If You Have The Content, Then Apache Has the Technology!
This is not a comprehensive guide
these are the projects Nick used
and that I had time to test!
Proliferation of Content Types

• By some accounts, 16K to 51K content types*  
  • What to do with content types?  
    • Parse them, but How?  
    • Extract their text and structure  
    • Index their metadata  
    • In an indexing technology like Lucene, Solr, ElasticSearch  
  • Identify what language they belong to  
    • Ngrams

* http://fileext.com
Importance: Content Types
IANA MIME Registry

• 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
The Information Landscape
Content Types are Important
Let's see what Apache is doing to deal with Content!
Apache POI

- File format reader and writer for Microsoft office file formats
- Support binary & ooxml formats
- Strong read edit write for .xls & .xlsx
- Read and basic edit for .doc & .docx
- Read and basic edit for .ppt & .pptx
- Read for Visio, Publisher, Outlook
- Continues growing/improving with time
- [http://poi.apache.org](http://poi.apache.org)
- Got it working; MS TNEF example; docs outdated

```java
```

```
DHS open source government policy templatev4.docx  How to FOSS Your Government Project - TEMPLATE.DOCX
OSSWkingGroupMembershipList.xlsx
Freedom to Collaborate.odt
OSSBrainstormSummary.docx  message.rtf
How to FOSS Your Government Project - NSA version.docx
OSSFramework.xlsx
```

[chipotle:ApacheConNA2015/content-talk/poi-3.12-beta1] mattmann%
Apache PDFBox

• Read, Write, Create and Edit PDFs
• Create PDFs from text
• Fill in PDF forms
• Extract text and formatting (Lucene, Tika etc)
• Edit existing files, add images, add text etc
• Continues to improve with each release!

http://pdfbox.apache.org/

[chipotle:Apache/ApacheConNA2015/content-talk] mattmann% emacs HI_APACHECON.txt
[chipotle:Apache/ApacheConNA2015/content-talk] mattmann% java -jar pdfbox-app-1.8.9.jar TextToPDF HI_APACHECON.pdf HI_APACHECON.txt
Apache ODF Toolkit

- File format reader and writer for ODF (Open Document Format) files
- A bit like Apache POI for ODF
- ODFDOM – Low level DOM interface for ODF Files
- Simple API – High level interface for working with ODF Files
- ODF Validator – Pure java validator
- Validator page says managed using Mercurial – needs updating?
- Time to graduate? 😊

ODF Validator Result Page

Result for 07-08-22-MetaData-Examples.odt

The document is conformant ODF1.0!

**Details:**

07-08-22-MetaData-Examples.odt: Info: ODF version of root document: 1.0
07-08-22-MetaData-Examples.odt/META-INF/manifest.xml: Warning: The directory 'Thu INF/manifest.xml' file of ODF package '07-08-22-MetaData-Examples.odt'
07-08-22-MetaData-Examples.odt/META-INF/manifest.xml: Warning: The directory 'Cor the 'META-INF/manifest.xml' file of ODF package '07-08-22-MetaData-Examples.odt'
Apache Tika

• Java (+app +server +OSGi) library for detecting and extracting content
• Identifies what a blob of content is
• Gives you consistent, structured metadata back for it
• Parses the contents into plain text, HTML, XHTML or sax events
• Growing fast!
• http://tika.apache.org/
• [chipotle:/~/Desktop/Apache/ApacheConNA2015] mattmann% tika -t ACNA15_Mattmann_IfYouHaveContent-v1.pptx | more
• Chris A. Mattmann, NASA JPL, USC & the ASF
• @chrismattmann
• ...

...
Apache Cocoon

• Component Pipeline framework
• Plug together “Lego-Like” generators, transformers and serialisers
• Generate your content once in your application, serve to different formats
• Read in formats, translate and publish
• Can power your own “Yahoo Pipes”
• Modular, powerful and easy

• Seems like a web framework?
Apache Xalan

• XSLT processor
• XPath engine
• Java and C++ flavours
• Cross platform
• Library and command line executables
• Transform your XML
• Fast and reliable XSLT transformation engine
  • Project rebooted in 2014!
• http://xalan.apache.org
• Tried to find a quickstart example, found a couple, but didn’t know how to get the jar files, etc. Another time! (it’s a library, so we’ll cut them slack)
Apache XMLGraphics FOP

• XSL-FO processor in Java
• Reads W3C XSL-FO, applies the formatting rules to your XML document, and renders it
• Output to Text, PS, PDF, SVG, RTF, Java Graphics2D etc
• Lets you leave your XML clean, and define semantically meaningful rich rendering rules for it

• Yay example worked (remove –awt and pass it an output file.pdf name)
Apache SIS

• Spatial Information System
• Java library for working with geospatial content
• Enables geographic content searching, clustering and archiving
• Supports co-ordination conversions
• Implements GeoAPI 3.0, uses ISO-19115 + ISO-19139 + ISO-19111
• http://sis.apache.org/
• Yay QuickStart command line works! (make sure you type ./bin/sis)
Apache UIMA

• Unstructured Information analysis
• Lets you build a tool to extract information from unstructured data
• Language Identification, Segmentation, Sentences, Entries etc
• Components in C++ and Java
• Network enabled – can spread work out across a cluster
• Helped IBM to win Jeopardy!
• http://uima.apache.org/
Apache OpenNLP

- Natural Language Processing
- Various tools for sentence detection, tokenization, tagging, chunking, entity detection etc
- Maximum Entropy and Perception Based machine learning
- OpenNLP good when integrating NLP into your own solution
- http://opennlp.apache.org

- Best example I have: My USC student’s GeoTopicParser that uses OpenNLP to train polar text data as a model for geographic places; the uses geonames to lat/lng tag data
  - https://github.com/AranyaLi/GeoParsingNSF
Apache cTAKES

- Clinical Text Analysis and Knowledge Extraction System – cTAKES
- NLP system for information extraction from clinical records free text in EMR
- Identifies named entities from various dictionaries, e.g. diseases, procedures
- Does subject, content, ontology mappings, relations and severity
- Built on UIMA and OpenNLP
- 485Mb download…will try later!
Apache Mahout

- Scalable Machine Learning Library
- Large variety of scalable, distributed algorithms
- Clustering – find similar content
- Classification – analyse and group
- Recommendations
- Formerly Hadoop based, now moving to a DSL based on Apache Spark
- [http://mahout.apache.org](http://mahout.apache.org)
- Couldn’t figure out how to do something with it (I’ve heard it’s cool)
- Did find that it does LDA topic modeling (bookmark for later)
Apache Any23

• Anything To Triples
• Library, Web Service and CLI Tool
• Extracts structured data from many input formats
• RDF / RDFa / HTML with Microformats or Microdata, JSON-LD, CSV
• To RDF, JSON, Turtle, N-Triples, N-Quads, XML

http://any23.apache.org

• Couldn’t figure out binary distribution, so tried to build from source
Apache Blur

• Search engine for massive amounts of structured data at high speed
• Query rich, structured data model
• US Census example: show me all of the people in the US who were born in Alaska between 1940 and 1970 who are now living in Kansas.
• Maybe? Content → Classify → Search
• Built on Apache Hadoop
• http://incubator.apache.org/blur/

• Hadoop and Passwordless SSH; will try later
Apache Stanbol

- Set of re-usable components for semantic content management
- Components offer RESTful APIs
- Can add semantic services on top of existing content management
- Content Enhancement – reasoning to add semantic information
- Reasoning – add more semantic data
- Storage, Ontologies, Data Models etc
- [http://stanbol.apache.org](http://stanbol.apache.org)
- Had to build from source and first mirror didn’t work
- Source build failed:
Apache Clerezza

- For management of semantically linked data available via REST
- Service platform based on OSGi
- Makes it easy to build semantic web applications and RESTful services
- Fetch, store and query linked data
- SPARQL and RDF Graph API
- Renderlets for custom output
- Source and Release binaries links broken on the website
- Found TDB on http://archive.apache.org
Apache Jena

- Java framework for building Linked Data and Semantic Web applications
- High performance Triple Store
- Exposes as SPARQL http endpoint
- Run local, remote and federated SPARQL queries over RDF data
- Ontology API to add extra semantics
- Inference API – derive additional data
- http://jena.apache.org/
- Tutorial data doesn’t come with bin distro: had to find it
- https://jena.apache.org/tutorials/sparql_data/
Apache Marmotta

- Open source Linked Data Platform
- W3C Linked Data Platform (LDP)
- Read-Write Linked Data
- RDF Triple Store with transactions, versioning and rule based reasoning
- SPARQL, LDP and LDPath queries
- Caching and security
- Builds on Apache Stanbol and Solr
- [http://marmotta.apache.org/](http://marmotta.apache.org/)
What I’m not going to cover

• Data Management
• Serving Up Content
• Generating Content
• Working with Hosted Content
What would be great to do next

• Test more of the content technologies at Apache!
• Beers on me to the person who finishes testing all of them
• Report JIRA issues for the things I found
Acknowledgements

• Nick Burch, http://www.slideshare.net/gagravarr/if-you-have-the-content-then-apache-has-the-technology
• The projects I tested
• Everyone for listening!
Thank you!

• Chris Mattmann
• @chrismattmann
• mattmann@apache.org

• http://memex.jpl.nasa.gov/
• http://trec-dd.org/

• http://nsf-polar-cyberinfrastructure.github.io/datavis-hackathon/