Making Sense of Massive Amounts of Scientific Publications:

The Scientific Knowledge Miner Project

Francesco Ronzano, **Ana Freire**, Diego Saez-Trumper, Horacio Saggion

upf. Universitat Pompeu Fabra Barcelona

20 seconds… 1 paper

The Rise of Open Access
Science 04 Oct 2013
Vol. 342, Issue 6154, pp. 58-59

Information Overload (scientific repositories)

PubMed

WEB OF KNOWLEDGE℠

THOMSON REUTERS

arXiv.org

ELSEVIER

Scopus®

Information Overload (scientific repositories)

24,6M PubMed

WEB OF KNOWLEDGE℠ THOMSON REUTERS 90M

1M arXiv.org

ELSEVIER Scopus® 57M

Sometimes between 2017 and 2021, more than half of the papers available globally are expected to be published as Open Access articles.

Lewis, David W. "**The inevitability of open access.**" College & Research Libraries 73.5 (2012): 493-506.

The peculiarities of research publications

Making Sense of Massive Amounts of Scientific Publications: the Scientific Knowledge Miner Project

TITLE

Francesco Ronzano, Ana Freire, Diego Saez-Trumper and Horacio Saggion

Department of Information and Communication Technologies
Universitat Pompeu Fabra
Carrer Tanger 122-140, Barcelona, Spain
first.last@upf.edu

Abstract. The World Wide Web has become the hugest repository ever for scientific publications and it continues to increase at an unprecedented rate. Nevertheless, this information overload makes the exploration of this content a very time-consuming task. In this landscape, the availability of text mining tools to characterize and explore distinctive features of the scientific literature is mandatory. We present the Scientific Knowledge Miner (SKM) Project, that aims to investigate new approaches and frameworks to facilitate the extraction of knowledge from scientific publications across different disciplines. More specifically, we will focus on citation characterization, recommendation and scientific document summarization.

ABSTRACT

Keywords: text mining, information extraction, recommender systems, indexing, crawling, online resources.

1 Introduction:

During the last decade the amount of scientific information available on-line increased at an unprecedented rate. Recent estimates reported that a new paper is published every 20 seconds [1]. PubMed[1], Elsevier' Scopus[2] and Thomson Reuther's ISI Web of Knowledge[3] respectively contain more than 24, 57 and 90 million papers. In this scenario, the exploration of scientific literature has turned into an extremely complex and time-consuming task. The availability of text mining tools able to extract, aggregate and turn scientific unstructured textual contents into well organized and interconnected knowledge is fundamental.

However, scientific publications are characterized by several structural (title, abstract, figures, citations...), linguistic and semantic peculiarities that make them difficult to analyze by relying on general purpose text mining tools. One of the special features of scientific papers is their network of citations, that are starting to be exploited in several context including opinion mining [2, 7] and scientific text summarization [3, 8]. Besides citations, the interpretation of the semantics of the actual textual contents of

(SUB)SECTION

[1] http://www.ncbi.nlm.nih.gov/pubmed
[2] http://www.scopus.com
[3] http://www.webofknowledge.com

Fig. 2. Web based visualization of the information extracted from a paper thanks to the DRI Framework. In particular, we can see highlighted in bold the sentences of the paper classified as approach.

CAPTION

References

1. The rise of open access. Science 342(6154), 58–59 (2013)
2. Abu-Jbara, A., Ezra, J., Radev, D.R.: Purpose and polarity of citation: Towards nlp-based bibliometrics. In: HLT-NAACL. pp. 596–606 (2013)
3. Abu-Jbara, A., Radev, D.: Coherent citation-based summarization of scientific papers. In: Proc. of 49th Annual Meeting of the ACL: Human Language Techologies. pp. 500–509. ACL (June 2011)
4. Balog, K., Ramampiaro, H., Takhirov, N., Nørvåg, K.: Multi-step classification approaches to cumulative citation recommendation. In: Proc. of the 10th Conference on Open Research Areas in Information Retrieval. pp. 121–128. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE (2013)
5. Fisas, B., Ronzano, F., Saggion, H.: A multi-layered annotated corpus of scientific papers. In: LREC Conference (2016)
6. He, Q., Kifer, D., Pei, J., Mitra, P., Giles, C.L.: Citation recommendation without author supervision. In: Proc. of the fourth ACM international conference on Web search and data mining. pp. 755–764. ACM (2011)
7. Nakov, P.I., Schwartz, A.S., Hearst, M.A.: Citances: Citation sentences for semantic analysis of bioscience text. In: In Proc. of the SIGIR'04 workshop on Search and Discovery in Bioinformatics (2004)
8. Ronzano, F., Saggion, H.: Taking advantage of citances: citation scope identification and citation-based summarization. In: Text Analytics Conference (2014)
9. Ronzano, F., Saggion, H.: Knowledge extraction and modeling from scientific publications. In: Semantics, Analytics, Visualisation: Enhancing Scholarly Data Workshop co-located with the 25th International World Wide Web Conference April 11, 2016 - Montreal, Canada (2016)
10. Saggion, H.: Summa: A robust and adaptable summarization tool. In: Traitement Automatique des Langues. vol. 49.2 (2008)
11. Teufel, S., Moens, M.: Summarizing scientific articles: experiments with relevance and rhetorical status. Computational linguistics 28(4), 409–445 (2002)
12. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: Proc. 2006 conference on empirical methods in NLP. pp. 103–110. ACL (2006)

BIBLIOGRAPHIC ENTRY

Scientific publications: claims

In order to take full advantage of the knowledge present in scientific publications proper **semantic indexing, search** and **content aggregation** approaches, are required.

Benefits:

- Search of new information on specific scientific problems
- Semi-automatic assessment of papers and research proposals
- Hypothesis formulation
- Tracking of scientific and technological advances
- Scientific intelligence
- Assisted report and review writing
- Question answering
- …

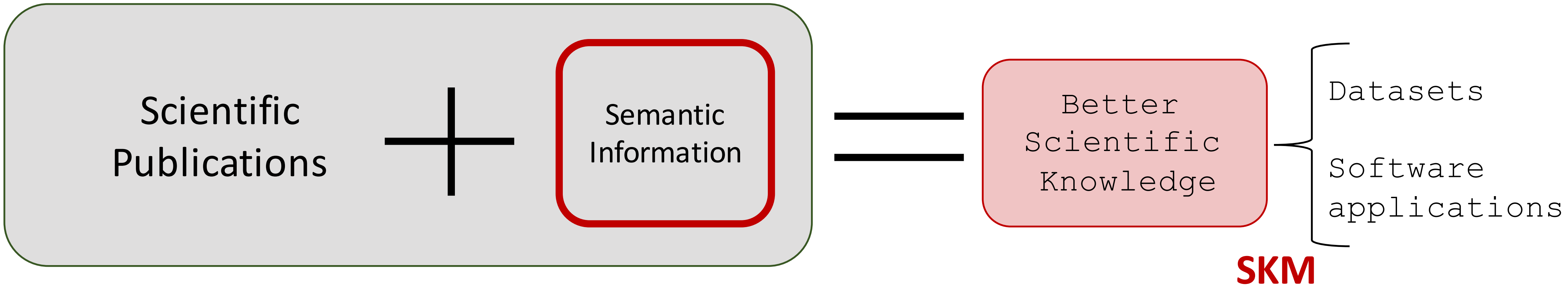
The Scientific Knowledge Miner Project (SKM)

Facilitate the extraction of knowledge from scientific publications across many disciplines.

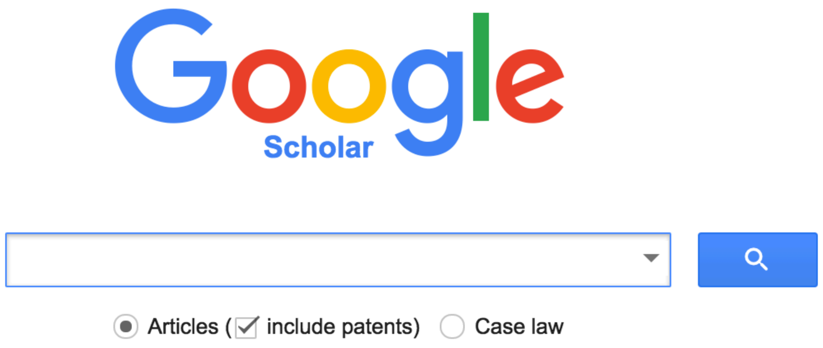
Improve a variety of use cases such as:

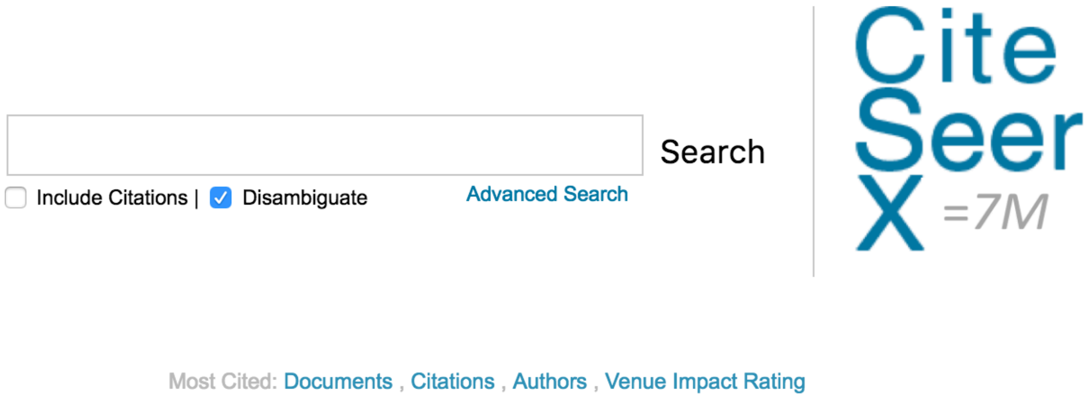
- Citation Characterization
- Citation Recommendation
- Summarization
- …

➢ KEY: Papers are enriched with **structural**, **linguistic** and **semantic information**



The Scientific Knowledge Miner Project (SKM)

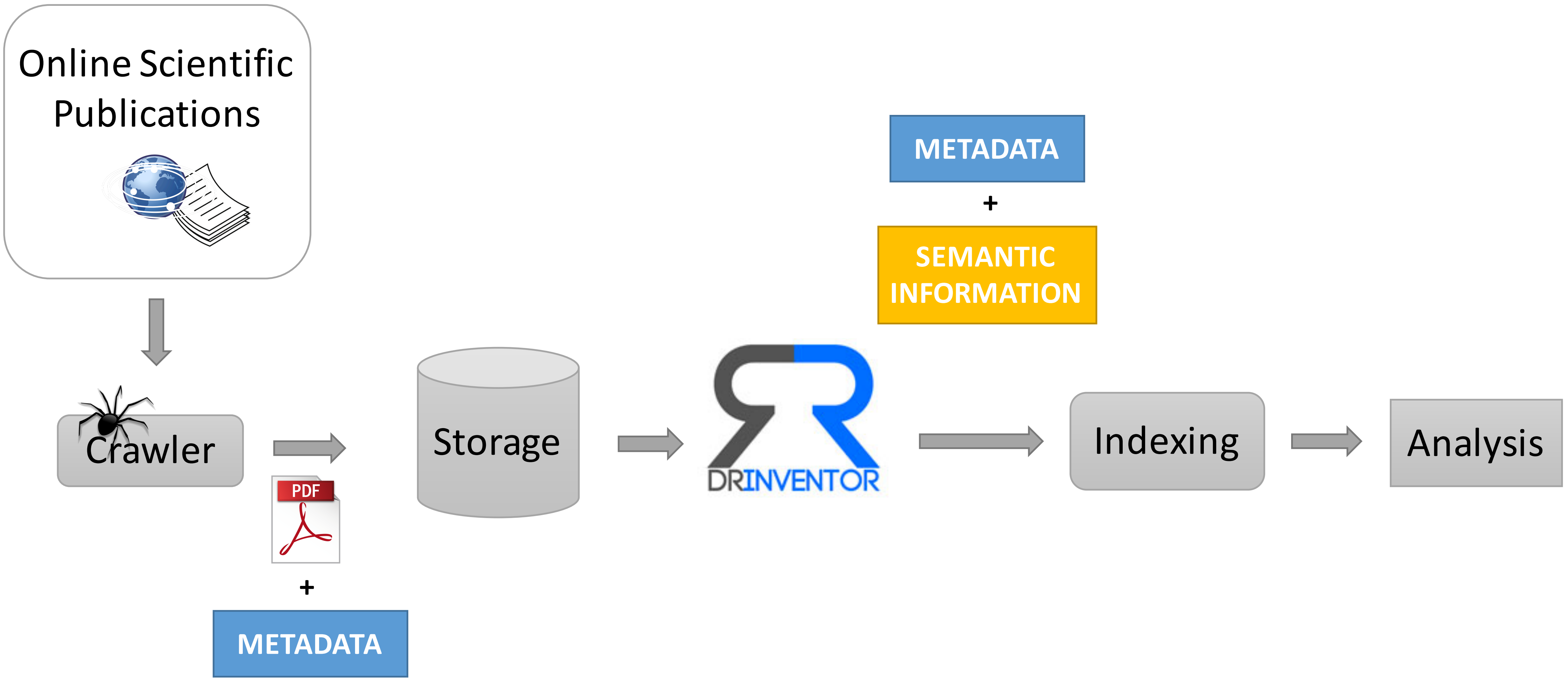




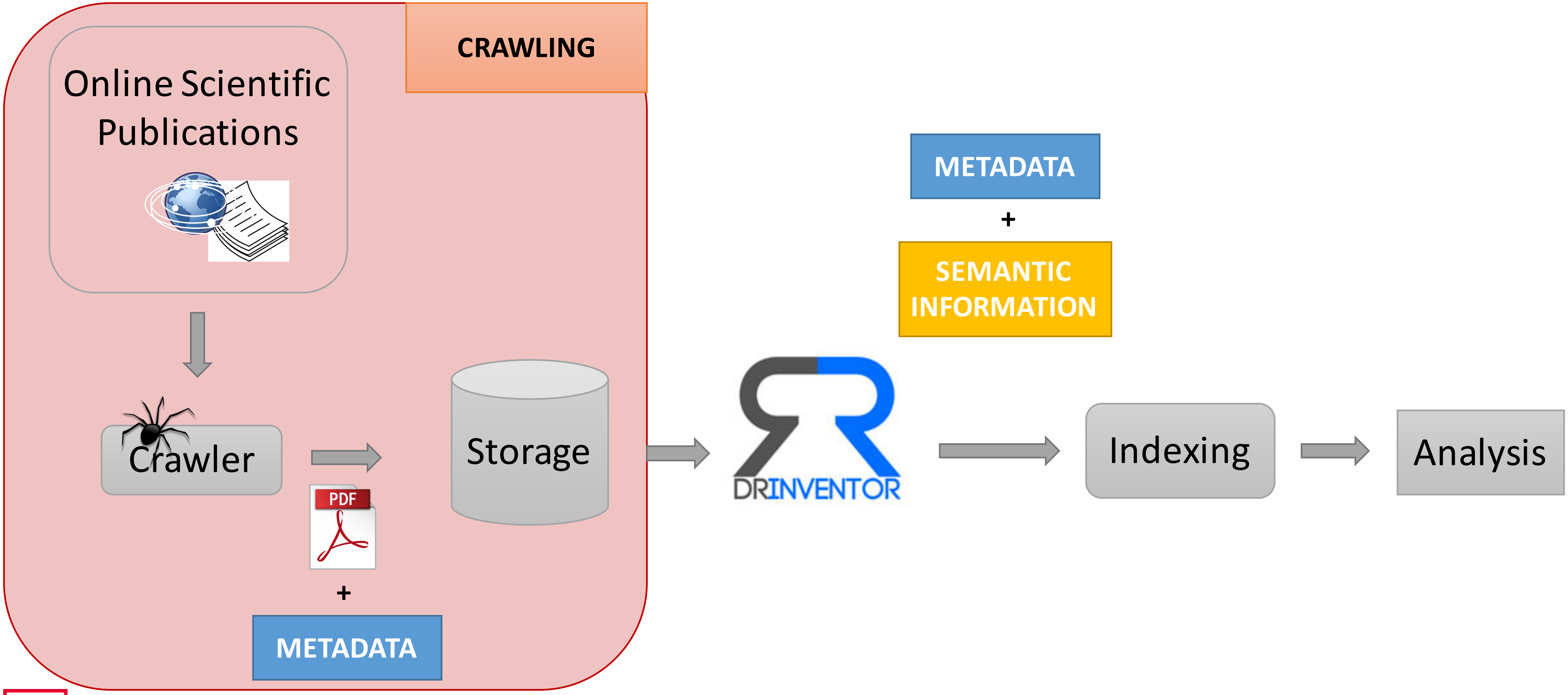
The SKM approach to the analysis of scientific literature:

- Relies on a finer-grained analysis of the contents of publications
- Is grounded on the automated characterization of a varied set of semantic aspects of papers, including the rhetorical structure or the purpose of citations.

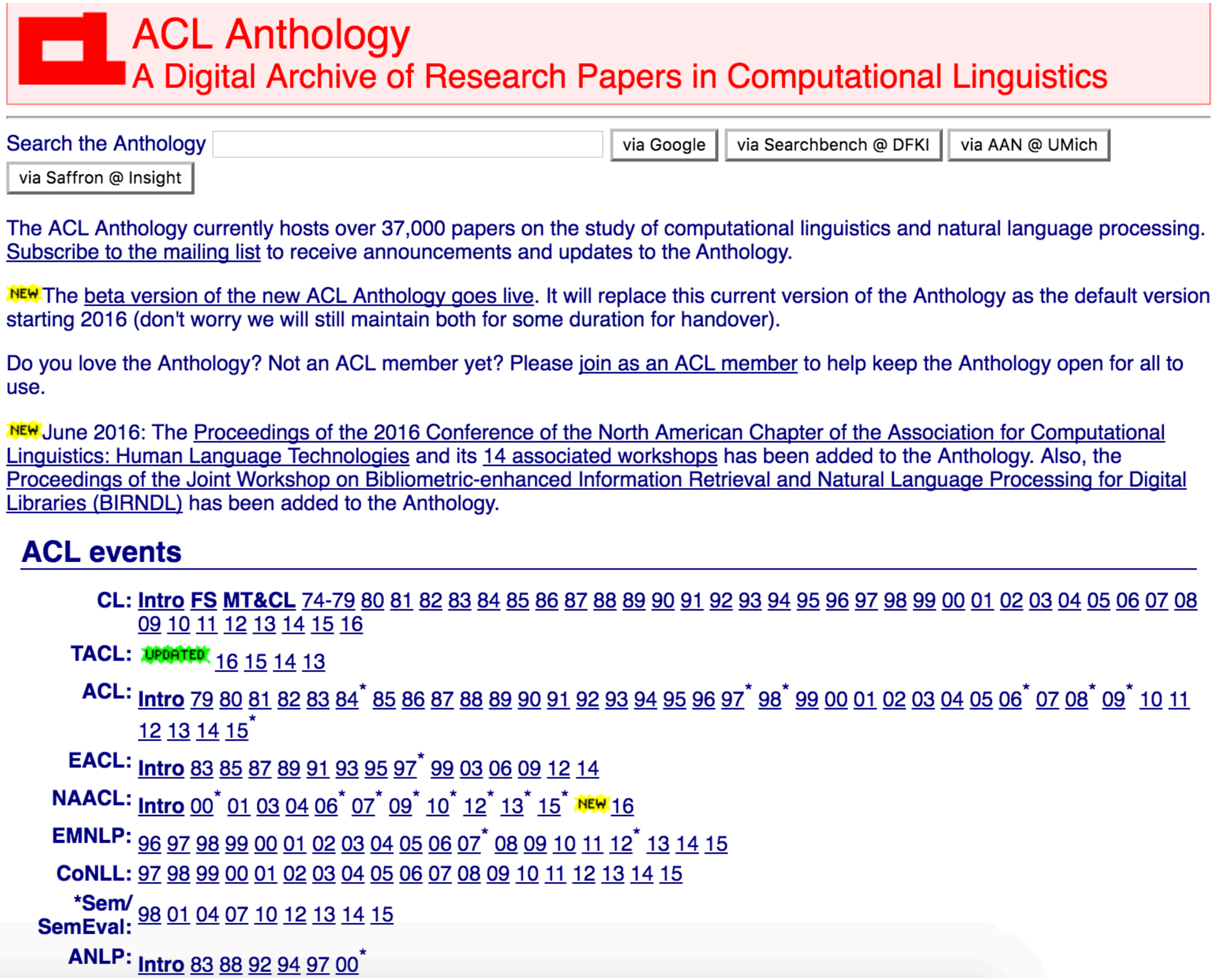
The Scientific Knowledge Miner Project (SKM)



The Scientific Knowledge Miner Project (SKM)



Crawling


ACL Anthology
A Digital Archive of Research Papers in Computational Linguistics

Search the Anthology [via Google] [via Searchbench @ DFKI] [via AAN @ UMich] [via Saffron @ Insight]

The ACL Anthology currently hosts over 37,000 papers on the study of computational linguistics and natural language processing. Subscribe to the mailing list to receive announcements and updates to the Anthology.

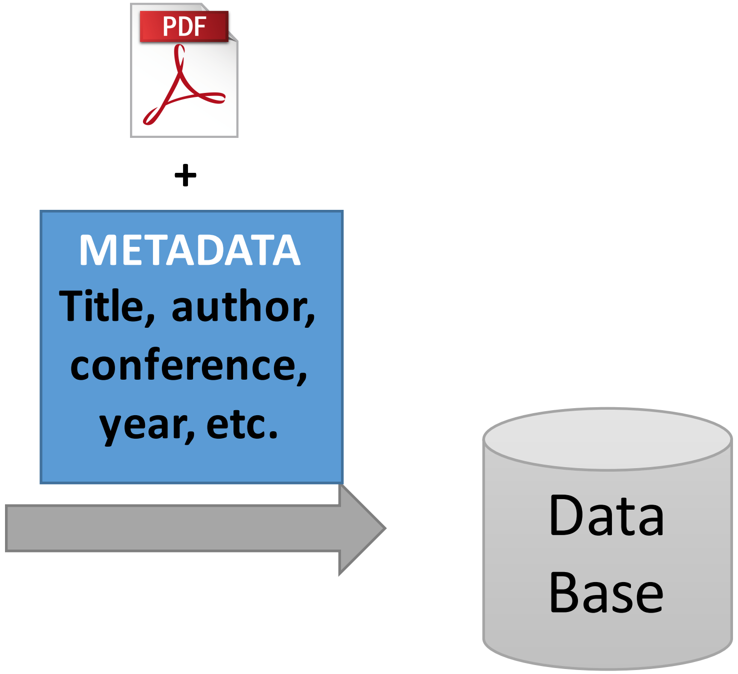
NEW The beta version of the new ACL Anthology goes live. It will replace this current version of the Anthology as the default version starting 2016 (don't worry we will still maintain both for some duration for handover).

Do you love the Anthology? Not an ACL member yet? Please join as an ACL member to help keep the Anthology open for all to use.

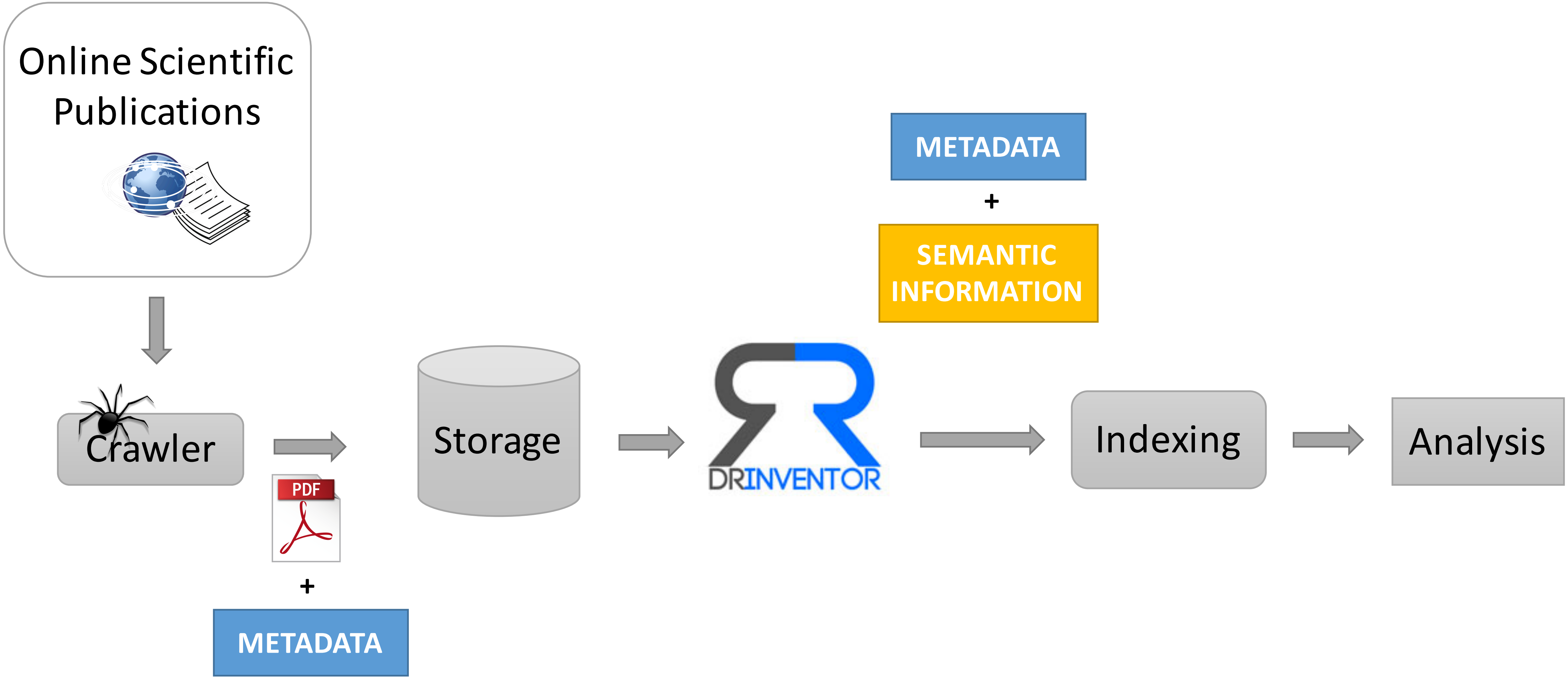
NEW June 2016: The Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies and its 14 associated workshops has been added to the Anthology. Also, the Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) has been added to the Anthology.

ACL events

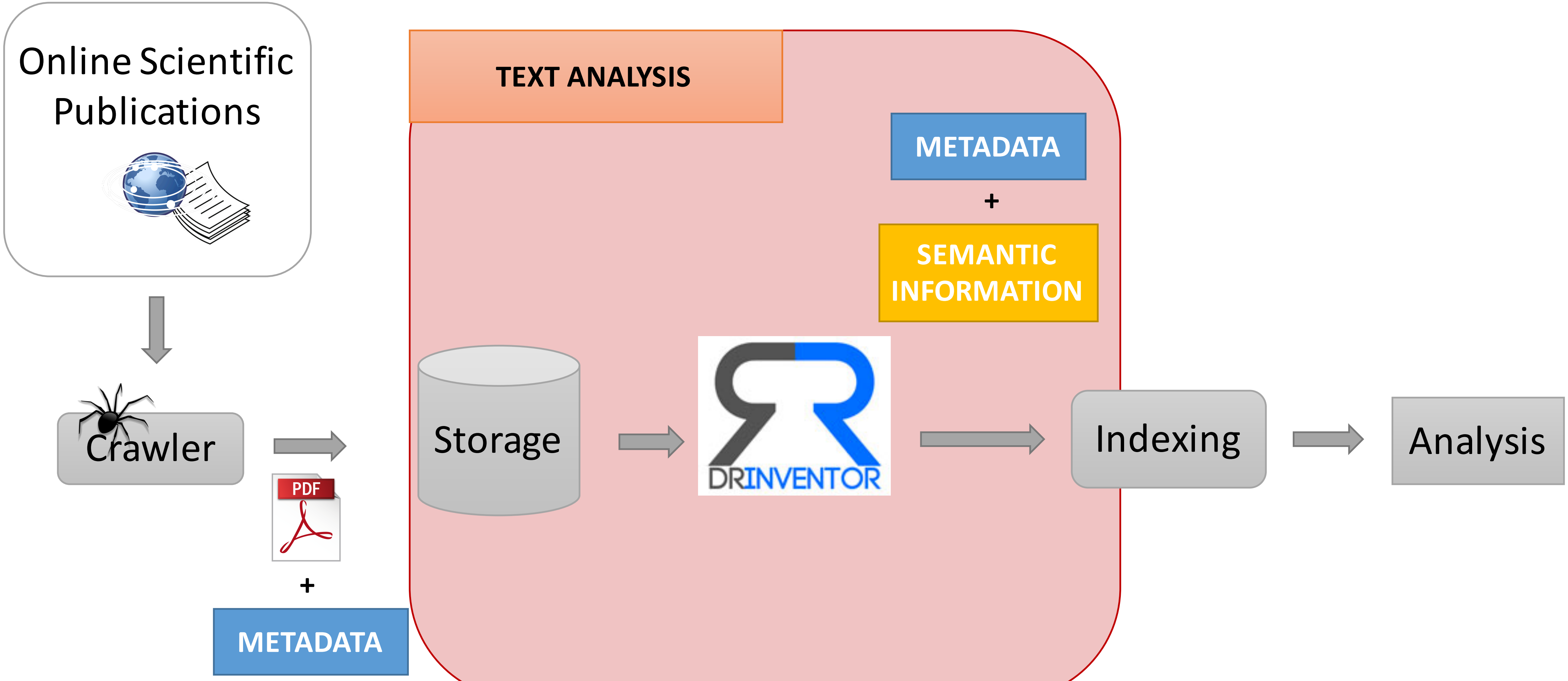
CL: Intro FS MT&CL 74-79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16
TACL: UPDATED 16 15 14 13
ACL: Intro 79 80 81 82 83 84* 85 86 87 88 89 90 91 92 93 94 95 96 97* 98* 99 00 01 02 03 04 05 06* 07 08* 09* 10 11 12 13 14 15*
EACL: Intro 83 85 87 89 91 93 95 97* 99 03 06 09 12 14
NAACL: Intro 00* 01 03 04 06* 07* 09* 10* 12* 13* 15* NEW 16
EMNLP: 96 97 98 99 00 01 02 03 04 05 06 07* 08 09 10 11 12* 13 14 15
CoNLL: 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15
***Sem/ SemEval:** 98 01 04 07 10 12 13 14 15
ANLP: Intro 83 88 92 94 97 00*



The Scientific Knowledge Miner Project (SKM)



The Scientific Knowledge Miner Project (SKM)



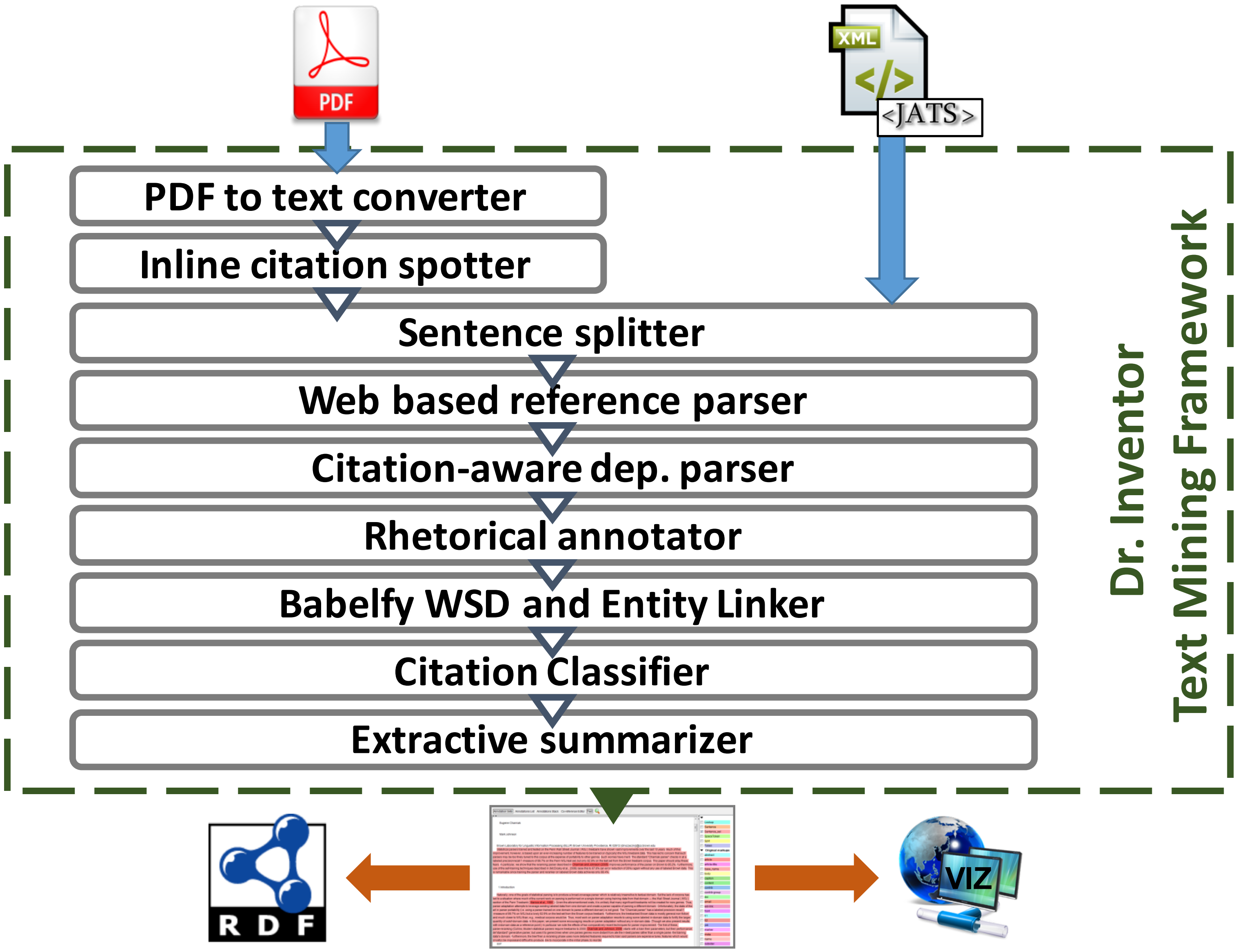
Dr. Inventor Text Mining Framework

- Integrate and customize **text mining tools** and **on-line services** to enable and ease a wide range of scientific publication analyses
- Papers are enriched with **structural**, **linguistic** and **semantic information**

http://backingdata.org/dri/library/

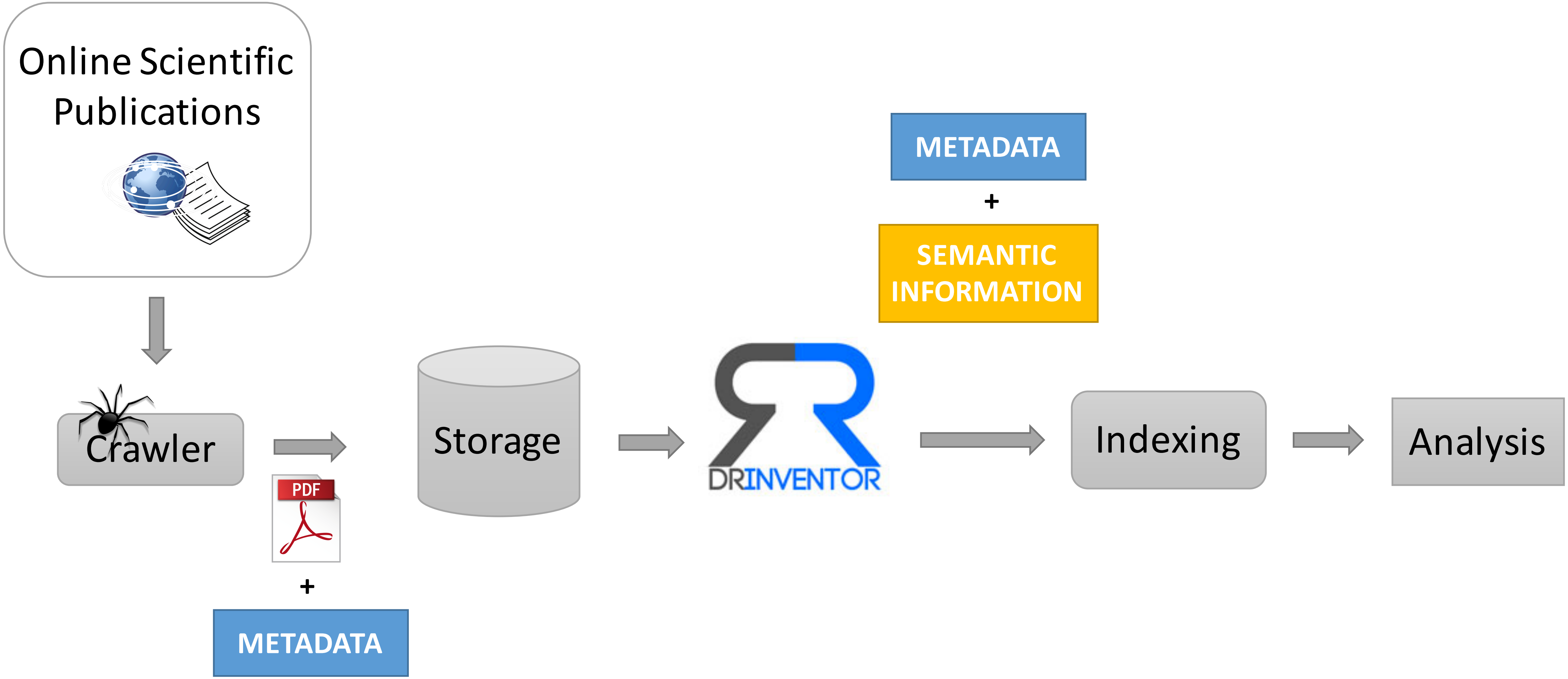
- Self-contained Java 8 library managed by Maven
- Focused on **textual content**
- Relying on a **shared data model** (java classes) to represent a paper
- Exposing a **convenient API** to access the mined information
- Based on GATE general architecture for text engineering to manage **textual annotations**

Dr. Inventor Text Mining Framework

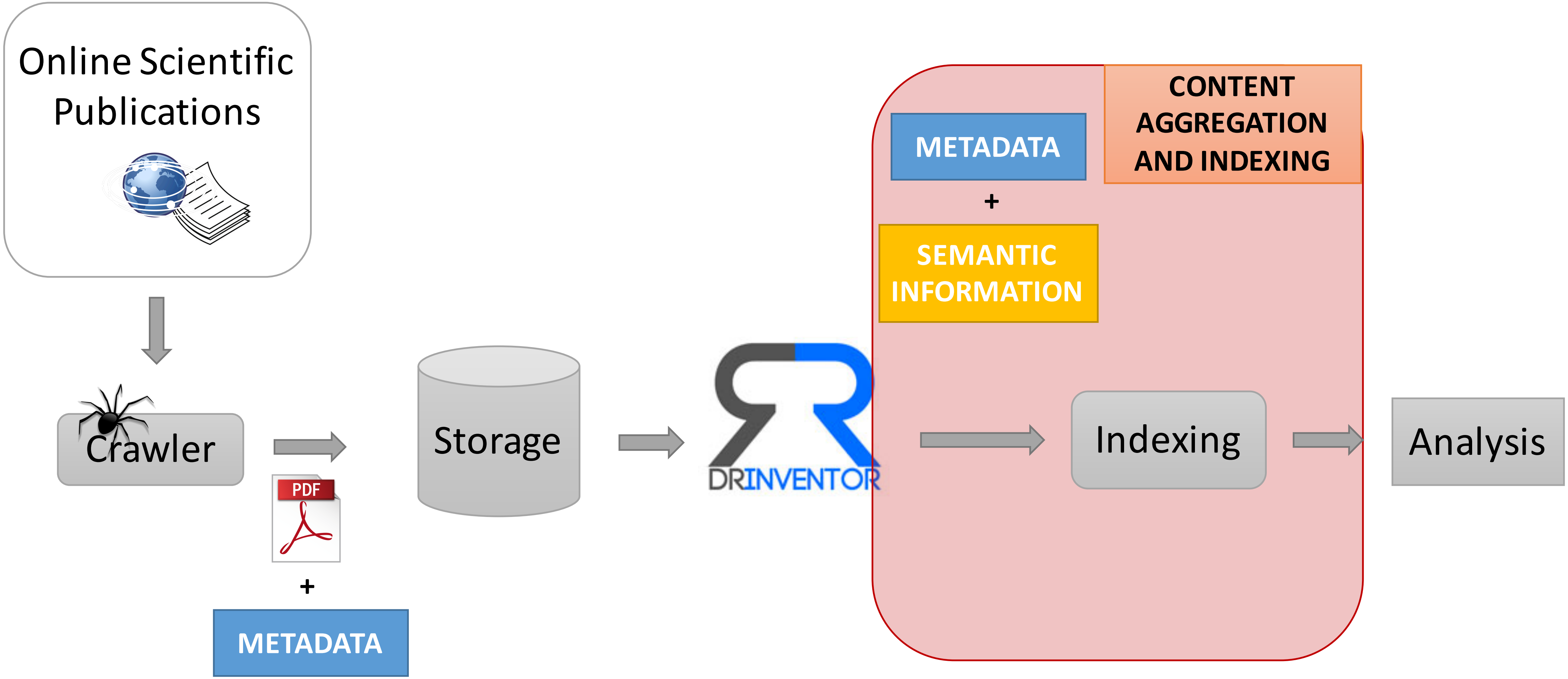


Universitat Pompeu Fabra Barcelona

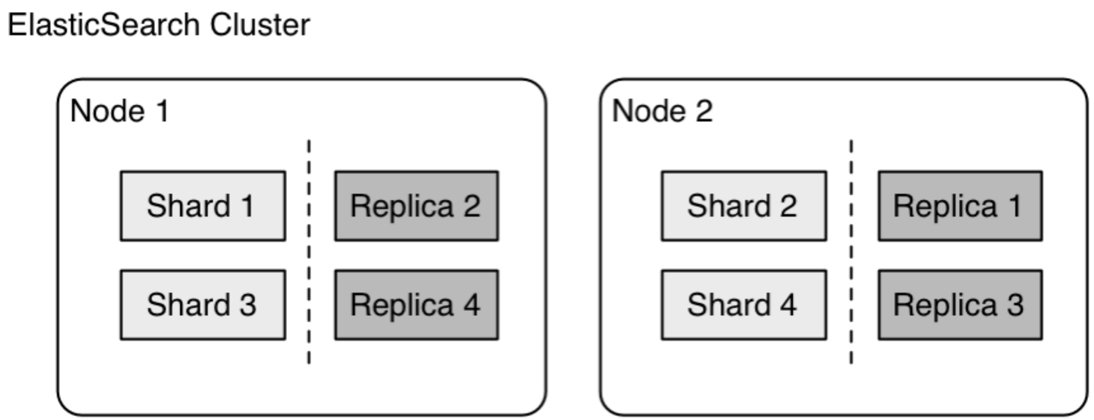
The Scientific Knowledge Miner Project (SKM)



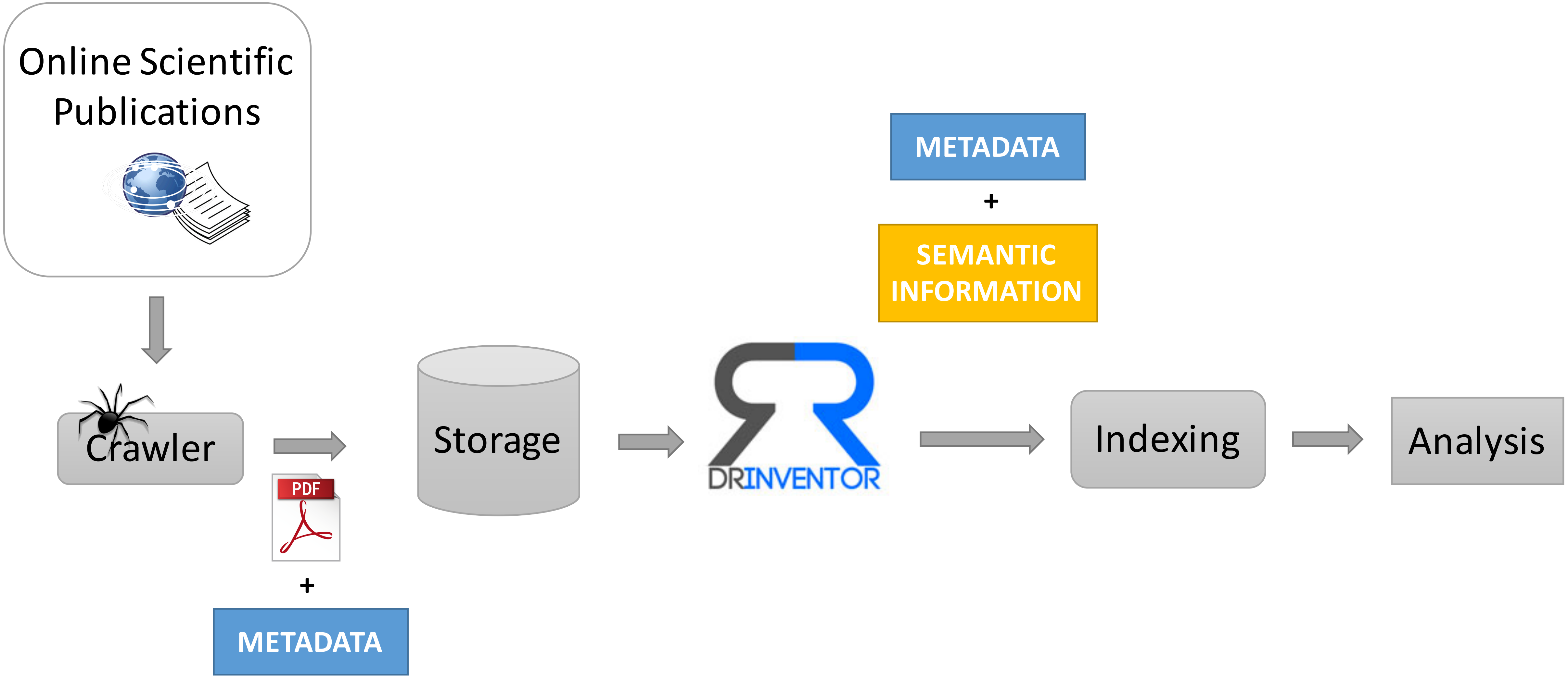
The Scientific Knowledge Miner Project (SKM)



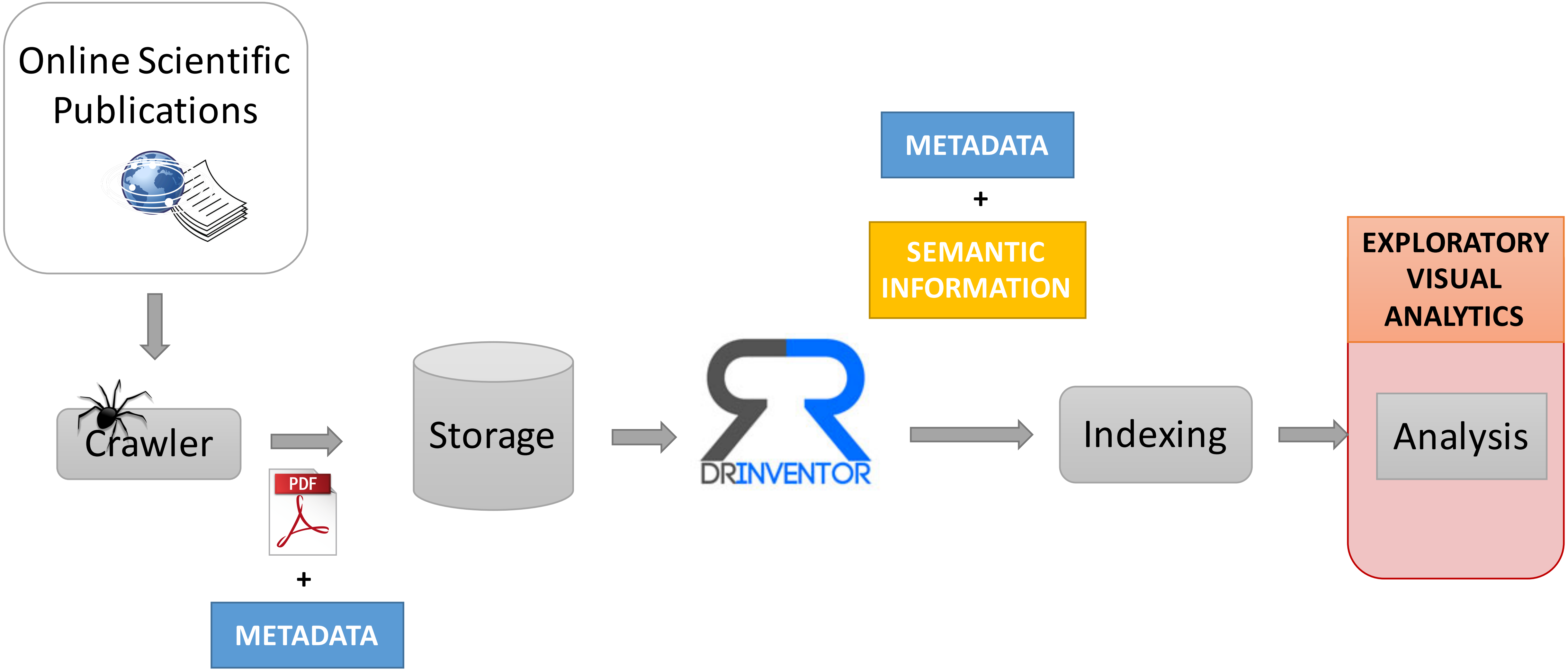
Indexing


ElasticSearch Cluster
Node 1
Shard 1
Replica 2
Shard 3
Replica 4
Node 2
Shard 2
Replica 1
Shard 4
Replica 3

The Scientific Knowledge Miner Project (SKM)



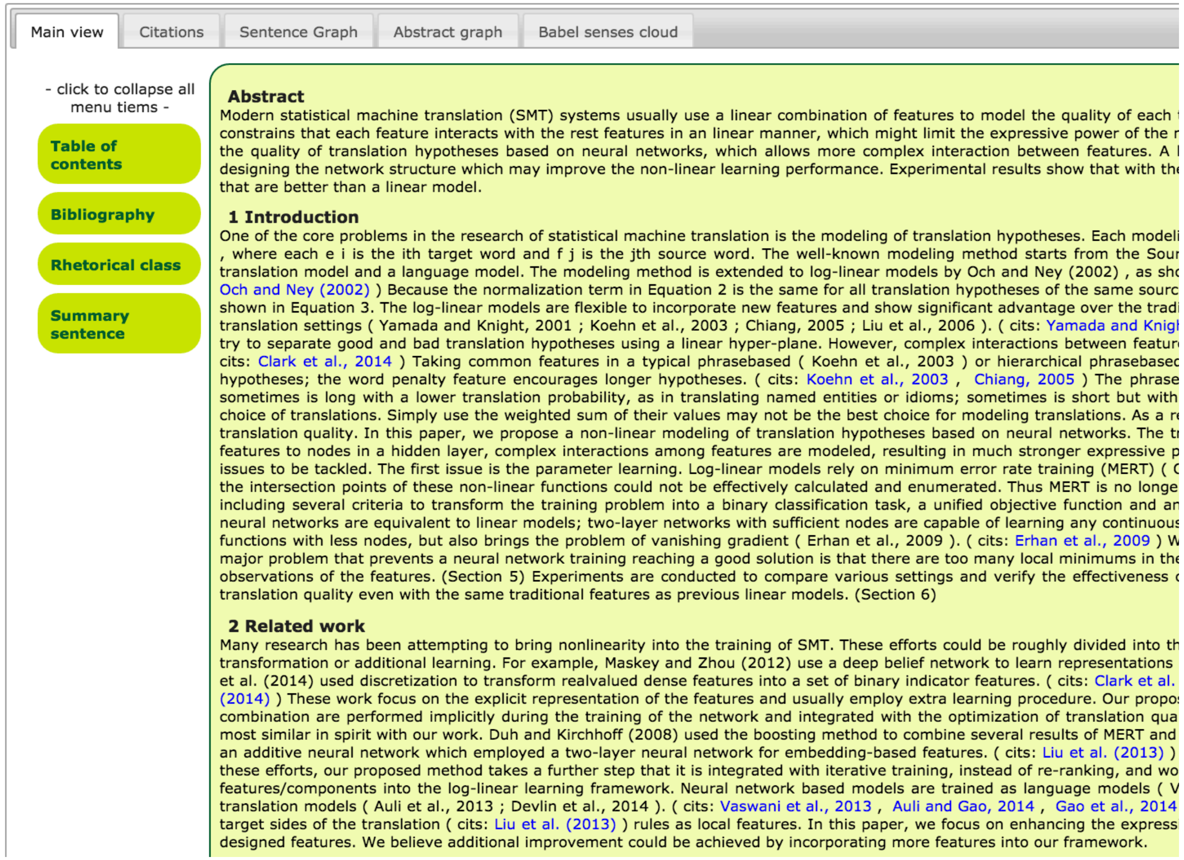
The Scientific Knowledge Miner Project (SKM)



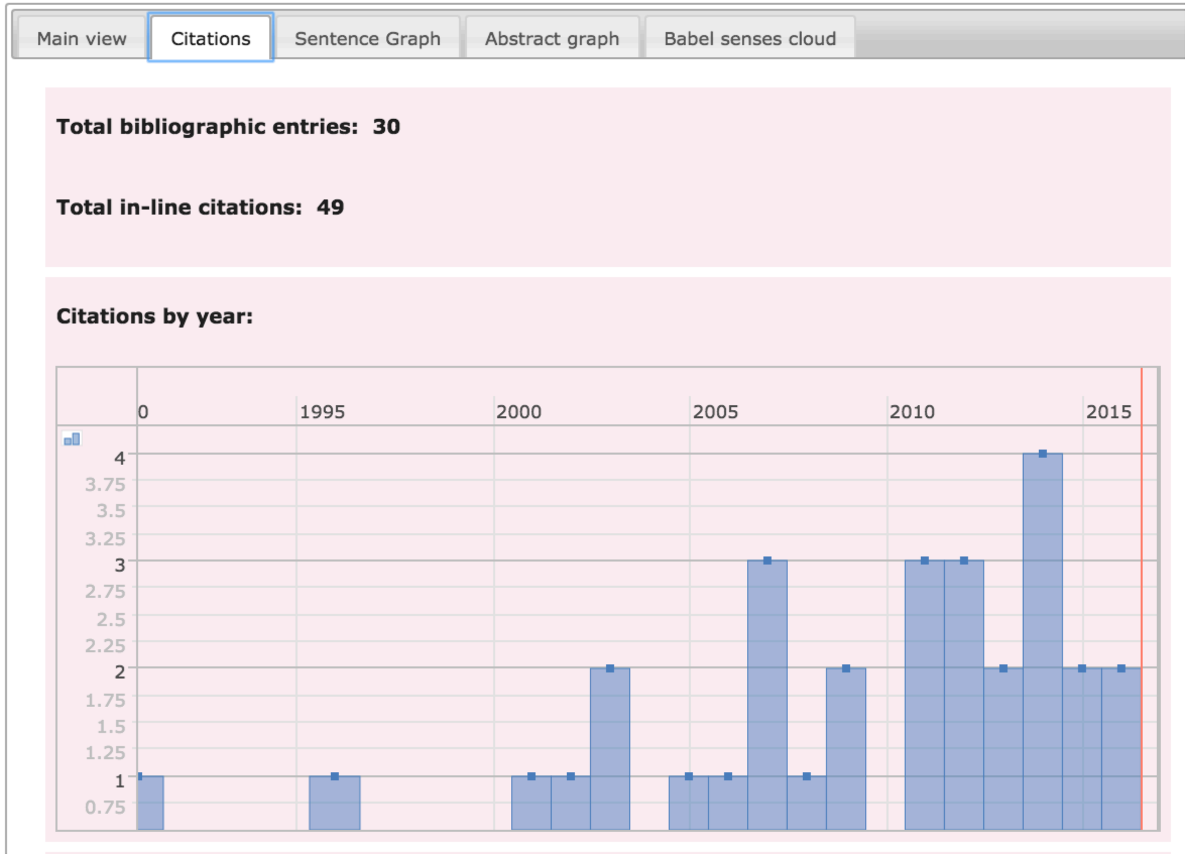
Universitat Pompeu Fabra Barcelona

Analysis

http://backingdata.org/dri/viz/

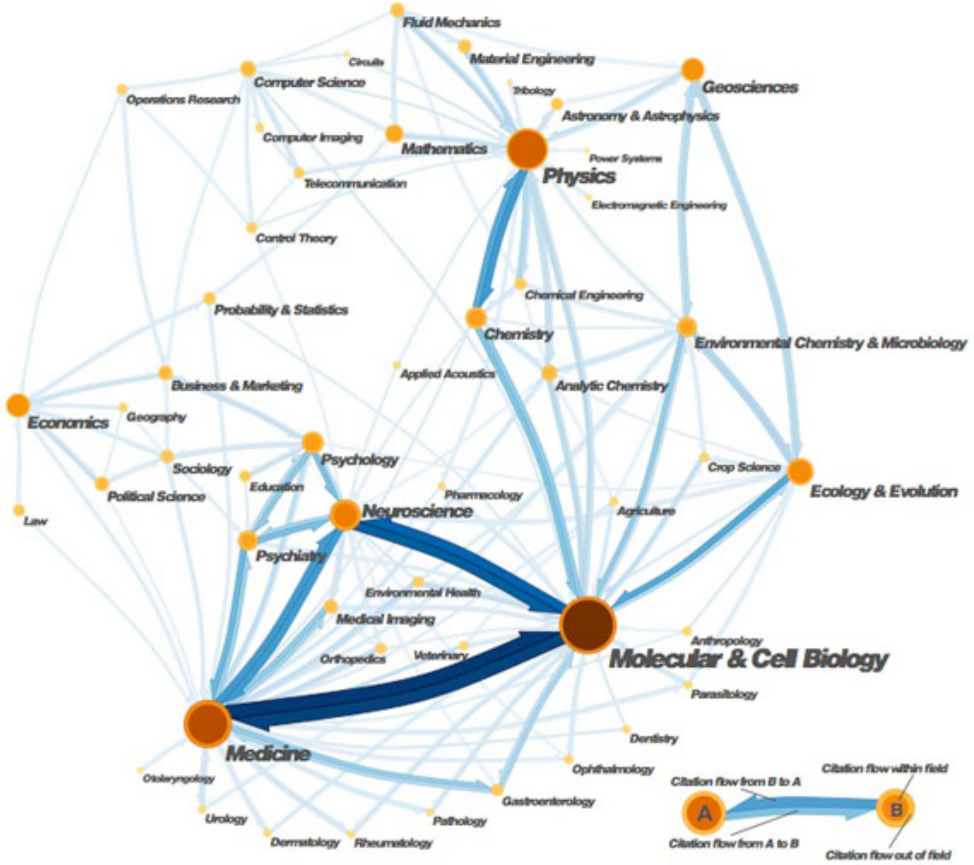


Non-linear Learning for Statistical Machine Translation
Huang Shujian - Chen Huadong - Dai Xinyu - Chen Jiajun
(2015) (URL: http://arxiv.org/abs/1503.00107)



Use Case 1: Citation Characterization

Experiment new metrics: what do others say about one paper?



Enrich citation counts with semantics

DRINVENTOR FRAMEWORK

CITATION PURPOSE

- **Criticism**
- **Comparison**
- **Use**
- **Substantiation**
- **Basis**
- **Neutral**

+ 17 sub-purposes

Use Case 2: Citation Recommendation

Recommend similar papers / authors

SENTENCE RHETORICAL CATEGORY
Background
Approach
Challenge
Outcome
Future Work
+ 3 sub-categories

Some alternative phrase alignment approaches have been developed, which do not rely on the Viterbi word alignment. Both (Marcu, 2002) and (Zhang, 2003) consider a sentence pair as different realizations of a sequence of concepts. These alignment approaches segment the sentences into a sequence of phrases.

CHALLENGE

BACKGROUND

HYPOTHESIS

Use Case 3: Scientific Document Summarization

Extractive summarization

Some alternative phrase alignment approaches have been developed, which do not rely on the Viterbi word alignment. Both (Marcu, 2002) and (Zhang, 2003) consider a sentence pair as different realizations of a sequence of concepts. These alignment approaches segment the sentences into a sequence of phrases.

Summary:
Some alternative phrase alignment approaches have been developed, which do not rely on the Viterbi word alignment.
These alignment approaches segment the sentences into a sequence of phrases.

SENTENCE SUMMARY RELEVANCE (1 to 5 ratings)

and

HAND-WRITTEN SUMMARY

Conclusions and future work

Scientific Knowledge Miner (SKM) aims at facilitating the extraction, aggregation and navigation of knowledge from scientific publications.

- Consolidate the SKM publication mining infrastructure
- Exploit the semantics of papers to perform large scale investigations of:
  - Alternative metrics to evaluate a paper based on citation semantics
  - Semantically motivated recommendation of scientific publications
  - Summarization of scientific literature

Acknowledgements

EXCELENCIA
MARÍA
DE MAEZTU

DRINVENTOR
FRAMEWORK

ACCIÓ

Generalitat
de Catalunya

GOBIERNO
DE ESPAÑA

MINISTERIO
DE ECONOMÍA
Y COMPETITIVIDAD

Financé avec l'aide du FEDER
Fonds Européen de
Développement Régional

Making Sense of Massive Amounts of Scientific Publications:

The Scientific Knowledge Miner Project

{francesco.ronzano, ana.freire, diego.saez, horacio.saggion}@upf.edu

upf. Universitat Pompeu Fabra Barcelona