Sourcerer: An Infrastructure for Large-scale Collection and Analysis of Open-source Code

Sushil Bajracharya, **Joel Ossher**, Cristina Lopes
Donald Bren School of Information and Computer Sciences
University of California, Irvine

University of California Irvine jossher@uci.edu

SOURCERER

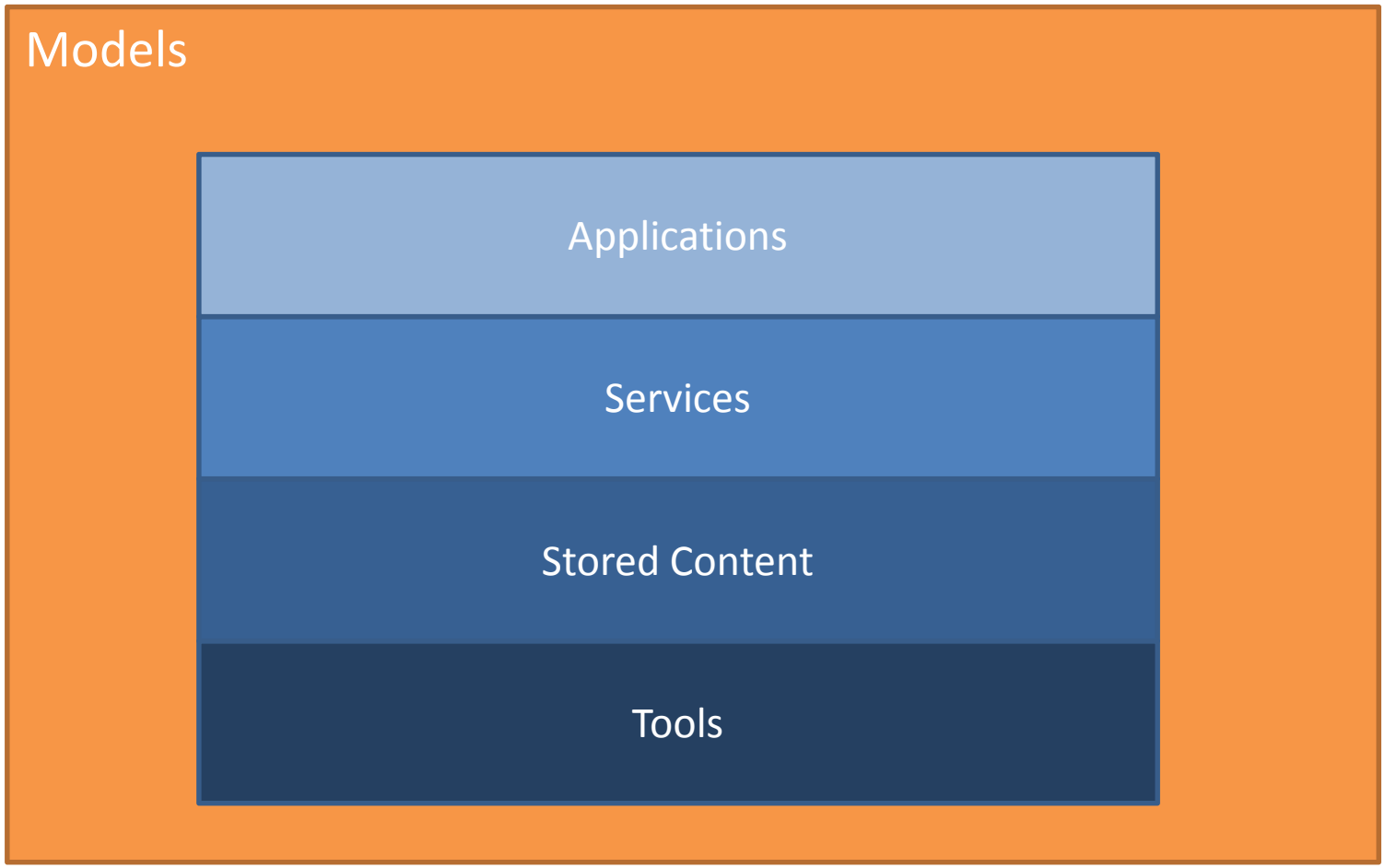
Sourcerer's Inception

- Started in 2005
- Motivation
  - Explore the use of structural information for code retrieval
  - Enable data mining on large quantities of source code
- Target: Open-source Java code
  - Open Source movement provides a large quantity of high quality code
  - Java is popular, and amenable to static analysis

Sourcerer Today

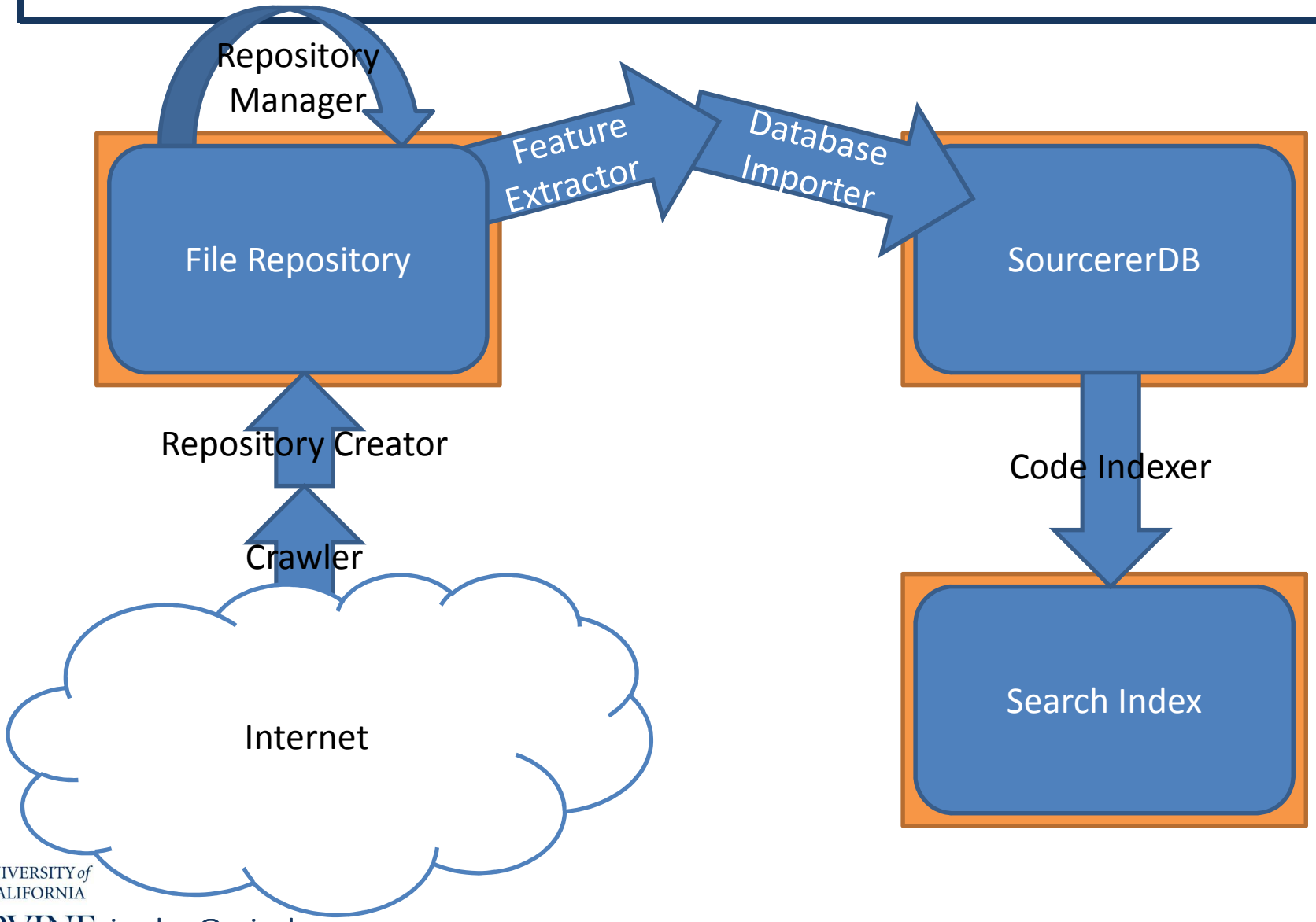
- Collection of loosely coupled Java tools
  - www.github.com/sourcerer/Sourcerer
- Aggregated repository of open source code
  - www.ics.uci.edu/~lopes/datasets/index.html
- Services
  - http://sourcerer.ics.uci.edu/services/
- Applications

Layered Architecture



UNIVERSITY of CALIFORNIA
IRVINE jossher@uci.edu

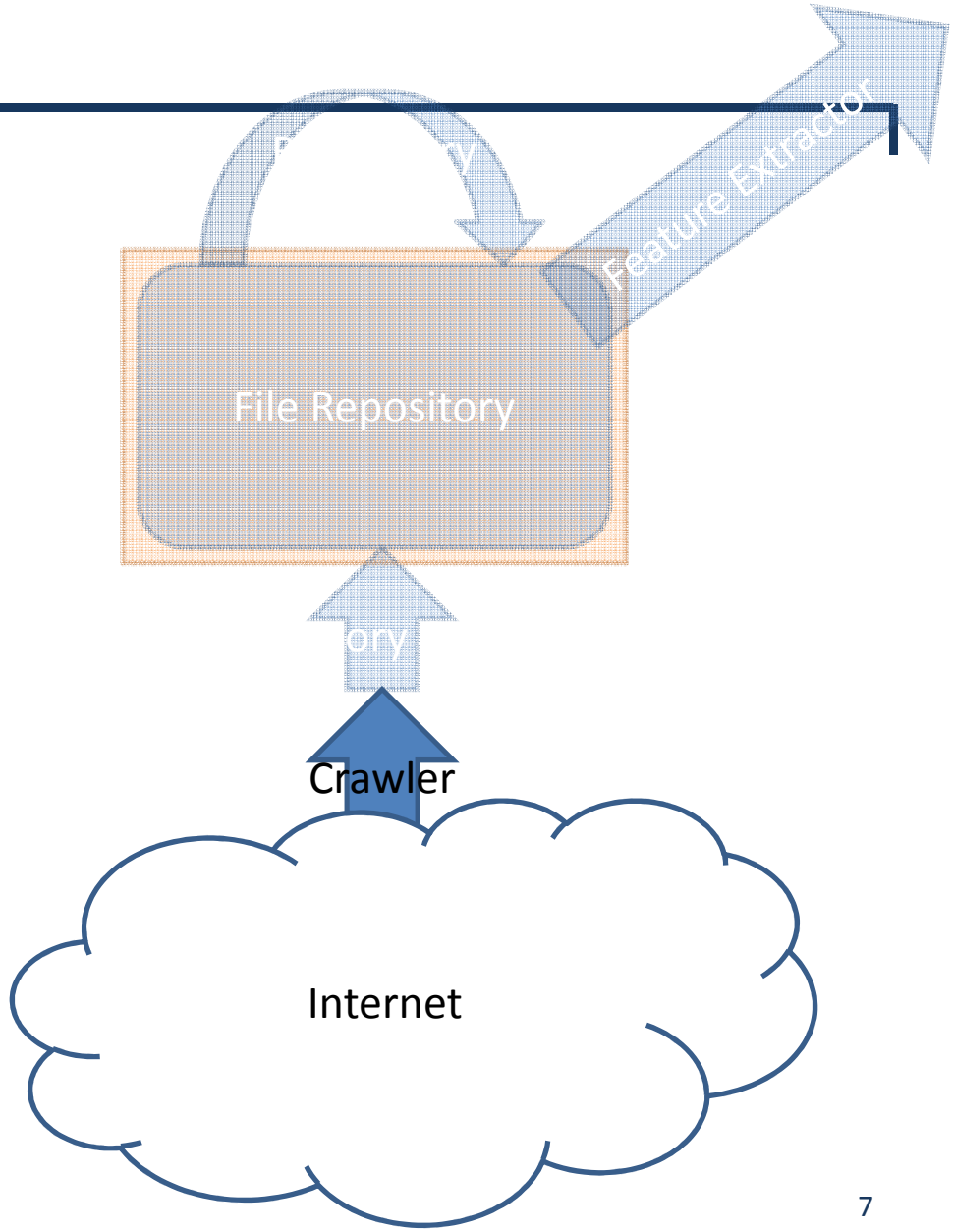
Tools and Stored Content



University of California Irvine jossher@uci.edu

Code Crawler

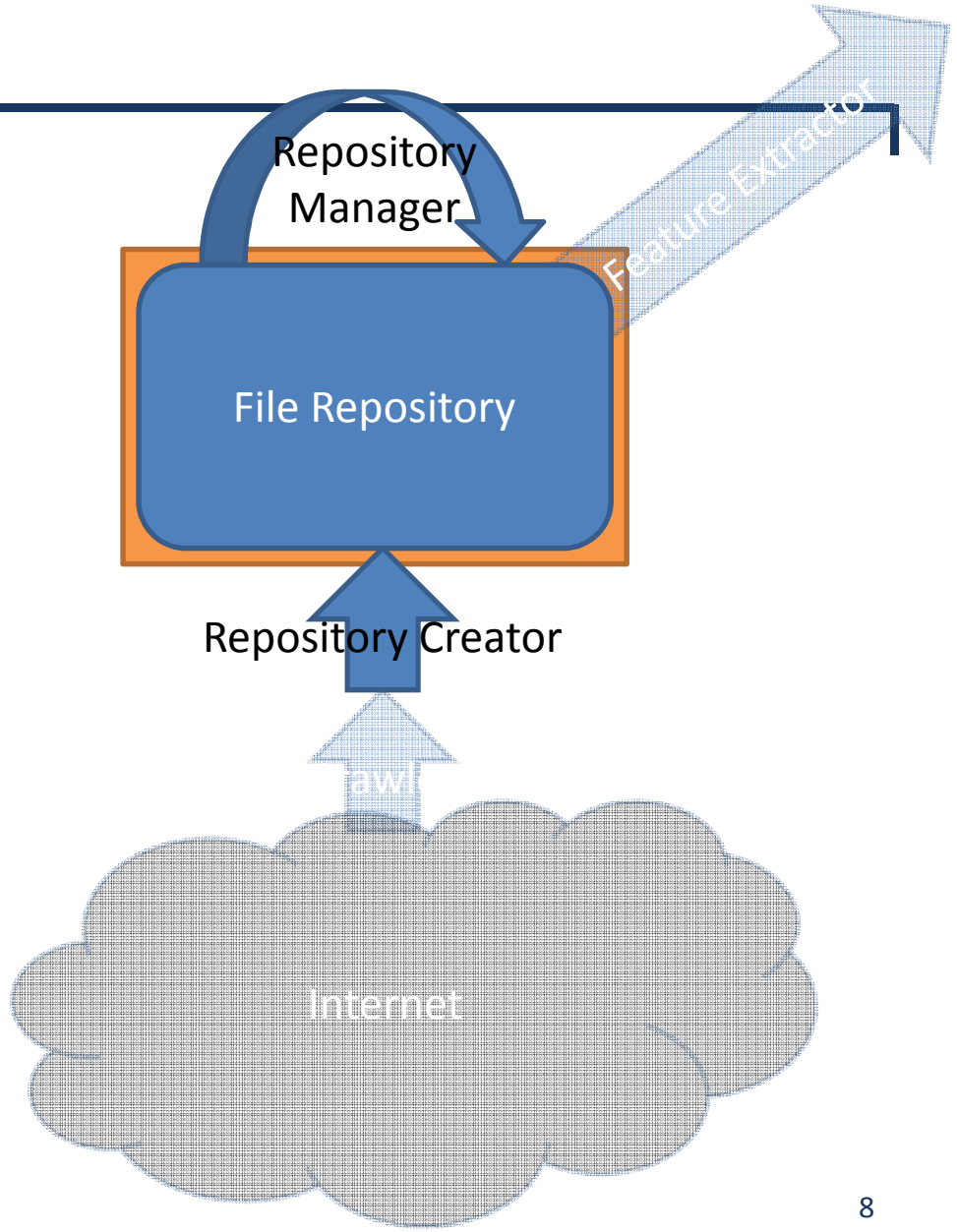
- Input
  - List of seed pages
- Output
  - List of project pages
- Plugin-based
  - Sourceforge
  - Java.net
  - Tigris
  - Google Code Hosting
  - Apache



University of California Irvine jossher@uci.edu

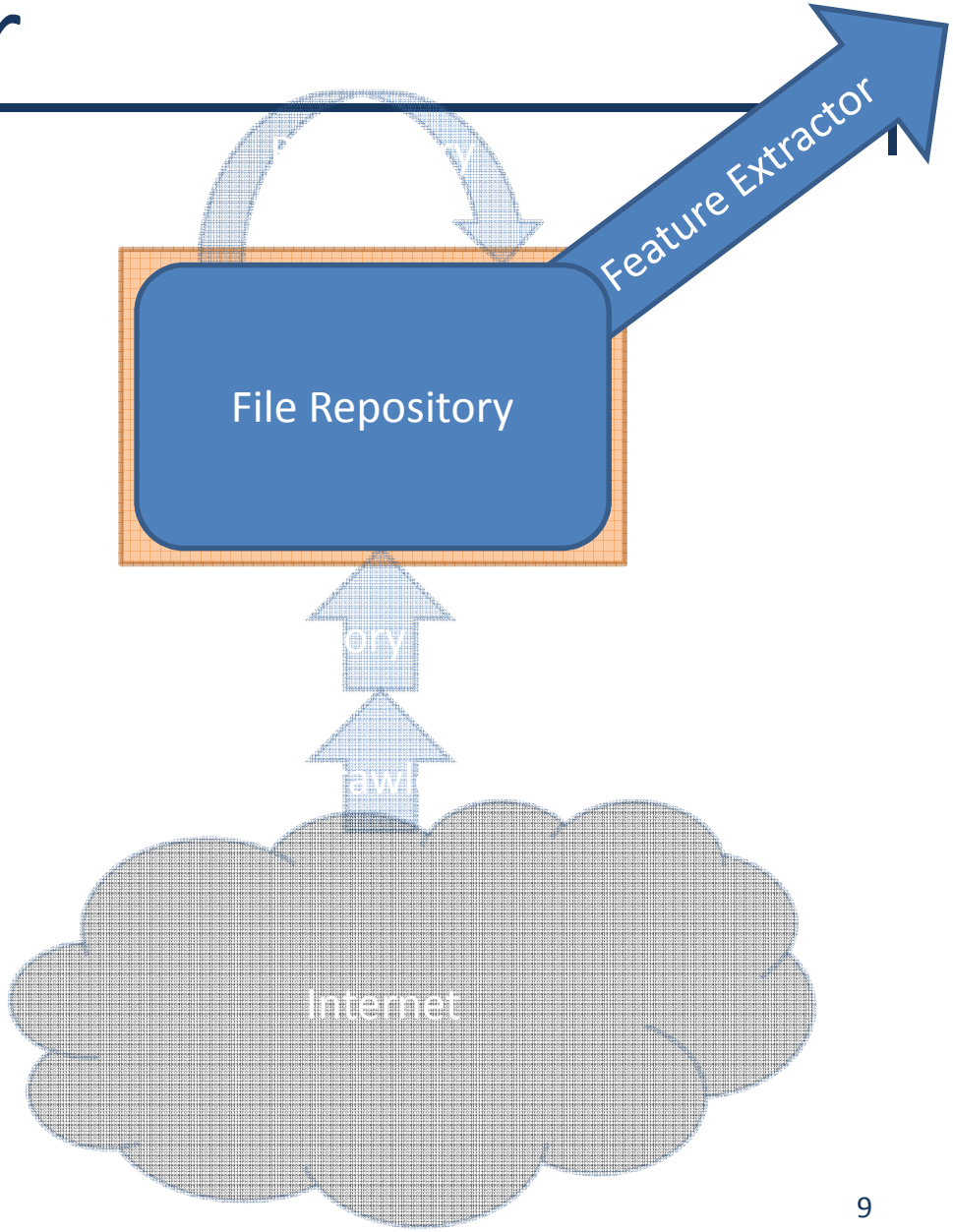
File Repository

- Local aggregated repository
- Repository Creator
  - Input
    - List of project pages
  - Output
    - Populated file repository
- Repository Manager
  - Housekeeping tasks



Feature Extractor

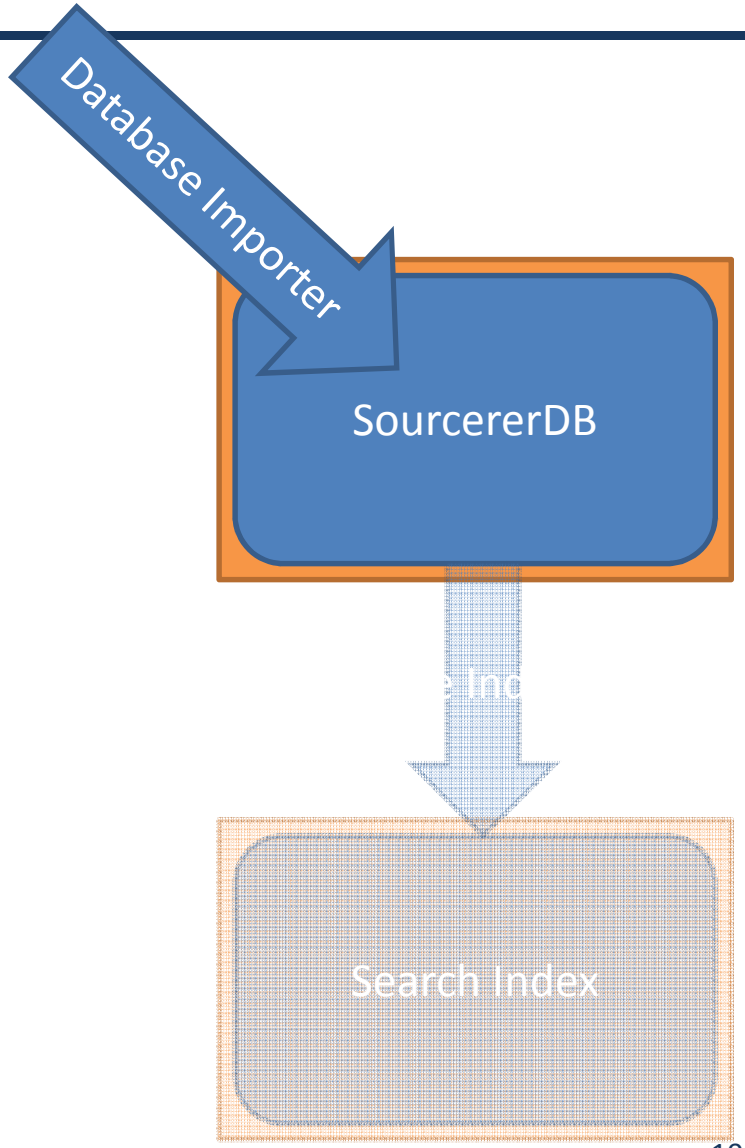
- Input
  - File Repository
- Output
  - Files containing entities and relations
- Entity-relationship metamodel
- Headless Eclipse plugin
  - Uses Eclipse Java development tools (JDT)



UNIVERSITY of CALIFORNIA
IRVINE jossher@uci.edu

SourcererDB

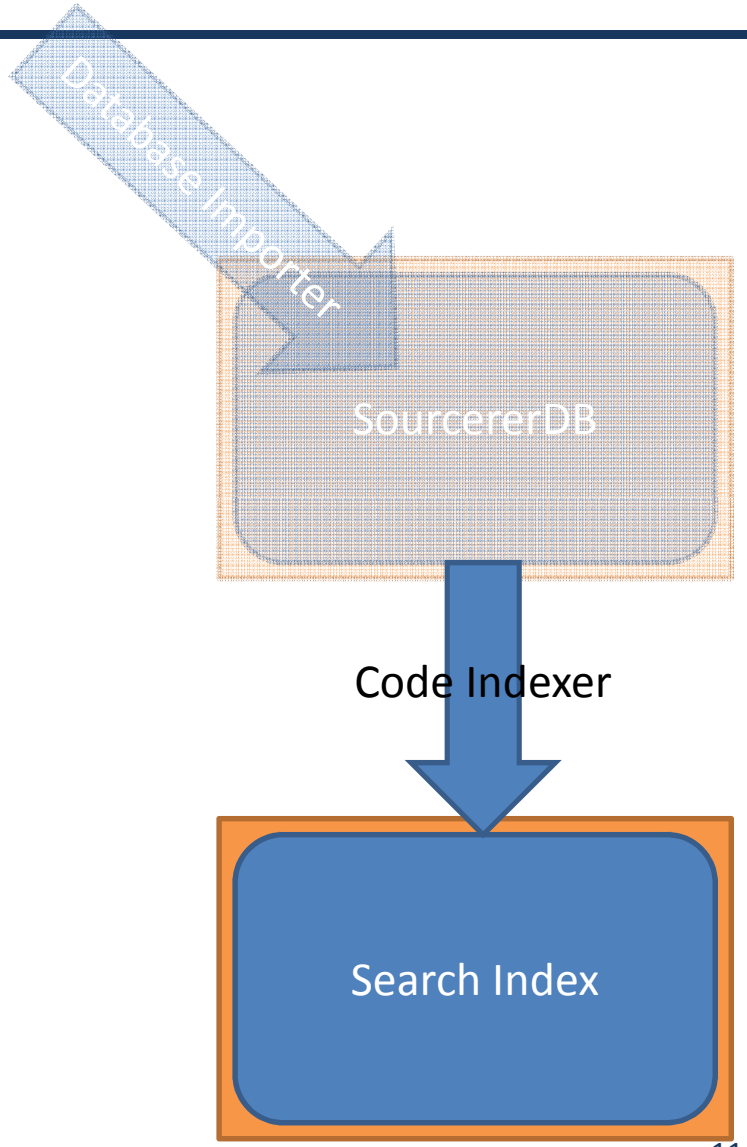
- MySQL database
- Database importer
  - Incremental
  - Parallel
  - Input
    - Feature extractor output
  - Output
    - SourcererDB



University of California Irvine jossher@uci.edu

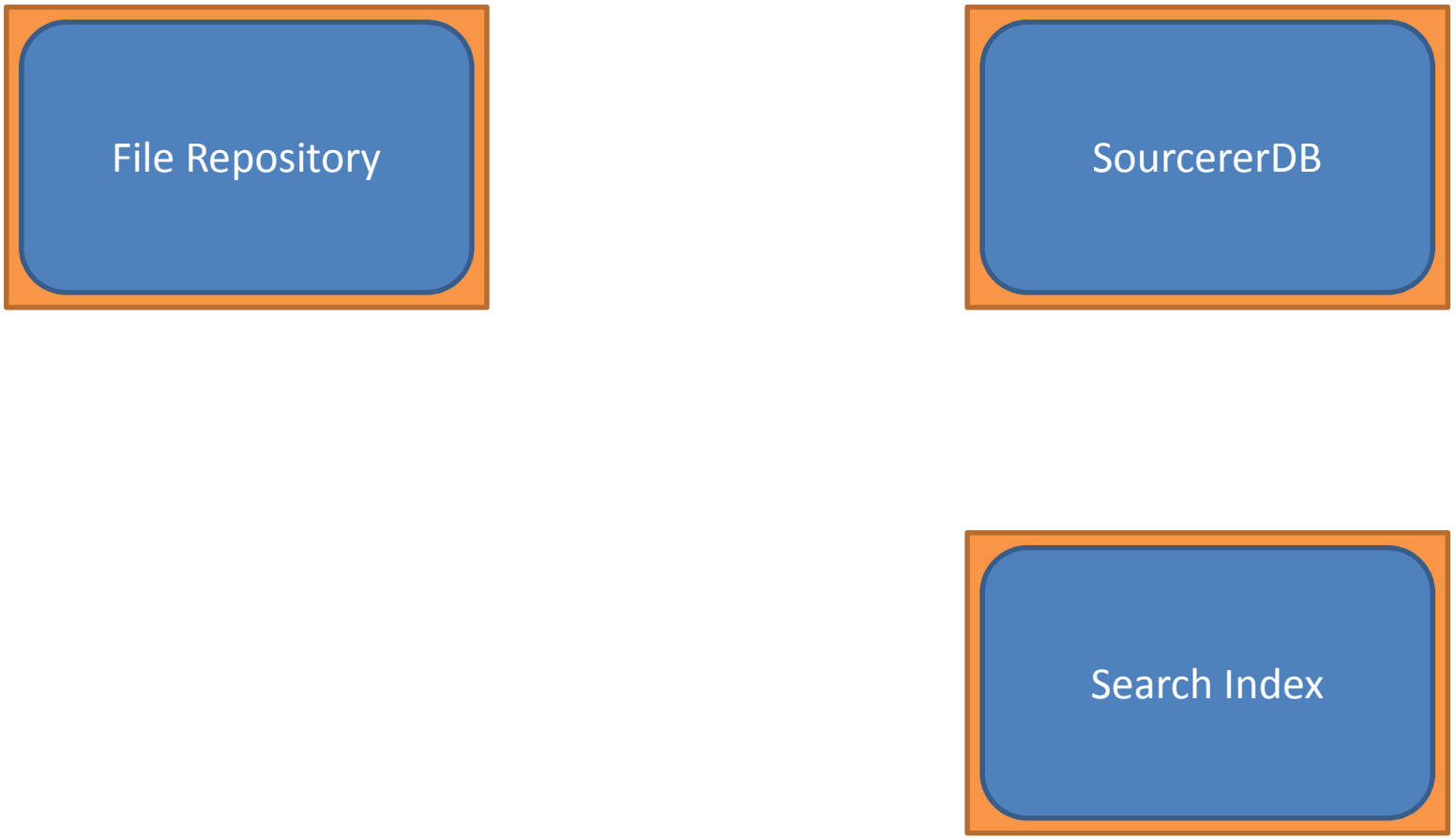
Search Index

- Text search for code entities
- Apache Solr
  - Search platform for Lucene
- Code Indexer
  - Heavily parallel



University of California Irvine jossher@uci.edu

Stored Contents Recap



Sourcerer Services

- Repository Access
  - Look up text matching SourcererDB entities / relations
- Relational Query
  - Direct access to SourcererDB
- Code Search Service
  - Access the Lucene index
- Dependency Slicing

Applications

- Sourcerer Code Search Engine
  - sourcerer.ics.uci.edu/sourcerer/search/index.jsp
- CodeGenie
  - Test-driven code search
- Sourcerer API Search
  - Demo!

LESSONS LEARNED

UNIVERSITY of CALIFORNIA
IRVINE jossher@uci.edu

Lesson One: Reuse

- Feature extractor 1.0
  - Corollary: javac
- Code crawler woes

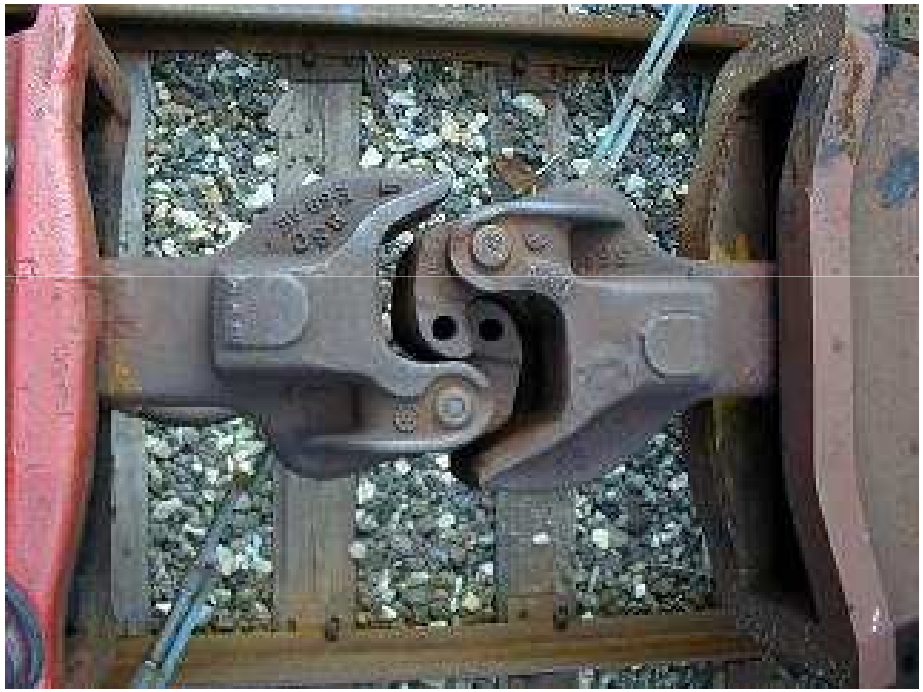
UNIVERSITY of CALIFORNIA
IRVINE jossher@uci.edu

Lesson Two:Performance & Scalability

- Research prototype
- Jars directory
- Repository migration

Lesson Three: Loose Coupling

- Sourcerer M1
- CASI



Lesson Four: YCMEH

- You can't make everyone happy
  - Why only Java?
  - Why no X project or Y repository?
  - Why no versioning information?
  - …
  - If you try, no one will be happy (since your tool will never be released)

Thank you!

- Contact: jossher@uci.edu