



SeerSuite: Developing a Scalable and Reliable Application Framework for Building Digital Libraries by Crawling the Web

Pradeep Teregowda*, Isaac Councill#, Juan Fernandez*, Shuyi Zheng*, Madian Khabisa*, C. Lee Giles*

* Pennsylvania State University

#Google

SeerSuite

- A framework for building digital libraries.
 - Reliable – around the clock service with minimal downtime
 - Robust – continue providing services, even while some components are constrained.
 - Scalable – support increasing user requests and documents.
 - Flexible (modular), Portable (across operating systems).
- Features
 - Automatic acquisition of new documents by focused web crawling
 - Full text indexing
 - Autonomous citation indexing, linking documents through citations.
 - Automatic metadata extraction for each document.
 - MyCiteSeer for personalization.
 - New features in development, e.g.
 - Table extraction and search
 - Algorithm extraction and search

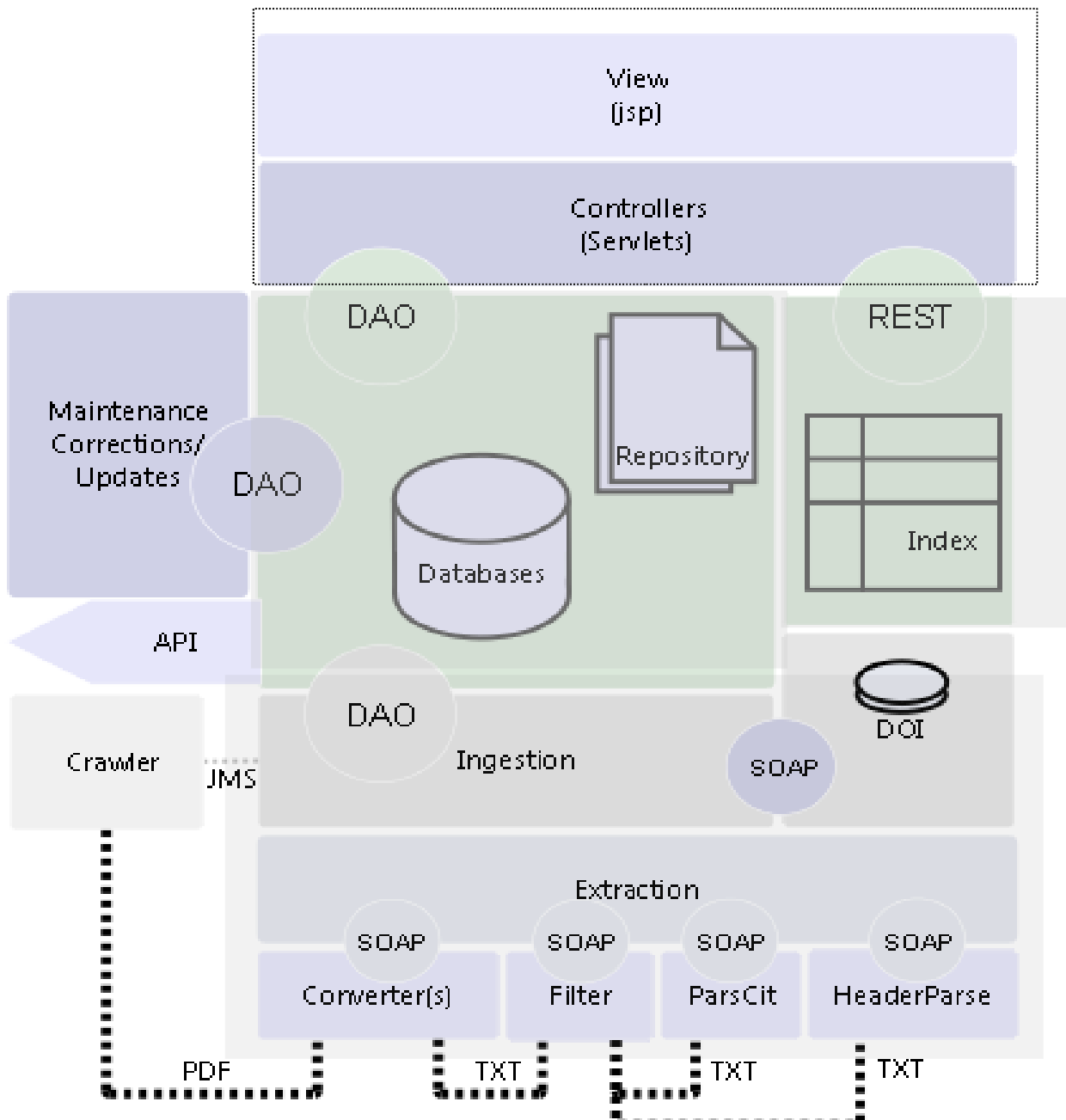
Outline

- Evolution
 - A brief discussion of history, features, advances.
- Architecture
 - Description of components, modules of SeerSuite.
- Workflow
 - Identify steps in adding documents
- Deployment
 - SeerSuite as CiteSeer^x- deployment, interface, federation and usage.

Digital Libraries

- Digital libraries (DLs) continue to grow and be used
 - Cyberinfrastructure for scientists and academics
 - Google Scholar is very popular & to some invaluable
 - Publisher collections
 - ACM portal, Scopus, etc.
 - Library of Congress (NDLP)
- Document acquisition
 - Author submissions
 - RePec (economics).
 - ArXiv (physics)
 - Web harvesting (Crawler based)
 - CiteSeer^x (mostly computer science)
 - crawls author homepages, not publishers
 - Google Scholar, considerable data acquired from publishers.

SeerSuite Architecture



Web Application (View, Controllers)

Data Storage
(Index, Database, Repository)

Metadata Extraction
(Extraction, Ingestion, DOI)

Architecture Details

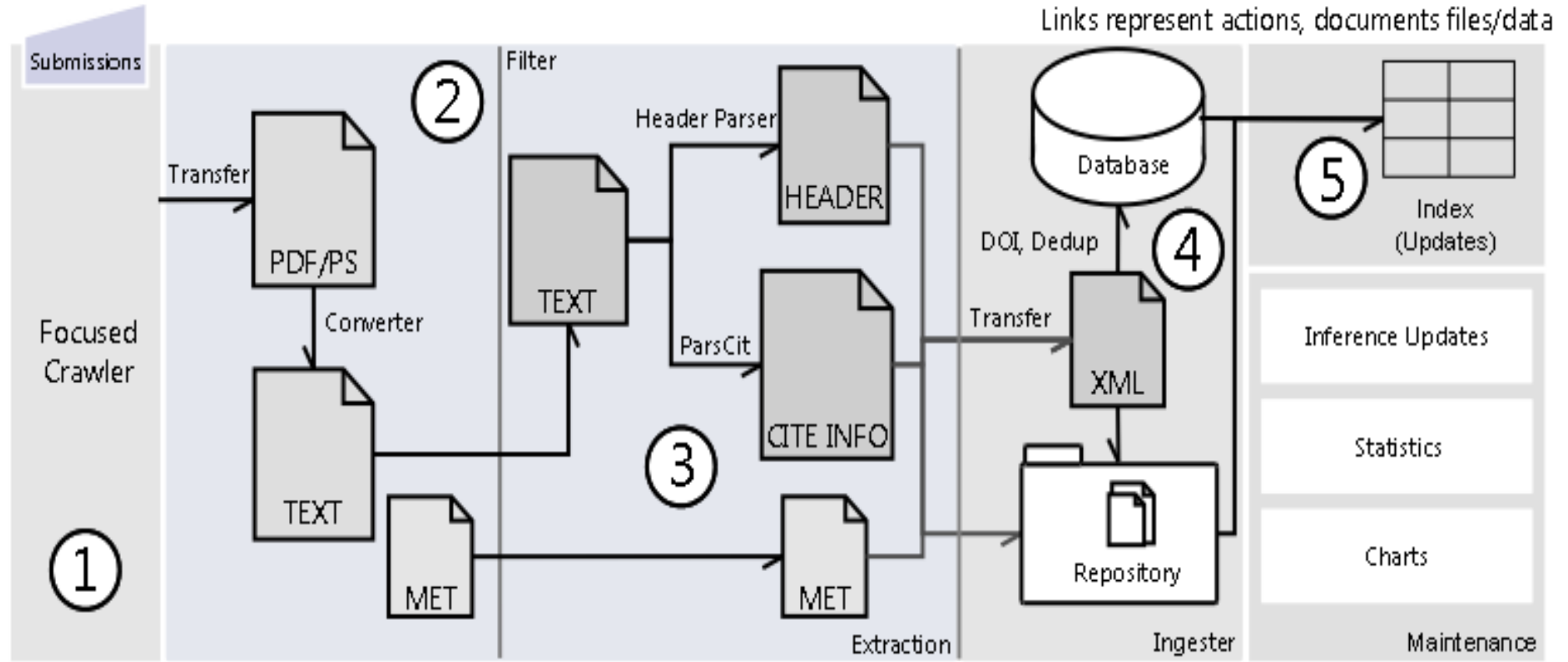
- Web Applications
 - Built using the Java Spring framework,
 - jsp, javascript (dojo, mootools) for presentation.
 - Servlets/Controllers
- Data Storage
 - Repository (files)
 - Index (fast search)
 - Database (graph, metadata)
- Extraction and Ingestion
 - PDF to Text conversion (pdfbox, TET).
 - Converted documents filtered.



Architecture Details

- Extraction and Ingestion
 - Support Vector Machines for document metadata, CRF for citation extraction.
 - DOI – Unique internal identification of documents
- Crawler
 - Heritrix with a Java Message Service based system over ActiveMQ.
- Maintenance
 - Keep graph, index, services updated, external links.

Workflow



http://uninterestingplace.edu

Seed

www.psu.edu

User Submission

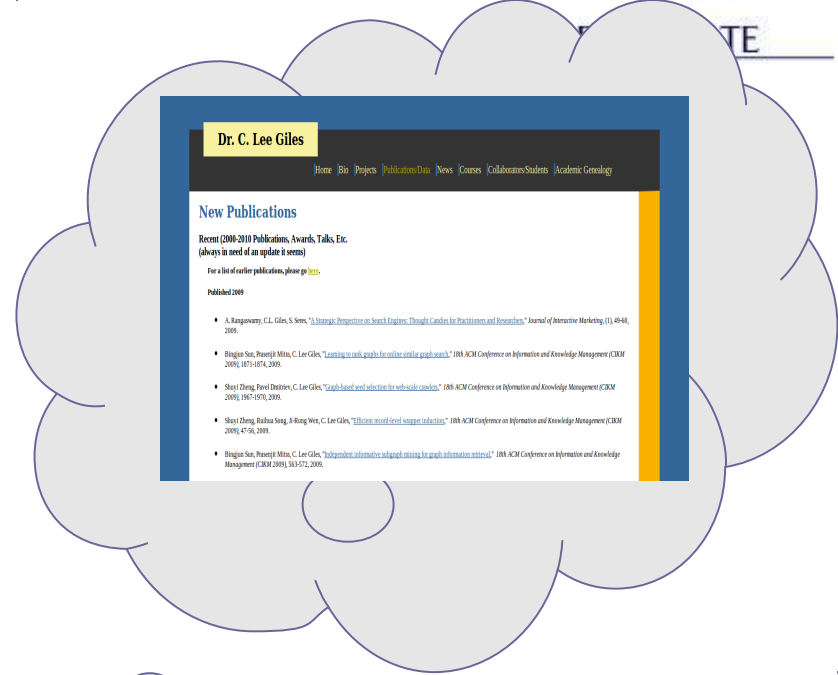
giles.ist.psu.edu/publications

Focused Crawler

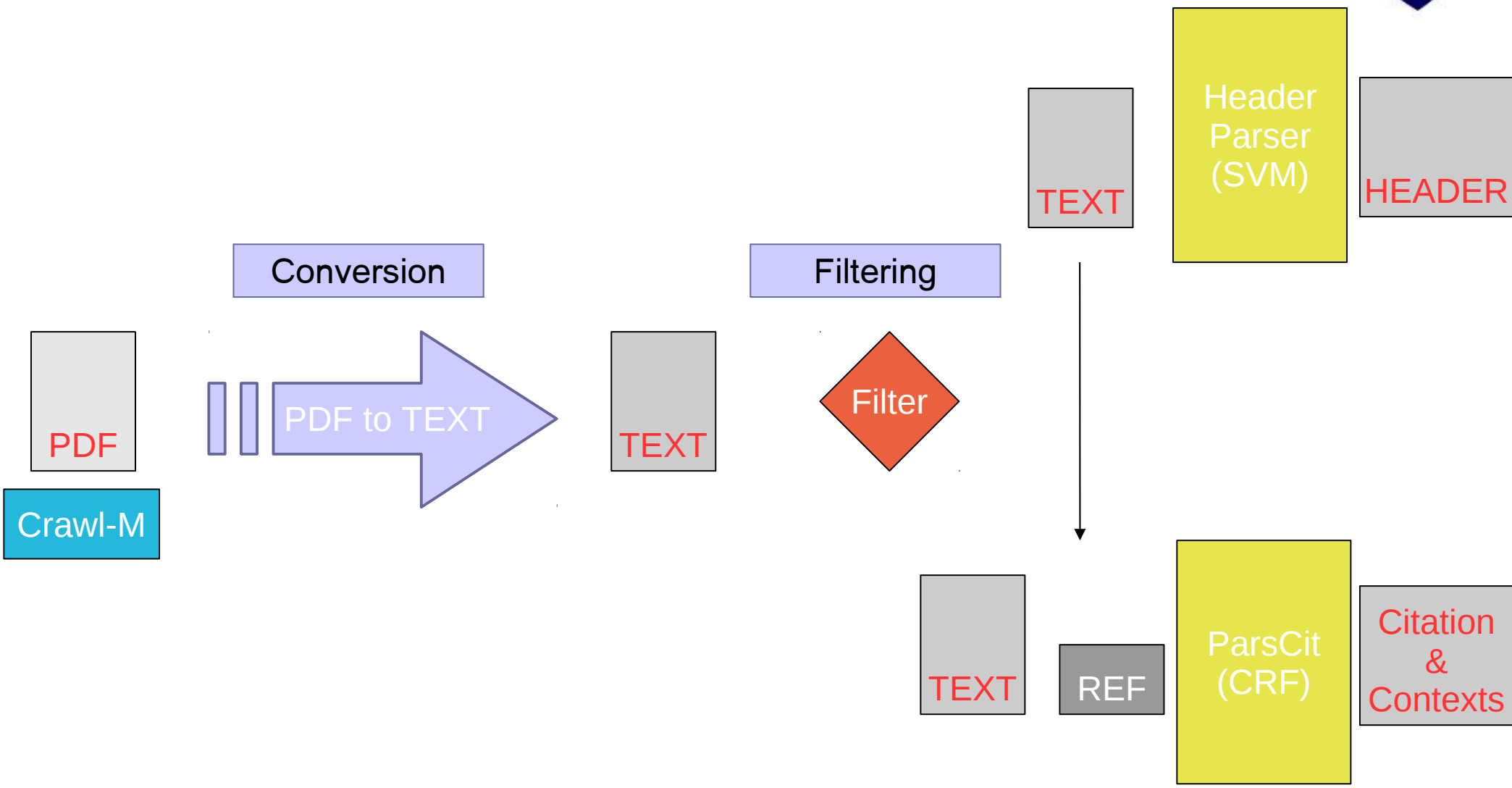
Fetch

PDF

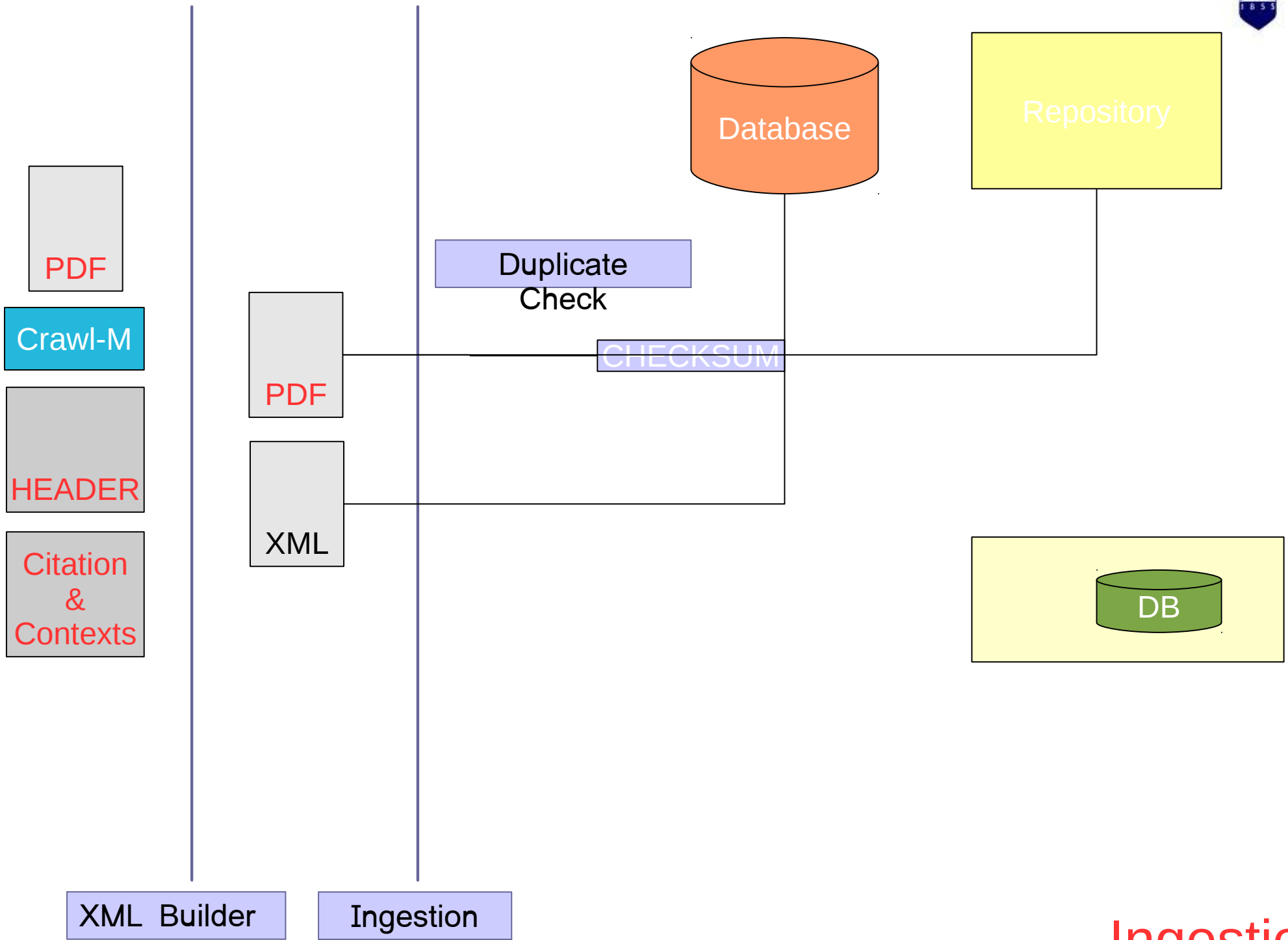
Crawl-M



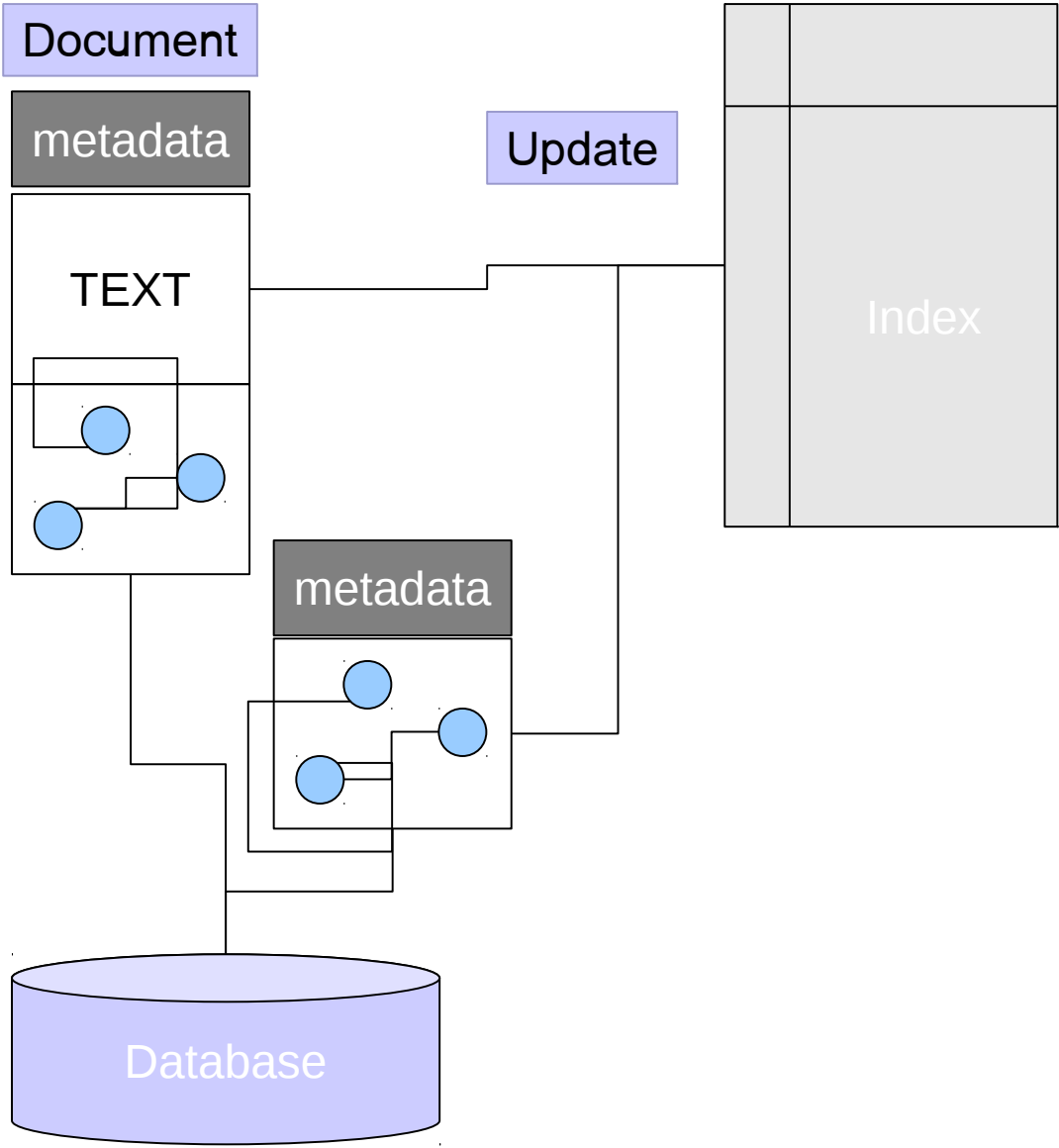
Focused Crawling



Metadata Extraction

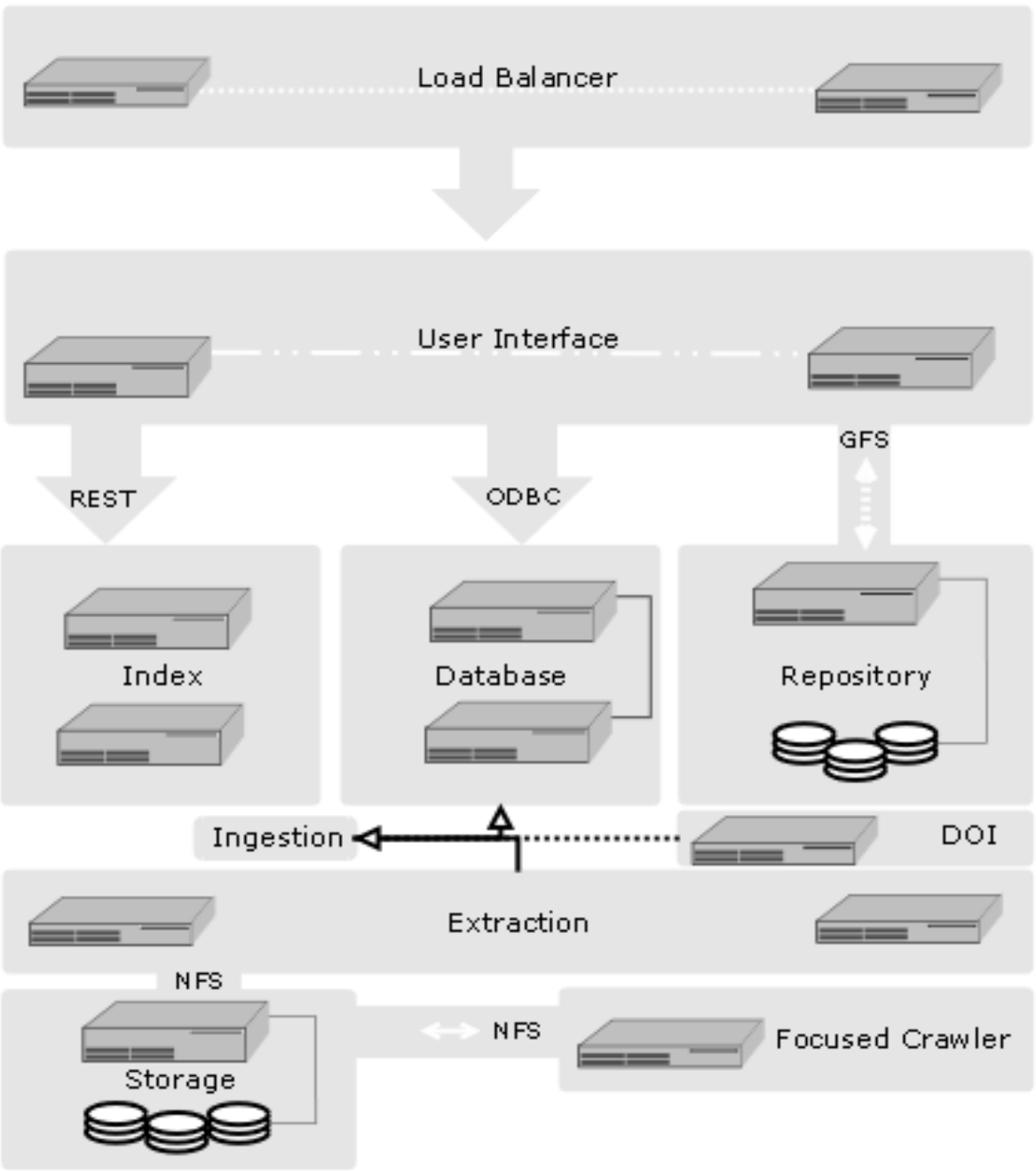


Ingestion



Maintenance: Indexing

Deployment: CiteSeer^x



- Off-the-shelf-hardware
 - x86 based servers, DAS storage
- Linux
 - Redhat Cluster Suite (GNBD/GFS)
- Tomcat platform
 - Web applications/
 - Interfaces (OAI/API)
- Database
 - MySQL RDBMS
- Indexing
 - Solr

User Interface

- Several interface views
 - Search
 - Access to the full text of all documents,
 - citations,
 - Authors.
 - Ranked by user criterion.
 - Document Summary
 - Presents document metadata,
 - Citations
 - Citation graphs,
 - Links to copies
 - Links to other bibliography sources.
 - Citation Relationships
 - Co-citations
 - Active bibliography

Search



Documents | Authors | Tables /

Search Bar

johnson Search
 Include Citations | [Advanced Search](#) | [Help](#)

Searching for johnson – sorted by Number of Citations

Order by: [Relevance](#) | [Year \(Descending\)](#) | [Year \(Ascending\)](#) | [Recency](#)
Try your query at: [Scholar](#) | [Yahoo!](#) | [Ask](#) | [Bing](#) | [CSB](#)

Criterion

22,310 documents found, showing 1 through 10. [Next 10](#) → [ATOM](#) [RSS](#)

Result

▼ [Handbook of Applied Cryptography](#)
by Alfred J. Menezes, Paul C. Van Oorschot, Scott A. Vanstone, R. L. Rivest — 1997
... Gustafson Darrel Hankerson Anwar Hasan Don **Johnson** Mike Just Andy Klapper Lars Knudsen Neal Koblitz Çetin...
Cited by 1642 (21 self) – [Add To MetaCart](#)

▼ [Dynamic source routing in ad hoc wireless networks](#)
by David B. Johnson, David A. Maltz — 1996 — Mobile Computing
...Dynamic Source Routing in Ad Hoc Wireless Networks David B. **Johnson** David A. Maltz Computer...
Cited by 1507 (32 self) – [Add To MetaCart](#)

▼ [The Entity-Relationship Model: Toward a Unified View of Data](#)
by Peter Pin-shan Chen — 1976 — ACM Transactions on Database Systems
... of attribute-value pairs. "3", "red", "Peter", and "**Johnson**" are values. Values are classified into different...
Cited by 1121 (3 self) – [Add To MetaCart](#)

▼ [WordNet: An on-line lexical database](#)
by George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, Katherine Miller — 1990 — International Journal of Lexicography
... of linguistic knowledge in general, and lexical knowledge in particular—Miller and **Johnson**-Laird (1976) have...
Cited by 1074 (6 self) – [Add To MetaCart](#)

▼ [A Performance Comparison of Multi-Hop Wireless Ad Hoc Network Routing Protocols](#)
by Josh Broch, David A. Maltz, David B. Johnson, Yih-chun Hu, Jorjeta Jetcheva — 1998
.... Maltz David B. **Johnson** Yih-Chun Hu Jorjeta Jetcheva Abstract An ad hoc network is a collection...
Cited by 977 (25 self) – [Add To MetaCart](#)

▼ [Affective Computing](#)
by Rosalind W. Picard, R. W. Picard, Marie Curie — 1995
... arguments for the essential role of emotion. **Johnson**-Laird and Shafir have recently reminded the cognition...
Cited by 792 (33 self) – [Add To MetaCart](#)

Document Summary

Home | Statistics | About | Bulletin | Submit Documents | Feedback | MetaCart | Sign in to MyCiteSeerX

CiteSeer^x beta

Documents | Authors | Tables /


Search []

Include Citations | Advanced Search | Help

Summary | Related Documents | Version History

WordNet: An on-line lexical database (1990) [1074 citations — 6 self]

by George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, Katherine Miller
International Journal of Lexicography
Add To MetaCart

DOWNLOAD:
http://www.cfil.itib.ac.in/archives/english_wordn
<http://l2r.cs.uiuc.edu/~danr/Teaching/CSS98-05/Pap>
<http://wordnet.princeton.edu/5papers.pdf>
 CiteULike
 CACHED: 

Add to Collection | Correct Errors | Monitor Changes

Abstract:
 WordNet is an on-line lexical reference system whose design is inspired by current

Citations

- 805 [How to Do Things with Words](#) - Austin - 1975
- 385 [The Case for Case](#) - Fillmore - 1968
- 343 [Semantics and cognition](#) - Jackendoff - 1983
- 293 [Word Meaning and Montague Grammar](#) - Dowty - 1979
- 282 [Semantic Interpretation in Generative Grammar](#) - Jackendoff - 1972
- 251 [Lexical Semantics](#) - Cruse - 1986
- 222 [Frequency Analysis of English Usage: Lexicon and Grammar](#) - Francis, Kucera - 1982
- 216 [Learnability and Cognition: The Acquisition of Argument Structure](#) - Pinker - 1989
- 147 [The Mathematics of Inheritance Systems](#) - Touretzky - 1986
- 146 [Cognitive representations of semantic categories](#) - Rosch - 1975
- 142 [Lexicalization patterns: Semantic structure in lexical forms](#) - Talmy - 1985
- 130 [Basic Color Terms: Their Universality and Evolution](#) - Berlin, Kay - 1991
- 129 [A taxonomy of part-whole relations](#) - Winston, Chaffin, et al. - 1987
- 126 [Retrieval Time From Semantic Memory](#) - Collins, Quillian - 1969
- 120 [A system for representing and using real-world knowledge. Unpublished](#) - Fahlman - 1977
- 118 [The measurements of meaning](#) - Osgood, Suci, et al. - 1957
- 115 [The Syntax and Semantics of Complex Nominals](#) - Levi - 1978
- 112 [The structure of semantic theory](#) - Katz, Fodor - 1963
- 90 [Word Formation in Generative Grammar](#) - Aronoff - 1976
- 88 [Conceptual dependency: A theory of natural language understanding](#) - SCHANK - 1972
- 73 [Meaning and Necessity](#) - Carnap - 1947
- 72 [Deep structure, surface structure and semantic interpretation](#) - Chomsky - 1971
- 61 [Word concepts: A theory of simulation of some basic semantic capabilities](#) - Quillian - 1967
- 48 [The Categories and Types of Present-Day English Word Formation. A Synchronic-Diachronic Approach.](#) München: Beck'sche Verlagsbuchhandlung - Marchand - 1969
- 35 [Universals of Color Naming and Memory](#) - Heider - 1972

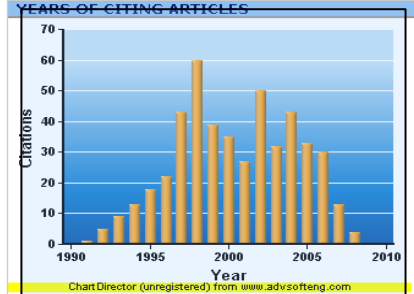
POPULAR TAGS
 Add a tag: [] Submit
 No tags have been applied to this document.

BIBTEX | ADD TO METACART


```
@ARTICLE{Miller90wordnet:an,
  author = {George A. Miller, and Richard Beckwith and Christiane Fellbaum and Derek Gross and Katherine Miller},
  title = {WordNet: An on-line lexical database},
  journal = {International Journal of Lexicography},
  year = {1990},
  volume = {3},
  pages = {235--244}}

```

YEARS OF CITING ARTICLES



ChartDirector (unregistered) from www.advsofteng.com

BOOKMARKS


OPENURL

Document Details

Downloads and External Links

myCiteSeer Launch Points

Citations

BibTeX

Citation Graph

Citation Relationships

Home | Statistics
MetaCart | Sign in to MyCiteSeerX

CiteSeer^x
beta

Documents
Authors
Tables !

Include Citations | [Advanced Search](#) | [Help](#)

Summary
Related Documents
Version History

Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and Model-Based Approach (2000) [104 citations — 7 self]

by David Pennock , Eric Horvitz , Steve Lawrence , C Lee Giles
 In Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence
 Add To MetaCart

DOWNLOAD:
<http://www.neci.nec.com/~lawrence/papers/collab-ua>
<http://dpennock.com/papers/pd-uai-00.pdf>

CACHED:
 |

Add to Collection
Correct Errors
Monitor Changes

Documents Related by Co-Citation

- 547 [Empirical Analysis of Predictive Algorithms for Collaborative Filtering](#) – John S. Breese, David Heckerman, Carl Kadie - 1998
- 665 [GroupLens: An Open Architecture for Collaborative Filtering of Netnews](#) – Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, John Riedl - 1994
- 277 [An algorithmic framework for performing collaborative filtering](#) – J Herlocker, J Konstan, A Borchers, J Riedl - 1999
- 126 [Application of Dimensionality Reduction in Recommender System -- A Case Study](#) – Badrul M. Sarwar , George Karypis, Joseph A. Konstan, John T. Riedl - 2000
- 620 [Social Information Filtering: Algorithms for Automating "Word of Mouth"](#) – Upendra Shardanand, Pattie Maes - 1995
- 183 [Learning collaborative information filters](#) – Daniel Billsus, Michael J Pazzani - 1998
- 159 [Recommendation as Classification: Using Social and Content-Based Information in Recommendation](#) – Chumki Basu, Haym Hirsh, William Cohen - 1998
- 140 [Combining Collaborative Filtering with Personal Agents for Better Recommendations](#) – Nathaniel Good, J. Ben Schafer, Joseph A. Konstan, Al Borchers, Badrul Sarwar, Jon Herlocker, John Riedl - 1999
- 393 [GroupLens: Applying collaborative filtering to Usenet news](#) – Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, John Riedl, High Volume - 1997
- 125 [Eigentaste: A Constant Time Collaborative Filtering Algorithm](#) – Ken Goldberg, Theresa Roeder, Dhruv Gupta, Chris Perkins - 2000
- 211 [Recommending and Evaluating Choices in a Virtual Community of Use. SIGCHI'95](#) – Will Hill, Larry Stead, Mark Rosenstein, George Furnas - 1995
- 91 [Clustering Methods for Collaborative Filtering](#) – Lyle Ungar, Dean Foster, Ellen Andre, Star Wars, Fred Star Wars, Dean Star Wars, Jason Hiver Whispers - 1998
- 92 [Combining Content-Based and Collaborative Filters in an Online Newspaper](#) – Mark Claypool, Anuja Gokhale, Tim Miranda, Pavel Murnikov, Dmitry Netes, Matthew Sartin - 1999
- 92 [Latent Class Models for Collaborative Filtering](#) – Thomas Hofmann - 1999
- 268 [Item-based Collaborative Filtering Recommendation Algorithms](#) – Badrul Sarwar, George Karypis, Joseph Konstan, John Riedl - 2001
- 70 [Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments](#) – Alexandrin Popescul , Lyle H. Ungar, David M. Pennock, Steve Lawrence - 2001
- 442 [Using collaborative filtering to weave an information tapestry](#) – David Goldberg, David Nichols, Brian M. Oki, Douglas Terry - 1992
- 297 [Fab: Content-based, collaborative recommendation](#) – Marko Balabanovic, Yoav Shoham - 1997
- 213 [Analysis of Recommendation Algorithms for E-Commerce](#) – Badrul Sarwar, George Karypis, Joseph Konstan, John Riedl - 2000

View or Download | [Add to My Collection](#) | [Correct Errors](#)

Related Documents: [Active Bibliography](#) | [Co-citation](#)

Citation Relationship - Co-Citation

Home | Statistics | [About CiteSeer^x](#) | [Bulletin](#) | [Submit Documents](#) | [Feedback](#) | [Privacy Policy](#) | [CiteSeer^x Data](#) | [Source Code](#)

© 2007 The Pennsylvania State University
 Developed at and hosted by [The College of Information Sciences and Technology](#) at Penn State



MyCiteSeer Interface

- A personal portal space for users
 - Track and Manage
 - User defined collections
 - Tags
 - Search queries
 - Correct document metadata.
 - Monitor documents.
 - Generate API keys.
- Planned features
 - New interface
 - More extensive metadata.

MyCiteSeer



[Documents](#) | [Authors](#) | [Tables !](#)

Include Citations | [Advanced Search](#) | [Help](#)

- ACCOUNT HOME**
- [Profile](#)
- [Collections](#)
- [Tags](#)
- [Monitoring](#)
- [Admin Console](#)

Menu

Hi
Welcome to your personal portal into CiteSeerX

Latest News (See All)

[New Features](#) [Tue Mar 17 13:16:00 EDT 2009]

[Read more...](#)

[OAI Service available](#) [Mon Mar 16 21:33:00 EDT 2009]

You can now download CiteSeer^x metadata through the OAI service interface

[Read more...](#)

[Service upgrades and Activations](#) [Mon Mar 16 21:17:00 EDT 2009]

The submission and corrections have now been activated on CiteSeer^x.

[Read more...](#)



Other Interfaces: OAI - PMH

- Programmatic Access – metadata is always in high demand.
- A low barrier mechanism, was supported by CiteSeer
- Extend the existing framework to support OAI.
- CGI with embedded database vs. Servlets with DAO, more efficient and simpler implementation.
- OAI-2 with Dublin Core format.
- Many harvesters available for OAI-2.

API

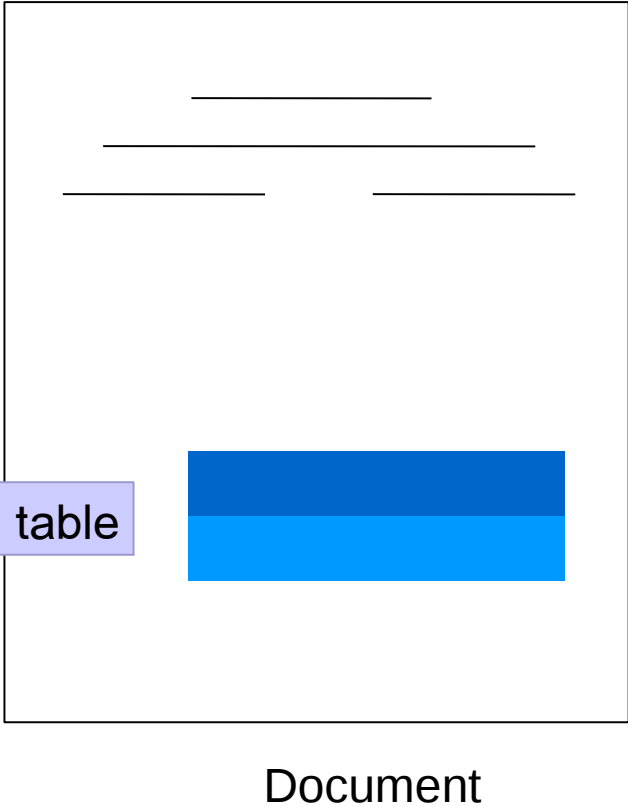
- API is central to programmatic access to SeerSuite.
 - Exposes relationships and data elements.
 - Implements a REST based service providing access to
 - Document metadata (docid)
 - Authors (aid),
 - Citations (cid),
 - Key-words, and citation contexts are provided.
- Built using the Jersey library (JAX-RS)
- Uses MyCiteSeer
 - Control access to API.
 - Limits number of queries per day.

Federation of Services

- CiteSeer^x provides services not part of SeerSuite
 - Consequence of constant research and development.
 - Infrastructure shared with SeerSuite
 - Web app framework, Data storage: Database, Repository.
- Service examples:
 - Table search – from TableSeer
 - Disambiguated author search
 - Future services: Algorithm search, Figure search, Citation recommendation, etc.

Table Search

- Table extraction
 - Table caption and content
- Table search
 - Ingestion extracted table
 - Database and Index.
 - Link table with document
- Index
 - Separate from document index.
- Other infrastructure part of SeerSuite
- Template for newer services

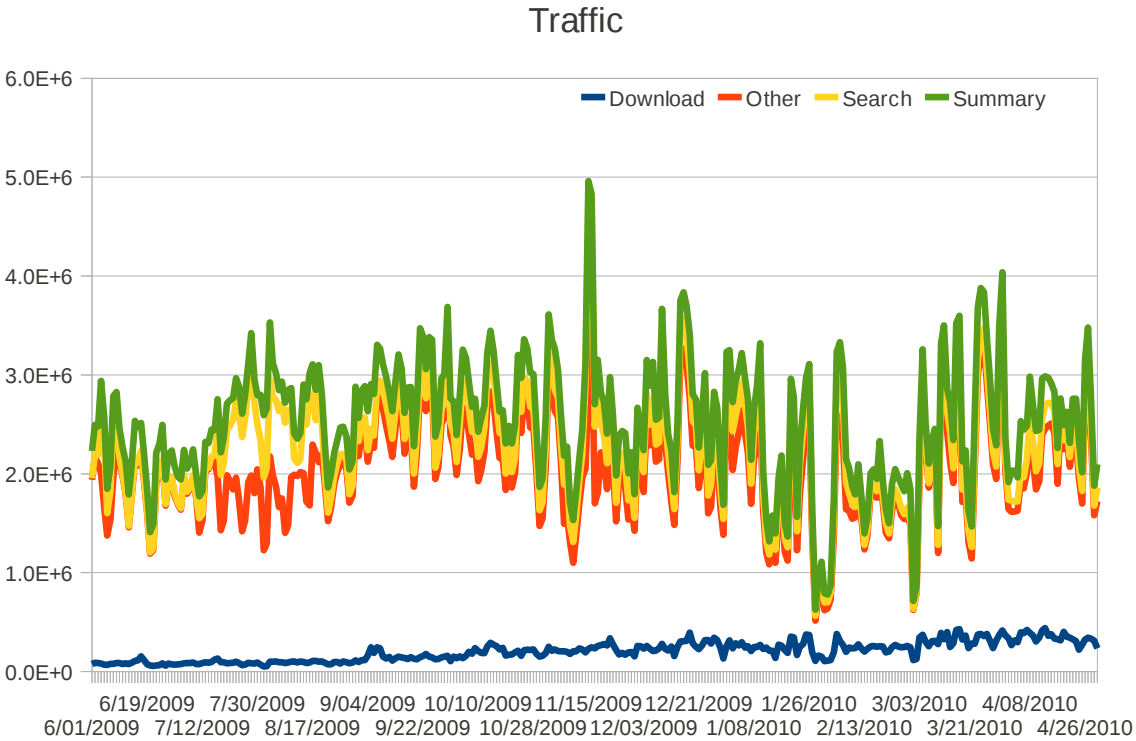


Disambiguated Author Search

- Author Disambiguation
 - Essential to identify and attribute records accurately.
 - Which M. Johnson to cite?.
- Algorithms constantly in development
 - DBSCAN and LASVM
 - Uses co-authorship, header information (address, affiliation)
 - Upcoming method includes Random Forests and is online.
- Separate index.
- Other infrastructure part of SeerSuite

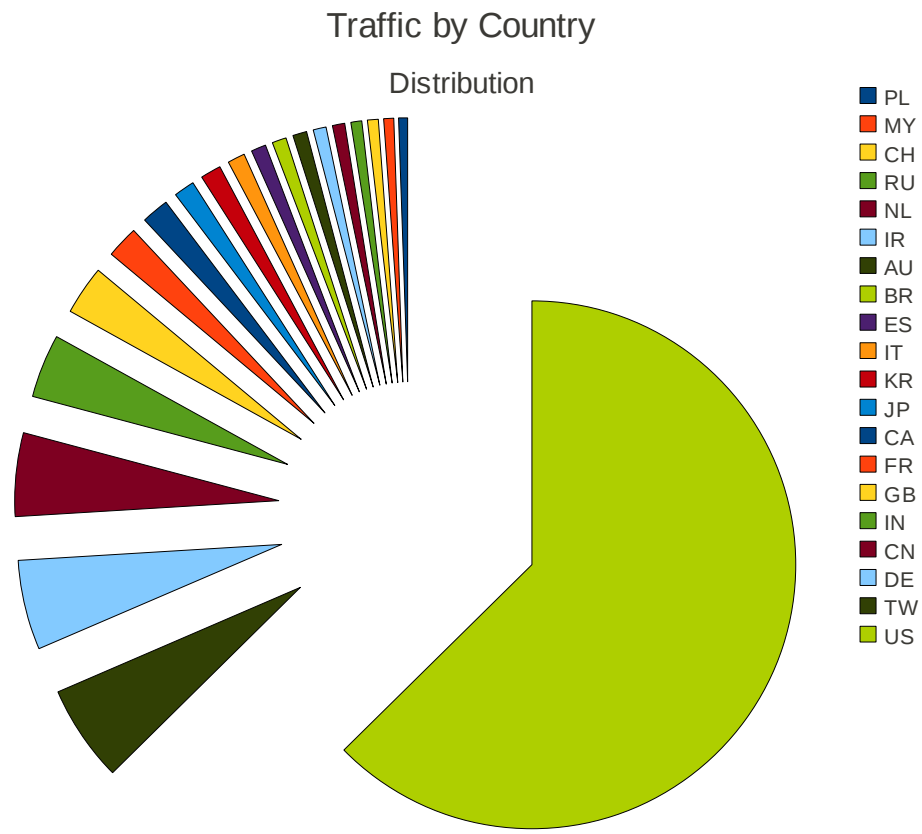
Usage - Traffic

- 2 million hits on average every day.
- Images, javascript dominate.
- Downloads and Document summaries are popular.
- Search has the highest variation.
- MyCiteSeer receives little traffic (< 1% of total.)



Usage – Country Distribution

- Traffic from all over the globe.
- US dominates
- Germany, China, India, Taiwan, UK are other sources of traffic.
- Most of the external referrals are from search engines – Google, Google Scholar, Yahoo, Bing.



Collaboration

- SeerSuite is a collaborative effort
 - Collaborators (no mirrors)
 - University of Arkansas, National University of Singapore, King Saud University host ***independent*** copies of CiteSeer^x.
 - Research directions
 - User interface
 - Metadata extraction and ranking
 - Information aggregation
 - Entity disambiguation
 - Trend monitoring
 - Citation recommendations
- CiteSeer^x data available upon request (rsync)
 - Documents, databases, anonymized logs.
 - Data sharing
 - Cornell, CMU, MIT, University College London, NSWC, others.

Lessons Learned

- Multi-tier architecture, open source applications can be used to build scalable, reliable and robust services.
- Need for virtualization – cost effective.
- Data requests – building API's important.
- Federated services make adopting new services possible.
- Metadata extraction – always room for improvement
- Optimizations implemented allow better performance.
- Several improvements such as UI and performance enhancements possible
- Heavily used but not heavily implemented (SeerSuite)

Conclusions and Summary

- Overview of SeerSuite
 - Architecture, Workflow, Deployment, UI, other interfaces including OAI, API
- Federation of services
 - Table search
 - Author disambiguation
 - Others planned
- Analysis of usage of CiteSeer^x
- Collaboration
- Lessons Learned
- [Download SeerSuite !](#)



Availability of Code

- Released under Apache Software Foundation License (version 2).
- Code for SeerSuite and related software available on Source forge
 - <http://sourceforge.net/projects/citeseerx>
- Virtual Machine with a deployment of SeerSuite
 - <http://singularity.ist.psu.edu:8080/seerlab.html>
- Support by the research group at Penn State



Q & A