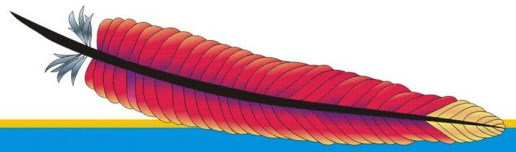
ApacheCon

If you have the Content, then Apache has the Technology!

A whistle-stop tour of the
Apache content related projects

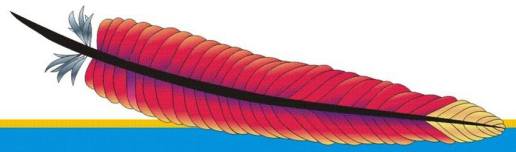


ApacheCon

Nick Burch

CTO

Quanticate

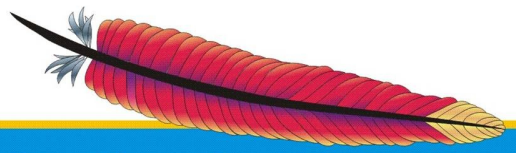


Apache Projects

- 154 Top Level Projects
- 33 Incubating Projects

- 46 “Content Related” Projects
- 8 “Content Related” Incubating Projects

(that excludes another ~30 fringe ones!)

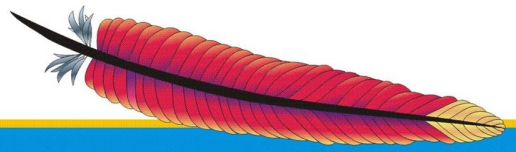


Picking the “most interesting” ones

36 Projects in 45 minutes

With time for questions...

This is not a comprehensive guide!



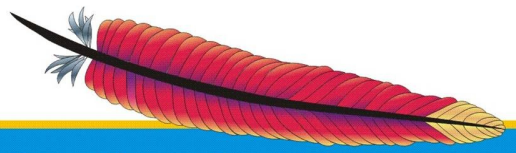
Active Committer - ~3 of these projects

Committer - ~6 of these projects

User - ~12 of these projects

Interested - ~24 of these projects

My experience levels / knowledge will vary from project to project!



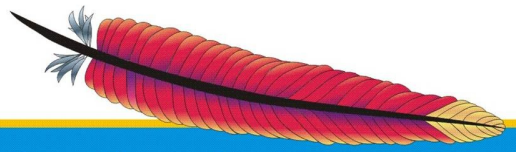
Different Technologies

- Transforming and Reading
- Text and Language Analysis
- RDF and Structured
- Data Management and Processing
- Serving Content
- Hosted Content

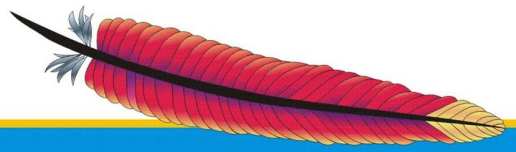
But not: Storing Content

What can we get in 45 mins?

- A quick overview of each project
- Roughly how they fit together / cluster into related areas
- When talks on the project are happening at ApacheCon
- The project's URL, so you can look them up and find out more!
- What interests ***me*** in the project



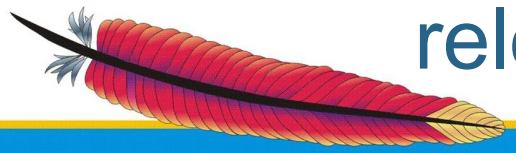
Transforming and Reading Content



Apache PDFBox

http://pdfbox.apache.org/

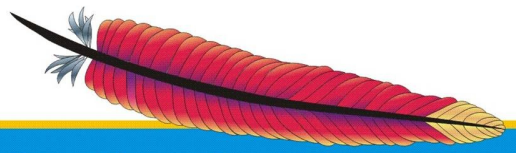
- Read, Write, Create and Edit PDFs
- Create PDFs from text
- Fill in PDF forms
- Extract text and formatting (Lucene, Tika etc)
- Edit existing files, add images, add text etc
- Continues to improve with each release!



Apache POI

http://poi.apache.org/

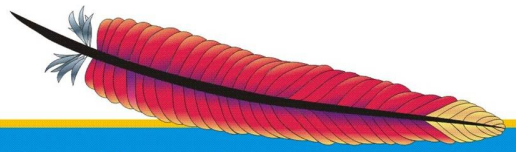
- File format reader and writer for Microsoft office file formats
- Support binary & ooxml formats
- Strong read edit write for .xls & .xlsx
- Read and basic edit for .doc & .docx
- Read and basic edit for .ppt & .pptx
- Read for Visio, Publisher, Outlook
- Continues growing/improving with time



ODF Toolkit (Incubating)

http://incubator.apache.org/odftoolkit/

- 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



Apache Tika

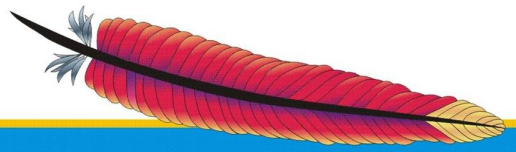
http://tika.apache.org/

- Talks – Tuesday + Wednesday
- 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!

Apache Cocoon

http://cocoon.apache.org/

- 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



Apache Xalan

http://xalan.apache.org/

- 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!

Apache XML Graphics: FOP

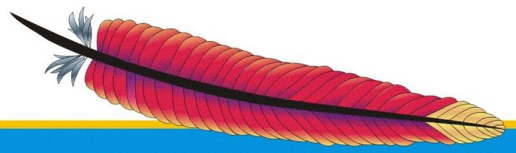
http://xmlgraphics.apache.org/fop/

- 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

Apache Commons: Codec

http://commons.apache.org/codec/

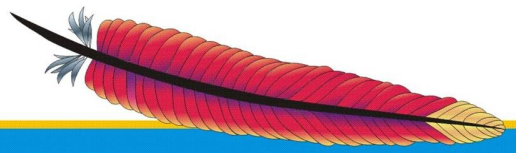
- Encode and decode a variety of encoding formats
- Base64, Base32, Hex, Binary Strings
- Digest – crypt(3) password hashes
- Caverphone, Metaphone, Soundex
- Quoted Printable, URL Encoding
- Handy when interchanging content with external systems



Apache Commons: Compress

http://commons.apache.org/compress/

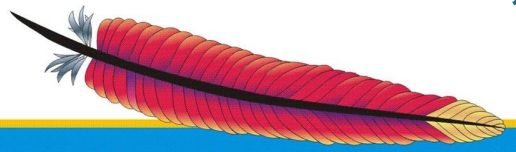
- Standard way to deal with archive and compression formats
- Read and write support
- zip, tar, gzip, bzip, ar, cpio, unix dump, XZ, Pack200, 7z, arj, lzma, snappy, Z
- Wider range of capabilities than java.util.Zip
- Common API across all formats



Apache Commons: Imaging

http://commons.apache.org/imaging/

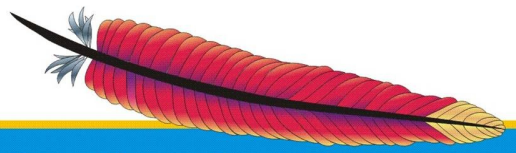
- Used to be called Commons Sanselan
- Pure Java image reader and writer
- Fast parsing of image metadata and information (size, color space, icc etc)
- Much easier to use than ImageIO
- Slower though, as pure Java
- Wider range of formats supported
- PNG, GIF, TIFF, JPEG + Exif, BMP, ICO, PNM, PPM, PSD, XMP



Apache SIS

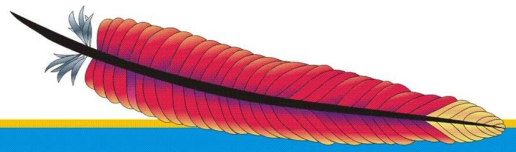
http://sis.apache.org/

- 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



Text and Language Analysis

Turing Content into Data



Apache UIMA

http://uima.apache.org/

- Unstructured Information analysis
- Lets you build a tool to extract information from unstructured data
- Language Identification, Segmentation, Sentences, Enties etc
- Components in C++ and Java
- Network enabled – can spread work out across a cluster
- Helped IBM to win Jeopardy!

Apache OpenNLP

http://opennlp.apache.org/

- 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
- UIMA wins for OOTB whole-solution

Apache cTAKES

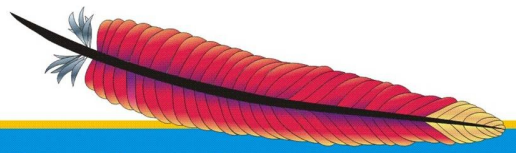
http://ctakes.apache.org/

- 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, eg diseases, procedues
- Does subject, content, ontology mappings, relations and severity
- Built on UIMA and OpenNLP

Apache Mahout

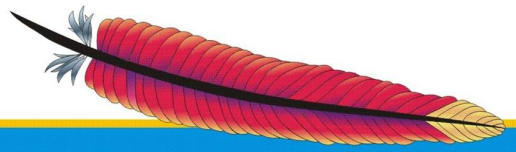
http://mahout.apache.org/

- 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



RDF, Structured and Linked Data

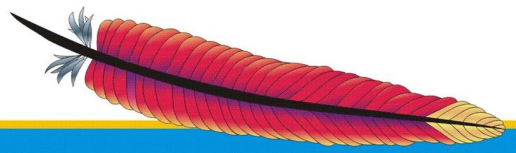
Track on Wednesday



Apache Any 23

http://any23.apache.org/

- Anything To Tripples
- 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



Apache Blur

http://incubator.apache.org/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

Apache Stanbol

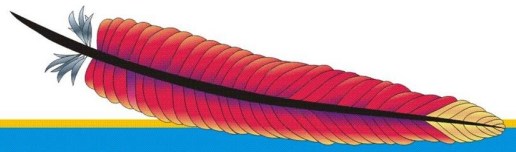
http://stanbol.apache.org/

- Set of re-usable components for semantic content management
- Components offer RESTful APIs
- Can add semantic services on top of existing content management systems
- Content Enhancement – reasoning to add semantic information to content
- Reasoning – add more semantic data
- Storage, Ontologies, Data Models etc

Apache Clerezza

http://clerezza.apache.org/

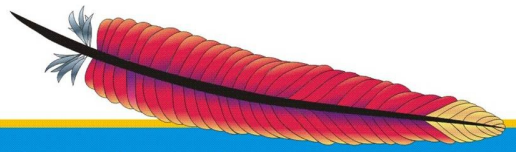
- For management of semantically linked data available via REST
- Service platform based on OSGi
- Makes it easy to build semantic web applications and RESTful services
- Fetch, store and query linked data
- SPARQL and RDF Graph API
- Renderlets for custom output



Apache Jena

http://jena.apache.org/

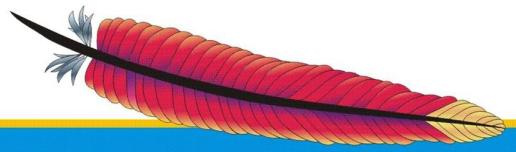
- Java framework for building Linked Data and Semantic Web applications
- High performance Tripple Store
- Exposes as SPARQL http endpoint
- Run local, remote and federated SPARQL queries over RDF data
- Ontology API to add extra semantics
- Inference API – derive additional data



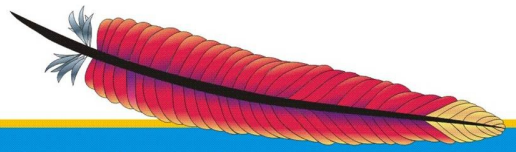
Apache Marmotta

http://marmotta.apache.org/

- Open source Linked Data Platform
- W3C Linked Data Platform (LDP)
- Read-Write Linked Data
- RDF Tripple Store with transactions, versioning and rule based reasoning
- SPARQL, LDP and LDPath queries
- Caching and security
- Builds on Apache Stanbol and Solr



Data Management and Processing



Apache Calcite (Incubating)

http://calcite.incubator.apache.org/

- Formerly known as Optiq
- Dynamic Data Management framework
- Highly customisable engine for planning and parsing queries on data from a wide variety of formats
- SQL interface for data not in relational databases, with query optimisation
- Complementary to Hadoop and NoSQL systems, esp. combinations of them

Apache MRQL (miracle)

http://mrql.apache.org/

- Large scale, distributed data analysis system, built on Hadoop, Hama, Spark
- Query processing and optimisation
- SQL-like query for data analysis
- Works on raw data in-situ, such as XML, JSON, binary files, CSV
- Powerful query constructs avoid the need to write MapReduce code
- Write data analysis tasks as SQL-like

Apache DataFu (Incubating)

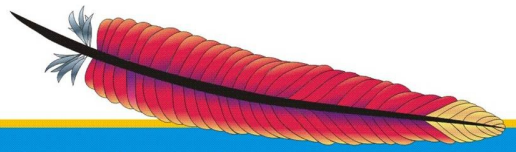
http://datafu.incubator.apache.org/

- Collection of libraries for working with large-scale data in Hadoop, for data mining, statistics etc
- Provides Map-Reduce jobs and high level language functions for data analysis, eg statistics calculations
- Incremental processing with Hadoop with sliding data, eg computing daily and weekly statistics

Apache Falcon (Incubating)

http://falcon.apache.org/

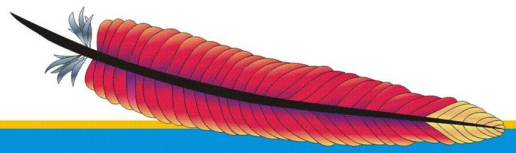
- Data management and processing framework built on Hadoop
- Quickly onboard data + its processing into a Hadoop based system
- Declarative definition of data endpoints and processing rules, inc dependencies
- Orchestrates data pipelines, management, lifecycle, motion etc



Apache Ignite (Incubating)

http://ignite.incuabtor.apache.org/

- Formerly known as GainGrid
- Only just entered incubation
- In-Memory data fabric
- High performance, distributed data management between heterogeneous data sources and user applications
- Stream processing and compute grid
- Structured and unstructured data

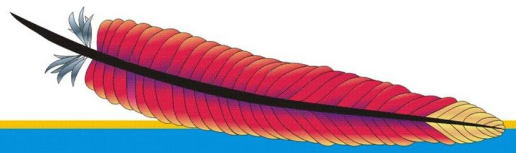


Serving up your Content

Apache HTTPD Server

http://httpd.apache.org/

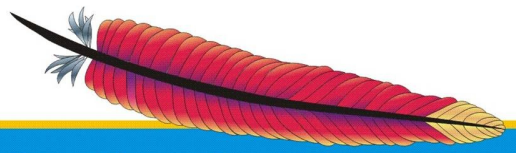
- Talks – All day today
- Very wide range of features
- (Fairly) easy to extend
- Can host most programming languages
- Can front most content systems
- Can proxy your content applications
- Can host code and content



Apache TrafficServer

http://trafficserver.apache.org/

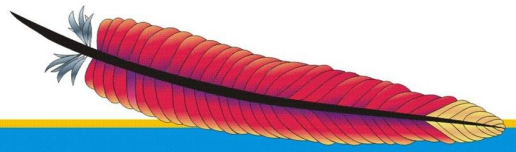
- High performance web proxy
- Forward and reverse proxy
- Ideally suited to sitting between your content application and the internet
- For proxy-only use cases, will probably be better than httpd
- Fewer other features though
- Often used as a cloud-edge http router



Apache Tomcat

http://tomcat.apache.org/

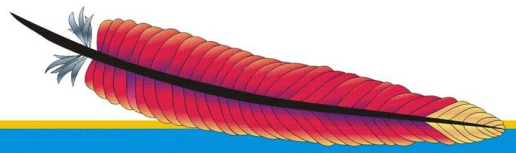
- Talks – Tuesday
- Java based, as many of the Apache Content Technologies are
- Java Servlet Container
- And you probably all know the rest!



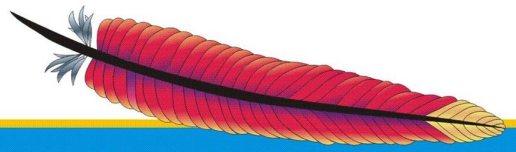
Apache Usergrid (Incubating)

http://usergrid.incubator.apache.org/

- Backend-as-a-Service “Baas” “mBaaS”
- Distributed NoSQL database + asset storage
- Mobile and server-side SDKs
- Rapidly build mobile and/or web applications, inc content driven ones
- Provides key services, eg users, queues, storage, queries etc



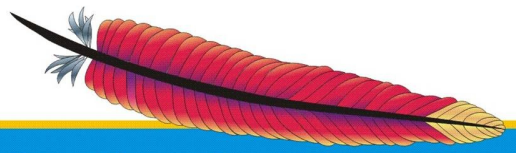
Generating Content



Apache OpenOffice

http://openoffice.apache.org

- Tracks – Tuesday and Wednesday
- Apache Licensed way to create, read and write your documents and content
- Our first big “Consumer Focused” project
- Can be used directly
- Or can be used as the upstream for other applications



Apache Forrest

http://forrest.apache.org/

- Document rendering solution build on top of cocoon
- Reads in content in a variety of formats (xml, wiki etc), applies the appropriate formatting rules, then outputs to different formats
- Heavily used for documentation and websites
- eg read in a file, format as changelog and readme, output as html + pdf

Apache Abdera

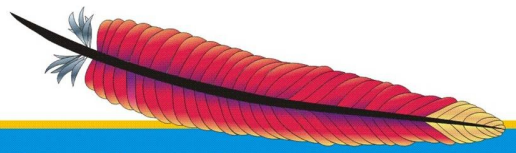
http://abdera.apache.org/

- Atom – syndication and publishing
- High performance Java implementation of RFC 4287 + 5023
- Generate Atom feeds from Java or by converting
- Parse and process Atom feeds
- Atompub server and clients
- Supports Atom extensions like GeoRSS, MediaRSS & OpenSearch

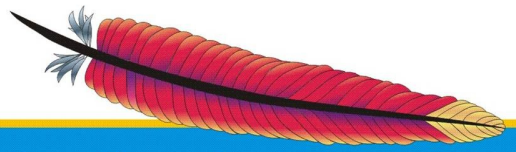
Apache JSPWiki

http://jspwiki.apache.org/

- Feature-rich extensible wiki
- Written in Java (Servlets + JSP)
- Fairly easy to extend
- Can be used as a wiki out of the box
- Provides a good platform for new wiki based application
- Rich wiki markup and syntax
- Attachments, security, templates etc



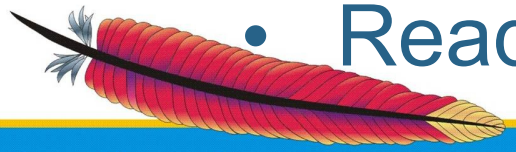
Working with Hosted Content



Apache Chemistry

http://chemistry.apache.org/

- Java, Python, .net, PHP, Mobile
- Atom, W*, Browser (JSON) interfaces
- OASIS CMIS (Content Management Interoperability Services)
- Client and Server bindings
- “SQL for Content”
- Consistent view on content across different repositories
- Read / Write / Manipulate content



Apache ManifoldCF

http://manifoldcf.apache.org/

- Name has changed a few times... (Lucene/Apache Connectors)
- Provides a standard way to get content out of other systems, ready for sending to Lucene etc
- Different goals to CMIS (Chemistry)
- Uses many parsers and libraries to talk to the different repositories / systems
- Analogous to Tika but for repos

Chemistry vs ManifoldCF

incubator /chemistry/ /connectors/

- ManifoldCF treats repo as nasty black box, and handles talking to the parsers
- Chemistry talks / exposes repo's contents through CMIS
- ManifoldCF supports a wider range of repositories
- Chemistry supports read and write
- Chemistry delivers a richer model
- ManifoldCF great for getting text out

Any Questions?

Any cool projects that
I happened to miss?