

---

# SeerSuite, Author Disambiguation and VIVO: Building the Semantic Web by Focused Crawling

Pradeep Teregowda, Pucktada Treeratpituk, C. Lee Giles  
Pennsylvania State University  
University Park, PA, USA

---

# Outline

---

- SeerSuite
    - CiteSeer<sup>x</sup>
    - Disambiguation
      - Disambiguation in CiteSeer<sup>x</sup>
    - Identifying Experts
      - SeerSeer
    - VIVO and SeerSuite
    - Future Work
-

# SeerSuite

---

- Goal: Effective dissemination of scientific and academic literature.
  - Approach: SeerSuite
    - Framework for building digital library search engines
      - Examples: CiteSeer<sup>x</sup>, Chem<sub>x</sub>Seer
    - Features:
      - Full text indexing, autonomous citation linking.
      - Collection built by crawling authors web sites.
      - Code open source, freely available, uses open source components.
      - Flexible, scalable, robust, reliable, portable.
      - Author disambiguation
-

# CiteSeer<sup>x</sup>

---

- Digital library search engine, instance of SeerSuite.
    - Document search and view, Citation search, Author search, Author disambiguation, Table search.
  - Indexes over
    - 1.6 million documents.
    - 31 million citations.
    - 1.5 million unique authors.
  - Crawls/harvests
    - 150K (avg) documents every month, processes 160K (avg) and ingests around 10K (avg) documents every month.
  - Access
    - Over 2 million hits a day.
  - Metadata repository of academic documents.
-

# CiteSeer<sup>x</sup> - Services

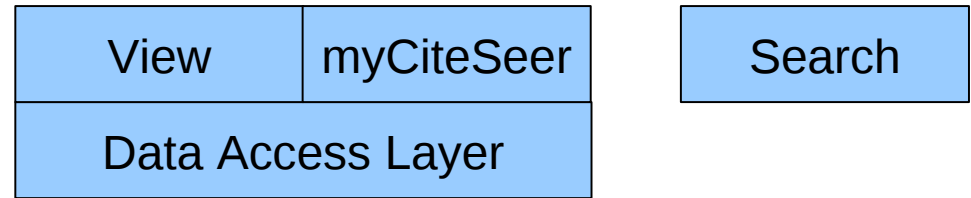
- Web application

- Search
- View
- myCiteSeer



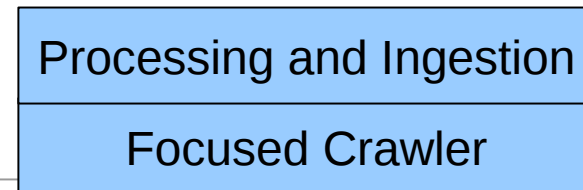
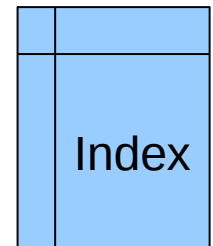
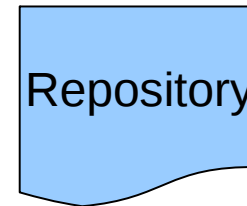
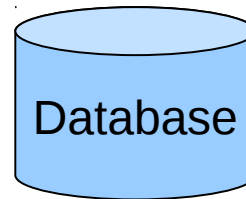
- Data Storage

- Database
- Index
- Repository.



- Acquisition

- Crawler
- Processors
- Ingestion.



# myCiteSeer<sup>x</sup>

---

- Personalization portal for users
  - Ability to
    - Correct documents
      - Crowd sourcing.
    - Tag documents
      - Collaborative user defined tags.
    - Create personal portfolios
      - Collections.
    - Monitor documents
      - Notifications on corrections, new citations.
    - Store queries.
-

# Author Disambiguation

---

- Disambiguation is essential to attribute author metadata accurately.
  - Issues:
    - Aliasing
    - Common name
    - Typographic errors.
  - Examples:
    - M. Johnson ( Aliasing – variation of the same name)
    - Mark Johnson (Common name – which mark johnson)
    - Merk Johnson (Typographic errors – Merk for Mark).
-

# Entity Disambiguation

---

- Long history of research among many communities:
    - Database community
      - Record Linkage
      - Database hardening
      - Duplicate Detection
    - Natural Language Processing
      - Co-reference resolution
      - Entity resolution
      - String matching.
    - Algorithms generally consist of two parts: pairwise linking and clustering.
-

# CiteSeer<sup>x</sup> - Disambiguation

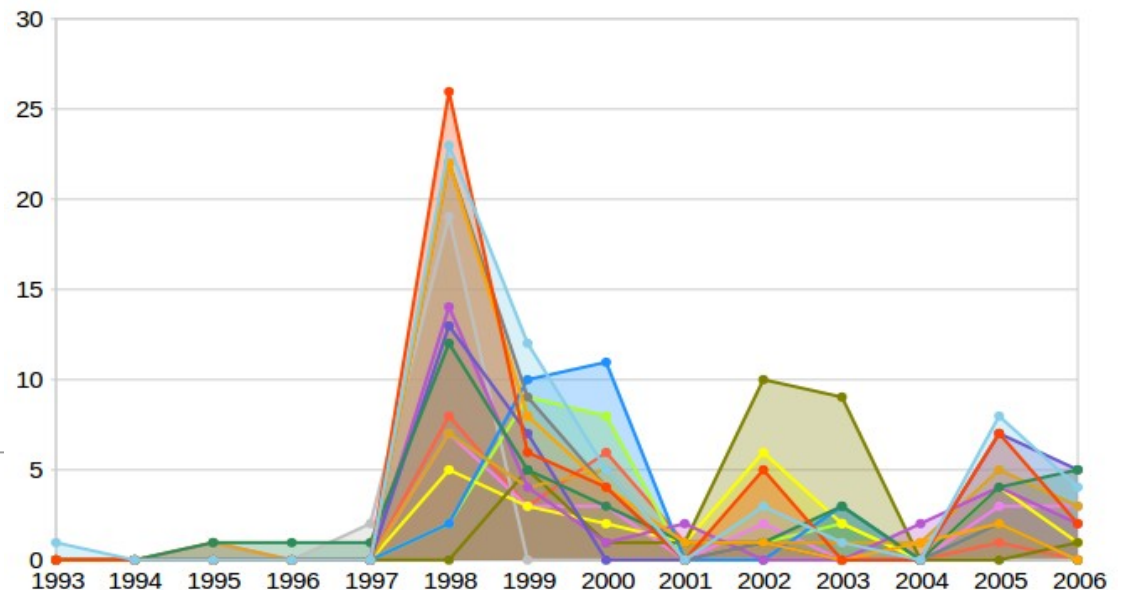
---

- CiteSeer<sup>x</sup> crawls documents from the web
    - Document metadata useful for disambiguation.
  - Approach:
    - Learn to estimate the likelihood that two author names from two different papers refer to the same person, based on metadata such as affiliation, paper title, coauthors, etc., then do the clustering.
      - SVM (distance) + DBSCAN (cluster) method.
      - Topic Modeling
      - Random Forests (MEDLINE)
-

# Identifying Experts

- Disambiguated author records and key phrase extraction of publications can be used with the citation graph information to accurately identify
  - Publication History
    - Author interests.
    - Trend analysis of author interest.
  - Most cited authors
  - Most relevant work
  - Author expertise.

Trends of author interests



# SeerSeer

- Identifying expertise & research interests for each disambiguated author based on the content of their publication records.

**Mark Johnson**

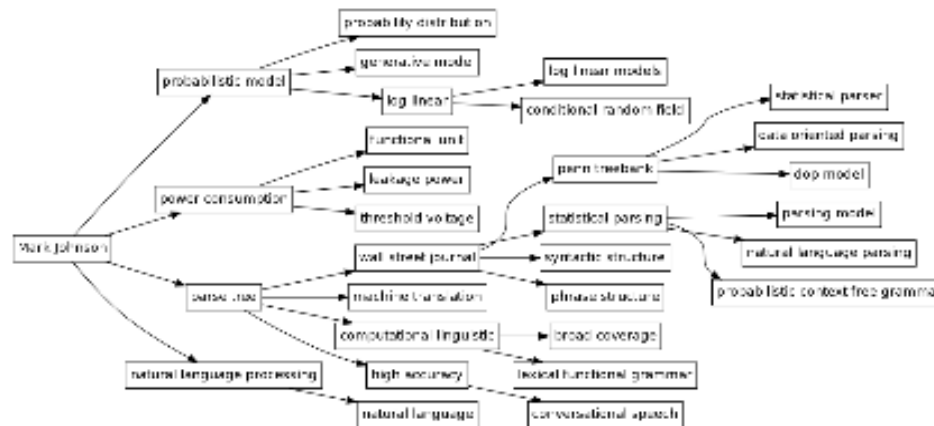
**Brown Laboratory for Linguistic Information Processing  
BLLIP Brown University**

## CiteSeerX Statistics

Publication years	1990-2008
Publication count	92
Citation count	922
H-Index	16

## >> Taxonomy of Expertise

Author Profile

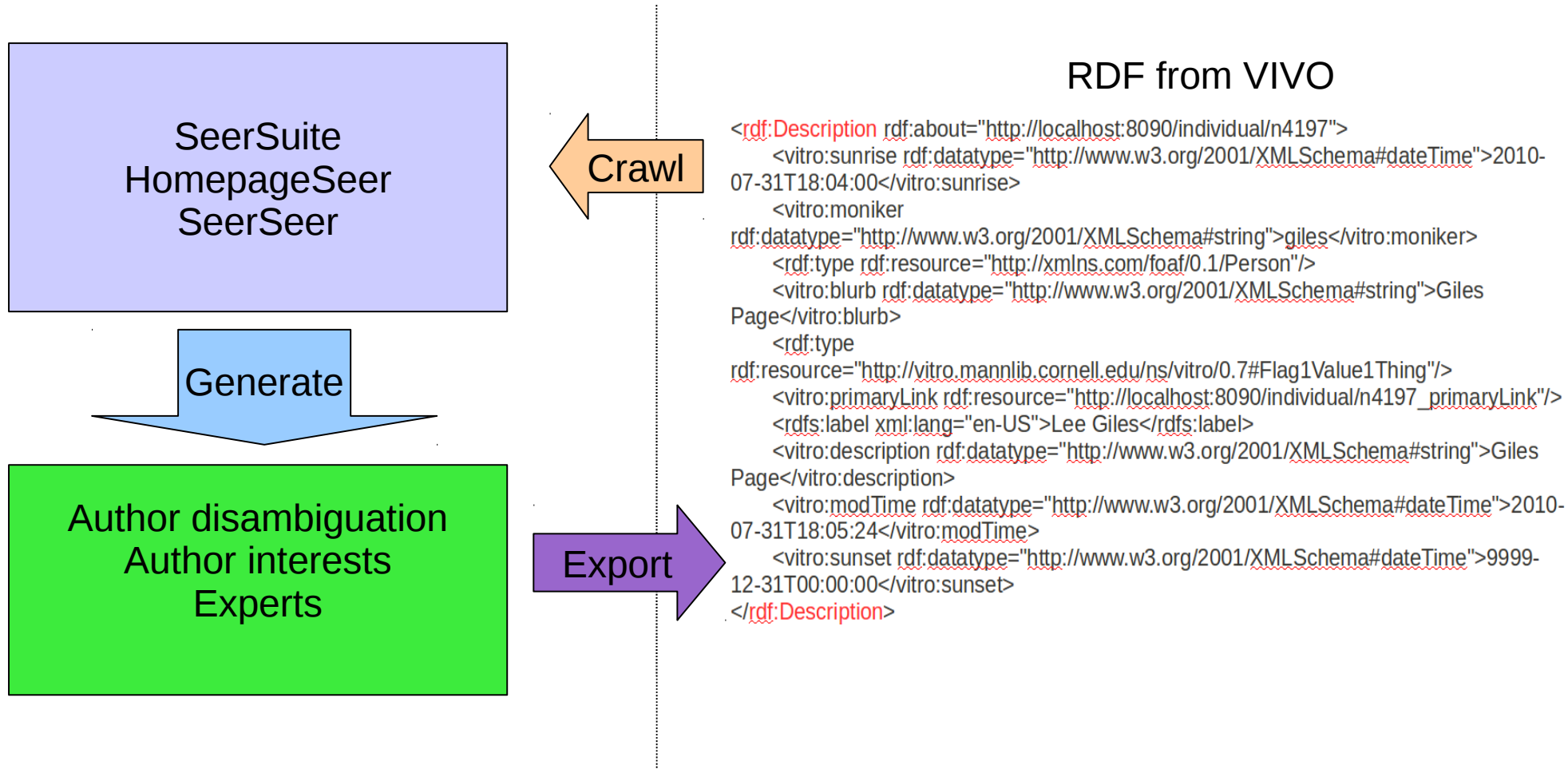


# Author metadata

---

- CiteSeer<sup>x</sup> author metadata
    - Name and variations.
    - Affiliation(s).
    - Publication history.
    - Home page information if available.
  - In addition, SeerSeer exposes
    - Interests and trends
    - Expertise
      - Arranged in hierarchy.
-

# VIVO and SeerSuite



# VIVO and SeerSuite interaction

---

- SeerSuite, HomepageSeer, SeerSeer can take advantage of the RDF information exposed by VIVO
    - Improve coverage of authors, publications.
    - Additional information during disambiguation and profile generation
  - Expose Seer metadata through VIVO and RDF
    - Disambiguated records
    - Citation graph based services – co-authors, ranking
    - Expertise and interest detection.
-

# Conclusions

---

- Many opportunities for VIVO and SeerSuite interactions
  - Increase metadata availability for both resources
  - Better author and entity disambiguation benefits the scientific community.
  - Expertise generation and metadata
    - Valuable beyond individual projects.
-

# Future Work

---

- Develop extensible consumers to work with VIVO sources.
    - Incorporate VIVO metadata into disambiguation, seer services.
  - Generate records from existing metadata for VIVO data consumers.
    - Ontologies based on the VIVO framework.
  - Extend existing SeerSuite framework to provide richer and extensive metadata.
-

---

Q & A

---