A Real-world Story of Converting a Database-based application to being Content-driven

Leading the Wave of Open Source

A Real-world Story of Converting a Database-based application to being Content-driven

The Longest and Most Unimaginative Title in the ApacheCon program

Brett Porter

Brett Porter

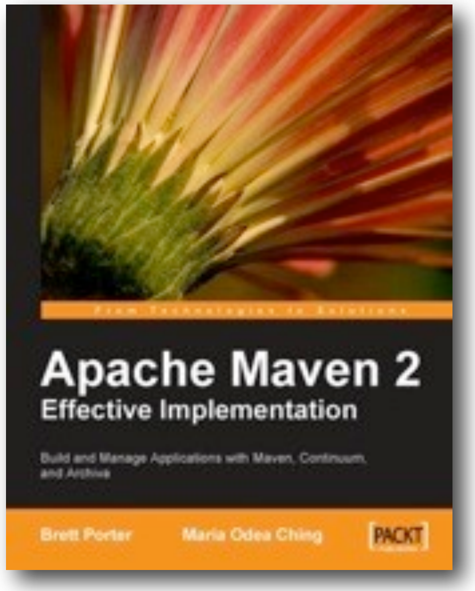
- Apache Software Foundation
  - Apache Maven, Archiva, Continuum, NPanday, Infrastructure, others
  - Member, former Director

Brett Porter

- Apache Software Foundation
  - Apache Maven, Archiva, Continuum, NPanday, Infrastructure, others
  - Member, former Director
- MaestroDev
  - VP, Product Development
  - Directing Maestro 3 development

Brett Porter

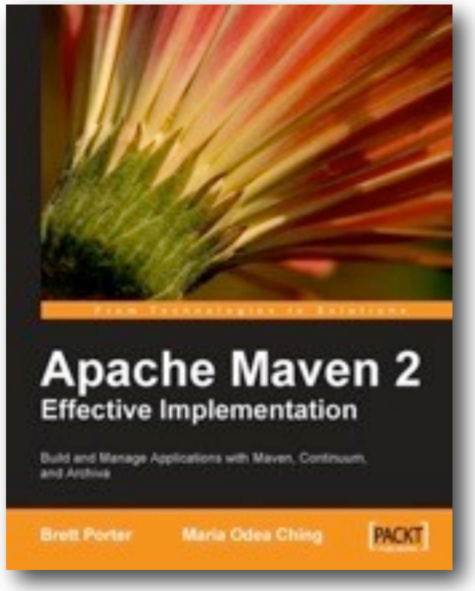
- Apache Software Foundation
  - Apache Maven, Archiva, Continuum, NPanday, Infrastructure, others
  - Member, former Director
- MaestroDev
  - VP, Product Development
  - Directing Maestro 3 development
- Co-author
  - Apache Maven 2: Effective Implementation
  - Better Builds with Maven



Leading the Wave of Open Source

Brett Porter

- Apache Software Foundation
  - Apache Maven, Archiva, Continuum, NPanday, Infrastructure, others
  - Member, former Director
- MaestroDev
  - VP, Product Development
  - Directing Maestro 3 development
- Co-author
  - Apache Maven 2: Effective Implementation
  - Better Builds with Maven
- Australian
  - Sydney



Content-related Experience

Leading the Wave of Open Source

Content-related Experience

- This page left intentionally blank

A Practical Example

- This is about a change we made in our project
- May apply to you if you have a similar challenge
- Only certain types of applications fit as a content application

A Practical Example

- This is about a change we made in our project
- May apply to you if you have a similar challenge
- Only certain types of applications fit as a content application

- Apache Archiva
  - You can check out the code for yourself, it's open source
  - http://svn.apache.org/repos/asf/archiva/trunk/

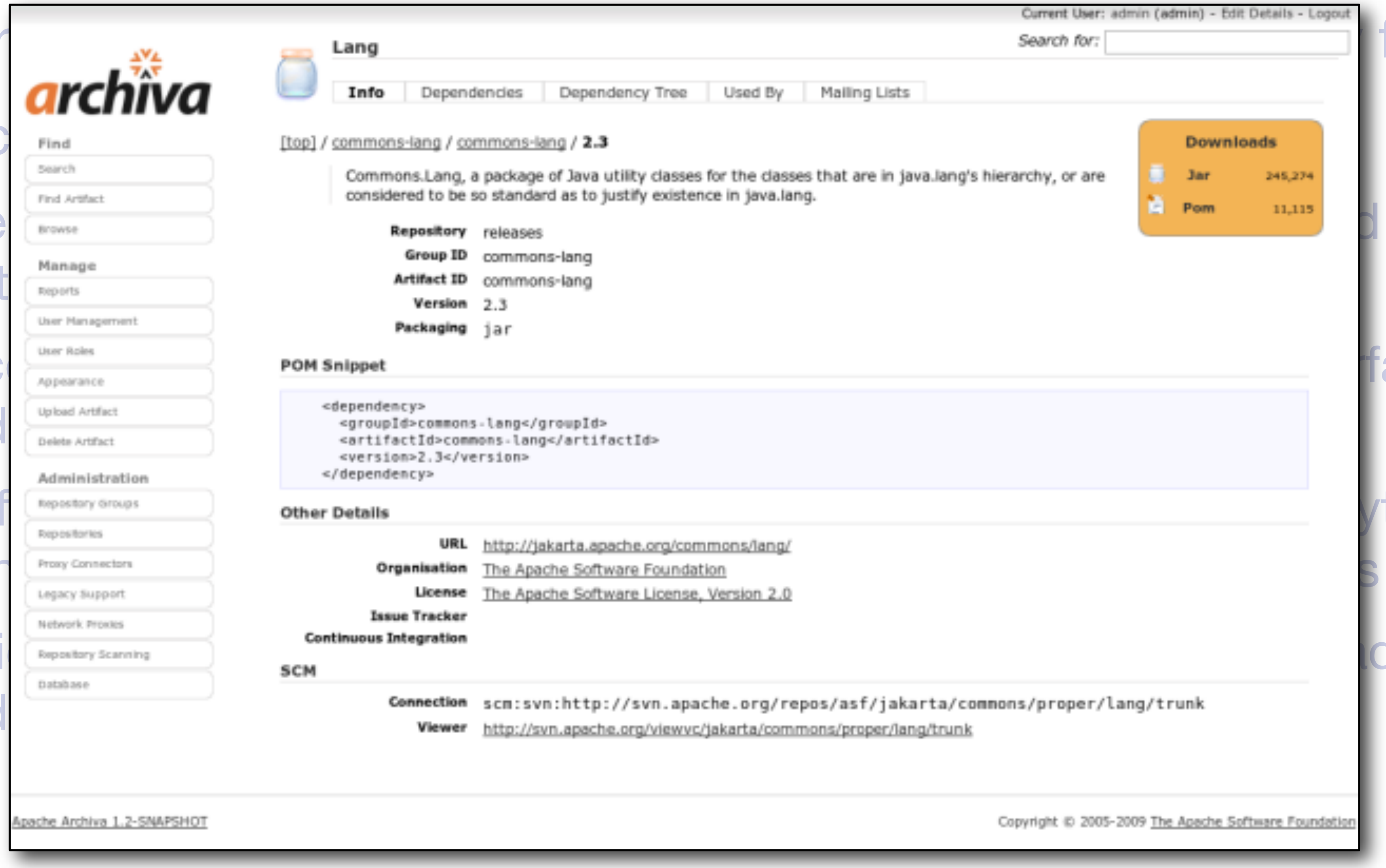
Leading the Wave of Open Source

Archiva: Some Background

- Repository Manager for Maven (and other similar tools)
- Naturally hierarchical content based on the Maven repository format
- Basically an artifact file server with a custom interface
  - rule-based retrieval and management of artifacts and associated metadata
  - access directly over HTTP and WebDAV, a user-driven web interface, and some web services
  - artifacts are typically binaries, ranging from small to multi-gigabyte, with information attached from the Maven POM or other sources
  - typically a large number of files, and rapid turnover as new are added and older development snapshots are purged

Archiva: Some Background

- Repository Manager for Maven (and other similar tools)
- Natur format
- Basic
  - rule d
    met
  - acc face,
    and
  - artif yte,
    with s
  - typi dded
    and



Current User: admin (admin) - Edit Details - Logout

Search for:

archiva

Find
- Search
- Find Artifact
- Browse

Manage
- Reports
- User Management
- User Roles
- Appearance
- Upload Artifact
- Delete Artifact

Administration
- Repository Groups
- Repositories
- Proxy Connectors
- Legacy Support
- Network Proxies
- Repository Scanning
- Database

Lang

Info | Dependencies | Dependency Tree | Used By | Mailing Lists

[top] / commons-lang / commons-lang / **2.3**

Commons.Lang, a package of Java utility classes for the classes that are in java.lang's hierarchy, or are considered to be so standard as to justify existence in java.lang.

Repository	releases
Group ID	commons-lang
Artifact ID	commons-lang
Version	2.3
Packaging	jar

Downloads

Jar	245,274
Pom	11,115

POM Snippet

```
<dependency>
  <groupId>commons-lang</groupId>
  <artifactId>commons-lang</artifactId>
  <version>2.3</version>
</dependency>
```

Other Details

URL	http://jakarta.apache.org/commons/lang/
Organisation	The Apache Software Foundation
License	The Apache Software License, Version 2.0
Issue Tracker	
Continuous Integration	

SCM

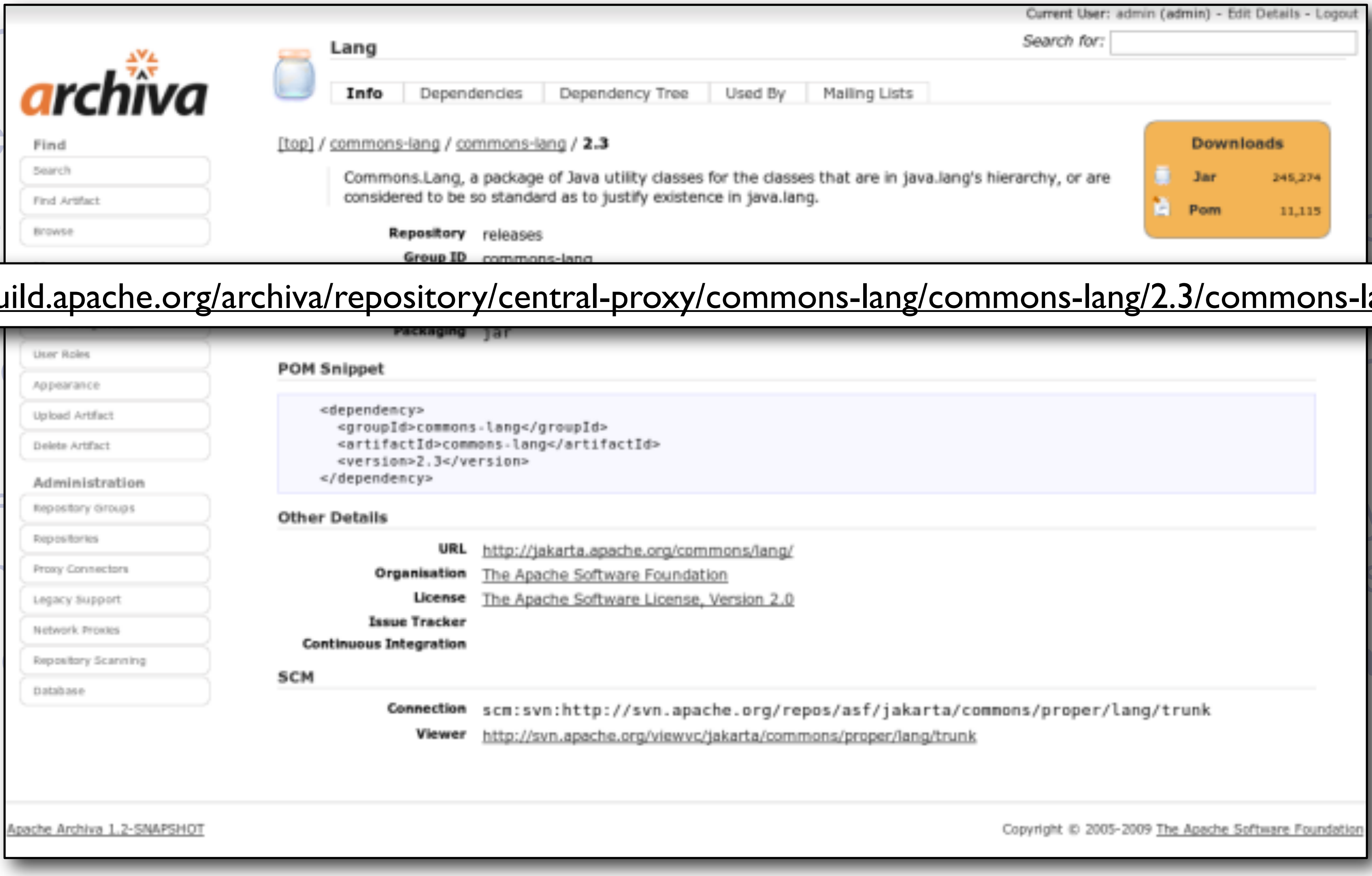
Connection	scm:svn:http://svn.apache.org/repos/asf/jakarta/commons/proper/lang/trunk
Viewer	http://svn.apache.org/viewvc/jakarta/commons/proper/lang/trunk

Apache Archiva 1.2-SNAPSHOT

Copyright © 2005-2009 The Apache Software Foundation

Archiva: Some Background

- Repository Manager for Maven (and other similar tools)
- Natur format
- Basic
  - rule d
  - acc face, and
  - artif yte, with s
  - typi dded and



Current User: admin (admin) - Edit Details - Logout

Search for:

Lang

Info | Dependencies | Dependency Tree | Used By | Mailing Lists

[top] / commons-lang / commons-lang / **2.3**

Commons.Lang, a package of Java utility classes for the classes that are in java.lang's hierarchy, or are considered to be so standard as to justify existence in java.lang.

Repository	releases
Group ID	commons-lang
Packaging	jar

Downloads	
Jar	245,274
Pom	11,115

POM Snippet

```
<dependency>
  <groupId>commons-lang</groupId>
  <artifactId>commons-lang</artifactId>
  <version>2.3</version>
</dependency>
```

Other Details

URL	http://jakarta.apache.org/commons/lang/
Organisation	The Apache Software Foundation
License	The Apache Software License, Version 2.0
Issue Tracker	
Continuous Integration	

SCM

Connection	scm:svn:http://svn.apache.org/repos/asf/jakarta/commons/proper/lang/trunk
Viewer	http://svn.apache.org/viewvc/jakarta/commons/proper/lang/trunk

Apache Archiva 1.2-SNAPSHOT

Copyright © 2005-2009 The Apache Software Foundation

http://vmbuild.apache.org/archiva/repository/central-proxy/commons-lang/commons-lang/2.3/commons-lang-2.3.jar

Other Repository Managers

Other Repository Managers

Nexus

- Uses Lucene and flat files for metadata
- Has an established anti-database stance
- Focuses on “self-healing” metadata to ensure integrity

Other Repository Managers

Nexus

- Uses Lucene and flat files for metadata
- Has an established anti-database stance
- Focuses on “self-healing” metadata to ensure integrity

artifactory

- Initially used JCR to store everything, including binary artifact data
- Claimed benefits of integrity
- Had reputation for wedging the database
- Harder to import/export the content
- Now seems to support a filesystem-only repository

Archiva History

- Around in part since March 2005
  - converting Maven 1 to Maven 2 repositories
  - relied heavily on scanning the repository on the filesystem and pulling out metadata
- Grew into a repository manager application as a Maven subproject
  - promoted to top level project Apache Archiva in March 2008
- Architecture was using Lucene as a “database”, but stored everything in its original form

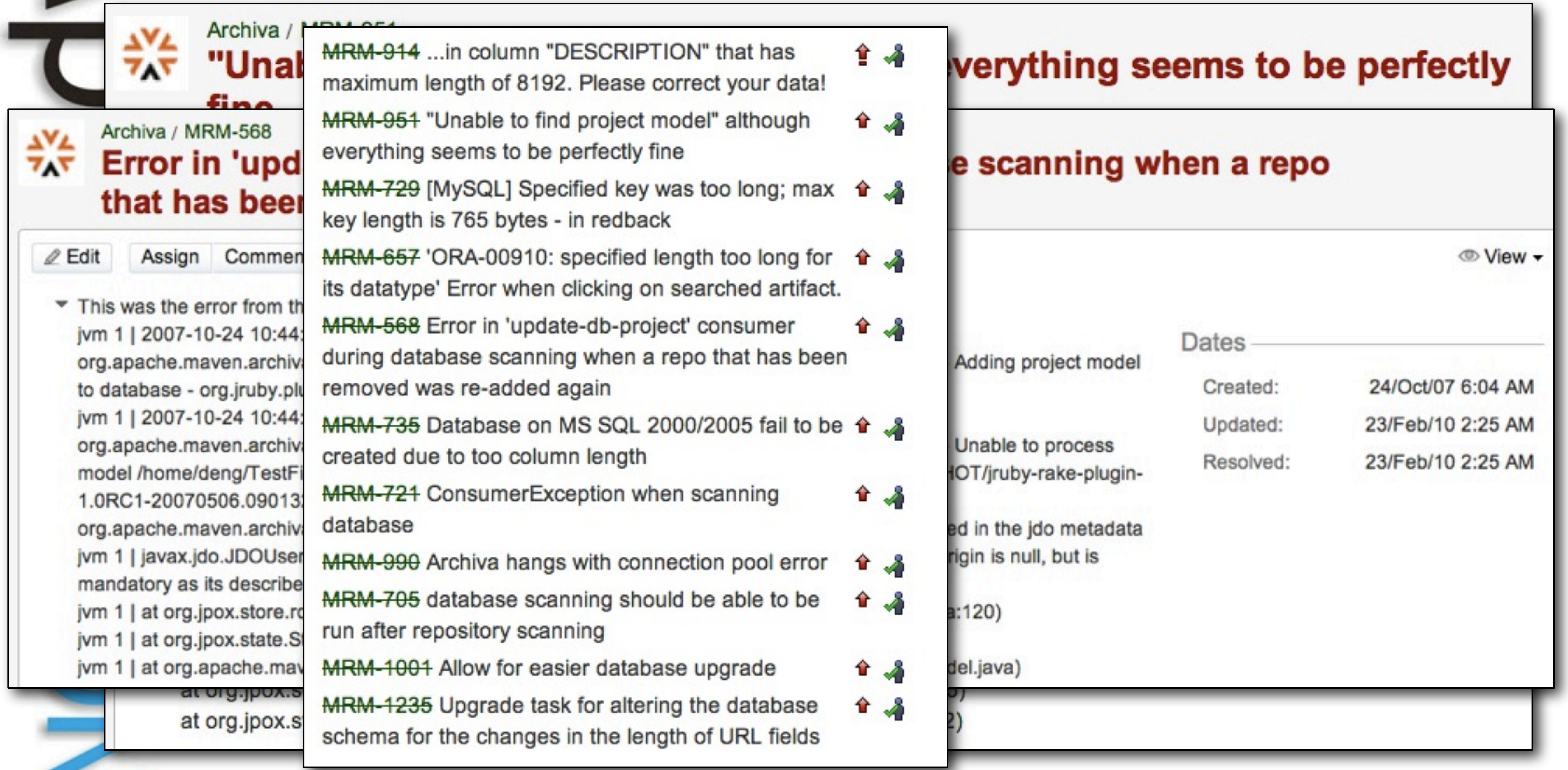
Archiva 1.0 Architecture

- Leap forward in functionality
- All of storage was re-done, partly using database
  - to be able to query easily and use persistence APIs
  - to do two phase scanning

This Didn’t Work Out So Well

- We used JDO 2 (JPOX) - not a great deal of resources for it
- Two-phase scanning could get out of sync
- Fell into the classic trap - only one person knew how it worked
- A few problems started to crop up
  - database exceptions deep in the stack that were hard to deal with
  - performance concerns
  - memory consumption (particularly with embedded database)
  - lack of extensibility for metadata
  - configuration for initial set up was not necessarily out of the box

This Didn't Work Out So Well



~~MRM-914~~ ...in column "DESCRIPTION" that has maximum length of 8192. Please correct your data!

~~MRM-951~~ "Unable to find project model" although everything seems to be perfectly fine

~~MRM-729~~ [MySQL] Specified key was too long; max key length is 765 bytes - in redback

~~MRM-657~~ 'ORA-00910: specified length too long for its datatype' Error when clicking on searched artifact.

~~MRM-568~~ Error in 'update-db-project' consumer during database scanning when a repo that has been removed was re-added again

~~MRM-735~~ Database on MS SQL 2000/2005 fail to be created due to too column length

~~MRM-721~~ ConsumerException when scanning database

~~MRM-990~~ Archiva hangs with connection pool error

~~MRM-705~~ database scanning should be able to be run after repository scanning

~~MRM-1001~~ Allow for easier database upgrade

~~MRM-1235~~ Upgrade task for altering the database schema for the changes in the length of URL fields

Archiva / MRM-568
Error in 'upd ... e scanning when a repo that has been
Edit
Assign
Commen
View
This was the error from th
Dates
Created: 24/Oct/07 6:04 AM
Updated: 23/Feb/10 2:25 AM
Resolved: 23/Feb/10 2:25 AM
"Unal ... everything seems to be perfectly fine

Leading the Wave of Open Source

Motivation for Change

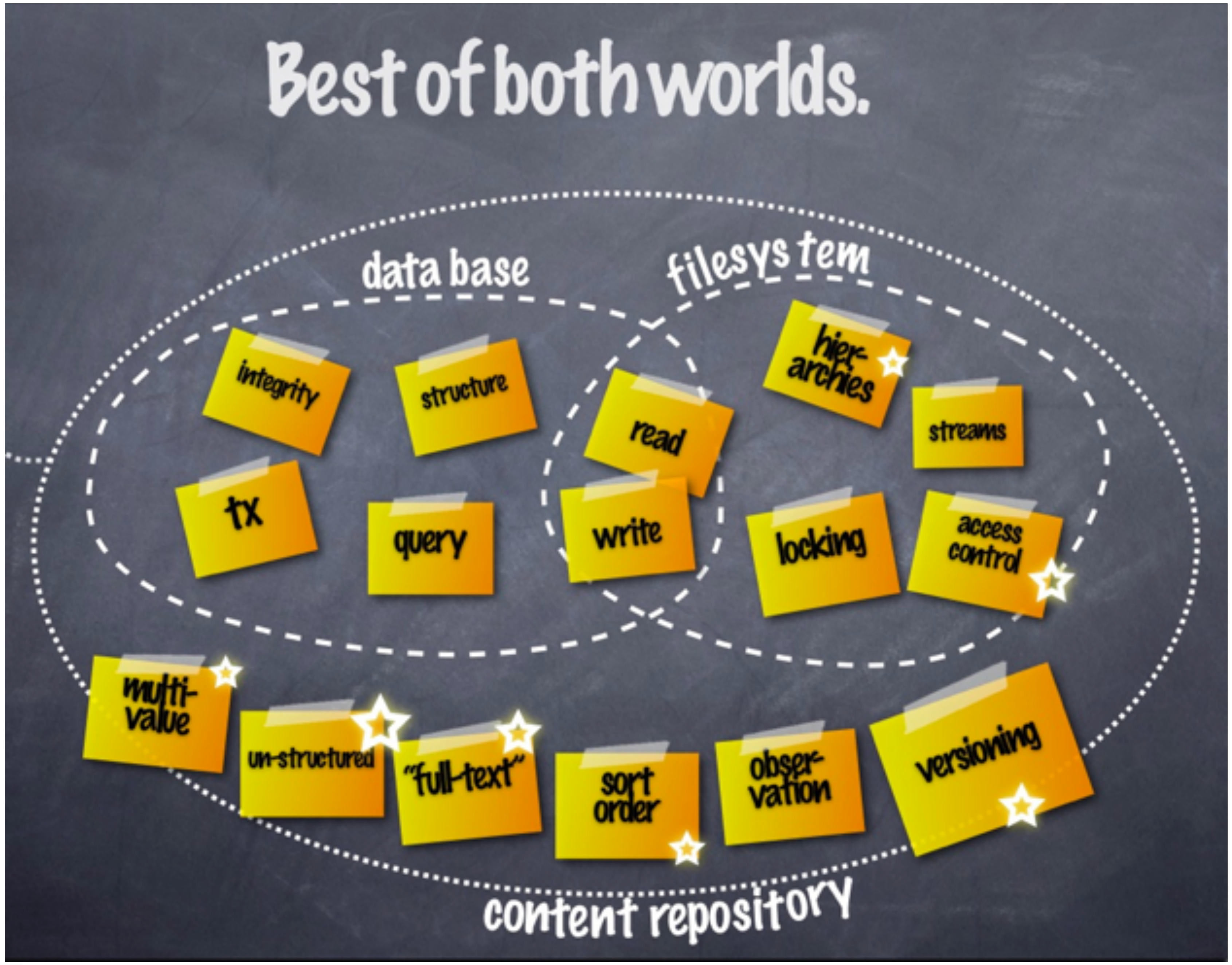
- The architecture was holding it back
- Wanted to implement extensible metadata for artifacts

Back to the Future

- Blend up the strengths of the original architecture, the newer feature set, and the added metadata
- Improve or solve the problems we'd seen
- Remove the database altogether
- Separate the metadata from the storage
- Pick up other improvements on the way, like lazy-loading content inside artifacts and proxying remote repositories
- Move toward a defined target architecture, and keep it working along the way
- Add a plugin architecture

Content vs. Database

- http://java.dzone.com/articles/java-content-repository-best



Hierarchical vs. Relational

- Hierarchy familiar for XML, DOM, filesystems
- How much structure is known in advance?
- What type of queries are needed?
  - hierarchy good at locating content, but not based on joined data
- Databases are not as good at transitive retrieval, or navigation / traversal of data

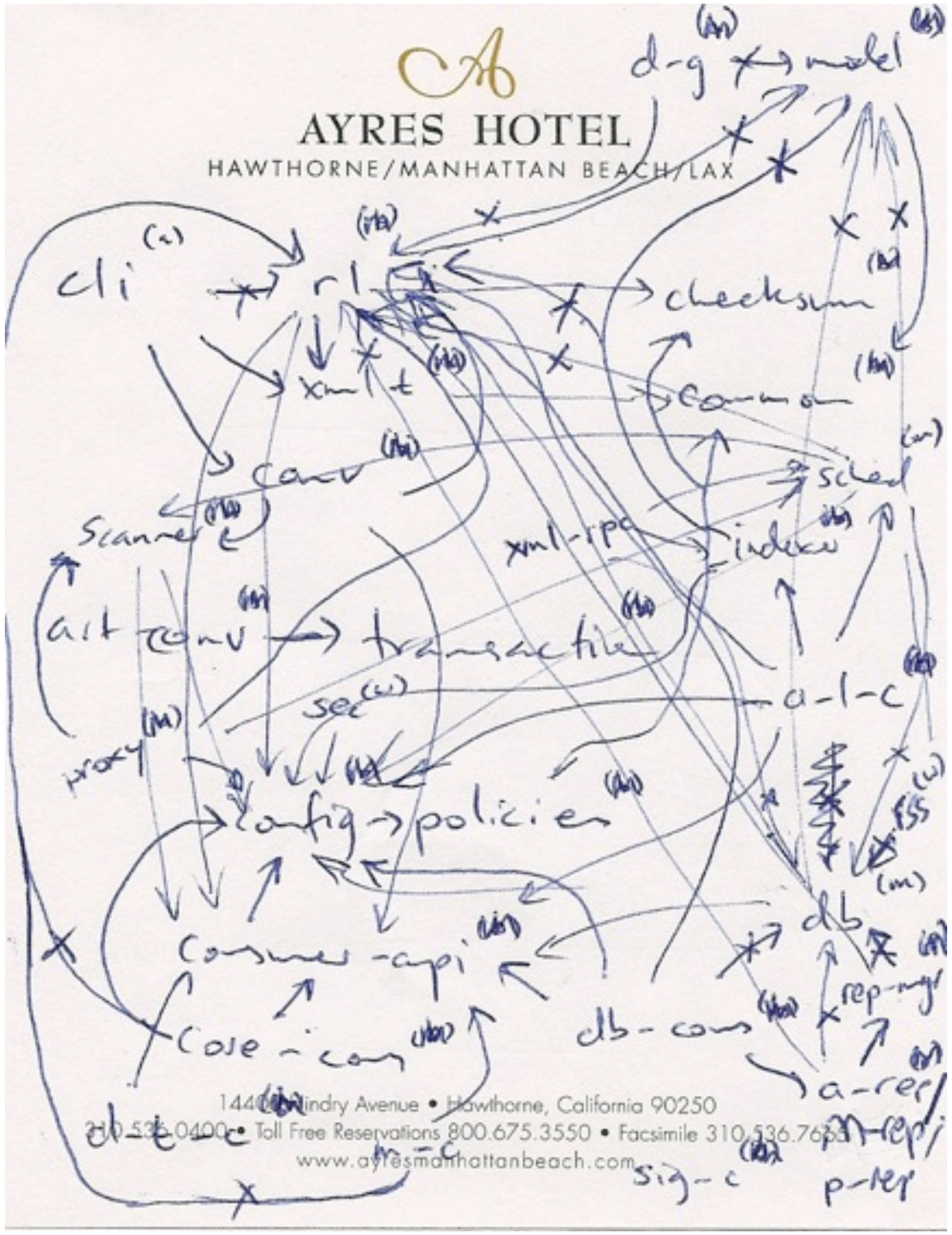
- Archiva is hierarchical
  - filesystem-like structure
  - POM inheritance & dependency relationships

Archiva: First Steps

- Reviewed the architecture

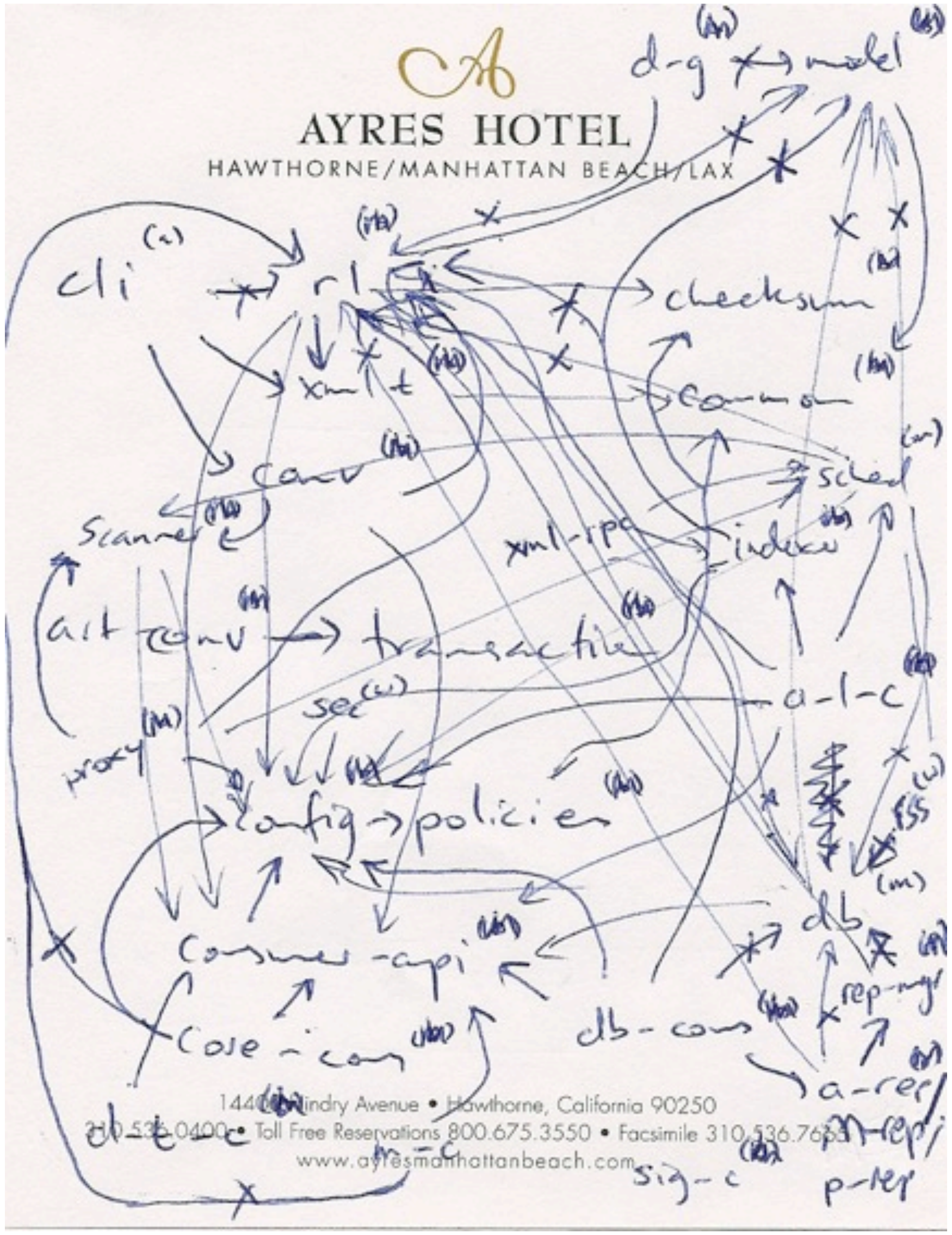
Archiva: First Steps

- Reviewed the architecture



Archiva: First Steps

- Reviewed the architecture

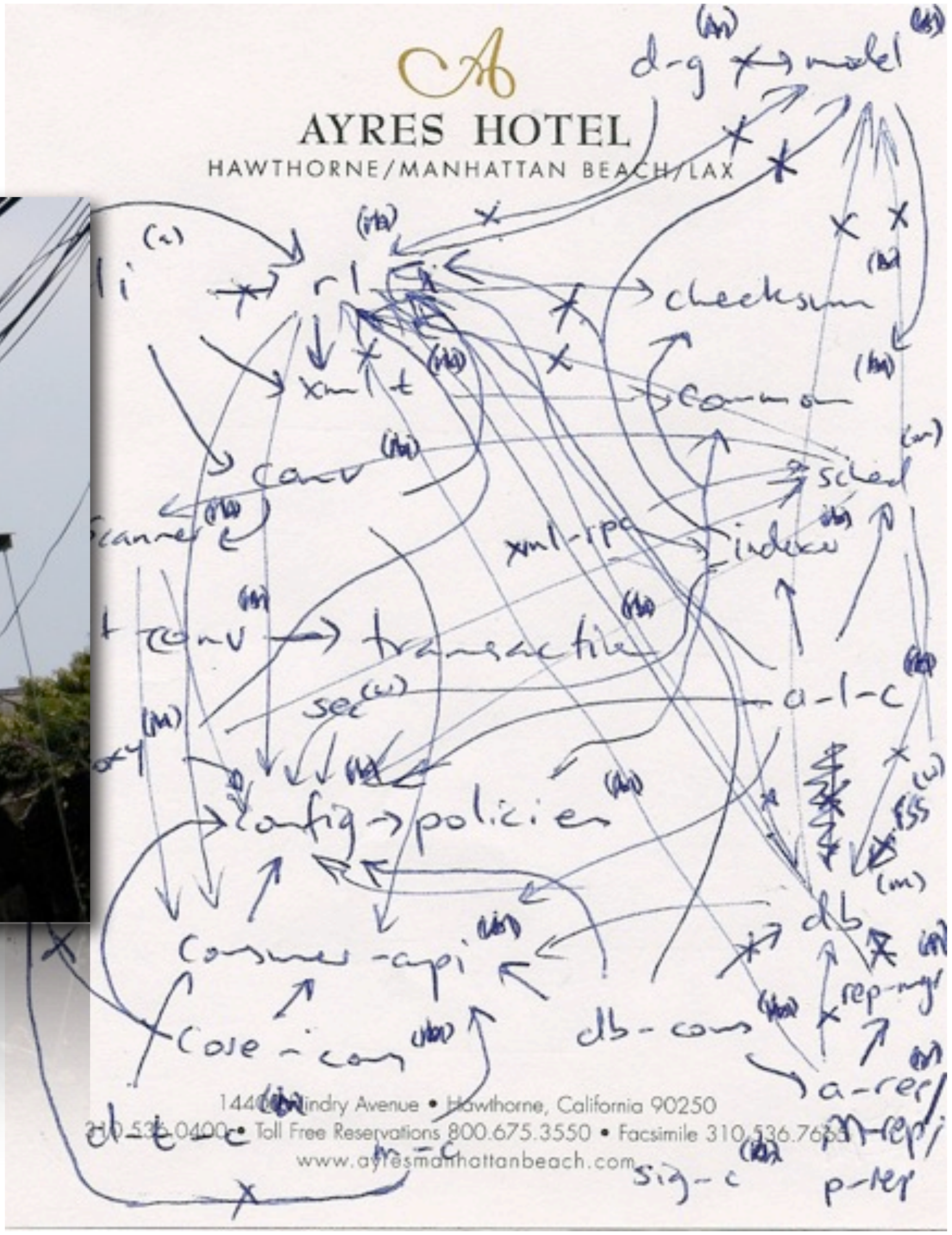


YIKES!

Archiva: First Steps

- Reviewed the architecture

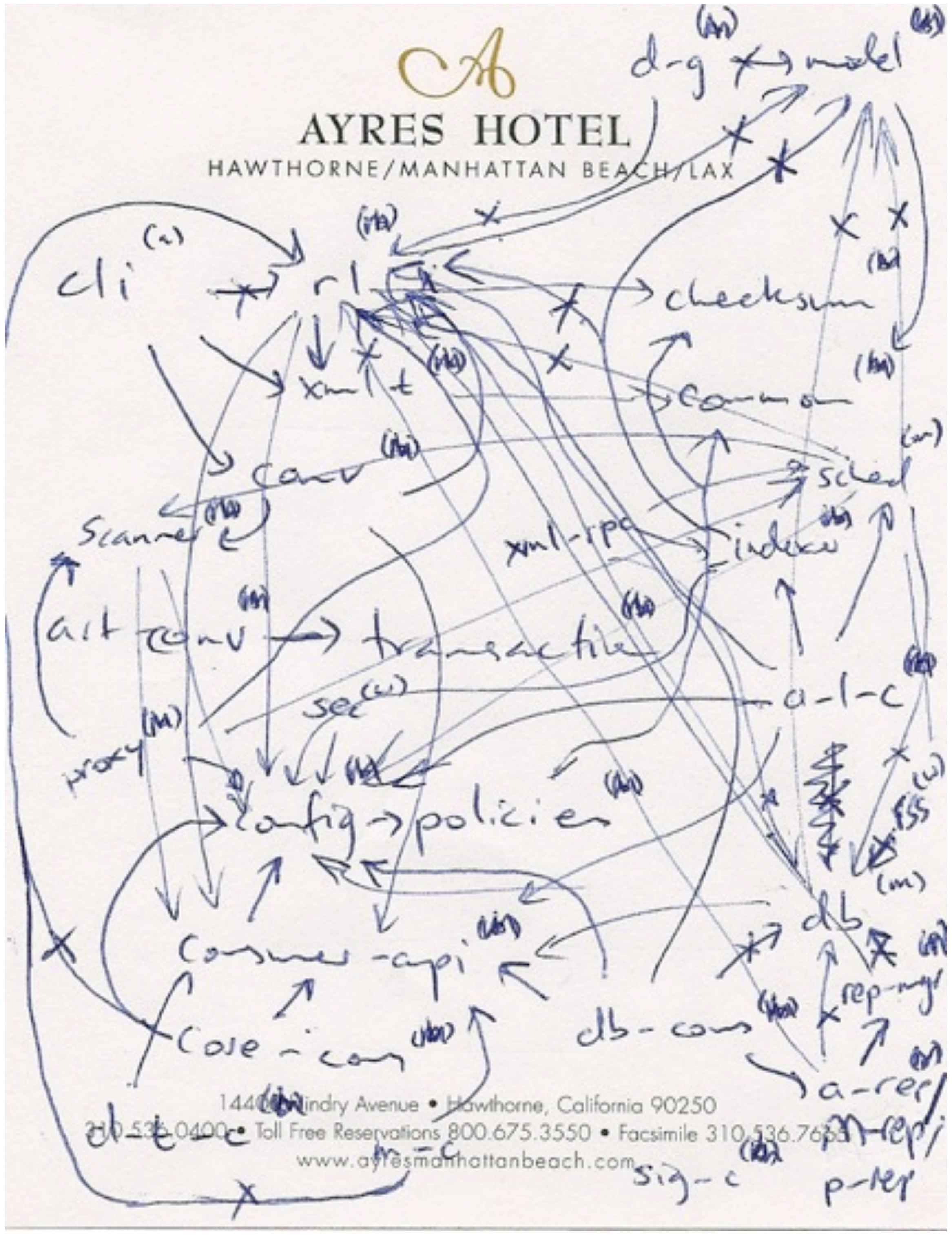
YIKES!


AYRES HOTEL
HAWTHORNE/MANHATTAN BEACH/LAX
cli
rl
xml-t
checksum
common
conv
sched
scanner
xml-rpc
index
conv
transaction
sec
a-l-c
config
policies
consumer-api
core-con
db-cons
db
rep-mgr
a-rep
p-rep
sig-c
d-g
model
14400 Hindry Avenue • Hawthorne, California 90250
310.536.0400 • Toll Free Reservations 800.675.3550 • Facsimile 310.536.7653
www.ayresmanhattanbeach.com

Thursday, 4 November 2010

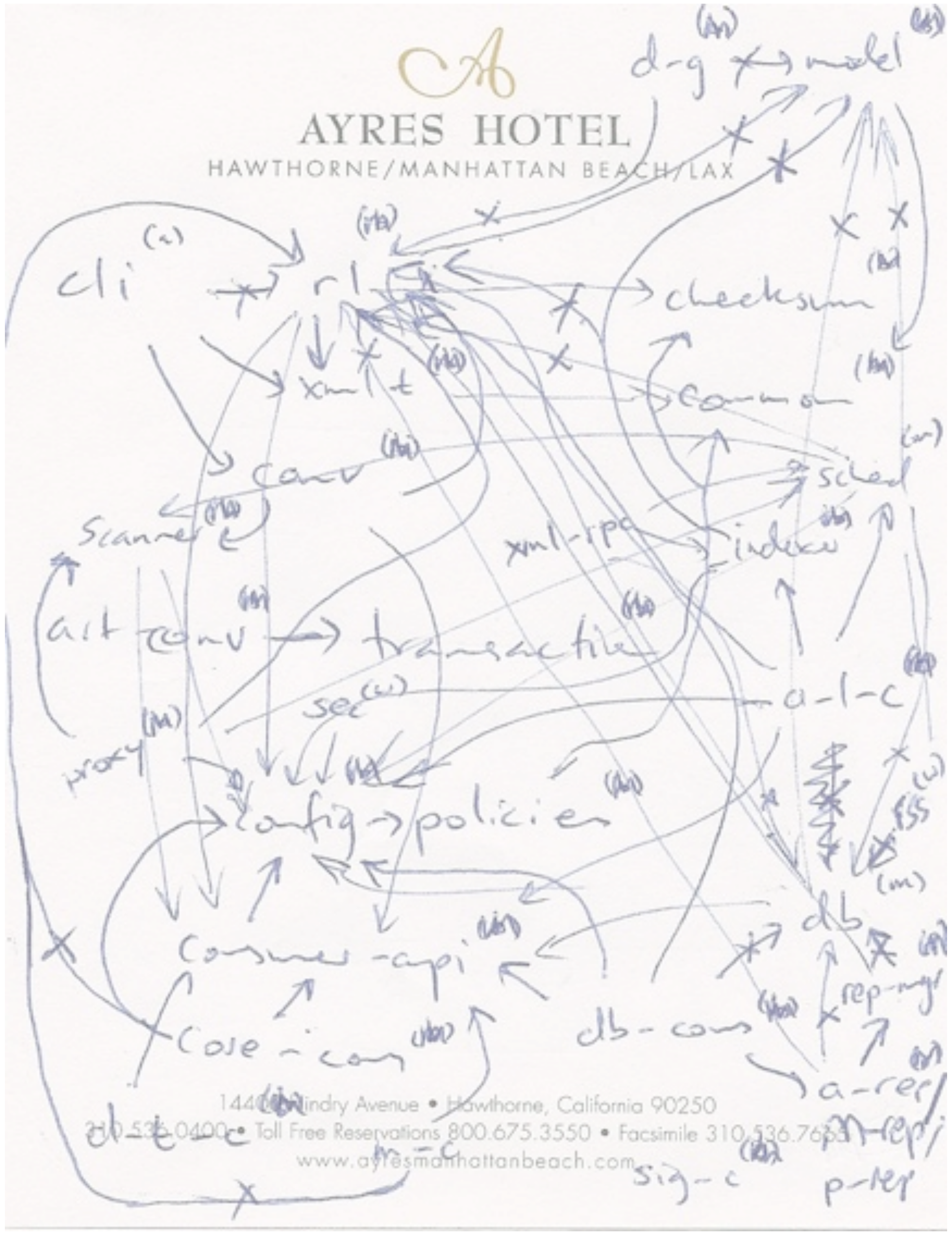
Target Restructuring

- Technically unrelated to the effort at hand
- Trying to remove the database showed how pervasive it was
  - model (JDO annotated classes) and database module were everywhere
  - repository-layer was the culprit
  - other modules grew up out of what was available
- Started to build the right abstraction
  - directly replace some uses
  - others were redirected via the old code

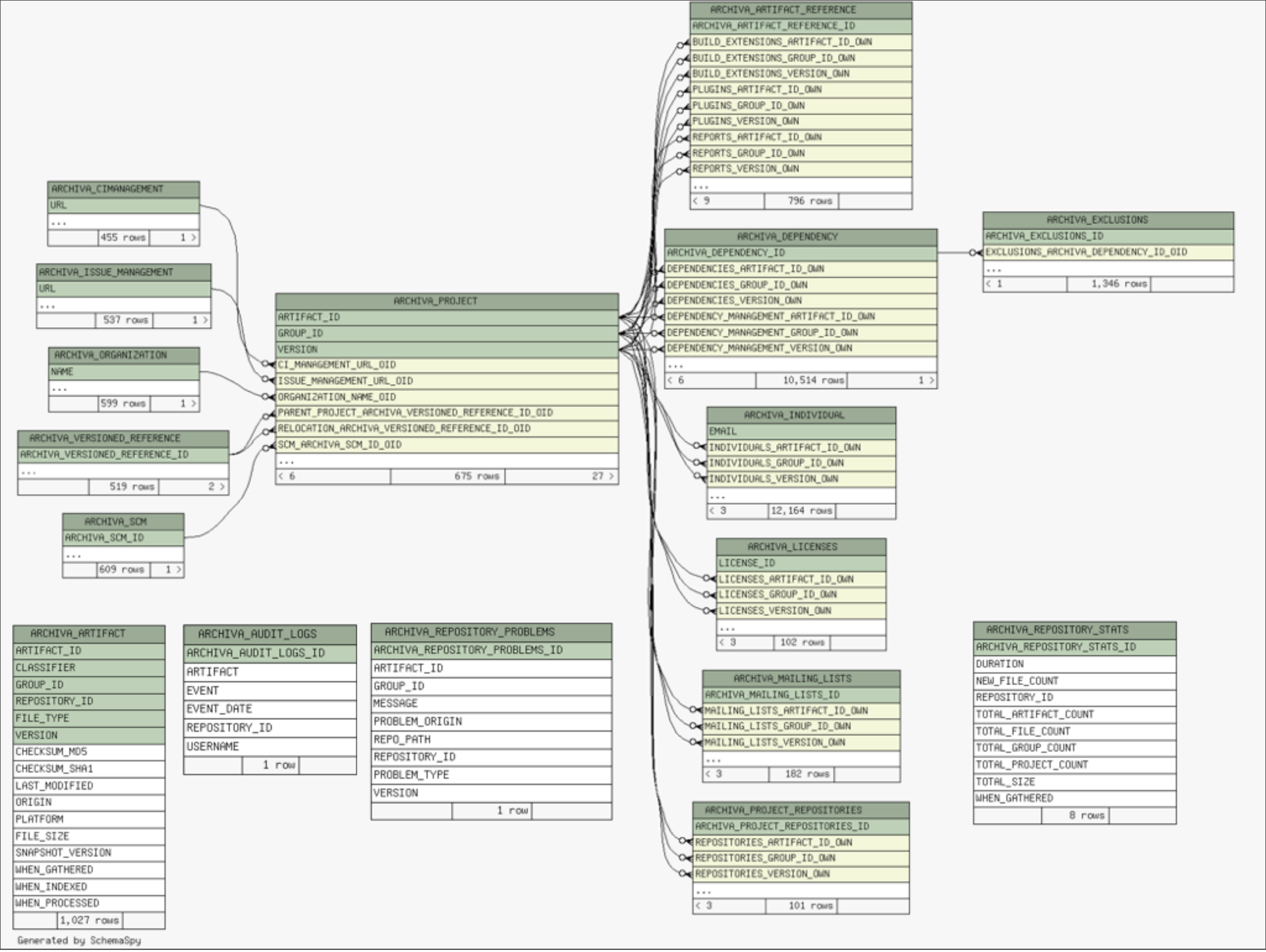


Target Restructuring

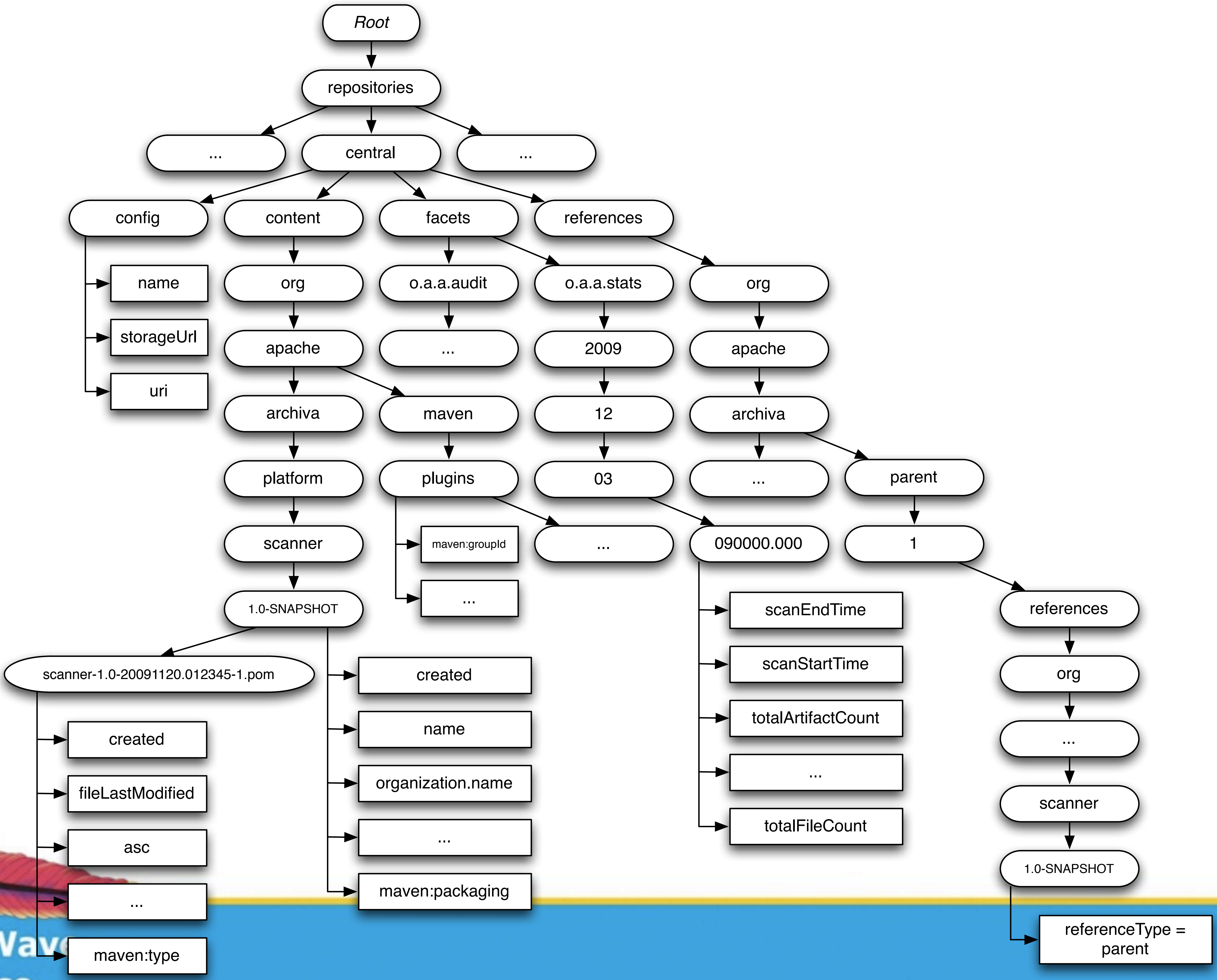
- Technically unrelated to the effort at hand
- Trying to remove the database showed how pervasive it was
  - model (JDO annotated classes) and database module were everywhere
  - repository-layer was the culprit
  - other modules grew up out of what was available
- Started to build the right abstraction
  - directly replace some uses
  - others were redirected via the old code



Thursday, 4 November 2010


ARCHIVA_ARTIFACT_REFERENCE
ARCHIVA_ARTIFACT_REFERENCE_ID
BUILD_EXTENSIONS_ARTIFACT_ID_OWN
BUILD_EXTENSIONS_GROUP_ID_OWN
BUILD_EXTENSIONS_VERSION_OWN
PLUGINS_ARTIFACT_ID_OWN
PLUGINS_GROUP_ID_OWN
PLUGINS_VERSION_OWN
REPORTS_ARTIFACT_ID_OWN
REPORTS_GROUP_ID_OWN
REPORTS_VERSION_OWN
...
< 9
796 rows
ARCHIVA_CIMANAGEMENT
URL
...
455 rows
1 >
ARCHIVA_ISSUE_MANAGEMENT
URL
...
537 rows
1 >
ARCHIVA_ORGANIZATION
NAME
...
599 rows
1 >
ARCHIVA_VERSIONED_REFERENCE
ARCHIVA_VERSIONED_REFERENCE_ID
...
519 rows
2 >
ARCHIVA_SCM
ARCHIVA_SCM_ID
...
609 rows
1 >
ARCHIVA_PROJECT
ARTIFACT_ID
GROUP_ID
VERSION
CI_MANAGEMENT_URL_OID
ISSUE_MANAGEMENT_URL_OID
ORGANIZATION_NAME_OID
PARENT_PROJECT_ARCHIVA_VERSIONED_REFERENCE_ID_OID
RELOCATION_ARCHIVA_VERSIONED_REFERENCE_ID_OID
SCM_ARCHIVA_SCM_ID_OID
...
< 6
675 rows
27 >
ARCHIVA_EXCLUSIONS
ARCHIVA_EXCLUSIONS_ID
EXCLUSIONS_ARCHIVA_DEPENDENCY_ID_OID
...
< 1
1,346 rows
ARCHIVA_DEPENDENCY
ARCHIVA_DEPENDENCY_ID
DEPENDENCIES_ARTIFACT_ID_OWN
DEPENDENCIES_GROUP_ID_OWN
DEPENDENCIES_VERSION_OWN
DEPENDENCY_MANAGEMENT_ARTIFACT_ID_OWN
DEPENDENCY_MANAGEMENT_GROUP_ID_OWN
DEPENDENCY_MANAGEMENT_VERSION_OWN
...
< 6
10,514 rows
1 >
ARCHIVA_INDIVIDUAL
EMAIL
INDIVIDUALS_ARTIFACT_ID_OWN
INDIVIDUALS_GROUP_ID_OWN
INDIVIDUALS_VERSION_OWN
...
< 3
12,164 rows
ARCHIVA_LICENSES
LICENSE_ID
LICENSES_ARTIFACT_ID_OWN
LICENSES_GROUP_ID_OWN
LICENSES_VERSION_OWN
...
< 3
102 rows
ARCHIVA_MAILING_LISTS
ARCHIVA_MAILING_LISTS_ID
MAILING_LISTS_ARTIFACT_ID_OWN
MAILING_LISTS_GROUP_ID_OWN
MAILING_LISTS_VERSION_OWN
...
< 3
182 rows
ARCHIVA_PROJECT_REPOSITORIES
ARCHIVA_PROJECT_REPOSITORIES_ID
REPOSITORIES_ARTIFACT_ID_OWN
REPOSITORIES_GROUP_ID_OWN
REPOSITORIES_VERSION_OWN
...
< 3
101 rows
ARCHIVA_ARTIFACT
ARTIFACT_ID
CLASSIFIER
GROUP_ID
REPOSITORY_ID
FILE_TYPE
VERSION
CHECKSUM_MD5
CHECKSUM_SHA1
LAST_MODIFIED
ORIGIN
PLATFORM
FILE_SIZE
SNAPSHOT_VERSION
WHEN_GATHERED
WHEN_INDEXED
WHEN_PROCESSED
1,027 rows
ARCHIVA_AUDIT_LOGS
ARCHIVA_AUDIT_LOGS_ID
ARTIFACT
EVENT
EVENT_DATE
REPOSITORY_ID
USERNAME
1 row
ARCHIVA_REPOSITORY_PROBLEMS
ARCHIVA_REPOSITORY_PROBLEMS_ID
ARTIFACT_ID
GROUP_ID
MESSAGE
PROBLEM_ORIGIN
REPO_PATH
REPOSITORY_ID
PROBLEM_TYPE
VERSION
1 row
ARCHIVA_REPOSITORY_STATS
ARCHIVA_REPOSITORY_STATS_ID
DURATION
NEW_FILE_COUNT
REPOSITORY_ID
TOTAL_ARTIFACT_COUNT
TOTAL_FILE_COUNT
TOTAL_GROUP_COUNT
TOTAL_PROJECT_COUNT
TOTAL_SIZE
WHEN_GATHERED
8 rows
Generated by SchemaSpy

Design a Content Model



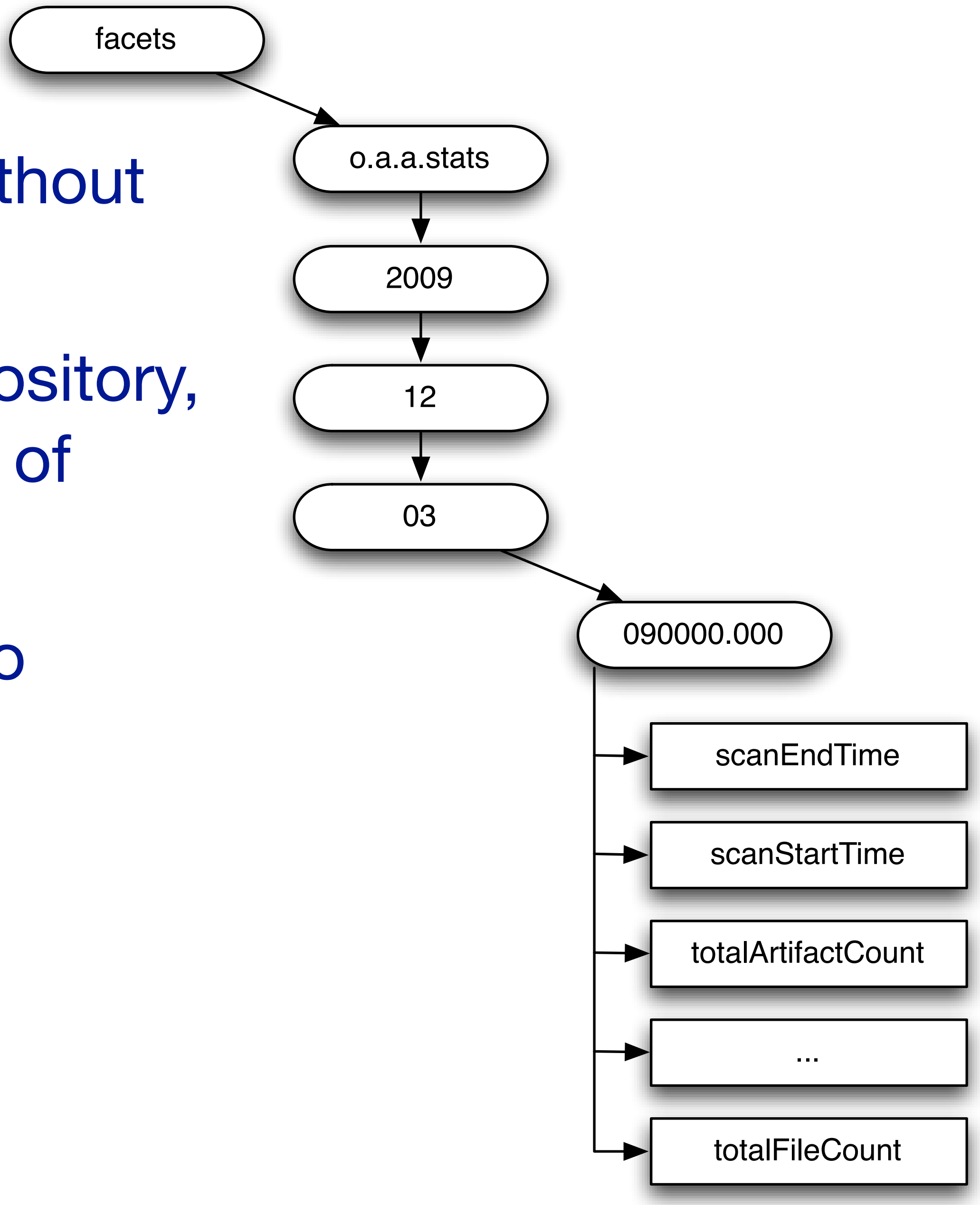
Leading the Wave of Open Source

Hierarchy

- Natural benefits for us due to the repository structure
- Can traverse to a part of a group without dealing with substrings
- Layout not identical to a Maven repository, and will be translated from a variety of input formats
- Unstructured data lends itself well to plugins and arbitrary metadata

Hierarchy

- Natural benefits for us due to the repository structure
- Can traverse to a part of a group without dealing with substrings
- Layout not identical to a Maven repository, and will be translated from a variety of input formats
- Unstructured data lends itself well to plugins and arbitrary metadata



Leading the Wave of Open Source

David’s Model

- Helpful reference
  - http://wiki.apache.org/jackrabbit/DavidsModel
- Rule #1: Data First, Structure Later. Maybe.
- Rule #2: Drive the content hierarchy, don't let it happen.
- Rule #3: Workspaces are for clone(), merge() and update().
- Rule #4: Beware of Same Name Siblings.
- Rule #5: References considered harmful.
- Rule #6: Files are Files are Files.
- Rule #7: ID's are evil.

Faceted Metadata

- Designed faceted metadata model and corresponding API
- Assuming a hierarchical content model, though not yet using JCR
- For most compatibility with existing code, fully mapped object model

Repository API

- Content access mechanism
- Individual coordinate paths passed in (group, artifact, version)
    - no need to construct the strings, avoids layout being spread out
- Resolvers to fill in metadata or obtain artifacts
    - layered access
    - track completeness
    - proxying remotely
- Resolver is a generally useful pattern if you need to load the data from an external source on the fly
- Metadata vs. Storage

Metadata Persistence

- To get it working, a simple hand-rolled file-based implementation
- With everything working, now saw what we could achieve with JCR

JCR - Java Content Repository

- Using Jackrabbit
- Very simple translation to the JCR API
- Initial memory usage is much higher
- Performance was still better than others
- Switched to file-based persistence of the content repository
  - <PersistenceManager class = "org.apache.jackrabbit.core.persistence.bundle.BundleFsPersistenceManager"/>
  - Yet to be proven at scale in Archiva, potentially not as performant
- Not yet tried OCM (Object Content Mapping - similar to ORM but for content repositories)

Filling Metadata

```
ProjectVersionMetadata versionMetadata =
  new ProjectVersionMetadata();

try
{
  Node root = session.getRootNode();

  Node node = root.getNode(
    "repositories/" + repositoryId + "/content/" +
    namespace + "/" + projectId + "/" + projectVersion );

  versionMetadata.setId( projectVersion );
  versionMetadata.setName(
    node.hasProperty( "name" ) ?
      node.getProperty( “name” ) :
      null );
...
```

Leading the Wave of Open Source

Adding Metadata

```
try
{
  Node root = session.getRootNode();

  Node node = root.getNode(
    "repositories/" + repositoryId + "/content/" +
    namespace + "/" + projectId );

  Node versionNode = node.addNode(
    versionMetadata.getId() );

  versionNode.setProperty( "name",
    versionMetadata.getName() );
  versionNode.setProperty( "description",
    versionMetadata.getDescription() );
...
```

Where the Changes Helped

- Scanning vs. On-demand
- Dependency structure more performant and reliable
- Database was removed
  - whole class of exceptions just disappeared
  - previously unreliable operations like reverse dependency tree fixed
  - memory usage reduced
  - configuration simplified
- Metadata is more extensible
  - generic metadata plugin
  - new plugins contribute metadata without changing code or schema

Leading the Wave of Open Source

Challenges

- Query by artifact properties
  - e.g. how to find an artifact with a given checksum
- Correctly configuring Jackrabbit

Opportunities

- Exposing JCR API directly to Archiva plugins
- Integration of existing WebDAV access
- Security access directly integrated into JCR
- JCR event model
- JCR version control
- Sling
- General design - lots more to do!

Tips

- Review whether data is hierarchical or structure derived from the data
- Centralise access, but don’t overdo the abstraction
- Align content model to natural usage
  - try not to deal with constructing and parsing paths
  - deal with content directly rather than translating to objects
- Huge value in having automated unit tests to keep it working as you make significant changes

Archiva: Help Wanted

- Looking for developers to get involved
- Maven users, OSGi users
- JCR integration, Sling integration, improved UI
- dev@archiva.apache.org

Thanks!

- http://archiva.apache.org/
- http://archiva.apache.org/ref/1.4-SNAPSHOT/
- http://wiki.apache.org/jackrabbit/DavidsModel
- http://www.scribd.com/doc/11163161/JCR-or-RDBMS-why-when-how
- http://wiki.apache.org/jackrabbit/JcrLinks
- http://jackrabbit.apache.org/
- http://sling.apache.org/
- http://maven.apache.org/

ApacheCon

Leading the Wave
of Open Source