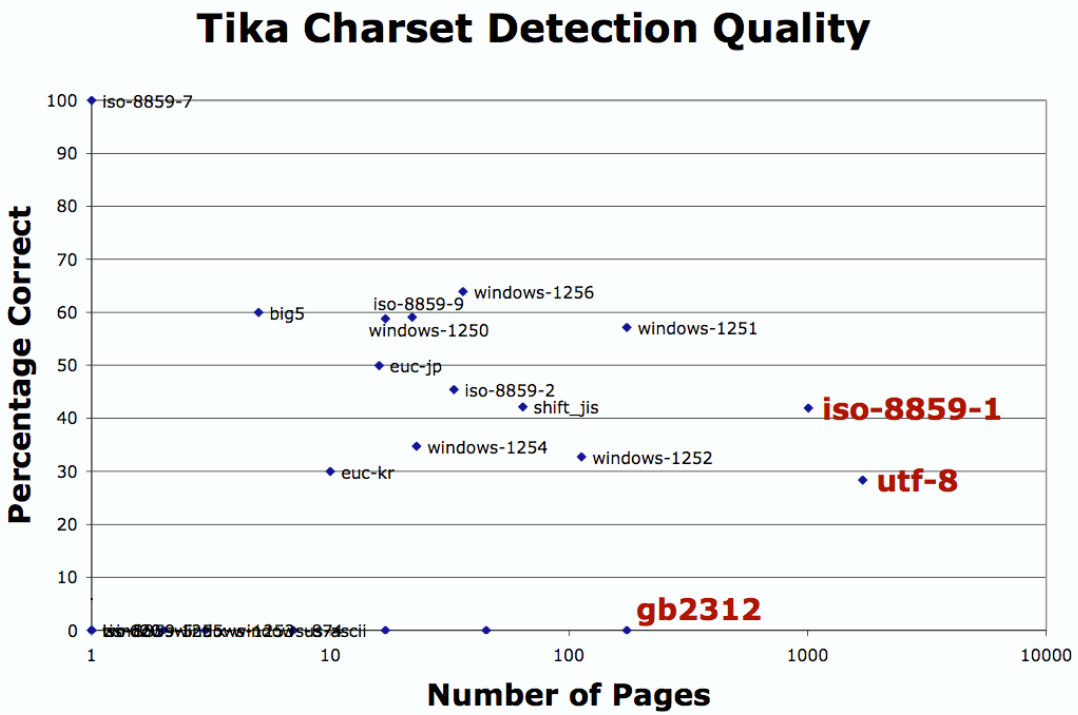
- Once you detect a language, automatically translating from a source language, to a destination language
- Field of statistical machine translation growing
- Many APIs and toolkits out there
  - APIS
    - Google Translate, Bing Translate, Lingo24
  - Toolkits
    - Moses, Joshua Decoder, etc.

Challenges

- Ability to uniformly extract and present metadata
- Scale
  - Extract on the fly, or extract during indexing?
  - Utility of content detection and analysis important both prior to indexing and after
- Integrating third-party parsing libraries is difficult
  - Many intrinsic dependencies
  - Non-uniform extraction interfaces
    - Some don't provide the same content
  - Slowdown

Challenges

- Language and charset detection is hard!



Challenges

- Maintenance of MIME type database as new MIMEs are constantly being identified
- Ensuring portability since content type detection and identification is becoming more and more needed even outside of the search engine
  - Firefox, Safari, HTTPD, etc., all must know about MIME types

Wrapup

- Content detection and analysis
  - MIME detection
  - Parsing and integration of parsing libraries
  - Language identification
  - Charset identification
  - Common Metadata models and formats
- Use in a number of areas within the domain of search engines