Making Sense of Massive Amounts of Scientific Publications:

The Scientific Knowledge Miner Project

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



# 20 seconds... 1 paper

The Rise of Open Access

Science 04 Oct 2013

Vol. 342, Issue 6154, pp. 58-59



## Information Overload (scientific repositories)



### WEB OF KNOWLEDGE™









# Information Overload (scientific repositories)

















Scopus



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

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.

**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<sup>1</sup>, Elsevier' Scopus<sup>2</sup> and Thomson Reuther's ISI Web of Knowledge<sup>3</sup> 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



**ABSTRACT** 

(SUB)SECTION



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.

#### References

- 1. The rise of open access. Science 342(6154), 58-59 (2013)
- Abu-Jbara, A., Ezra, J., Radev, D.R.: Purpose and polarity of citation: Towards nlp-based bibliometrics. In: HLT-NAACL. pp. 596–606 (2013)
- 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)
- 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 INTER-NATIONALES D'INFORMATIQUE DOCUMENTAIRE (2013)
- Fisas, B., Ronzano, F., Saggion, H.: A multi-layered annotated corpus of scientific papers. In: LREC Conference (2016)
- 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)
- 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)
- Ronzano, F., Saggion, H.: Taking advantage of citances: citation scope identification and citation-based summarization. In: Text Analytics Conference (2014)
- 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)
- Saggion, H.: Summa: A robust and adaptable summarization tool. In: Traitement Automatique des Langues. vol. 49.2 (2008)
- Teufel, S., Moens, M.: Summarizing scientific articles: experiments with relevance and rhetorical status. Computational linguistics 28(4), 409

  –445 (2002)
- 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)

**CAPTION** 

BIBLIOGRAPHIC ENTRY

<sup>1</sup> http://www.ncbi.nlm.nih.gov/pubmed

http://www.scopus.com

<sup>3</sup> http://www.webofknowledge.com

## 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
- **...**



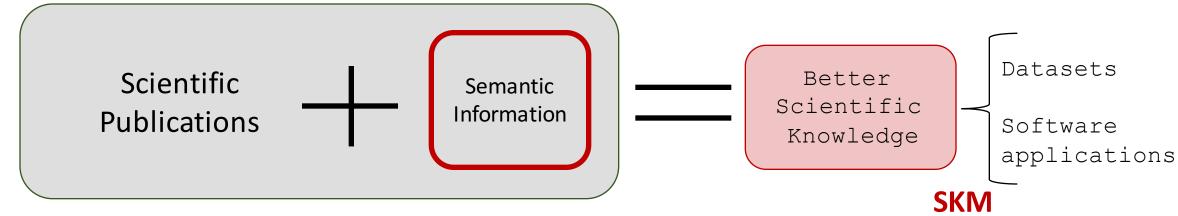
Facilitate the extraction of knowledge from scientific publications across many disciplines.

Improve a variety of use cases such as:

- Citation Characterization
- Citation Recommendation
- Summarization

**–** ...

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





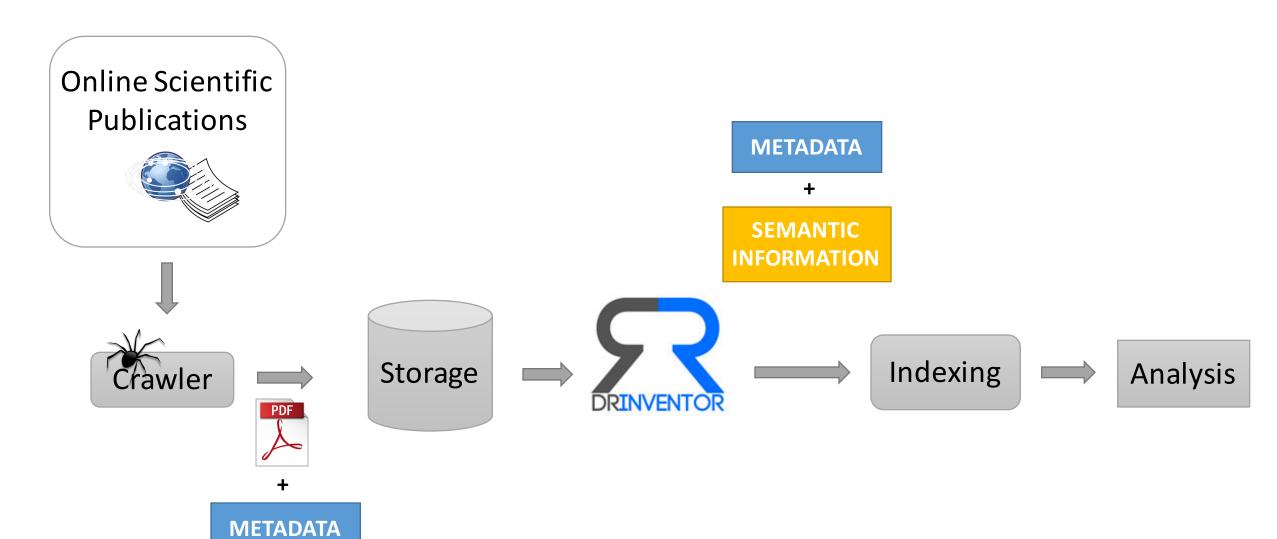


Most Cited: Documents, Citations, Authors, Venue Impact Rating

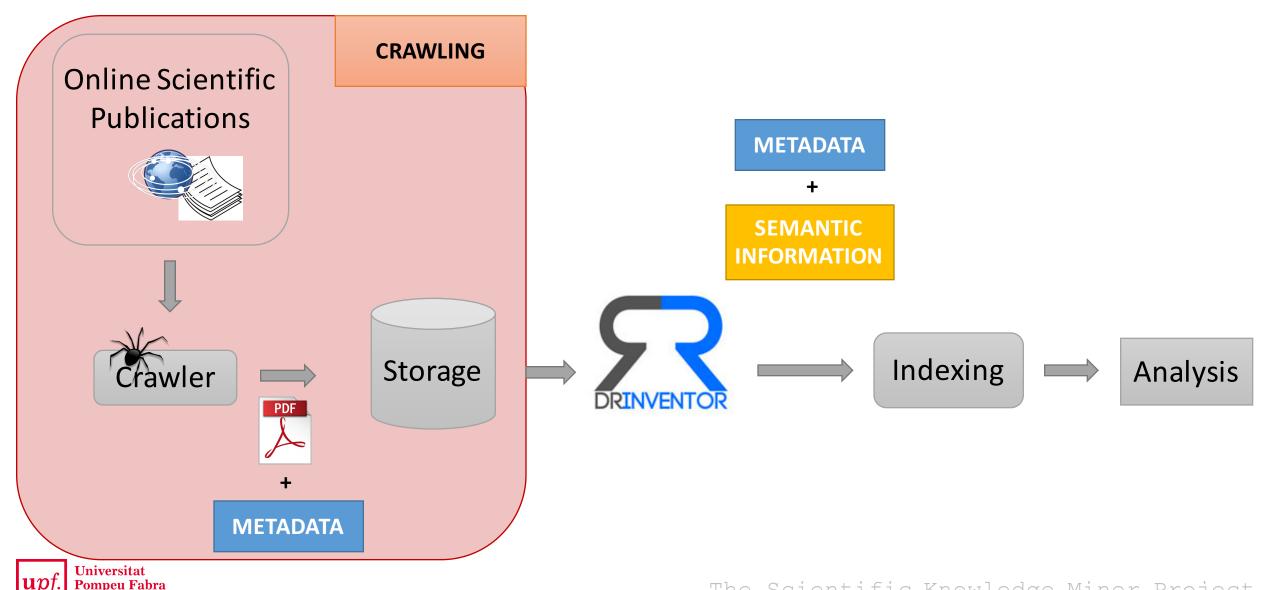
The SKM approach to the analysis of scientific literature:

- 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.









Barcelona

## Crawling



Search the Anthology 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.

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

MEW 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: **PROFITED** 16 15 14 13

ACL: <u>Intro</u> 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<sup>\*</sup>

EACL: <u>Intro</u> 83 85 87 89 91 93 95 97 99 03 06 09 12 14

NAACL: <u>Intro</u> 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

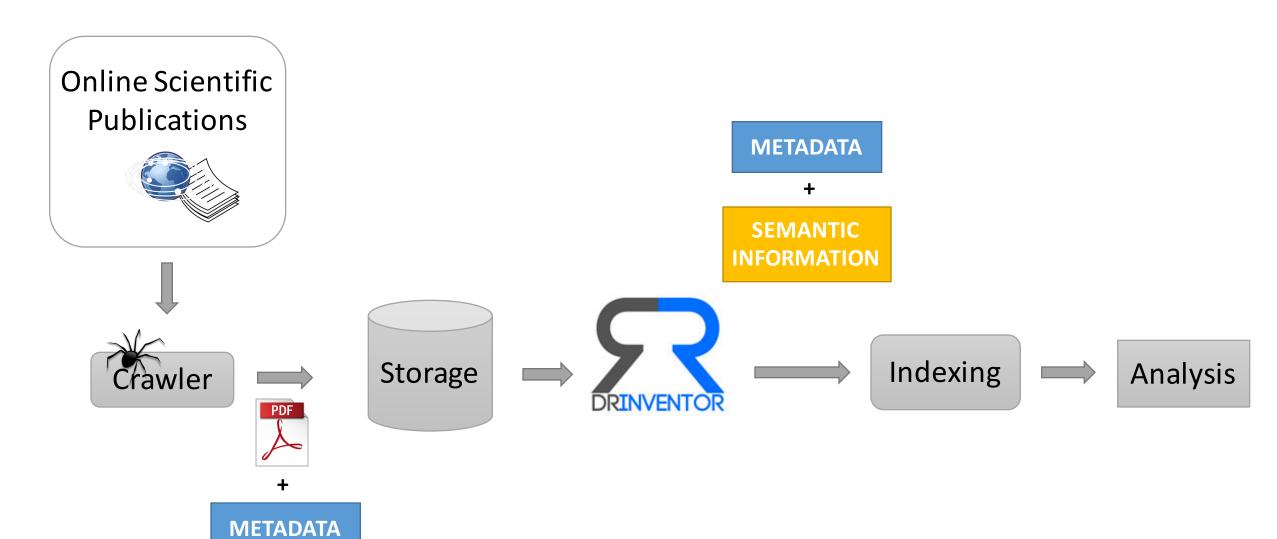
\*Sem/ 98 01 04 07 10 12 13 14 15 SemEval:

**ANLP:** Intro 83 88 92 94 97 00\*

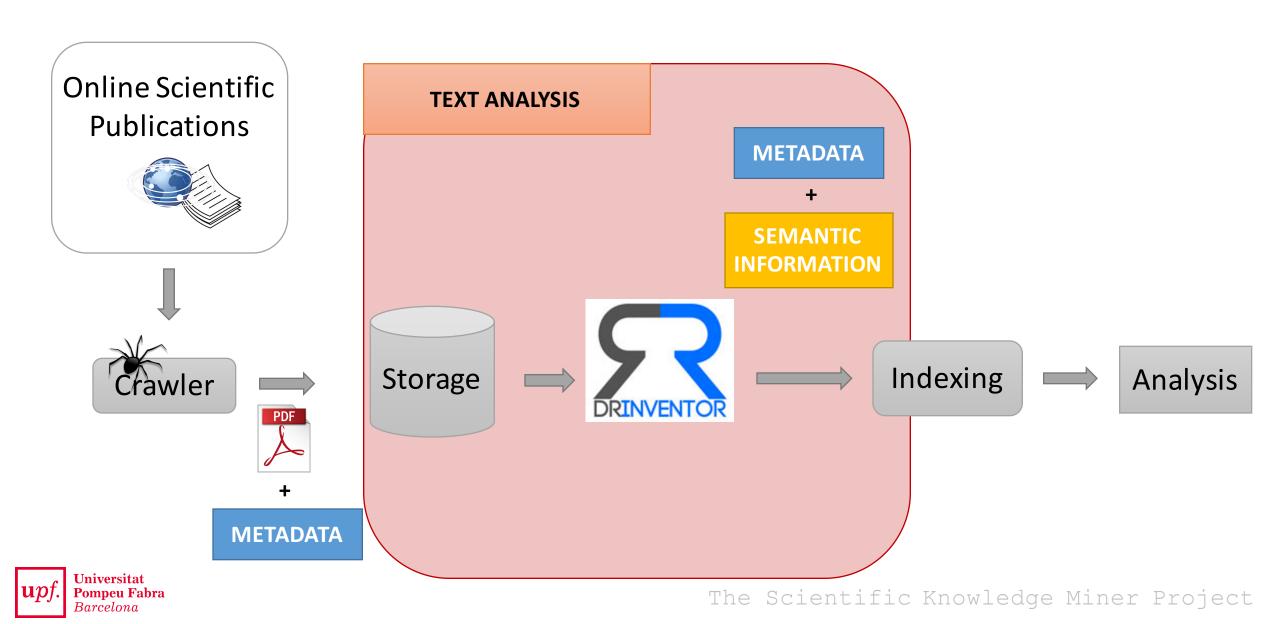


Data Base









## 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 library managed by **Mayen** 

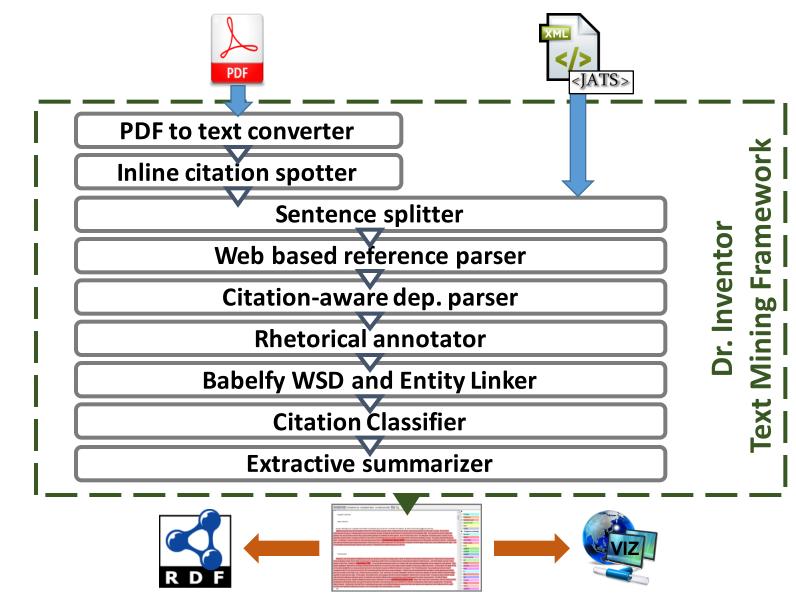
- Focused on textual content
- Relying on a shared data model (java classes) to represent a paper
- Exposing a convenient API to access the mined information
- Based on GA



to manage textual annotations

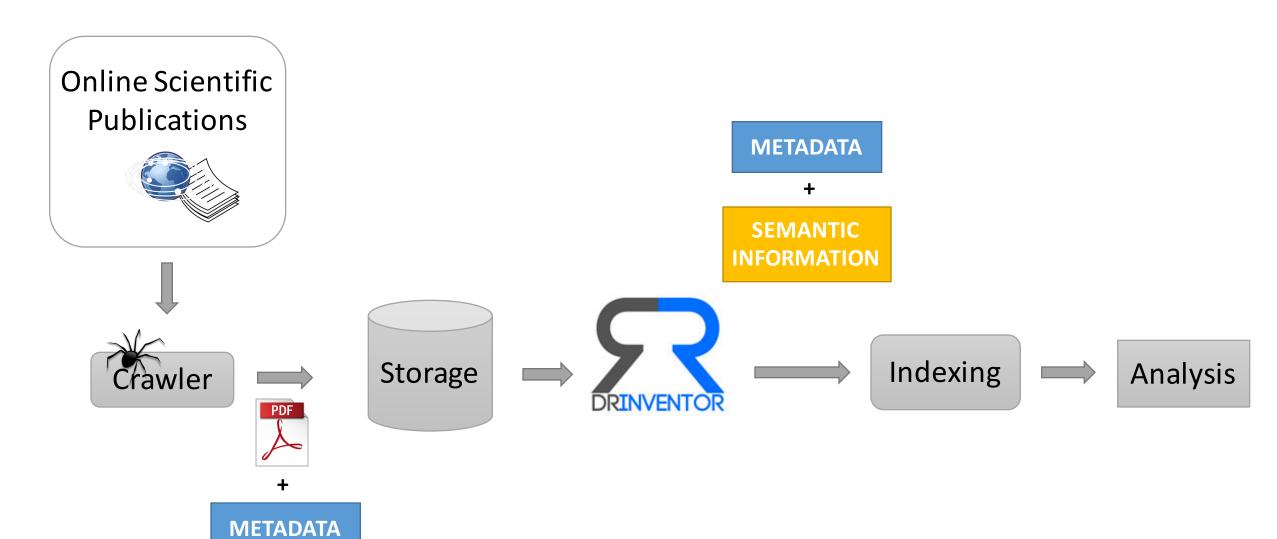


## Dr. Inventor Text Mining Framework

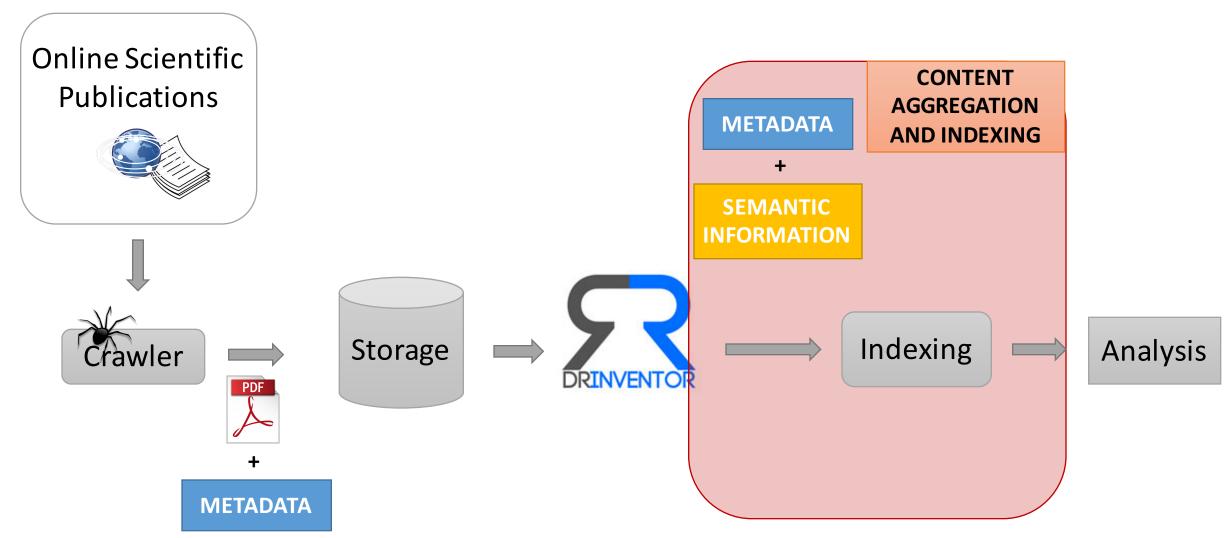




DRINVENTOR FRAMEWORK





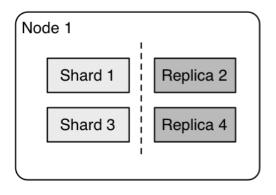


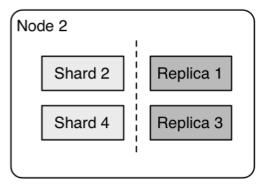


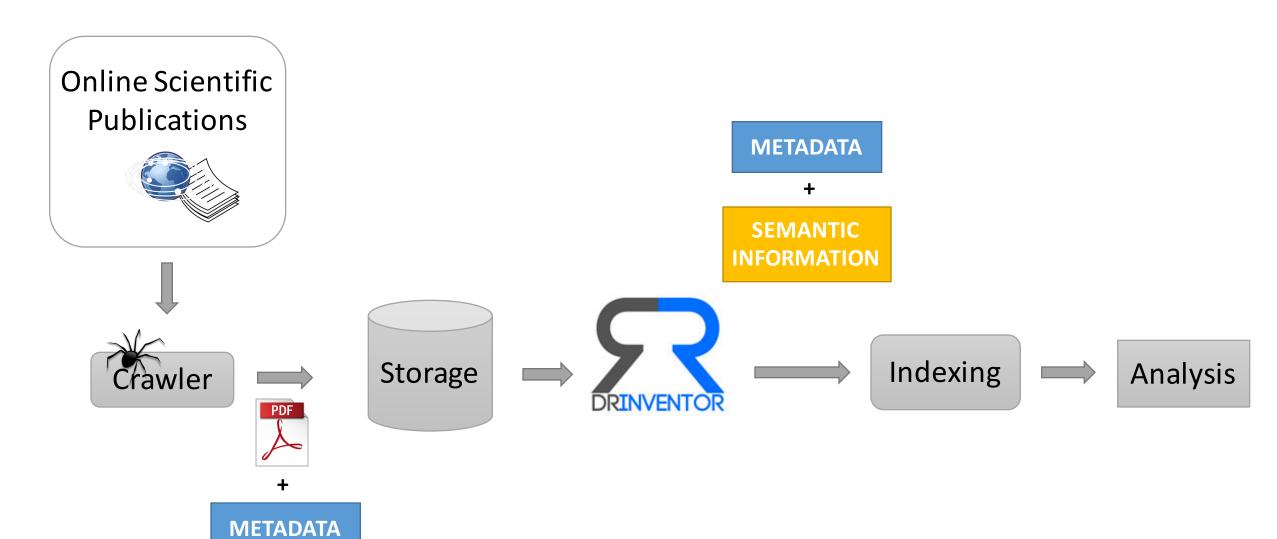
# Indexing



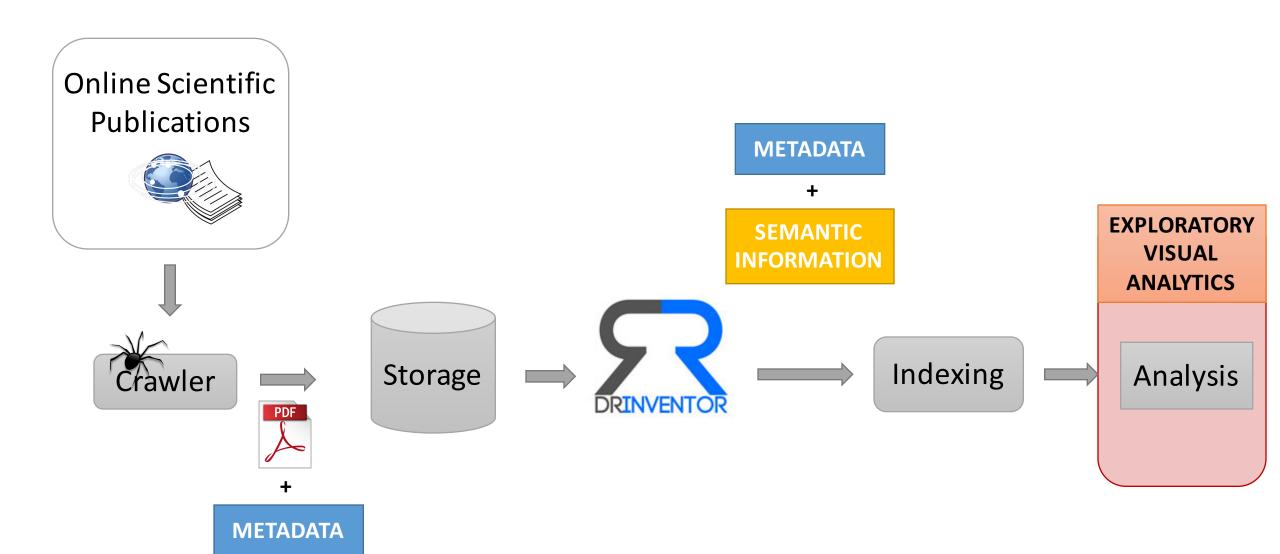
### ElasticSearch Cluster













## Analysis

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

#### Citations Sentence Graph Abstract graph Babel senses cloud Main view - click to collapse all menu tiems Modern statistical machine translation (SMT) systems usually use a linear combination of features to model the quality of each constrains that each feature interacts with the rest features in an linear manner, which might limit the expressive power of the r Table of the quality of translation hypotheses based on neural networks, which allows more complex interaction between features. A l contents designing the network structure which may improve the non-linear learning performance. Experimental results show that with the that are better than a linear model. **Bibliography** 1 Introduction One of the core problems in the research of statistical machine translation is the modeling of translation hypotheses. Each modeli where each e i is the ith target word and f j is the jth source word. The well-known modeling method starts from the Sour Rhetorical class translation model and a language model. The modeling method is extended to log-linear models by Och and Ney (2002), as sho Och and Ney (2002) ) Because the normalization term in Equation 2 is the same for all translation hypotheses of the same sourc shown in Equation 3. The log-linear models are flexible to incorporate new features and show significant advantage over the tradi **Summary** translation settings ( Yamada and Knight, 2001; Koehn et al., 2003; Chiang, 2005; Liu et al., 2006). ( cits: Yamada and Knight sentence try to separate good and bad translation hypotheses using a linear hyper-plane. However, complex interactions between feature cits: Clark et al., 2014 ) Taking common features in a typical phrasebased ( Koehn et al., 2003 ) or hierarchical phrasebased hypotheses; the word penalty feature encourages longer hypotheses. ( cits: Koehn et al., 2003, Chiang, 2005) The phrase sometimes is long with a lower translation probability, as in translating named entities or idioms; sometimes is short but with choice of translations. Simply use the weighted sum of their values may not be the best choice for modeling translations. As a retranslation quality. In this paper, we propose a non-linear modeling of translation hypotheses based on neural networks. The ti features to nodes in a hidden layer, complex interactions among features are modeled, resulting in much stronger expressive p issues to be tackled. The first issue is the parameter learning. Log-linear models rely on minimum error rate training (MERT) ( ( the intersection points of these non-linear functions could not be effectively calculated and enumerated. Thus MERT is no longe including several criteria to transform the training problem into a binary classification task, a unified objective function and an neural networks are equivalent to linear models; two-layer networks with sufficient nodes are capable of learning any continuous functions with less nodes, but also brings the problem of vanishing gradient ( Erhan et al., 2009 ). ( cits: Erhan et al., 2009 ) W major problem that prevents a neural network training reaching a good solution is that there are too many local minimums in the observations of the features. (Section 5) Experiments are conducted to compare various settings and verify the effectiveness of translation quality even with the same traditional features as previous linear models. (Section 6) Many research has been attempting to bring nonlinearity into the training of SMT. These efforts could be roughly divided into th transformation or additional learning. For example, Maskey and Zhou (2012) use a deep belief network to learn representations et al. (2014) used discretization to transform realvalued dense features into a set of binary indicator features. ( cits: Clark et al. (2014) These work focus on the explicit representation of the features and usually employ extra learning procedure. Our proposition combination are performed implicitly during the training of the network and integrated with the optimization of translation qua most similar in spirit with our work. Duh and Kirchhoff (2008) used the boosting method to combine several results of MERT and an additive neural network which employed a two-layer neural network for embedding-based features. (cits: Liu et al. (2013)) these efforts, our proposed method takes a further step that it is integrated with iterative training, instead of re-ranking, and wo features/components into the log-linear learning framework. Neural network based models are trained as language models ( V translation models ( Auli et al., 2013; Devlin et al., 2014). ( cits: Vaswani et al., 2013, Auli and Gao, 2014, Gao et al., 2014 target sides of the translation (cits: Liu et al. (2013)) rules as local features. In this paper, we focus on enhancing the express

designed features. We believe additional improvement could be achieved by incorporating more features into our framework.

### Non-linear Learning for Statistical Machine Translation

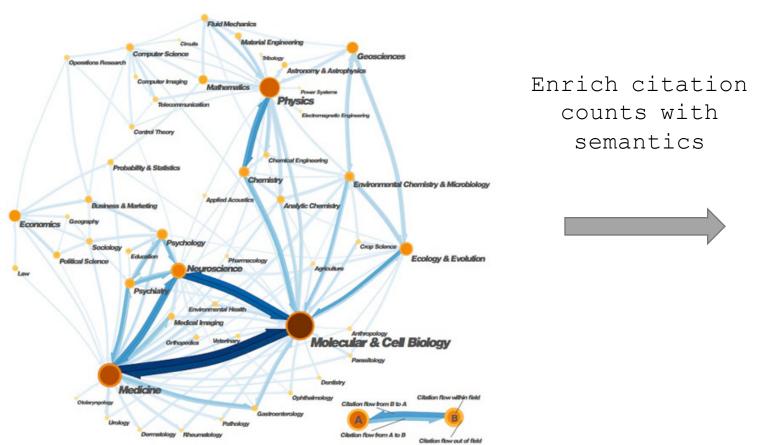
Huang Shujian - Chen Huadong - Dai Xinyu - Chen Jiajun (2015) (URL: <a href="http://arxiv.org/abs/1503.00107">http://arxiv.org/abs/1503.00107</a>)

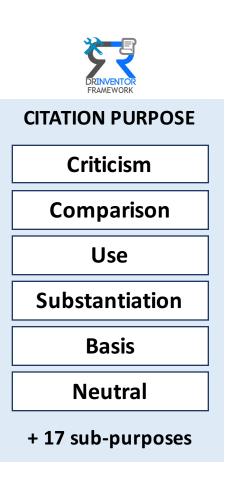




## Use Case 1: Citation Characterization

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







## 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



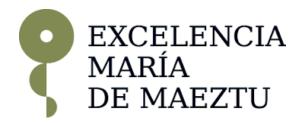
## Conclusions and future work

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:
  - o Alternative metrics to evaluate a paper based on citation semantics
  - o Semantically motivated recommendation of scientific publications
  - o Summarization of scientific literature



# Acknowledgements















Making Sense of Massive Amounts of Scientific Publications:

The Scientific Knowledge Miner Project

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

