

The background of the slide is a photograph of the Arizona State Capitol building in Phoenix, Arizona. The building is a large, classical-style structure with a prominent central dome. In the foreground, to the left, is the Pioneer Monument, which features several bronze statues on a tiered stone base. The sky is blue with scattered white clouds. The overall image has a slightly desaturated, cyan-like color cast.

Chris A. Mattmann, NASA JPL,  
USC & the ASF

[@chrismattmann](https://twitter.com/chrismattmann)

[mattmann@apache.org](mailto:mattmann@apache.org)



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

The image shows a Google search for "language identification" with approximately 6,620,000 results. The search results list several websites, including basistech.com, translation-guide.com, faganfinder.com, and wikipedia.org. A file manager window is open in the foreground, displaying a list of content types and their associated actions. The content types listed include 3GPP Movie (audio/3gpp), 3GPP2 Movie (audio/3gpp2), AIFF Audio File (audio/aiff), and ASF (audio/x-asf). The actions listed include "Use QuickTime Plug-in 7.6.6 (in Fire...)", "Always ask", and "Use Flip4Mac Windows Media Plugin 2...". A red circle highlights the "[PDF] Language Identific..." result in the search results.

Web Images Videos Maps News Shopping Gmail more

Google

language identification

About 6,620,000 results (0.25 seconds)

Language Identification  
www.basistech.com Detect

Language Identification  
How to find out what language  
www.translation-guide.com/lar

Translation Wizard > Lar  
Aug 25, 2003 ... Identify the la  
of something, this page will ide  
www.faganfinder.com > Transl

Language identification  
Language identification is th  
is in. Traditionally, identifier  
en.wikipedia.org/wiki/Languag

Language Identification  
Language identification tools:  
This collection of language id  
genealogy.about.com/.../langu

Language Identification  
Determine the language and  
www.basistech.com/language

[PDF] Language Identific  
File Format: PDF/Adobe Acrot  
by CV Wright - Cited by 24 - R

University of Southern California

cars - Google Search

Content Type Action

- 3GPP Movie (audio/3gpp) Use QuickTime Plug-in 7.6.6 (in Fire...
- 3GPP Movie (video/3gpp) Use QuickTime Plug-in 7.6.6 (in Fire...
- 3GPP2 Movie (audio/3gpp2) Use QuickTime Plug-in 7.6.6 (in Fire...
- 3GPP2 Movie (video/3gpp2) Use QuickTime Plug-in 7.6.6 (in Fire...
- AIFF Audio File (audio/aiff) Use QuickTime Plug-in 7.6.6 (in Fire...
- aim Always ask
- asf Use Flip4Mac Windows Media Plugin 2...
- asx (video/x-ms-asx) Use Flip4Mac Windows Media Plugin 2...
- asx (video/x-ms-wmx) Use Flip4Mac Windows Media Plugin 2...
- AVI Movie (video/avi) Use QuickTime Plug-in 7.6.6 (in Fire...
- AVI Movie (video/mxvideo) Use QuickTime Plug-in 7.6.6 (in Fire...
- AVI Movie (video/x-msvideo) Use QuickTime Plug-in 7.6.6 (in Fire...

CARS.gov - Car Allowance Rebate System - Home - Formerly Referred ...  
Feb 22, 2010 ... The official website for the CARS Car Allowance Rebate System.  
www.cars.gov/ - Cached - Similar

Disney/Pixar Cars - The Official Site  
The latest Cars Toons and movie clips, character biographies, games, photos, toys and  
downloads from the Disney/Pixar movie Cars.  
disney.go.com/cars/ - Cached - Similar

Images for cars - Report images

Find: language identification Next Previous Highlight all Match case

Done

Search

Advanced search

Sponsored links

Used Car Listings  
Research & Compare Cars, Get Quotes  
By Zip Code & Find One Near You.  
Autos.AOL.com

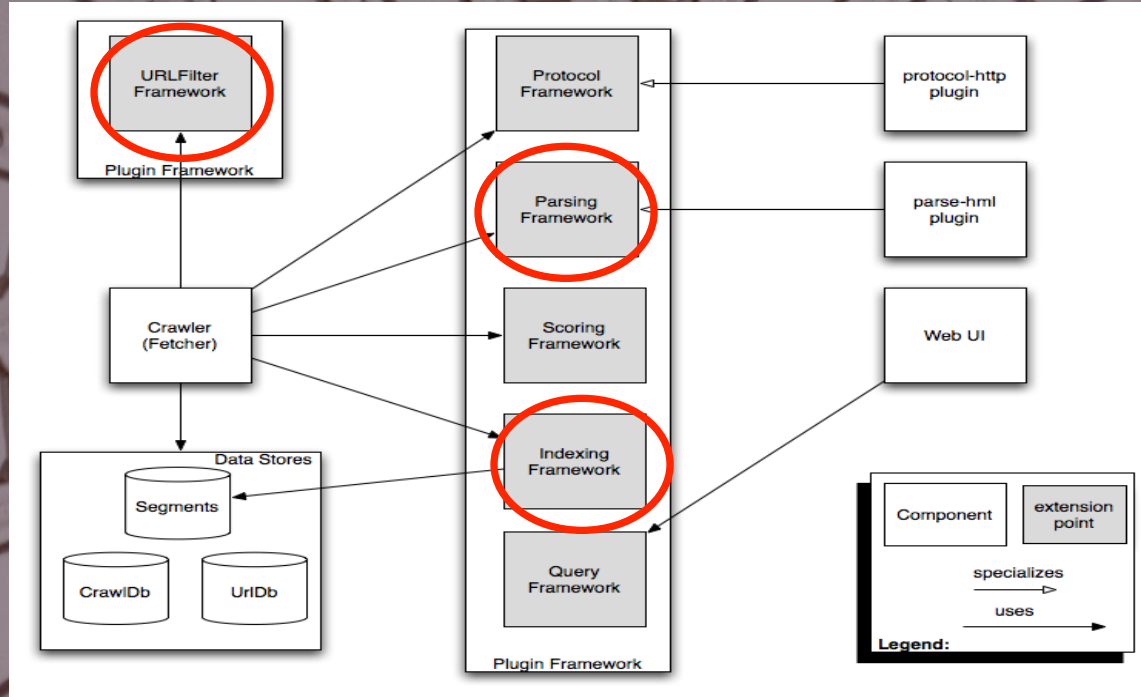
New and Used Car Research  
Free Car Price Quotes  
Research Cars at Edmunds.com  
www.Edmunds.com

Buy Cheap New Cars  
Don't Buy Retail New Cars!  
Find Dealers Offering Big Discounts  
dcyw.com/DriveCarsYouWant

Top Prices on New Cars  
Find out our Lowest Possible Price  
on New Cars, Trucks, and SUVs!  
CarPriceSecrets.com

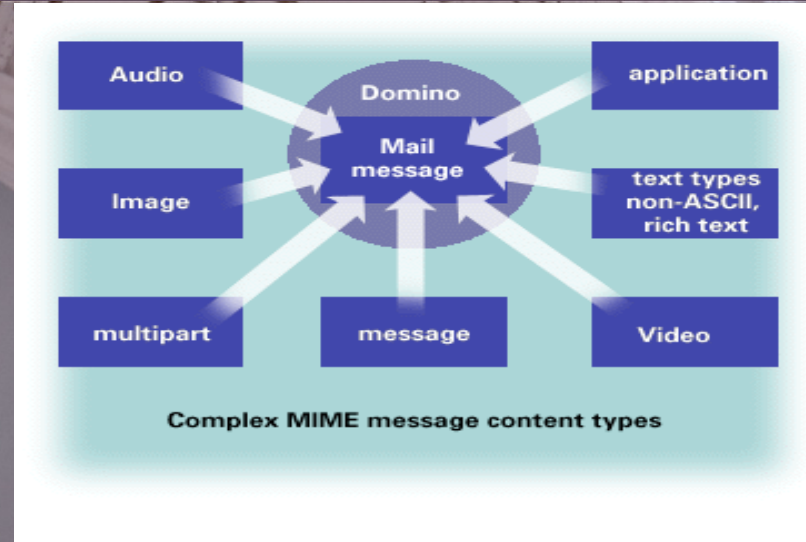
California - Cars  
Looking for Cars  
in California? Find it here!  
www.local.com  
California

# Importance: Content Types

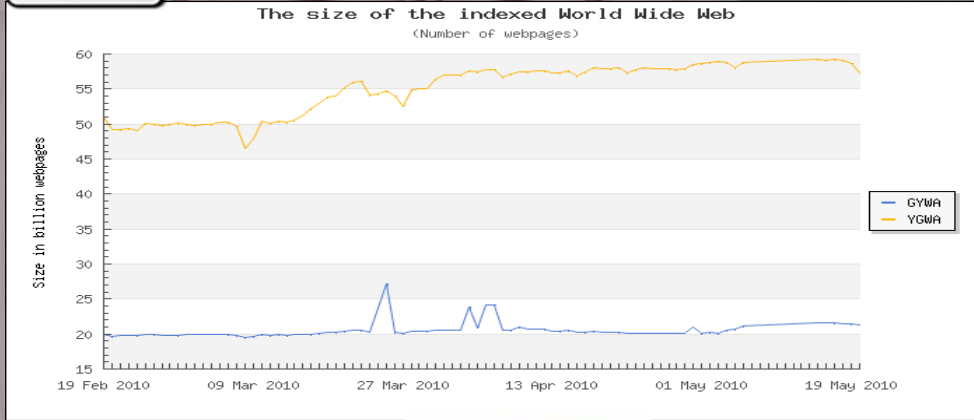


# IANA MIME Registry

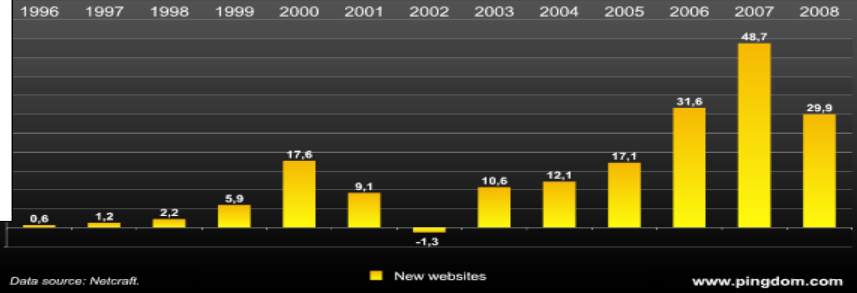
- Identify and classify file types
  - MIME detection
  - Glob pattern
    - \*.txt
    - \*.pdf
  - URL
    - http://...pdf
    - ftp://myfile.txt
  - Magic bytes
  - Combination of the above means
- Classification means reaction can be targeted



# The Information Landscape



**Increase in websites per year (in millions)**







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>
- Got it working; MS TNEF example; docs outdated

```
[chipotle:ApacheConNA2015/content-talk/poi-3.12-beta1] mattmann%  
-classpath poi-3.12-beta1-20150228.jar:poi-scratchpad-3.12-  
beta1-20150228.jar  
org.apache.poi.hmf.extractor.HMEFContentsExtractor /Users/mattman  
Desktop/STUFF/JPL/DARPA/DARPA\ XDATA/Open\ Source\ Program\  
Office/OSSInfo-Catalog/winmail.dat out  
Extracting...  
Extraction completed  
[chipotle:ApacheConNA2015/content-talk/poi-3.12-beta1] mattmann%  
out  
DHS open source government policy templatev4.docx How to FO  
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 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
- <http://incubator.apache.org/odftoolkit/>
- 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'!

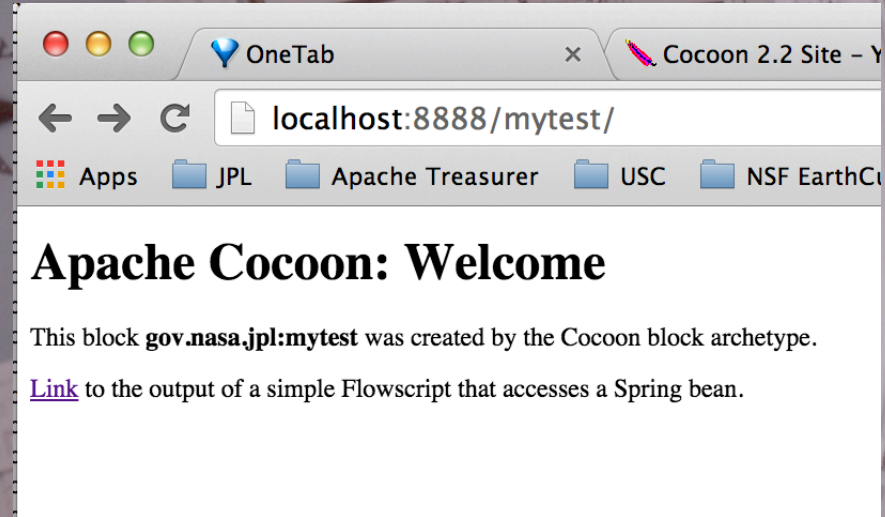
**07-08-22-MetaData-Examples.odt/META-INF/manifest.xml: Warning:** The directory 'Cor  
the 'META-INF/manifest.xml' file of ODF package '07-08-22-MetaData-Examples.odt'!

# 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
- <http://cocoon.apache.org/>
- 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
- <http://xmlgraphics.apache.org/fop/>
- Yay example worked (remove `-awt` and pass it an output file.pdf name)

This is not the latest Fop documentation, but just an fo example. FOP - p. 1

## FOP: An Open-Source XSL Formatter and Renderer

### A) What is FOP?

FOP is the world's first print formatter driven by XSL formatting objects. It is a Java 1.1 application that reads a formatting object tree and then turns it into a PDF document. The formatting object tree, can be in the form of an XML document (output by an XSLT engine like XT or Xalan) or can be passed in memory as a DOM Document or (in the case of XT) SAX events.

# 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)

```
[chipotle:ApacheConNA2015/content-talk/apache-sis-0.5] mattmann% ./bin/sis metadata https://github.com/opengeospatial/geoapi/raw/master/geoapi-netcdf/src/test/resources/org/opengis/wrapper/netcdf/NC
```

```
c  
WARNING This operation requires the "sis-temporal" module.
```

```
WARNING Can not assign units "degK" to dimension "SST".
```

```
Dimension of K is [Q], but the dimension of quantity of type javax.measure.quantity.Length is [L]
```

```
Metadata
```

```
├─Contact  
│   └─Role..... Point of contact  
│   └─Party  
│       └─Name..... NOAA/NWS/NCEP  
└─Spatial representation info  
    └─Number of dimensions..... 2
```



# Apache UIMA

- 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!
- <http://uima.apache.org/>

Annotation Results for Apache\_UIMA.txt.xml in examples/data/processed

efforts.

UIMA is a component framework for analysing unstructured content such as text, audio and video.  
It comprises an SDK and tooling for composing and running analytic components written in Java and C++, with some support for Perl, Python and TCL.

Apache UIMA mailing lists:

Users - uima-user@incubator.apache.org  
Developers - uima-dev@incubator.apache.org  
Commits - uima-commits@incubator.apache.org

Apache UIMA project committers:

Michael Baessler  
Edward Epstein  
Thilo Goetz  
Adam Lally  
Marshall Schor

Apache UIMA project Mentors:

Ken Coar (ASF member and Vice President)  
Sam Ruby (ASF member)

Legend

DocumentA...  EmailAddress  Name  PersonTitle  Sentence  
 Token

Click In Text to See Annotation Details

Annotations

- ▼ Name ("Michael Baessler")
  - begin = 2267
  - end = 2283
- ▼ Sentence

# 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/Ing 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, eg diseases, procedues
- Does subject, content, ontology mappings, relations and severity
- Built on UIMA and OpenNLP
- <http://ctakes.apache.org/>
- 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>
- 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)
- <http://mahout.apache.org/users/clustering/latent-dirichlet-allocation.html>

# 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

```
[chipotle:core/target/appassembler] mattmann% ./bin/any23 rover "https://git-wip-us.apache.org/repos/asf?p=tez.git;a=blog_plain;f=docs/src/site/resources/pmc/tez.rdf;hb=15857cb3"  
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".  
SLF4J: Defaulting to no-operation (NOP) logger implementation  
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
```

```
-----  
Apache Any23 :: rover  
-----
```

```
{ "quads" : [[ { "type" : "uri", "value" : "https://git-wip-us.apache.org/repos/tez"}, {"http://www.w3.org/1999/02/22-rdf-syntax-ns#type", { "type" : "uri", "value" : "http://projects.ap  
"}, null], [ { "type" : "uri", "value" : "https://git-wip-us.apache.org/repos/tez"}, {"http://projects.apache.org/ns/asfext#name", {"type" : "literal", "value" : "Apache Tez", "lang" : "  
"}, null], [ { "type" : "uri", "value" : "https://git-wip-us.apache.org/repos/tez"}, {"http://xmlns.com/foaf/0.1/homepage", { "type" : "uri", "value" : "http://tez.apache.org/"}, null], [   
ue" : "https://git-wip-us.apache.org/repos/tez"}, {"http://projects.apache.org/ns/asfext#chair", { "type" : "bnode", "value" : "node19is3lbfmx1"}, null], [ { "type" : "bnode", "value" : "  
http://www.w3.org/1999/02/22-rdf-syntax-ns#type", { "type" : "uri", "value" : "http://xmlns.com/foaf/0.1/Person"}, null], [ { "type" : "bnode", "value" : "node19is3lbfmx1"}, {"http://xmln  
{"type" : "literal", "value" : "Hitesh Shah", "lang" : "en", "datatype" : null}, null], [ { "type" : "uri", "value" : "https://git-wip-us.apache.org/repos/tez"}, {"http://projects.apache  
"}, {"type" : "literal", "value" : "Apache Tez is an effort to develop a generic application framework^M
```



# 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>
- Had to build from source and first mirror didn't work
- Source build failed:

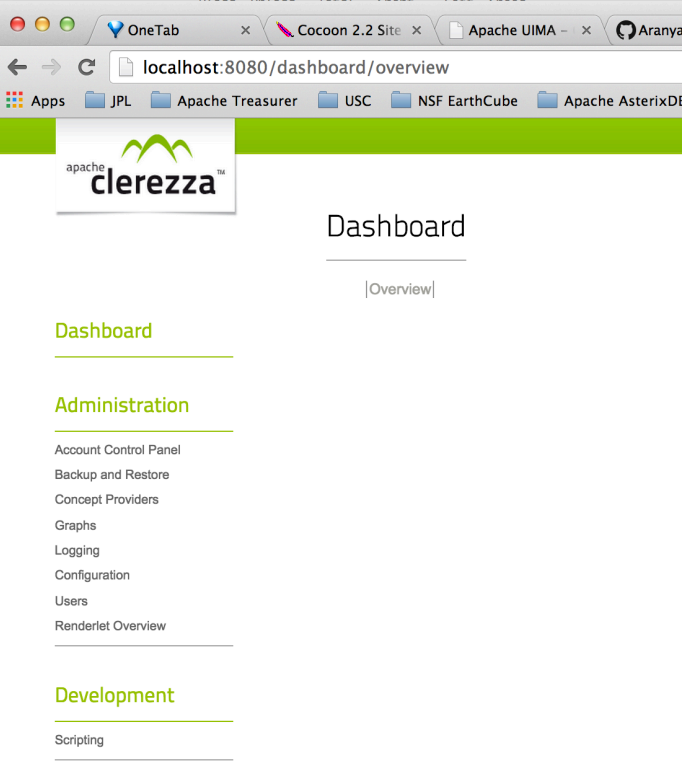
[ERROR] Failed to execute goal org.apache.maven.plugins:maven-antrun-plugin:1.6:run (download) on project org.apache.stanbol.data.sites.dbpedia: An Ant BuildException has occurred: The following error occurred while executing this line:

[ERROR] /Users/mattmann/Desktop/Apache/ApacheConNA2015/content-talk/apache-stanbol-0.12.0/data/sites/dbpedia/download\_index.xml:28: Failed to copy http://dev.iks-project.eu/downloads/stanbol-indices/\_26k.solrindex.bz2 to /Users/mattmann/Desktop/Apache/ApacheConNA2015/content-talk/apache-stanbol-0.12.0/data/sites/dbpedia/downloads/resources/org/apache/stanbol/data/site/dbpedia/default/index/dbpedia\_solrindex.bz2 due to Server returned HTTP response code: 504 for URL: http://dev.iks-project.eu/downloads/stanbol-indices/dbpedia\_26k.solrindex.bz2

# 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>



The screenshot shows a web browser window displaying the Apache Clerezza dashboard. The browser's address bar shows the URL `localhost:8080/dashboard/overview`. The dashboard has a green header with the Apache Clerezza logo. Below the header, the word "Dashboard" is centered, followed by a link for "[Overview]". The main content area is divided into three sections: "Dashboard", "Administration", and "Development". The "Administration" section contains a list of links: "Account Control Panel", "Backup and Restore", "Concept Providers", "Graphs", "Logging", "Configuration", "Users", and "Renderlet Overview". The "Development" section contains a link for "Scripting".

# 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/](https://jena.apache.org/tutorials/sparql_data/)

```
[chipotle:ApacheConNA2015/content-talk/apache-jena-2.13.0] mattmann% bin/sparql --data=doc/Tutorial/vc-db-1.rdf --query=doc/Tutorial/q1.rq
```

```
-----  
| x |
```

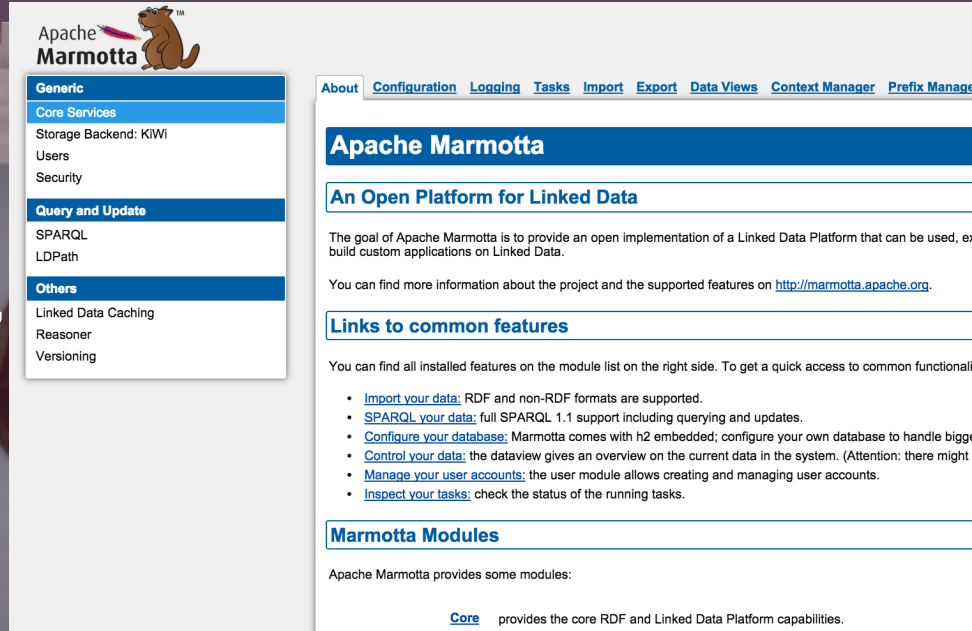
```
-----  
| <http://somewhere/JohnSmith/> |
```

```
-----  
[chipotle:ApacheConNA2015/content-talk/apache-jena-2.13.0] mattmann% █
```



# Apache Marmotta

- 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 LDPPath queries
- Caching and security
- Builds on Apache Stanbol and Solr
- <http://marmotta.apache.org/>



The screenshot shows the Apache Marmotta website. At the top left is the Apache Marmotta logo, featuring a brown bear holding a red arrow. Below the logo is a navigation menu with categories: Generic, Core Services, Query and Update, and Others. The 'Core Services' section lists 'Storage Backend: KiWi', 'Users', and 'Security'. The 'Query and Update' section lists 'SPARQL' and 'LDPPath'. The 'Others' section lists 'Linked Data Caching', 'Reasoner', and 'Versioning'. To the right of the navigation menu is a main content area with a blue header 'Apache Marmotta' and a sub-header 'An Open Platform for Linked Data'. Below this is a paragraph stating the goal of Apache Marmotta and a link to the project page. Further down is a section titled 'Links to common features' with a list of links: 'Import your data', 'SPARQL your data', 'Configure your database', 'Control your data', 'Manage your user accounts', and 'Inspect your tasks'. At the bottom is a section titled 'Marmotta Modules' with a paragraph stating 'Apache Marmotta provides some modules:' and a link for 'Core'.

Apache Marmotta

Generic

Core Services

Storage Backend: KiWi

Users

Security

Query and Update

SPARQL

LDPPath

Others

Linked Data Caching

Reasoner

Versioning

About Configuration Logging Tasks Import Export Data Views Context Manager Prefix Manager

## Apache Marmotta

### An Open Platform for Linked Data

The goal of Apache Marmotta is to provide an open implementation of a Linked Data Platform that can be used, extended, and build custom applications on Linked Data.

You can find more information about the project and the supported features on <http://marmotta.apache.org>.

### Links to common features

You can find all installed features on the module list on the right side. To get a quick access to common functionalities:

- [Import your data](#): RDF and non-RDF formats are supported.
- [SPARQL your data](#): full SPARQL 1.1 support including querying and updates.
- [Configure your database](#): Marmotta comes with h2 embedded; configure your own database to handle bigger datasets.
- [Control your data](#): the dataview gives an overview on the current data in the system. (Attention: there might be some data loss when using this feature.)
- [Manage your user accounts](#): the user module allows creating and managing user accounts.
- [Inspect your tasks](#): check the status of the running tasks.

### Marmotta Modules

Apache Marmotta provides some modules:

[Core](#) provides the core RDF and Linked Data Platform capabilities.

# What I'm not going to cover

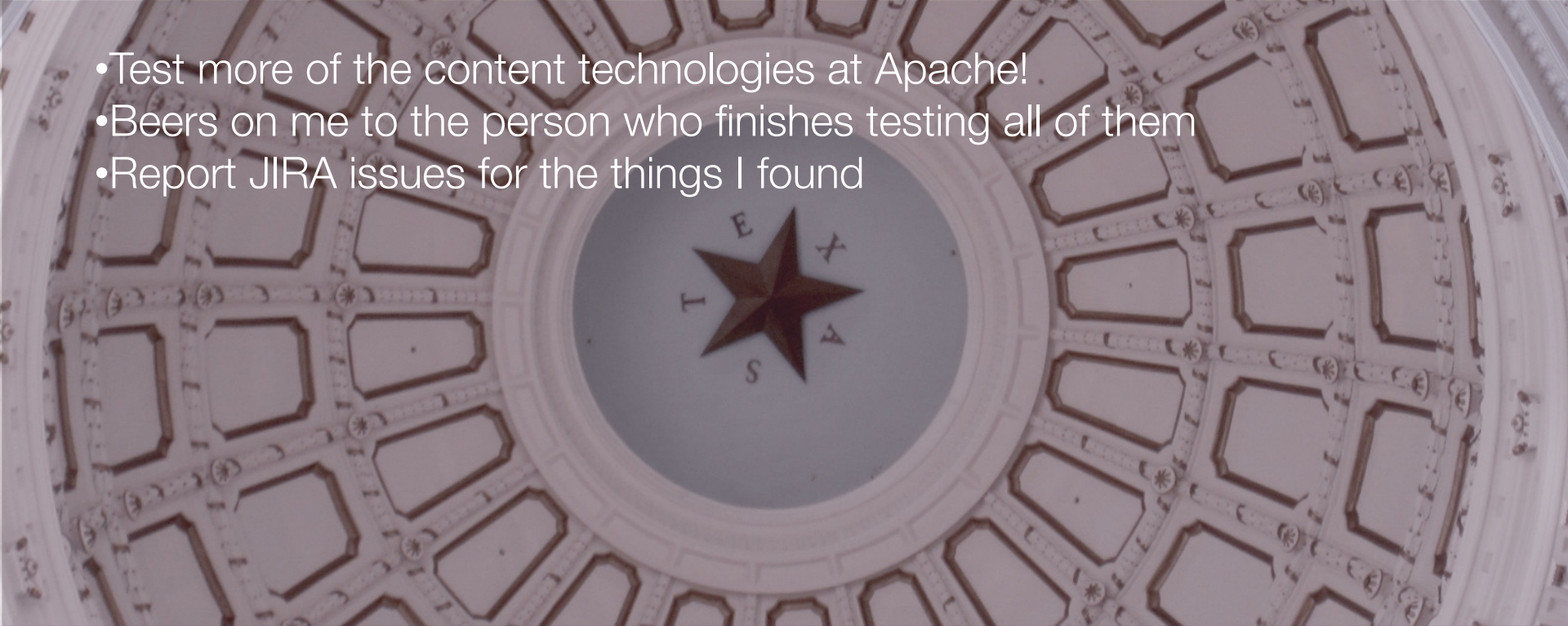
- Data Management
- Serving Up Content
- Generating Content
- Working with Hosted Content



# What would be great to do next

APACHE CON  
NORTH AMERICA

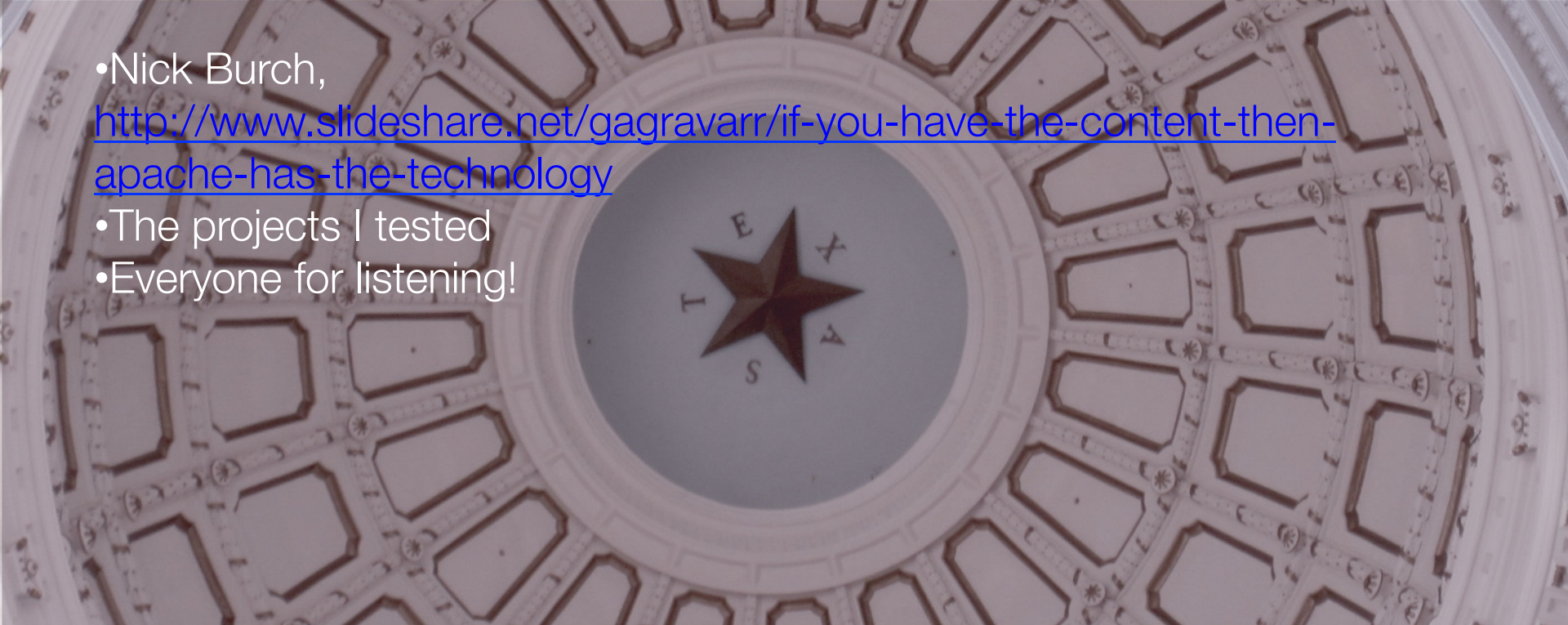
- Test more of the content technologies at Apache!
- Beers on me to the person who finishes testing all of them
- Report JIRA issues for the things I found





# 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](mailto:mattmann@apache.org)
- <http://memex.jpl.nasa.gov/>
- <http://trec-dd.org/>



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



The background of the slide is a photograph of the Arizona State Capitol building in Phoenix, Arizona. The building is a large, classical-style structure with a prominent central dome. In the foreground, to the left, is the Pioneer Monument, which features several bronze statues on a tiered stone base. The sky is blue with scattered white clouds. The overall image has a slightly desaturated, cyan-like tint.

Chris A. Mattmann, NASA JPL,  
USC & the ASF

[@chrismattmann](https://twitter.com/chrismattmann)

[mattmann@apache.org](mailto:mattmann@apache.org)