TEXT AND DATA MINING & MACHINE LEARNING MODELS TO BUILD AN ASSISTED LITERATURE REVIEW WITH RELEVANT PAPERS

Ronald Brisebois[1], Alain Abran[1], Apollinaire Nadembega[2*], Philippe N'techobo[3]

[1] École technologie supérieure, Université du Québec, Canada

[2] Network Research Lab., University of Montreal, Canada

[3] École Polytechnique de Montréal, Canada

[*]Corresponding author E-mail: apollinaire.nadembega@umontreal.ca

Abstract.

In the process of literature review writing, researchers need to search and read several papers to find those which are relevant to their research. This paper proposes an assisted literature review prototype (STELLAR – Semantic Topics Ecosystem Learning-based Literature Assistant Review) based on (1) text and data mining models that learn from researchers' annotated data and semantic enriched metadata, (2) machine learning models (MLM) and (3) a semantic metadata ecosystem (SMESE) (i) to discover papers and recommend relevant of them for a specific topic using ranking algorithm and (ii) to identify papers according to researchers' selections parameters and his annotations. Notice that SMESE is our prototype that semantically harvests papers from different sources.

Specifically, STELLAR allows to:

1. Identify the relevant papers from SMESE thanks to the computation of a new ranking index (called, DTb Index) based on paper's semantic and contextual metadata such as discipline, topic, venue, authors in order to define the Literature Corpus of a specific topic or area of research.
2. Define the Literature Corpus Radius making use of value of the similarity between each paper and a specific research area, topic, title and description (called LCR Index).
3. Assist the researcher in refining the list of papers relevant for the literature review. To narrow down the search for relevant papers, many views and relationships of the list of candidate papers are made available.

Using various types of datasets and a simulation prototypes, the STELLAR performance was evaluated and compared to two existing approaches.

Keywords: assisted literature review, literature review, machine learning, literature review enrichment, semantic topic detection, text and data mining.

◆

1. INTRODUCTION

With the evolving, interdisciplinary and digital nature of research, there are more and more scientific publications; which increases enormously the volume of scientific papers. However, the huge volume of scientific publications available is becoming an issue for researchers (Boote & Beile, 2005; Mayr, Scharnhorst, Larsen, Schaer, & Mutschke, 2014): given that their time is limited, it is becoming impossible for researchers to read and carefully evaluate every publication within their own specialized field. Whether a short review as an assignment in a Master's program, or a LR for a PhD thesis, students find it difficult to produce a literature review (LR).

To obtain a manual LR, the researchers must dedicate to searching for literature will vary according to their research topic; which is very labor intensive. For instance, Gall et al. (Gall, Borg, & Gall, 1996) estimate that a decent LR for a dissertation takes three to six months to complete. Researchers also have to stay aware of newly published papers on related topics to produce a meaningful LR. In (Carlos & Thiago, 2015; Gulo, Rubio, Tabassum, & Prado, 2015), authors claim that an LR must address a research question and identify primary sources and references. An ideal LR should retrieve all relevant papers for inclusion and exclude all irrelevant papers (Carlos & Thiago, 2015; Gulo et al., 2015).

In the context of scientific research, the ranking algorithms for papers evaluation are referred to as scientometrics or bibliometrics (Beel et al., 2013; Bornmann, Stefaner, Anegón, & Mutz, 2014, 2015; Cataldi, Di Caro, & Schifanella, 2016; Dong, Johnson, & Chawla, 2016; Franceschini, Maisano, & Mastrogiacomo, 2015; Hasson, Lu, & Hassoon, 2014; Madani & Weber, 2016; Marx & Bornmann, 2016; MASIC & BEGIC, 2016; Packalen & Bhattacharya, 2015; Rúbio & Gulo, 2016; Wan & Liu, 2014; Wang et al., 2014; Zhang, Zhang, & Hu, 2015). According to literature, semantic metadata can be extracted from papers using text and data mining (TDM) algorithms while machine learning models (MLM) learn from papers and researchers' annotated papers in order to identify relevant papers for a specific topic and research field.

In this view, this paper proposes a new ecosystem prototype called STELLAR (Semantic Topics Ecosystem Learning-based Literature Assistant Review), that defines and builds an assisted literature review (ALR). The ALR is designed to reduce the load of searching and reading of papers by pointing the researcher to a recommended selection of documents. To do that, STELLAR computes the ranking index, called Dynamic Topic based Index (DTb Index) that evaluates the relevancy of each harvested paper. The DTb Index allows identifying the relevant papers for a specific research area, discipline, topic, title and description. To compute the DTb Index, STELLAR makes use of paper's contextual and semantic metadata related to (1) paper's venue, (2) paper's authors and their affiliation institutes, (3) paper's references and (4) paper's citations analysis. Specifically, STELLAR papers relevance ranking algorithm considers several papers' features such as venue age, type and impact, citations category and polarity, researchers' annotated data, authors' impact and their affiliation institute. To assist the researcher, STELLAR selects the papers from SMESE, ordered according to their relevance thanks to DTb Index, for the literature corpus definition that should be use to build the literature review. The selection process takes into account the researcher's (1) researcher discipline and language, (2) researcher main topic, (3) his research title and (4) his research description. Indeed, STELLAR computes the literature corpus radius index (LCR Index) that represents the similarity between researcher's selection parameters and each paper located in SMESE. To give a visual representation, this similarity is called radius where the center of circle is the researcher's selection parameters; more a paper matches with researcher's selection parameters, more its LCR Index tends to be equal to zero and more it gets closer to the center of the circle.

Notice that the prototype of STELLAR has been implemented using our software ecosystem described in SMESE (Brisebois, Abran, & Nadembega, Unpublished results) and SMESE V3 (Brisebois, Abran, Nadembega, & N'techobo, Unpublished results). SMESE allows controlling the access of the sources and harvesting scientific papers while SMESE V3 allows enriching the harvested papers metadata in term of topics.

The remainder of this paper is organized as follows. Section 2 presents some related work while Section 3 describes the proposed ecosystem (STELLAR) multi-platform architectural model. Section 4 describes STELLAR processes to compute DTb Index and LCR Index based on MLM and TDM concepts. Section 5 evaluates the STELLAR algorithms via simulations and shows the STELLAR prototype for LCR representation. Section 6 concludes this paper and introduces the future work.

2. RELATED WORKS

The related works analysis focuses on two research sub domain of scientific assisted literature review:

i. Machine learning models

ii. Ranking of scientific papers

MLMs are much exploited by scientific papers relevance ranking algorithms.

2.1. Machine learning models

To extract hidden knowledge form the scientific papers, literature recommends making use of text and data mining technique. Indeed, TDM is a sub domain of artificial intelligence (AI) which uses machine learning models to perform human tasks in terms of text analysis. A MLM explores the definition and study of algorithms that can learn from and make predictions on data. In the context of TDM, MLM is used mainly for document's metadata enrichment and literature review refinement in the assisted literature review (ALR) process. For example, in the scientific text summarization, two main MLM trends are identified:

i. Supervised systems that rely on MLM algorithms trained on pre-existing document- summary pairs.
ii. Unsupervised techniques based on properties and heuristics derived from the text. The unsupervised summarization methods (He et al., 2015) are mainly based on the weight of words in sentences, as well as the sentence position in a document.

Carlos and Thiago (Carlos & Thiago, 2015) developed a supervised MLM-based solution for text mining scientific articles using the R language in "Knowledge Extraction and Machine Learning" based on social network analysis, topic models and bipartite graph approaches. Indeed, they defined a bipartite graph between documents and topics that makes use of the Latent Dirichlet Allocation topic model.

2.2. Ranking of scientific papers

Two means of quantitatively evaluating scientific research output are discussed in the literature: peer-review and citation-based bibliometrics indicators. The main limitation of citations-based approaches have been criticized for having a scope limited to academia (Marx & Bornmann, 2016).

Citation analysis is widely used to measure impact of scientific papers. Scientific paper ranking should also depend on the venue, the location of publication, the year, the author and the citation index. Some works in the field of scientific impact evaluation (Bornmann et al., 2014, 2015; Cataldi et al., 2016; Zhang et al., 2015) address the ranking of universities, institutions and research teams. For instance, M. Zhang et al. (Zhang et al., 2015) propose a method to discover and rank collaborative research teams.

For this research, many existing approaches for scientific paper ranking have been evaluated (Bornmann et al., 2014, 2015; Gulo et al., 2015; Hasson et al., 2014; Madani & Weber, 2016; Marx & Bornmann, 2016; Rúbio & Gulo, 2016; Wan & Liu, 2014; Wang et al., 2014). They suffer from a number of limitations:

i. Most existing approaches focus on the researcher index or journal index to evaluate scientific research impact, ignoring the papers index.
ii. Most only use the citations count; do not consider the age of papers.
iii. Do not take into account the Social Level Metric, and the polarity of citations.
iv. They do not consider the other types of venues, such as conference proceedings, workshops or unpublished documents.
v. Several approaches make use of MLM but with large manual contribution.

A comparison of two approaches proposed in the literature for scientific paper ranking is presented in Table 1: PTRA (Hasson et al., 2014) and ID3 (Rúbio & Gulo, 2016).

i. PTRA: Hasson et al. (Hasson et al., 2014) propose a ranking algorithm, called Paper Time Ranking Algorithm (PTRA).
ii. ID3: Rúbio and Gulo (Rúbio & Gulo, 2016) propose recommending papers based on known models, including the paper's content and bibliometric features.

It can be seen from Table 1 that in ranking and identifying relevant contributions, neither of these two approaches takes into account author impact, citation category, venue impact, authors' institutes or citing documents (the six rightmost columns).

Table 1. The PTRA and ID3 approaches for ranking papers

Approaches	Year of publication	Citation number	Reference	Venue type	Venue age	Authors' impact	Citation category	Venue impact	Authors' institutes	Citing document of cited document
PTRA (Hasson et al., 2014)	X	X		X						
ID3 (Rúbio & Gulo, 2016)	X	X	X	X						

3. STELLAR MULTI-PLATFORM ARCHITECTURAL MODEL

In this section, an overview of the STELLAR (Semantic Topics Ecosystem Learning-based Literature Assisted Review) architectural model and its prototype based on SMESE is presented. The three main processes of STELLAR are:

i. Discovery ALR
ii. Search & Refine ALR
iii. Assist & Recommend ALR.

3.1. Workflows of assisted literature reviews

An ALR process, as illustrated in Fig. 1, should allow using MLM for automated activities. In addition, it alerts the researchers about new relevant papers, or related publications. Fig. 1 shows that STELLAR assists researchers to:

i. Discover or find relevant papers for his research topic,
ii. Search or refine his search parameters,
iii. Evaluate exiting cited papers.

In the rest of this section, the STELLAR prototype is described in more detail.

3.2. Overview of the STELLAR prototype of an assisted literature review

A LR has to be systematic: it should assess each paper to determine its ranking and whether or not it is worth including in the LR. One of the aims of an ALR is to reduce the reading load by enabling the researcher to read only relevant papers. The STELLAR prototype (see Fig. 2) uses as inputs:

i. A universal research document repository (URDR) that is made possible thanks to SMESE architecture.
ii. The enriched metadata of papers such as researchers' annotations.

STELLAR MLM algorithm learns from researchers' annotated papers and the URDR papers' metadata to recommend relevant papers for a specific research field and topic.

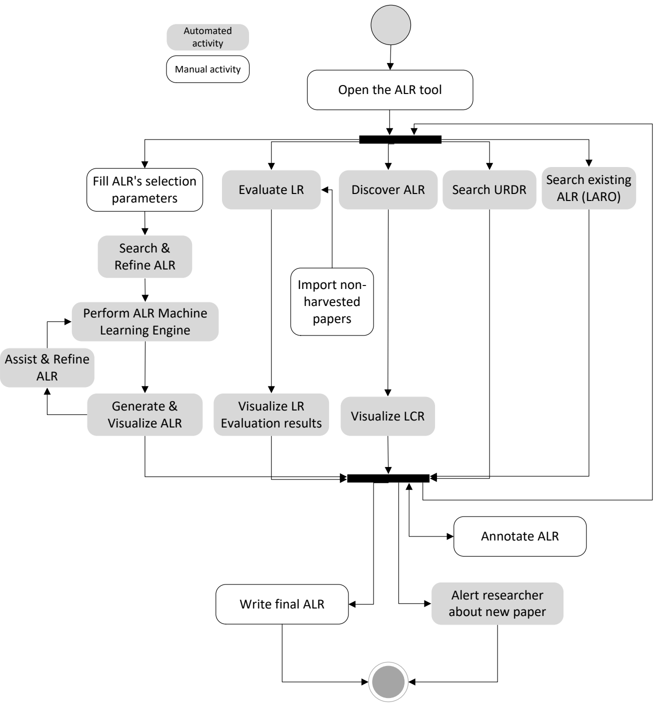


Fig. 1: Workflow of an assisted literature review

STELLAR first version prototype (STELLAR V1) architecture consists to four main parts as presented in Fig. 2:

A. Search & Refine ALR (Block A in the middle)

B. Assist & recommend ALR (Block B at the top-right)
C. Discover ALR Knowledge (Block C at the bottom)
D. Semantic Metadata Enrichments Software Ecosystem – SMESE V3; see (Brisebois, Abran, Nadembega, et al., Unpublished results) (top-left in Fig. 2 – see also Fig. 4).

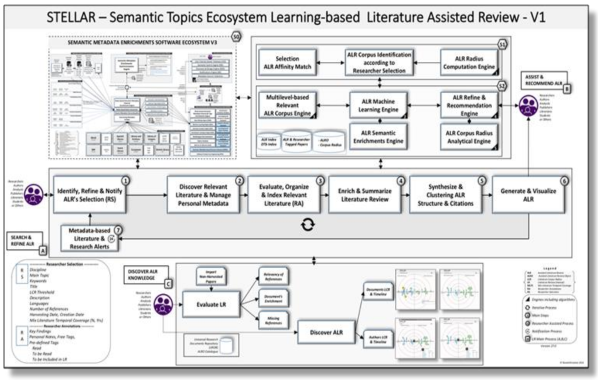

STELLAR – Semantic Topics Ecosystem Learning-based Literature Assisted Review - V1
SEMANTIC METADATA ENRICHMENTS SOFTWARE ECOSYSTEM V3
S0
S1
Selection ALR Affinity Match
ALR Corpus Identification according to Researcher Selection
ALR Radius Computation Engine
S2
Multilevel-based Relevant ALR Corpus Engine
ALR Machine Learning Engine
ALR Refine & Recommendation Engine
ALR Index DTb Index
ALR & Researcher Tagged Papers
ALRO Corpus Radius
ALR Semantic Enrichments Engine
ALR Corpus Radius Analytical Engine
ASSIST & RECOMMEND ALR
B
Identify, Refine & Notify ALR's Selection (RS)
1
Discover Relevant Literature & Manage Personal Metadata
2
Evaluate, Organize & Index Relevant Literature (RA)
3
Enrich & Summarize Literature Review
4
Synthesize & Clustering ALR Structure & Citations
5
Generate & Visualize ALR
6
Metadata-based Literature & Research Alerts
7
SEARCH & REFINE ALR
A
Researcher Selection
Discipline
Main Topic
Keywords
Title
LCR Threshold
Description
Languages
Number of References
Harvesting Date, Creation Date
Min Literature Temporal Coverage (%, Yrs)
Researcher Annotations
Key Findings
Personal Notes, Free Tags,
Pre-defined Tags
Read
To be Read
To be Included in LR
DISCOVER ALR KNOWLEDGE
C
Import Non-Harvested Papers
Evaluate LR
Relevancy of References
Document's Enrichment
Missing References
Discover ALR
Documents LCR & Timeline
Authors LCR & Timeline
Legend
Engines including algorithms
Iterative Process
Main Steps
Researcher Assisted Process
Notification Process
LR Main Process (A,B,C)

Fig. 2: STELLAR – Semantic Topics Ecosystem Learning-based Literature Assisted Review[1]

3.3. SEARCH & REFINE ALR

The Search & Refine ALR (block A in Appendix A) consists of seven steps. The first step, called **Identify, Refine & Notify ALR** allows identifying and refining the researcher selection (RS) metadata. These metadata are classified into two categories: Document Common Metadata section (top part of Table 2) and Researcher Annotations section (bottom part of Table 2). The second step is **Discover Relevant Literature & Manage Personal Metadata** that allows measuring the paper relevancy making use of the dynamic topic based index (DTb index); DTb index is computed making used of TDM approach. The third step, called **Evaluate, Organize & Index the Relevant Literature**, allows selecting the relevant papers that matches with the researcher requirement for his ALR. In contrast to Literature Corpus which denotes all the papers of a specific research topic, the ALR Corpus denotes only the papers of a Literature Corpus which meets RS metadata for an ALR. The next step, called **Enrich & Summarize the Literature Review** makes use of TDM and MLM approaches: to extract papers' subject, to detect papers' citation category en polarity, to extract papers' citation text and to performed abstract conformity. All these enrichments form the enriched metadata of paper that may be used to provide accurate summarization. **Synthesize & Clusterize the ALR Structure & Citations** step aims to synthesize and organize the relevant documents into clusters related to the LCR index while **Generate & Visualize the ALR** step aims to generate and visualize recommended papers in the Literature Corpus. Finally, **Metadata-based Literature & Research Alerts** allows detecting new relevant papers or new metadata related to the ALR.

Table 2. Researcher selection (RS) metadata

Number	Metadata	Description
	A. Document Common Metadata	
1	Discipline	Selection of the discipline related to the ALR
2	Main Topic	The main topic is one of the most important metadata for building the ALR. It should be as specific as possible.
3	Literature Corpus Radius	It is the main concept that makes it possible to refine the selection of research documents to be included in the ALR.
4	Keywords	The researcher has to identify keywords representative of the ALR.
5	Harvesting Date	Date of document harvesting
6	Creation Date	Date of document creation

[1] See Appendix A for a more readable version of Fig. 2.

7	Title	Title of the ALR
8	MLTC - Mix of Literature Temp. Cov. (Yrs, %)	MLTC is crucial to building and refining the ALR. It has two indicators: 1 - Number of years covered by the search 2 - Percentage of documents outside time-range to be included in ALR.
9	Description	A brief description of the research project of the ALR
10	Languages	The researcher has to choose the language of the papers.
11	Nb of References	The number of references that the ALR should consider.
	B. Researcher Annotations Metadata	
12	Key Findings	The Key Findings are annotations regarding important findings in the document identified by the researcher.
13	Free Tags	The researcher may place tags on a document in order to remember some information about it.
14	Personal Notes	The researcher may attach notes to a document in order to remember relative information. These notes can be used by STELLAR or the researcher to help specify the ALR.
15	Pre-defined Tags	These are predefined metadata to help the researcher. Examples: Read, To be read, To be included in the ALR

3.4. ASSIST & RECOMMEND ALR

Assist & recommend ALR (block B in Appendix A) represents the STELLR core that allows refining the ALR through two sets of steps (S1 and S2). It consists of the STELLAR MLM engines (engine 1 to 5) designed to identify a specific corpus, evaluate papers relevancy or define learning models. The Literature Corpus contains all the papers regardless of their LCR index and the type of selection metadata (i.e., RSs or RAs). The papers within corpus radius are those located at the surface (forming a disc) of a circle with the specific corpus radius – see Fig. 3.

Based on the definitions above, the Corpus Radius may be defined as the delimiter of the Literature Corpus suggested to the researcher for the ALR on the basis of the researcher's selections and annotations. The RS selection criteria are the researcher's metadata while the RA selection criteria consist of notes, tags and key findings.

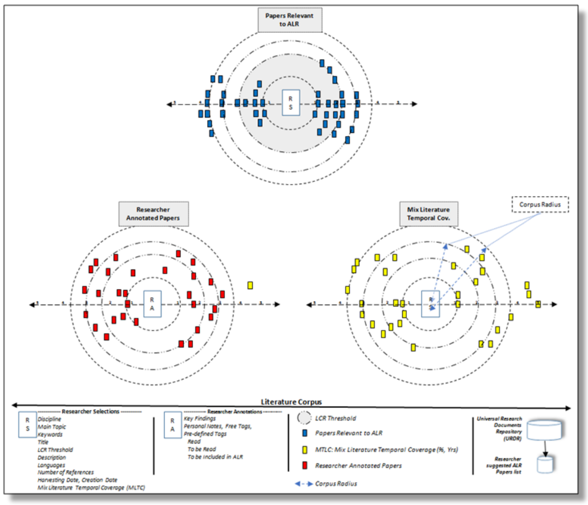


Fig. 3: Sources used to build the suggested list of ALR papers[2]

To illustrate, consider the papers in the corpus radius called ***"Papers relevant to ALR"*** (disk with blue dots at the top of Fig. 3): all the papers within the

[2] See Appendix B for a more readable version of Fig. 3

gray disc are whose LCR index is less than or equal to 2; in this case, the LCR threshold is set at 2.

3.5. Discover ALR Knowledge

The 'Discover ALR Knowledge' (Block C in Appendix A) has two main features. First, it allows unveiling the content of the ALR, discovering the papers harvested by SMESE and to explore the metadata generated by STELLAR MLM algorithms. Secondly, it analyses the references of manual LR in order to evaluate their relevance according to the research topic.

More specifically, the first feature "Evaluate LR" consists in an assisted evaluation of an already published LR. To evaluate an existing LR, this feature compares the existing LR to the one from STELLAR's MLM to quantify their similarity.

The tags created by the researchers are used to enrich the ALR metadata. The process 'Discover ALR Knowledge' makes it possible to drill down through different types of visualization of the corpus.

3.6. Semantic Metadata Enrichments Software Ecosystem SMESE V3

The SMESE V3 platform presented in Fig. 4 (Brisebois, Abran, Nadembega, et al., Unpublished results) is our semantic metadata enrichment software ecosystem for metadata aggregation and enrichment in order to create a semantic master metadata catalogue (SMMC). Notice that SMESE V3 includes SMESE V1 features; SMESE V3 checks continuously the access to the sources of scientific papers and analyses the data structures in order to adapt the harvesting algorithms. SMESE V3 also analyses the papers texts taking into account the documents organization and extracts the paper's research topics.

The SMESE V3 platform allows enrichment from different sources including linked open data. SMESE V3 is used by STELLAR to build its URDR (its base repository of harvested available papers at a given time *t*).

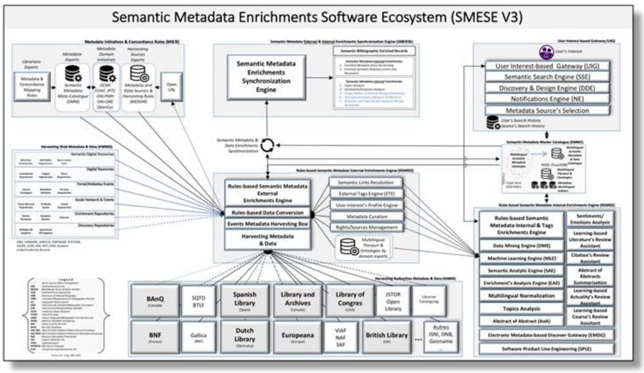

Semantic Metadata Enrichments Software Ecosystem (SMESE V3)
Metadata Initiatives & Concordance Rules (MICR)
Librarians Experts
Metadata & Concordance Mapping Rules
Metadata Experts
Semantic Metadata Meta-Catalogue (SMM)
Metadata Domain Initiatives
DCMI FOAF, IPTC OAI-PMH OAI-ORE OpenCyc
Harvesting Sources Experts
Metadata and Data Sources & Harvesting Rules (MDSHR)
Open URL
Semantic Metadata External & Internal Enrichments Synchronization Engine (SMEIESE)
Semantic Metadata Enrichments Synchronization Engine
Semantic Bibliographic Enriched Records
User Interest-based Gateway (UIG)
User's Interest
User Interest-based Gateway (UIG)
Semantic Search Engine (SSE)
Discovery & Design Engine (DDE)
Notifications Engine (NE)
Metadata Source's Selection
User's Search History
Source's Search History
Semantic Metadata & Data Enrichments Synchronization
Harvesting Web Metadata & Data (HWMD)
Semantic Digital Resources
Digital Resources
Portal/Websites Events
Social Network & Events
Enrichment Repositories
Discovery Repositories
Semantic Metadata Master Catalogue (SMMC)
Rules-based Semantic Metadata External Enrichments Engine (RSMEEE)
Rules-based Semantic Metadata External Enrichments Engine
Rules-based Data Conversion
Events Metadata Harvesting Box
Harvesting Metadata & Data
Semantic Links Resolution
External Tags Engine (ETE)
User Interest's Profile Engine
Metadata Curation
Rights/Sources Management
Multilingual Thesauri & Ontologies by domain experts
Rules-based Semantic Metadata Internal Enrichments Engine (RSMIEE)
Rules-based Semantic Metadata Internal & Tags Enrichments Engine
Data Mining Engine (DME)
Machine Learning Engine (MLE)
Semantic Analytic Engine (SAE)
Enrichment's Analysis Engine (EAE)
Multilingual Normalization
Topics Analysis
Abstract of Abstract (AoA)
Sentiments/ Emotions Analysis
Learning-based Literature's Review Assistant
Citation's Review Assistant
Abstract of Abstracts Summarization
Learning-based Actuality's Review Assistant
Learning-based Course's Review Assistant
Electronic Metadata-based Discover Gateway (EMDG)
Software Product Line Engineering (SPLE)
Harvesting Authorities Metadata & Data (HAMD)
Legend
BAnQ (Canada)
SQTD BTLF
Spanish Library (Spain)
Library and Archives (Canada)
Library of Congres (USA)
JSTOR Open Library
Librarian Cataloguing
BNF (France)
Gallica (BNF)
Dutch Library (Germany)
Europeana (Europe)
VIAF NAF SAF
British Library (UK)
...
Autres ISNI, DNB, Geoname

Fig. 4 SMESE V3 - Semantic Metadata Enrichments Software Ecosystem[3]

4. *STELLAR PROCESSES DESCRIPTION*

In this section, the MLM approach used by STELLAR to define its core of processes is presented. The core of STELLAR processes consists of five engines located in the bloc B (S1 and S2) of the architectural model of STELLAR. Fig. 5 shows these five engines of the core of STELLAR processes and the interaction between them to assist researchers for their ALR corpus selection. From now in this paper, the following terms are used interchangeably: document, paper and scientific paper.

Each one of these five core engines for STELLAR processes is described in detail in the following sub-sections. Indeed, using as inputs the URDR that contains existing papers, researcher annotations (RAs) and researcher selection (RS), the ***ALR radius computation engine (engine #1)*** computes the LCR index. Next, using as inputs the ALR Corpus and the training models built by researchers, ***ALR Machine Learning engine (engine #2)*** provides the ALR learning model used by the ***Multilevel-based Relevant ALR Corpus (engine #3)***. Indeed, when a new paper is harvests by SMESE, the Multilevel-based Relevant ALR Corpus of STELLAR computes the DTb Index that measures the relevancy of this paper and saves this DTb Index as new enriched metadata of the paper. The ***ALR Refine & Recommendation engine (engine #4)*** suggests the ALR references list to the researchers and assists them

[3] See Appendix C for a more readable version of Fig. 4

to refine this list while the ***ALR Corpus Radius Analytical engine (engine #5)*** builds dynamic graphical representations of the quantitative and qualitative metadata about selected ALR corpus.

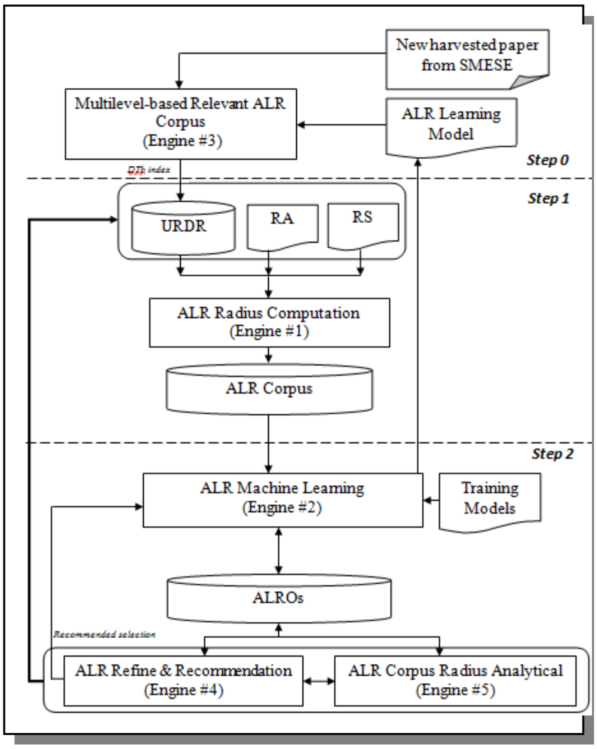

Newharvested paper from SMESE
Multilevel-based Relevant ALR Corpus (Engine #3)
ALR Learning Model
Step 0
DTb index
Step 1
URDR
RA
RS
ALR Radius Computation (Engine #1)
ALR Corpus
Step 2
ALR Machine Learning (Engine #2)
Training Models
ALROs
Recommended selection
ALR Refine & Recommendation (Engine #4)
ALR Corpus Radius Analytical (Engine #5)

Fig. 5: Interoperability of the core engines of STELLAR processes

In the rest of the section, we focus on the first four engines.

4.1. Multilevel-based relevant ALR Corpus

The multilevel-based relevant ALR Corpus (in Step 0 and 2) is presented here. It is used to evaluate the relevancy of a paper based on a number of scientometric measurements. The measurement of relevance is referred as the ALR Index. Three types of ALR Index are defined in STELLAR: personal, collaborative and dynamic topic-based (DTb). With the personal index, the ALR corpus can be restricted to documents tagged by the researcher as “To be included in the ALR” while collaborative index restricts the ALR corpus to the documents tagged as “To be included in the ALR” by the others researchers who are selected by researcher who requests the ALR corpus. The dynamic topic-based index (DTb index) selects documents for the ALR corpus when the researcher has not requested a personal or collaborative index. The DTb index is a weighted sum of the values that denote the importance of the different inputs considered.

4.2. ALR radius computation

ALR radius computation is used to select the relevant papers to be included in the ALR, according to the researcher selection (RS) and researcher annotations (RAs). The main factor of the ALR radius computation is the LCR Index. LCR index computation is defined as a sub-algorithm of the semantic ALR selection search that identifies the ALR corpus according to the RS and RAs defined in Fig. 5; in other word, LCR Index measures the similarity between a paper, considering its text and its metadata, and the RS and RAs parameters. To identify an ALR corpus as shown in the Step 1 of Fig. 5, the selection parameters (RA and RS) are classified into three categories (see Table 3).

In the following Fig. 6, the ALR selection search using the three categories of selection parameters is explained in detail.

Table 3. STELLAR classification of researcher’s selection (RS) and annotations (RAs) parameters

Evaluation-based	Selection-based	Sort-based
Main Topic	Discipline	Literature Corpus Radius (LCR)
Keywords	Languages	Mix of the Literature Temporal Coverage (MLTC)
Title	Document Researcher Annotations	Number of References
Description		

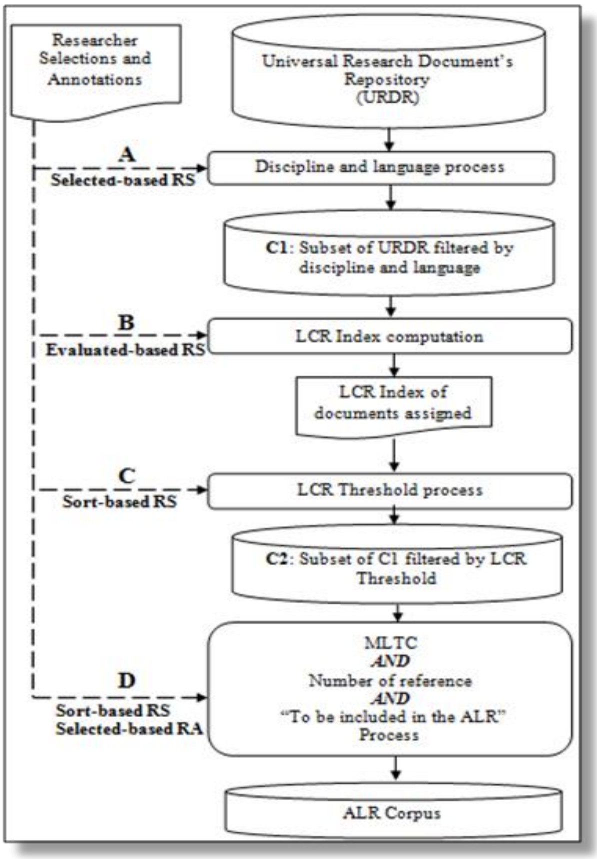

Researcher Selections and Annotations
Universal Research Document's Repository (URDR)
A
Selected-based RS
Discipline and language process
C1: Subset of URDR filtered by discipline and language
B
Evaluated-based RS
LCR Index computation
LCR Index of documents assigned
C
Sort-based RS
LCR Threshold process
C2: Subset of C1 filtered by LCR Threshold
D
Sort-based RS
Selected-based RA
MLTC
AND
Number of reference
AND
"To be included in the ALR"
Process
ALR Corpus

Fig. 6: Steps in a semantic ALR corpus selection search

A. Discipline and language researcher selections step

In step A in Fig. 6, volume of documents to be considered may be reduced, based on: discipline selection and language selection.

Let *DC* be the chosen discipline, let *LG* be the given language, let DISCIPLINE be the metadata that records the discipline of the documents in URDR, let LANGUAGE be the metadata that records the language of the documents in URDR and let *DiscLan_Corpus(DC,LG)* be the set of documents in the language ***LG*** that are in the discipline ***DC***. *DiscLan_Corpus(DC,LG)* is obtained as follows:

DiscLan_Corpus(DC, LG) = *[**select in** URDR the Documents **where** DISCIPLINE is "DC" **and** LANGUAGE is "LG"]*

This query to the URDR extracts only papers in the specified discipline and language. Let C_1 be the corpus of papers obtained in step A.

B. LCR index computation step

Based on the set of papers selected in step A, the LCR index is computed in step B making use of the evaluation-based selections (see Table 3). The LCR index computation step consists of five sub-steps as follows:

i. Similarity matching of researcher main topic with topics extracted from document abstracts

This sub-step process, the topic detection ML model called BM-Scalable Annotation-based Topic Detection (BM-SATD) (Brisebois, Abran, Nadembega, et al., Unpublished results) is used. BM-SATD combines semantic relations between terms with co-occurrence relations across the document, by making use of the document annotations.

Here, the similarity matching is based on the n-gram approach where the value n is used as the weight (Bertin, Atanassova, Sugimoto, & Lariviere, 2016): when the i-gram expression in the researcher main topic parameter is found in the abstract, the weight *i* is associated with this expression.

Making use of the weight i_p of each paper *p* of the set C_1, the normalisation of i_p N(i_p) is performed in order that N(i_p) value be between 0 and 1. Let $\boldsymbol{MT_p}$ be the N(i_p) of the paper *p*.

ii. Similarity matching of researcher keywords with document keywords

The weight j_p of the similarity matching of the researcher keywords parameter associated to paper *p* is the number keywords of paper *p* that are found in the set of researcher keywords parameter.

Making use of the weight j_p of each paper *p* of the set C_1, the normalisation of j_p N(j_p) is performed in order that N(j_p) value be between 0 and 1. Let $\boldsymbol{K_p}$ be the N(j_p) of the paper *p*.

iii. Similarity matching of researcher title with document titles

The researcher title and papers titles are pre-processed to filter noise. This consists in stemming, phrase extraction, part-of-speech filtering and removal of stop-words. Next, based on the terms obtained, the maximum n-gram of the researcher title which is met in the paper *p* title is used as the title selection impact value k_p.

Making use of the value k_p of each paper *p* of the set C_1, the normalisation of k_p N(k_p) is performed in order that N(k_p) value be between 0 and 1. Let $\boldsymbol{T_p}$ be the N(k_p) of the paper *p*.

iv. Similarity matching of researcher research topic description with document abstracts

The value l_p of The similarity matching of researcher research topic description is semantically compared with the paper *p* abstract using WordNet::Similarity (Pedersen, Patwardhan, & Michelizzi, 2004).

Making use of the value l_p of each paper *p* of the set C_1, the normalisation of l_p N(l_p) is performed in order that N(l_p) value be between 0 and 1. Let $\boldsymbol{D_p}$ be the N(l_p) of the paper *p*.

v. LCR index computation

Finally, when the similarity matching of each evaluation-based selection has been completed through sub-steps 1 to 4, the LCR index within the [0,1] range can be computed. Note that the LCR index is a weighted sum of the computed value of each evaluation-based selection. The difference in weight between two consecutive evaluation-based selections (i.e., α_i and α_{i+1}) is a predefined constant value.

$$LCR \quad Index(p) = \frac{(\alpha_1 \times MT_p) + (\alpha_2 \times K_p) + (\alpha_3 \times T_p) + (\alpha_4 \times D_p)}{\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4} \qquad (1)$$

vi. Literature Corpus Radius (LCR) threshold selection step

In this step, a set of documents is sorted or selected according LCR index value. For example, a researcher may indicate that the LCR threshold is 0.7; the output will then be a subset of corpus C whose LCR index is greater than or equal to 0.7. Let C_2 be the corpus of documents obtained in step C.

vii. MLTC AND Number of references AND "To be included in the ALR" step

MLTC is the Mix Literature Temporal Coverage. Let MLTC (x, y) with its number of selections equal N: this means the researcher expects to have at most ***N*** documents, with a maximum of ***(100-x)%*** (i.e., $\frac{N}{100} \times (100 - x)$) that are at most ***y*** years old, and including all the documents tagged "To be included in the ALR". Note that the latter documents have priority.

First, a list (in descending order) is created based on the LCR index applied to corpus C_1 where the documents tagged "To be included in the ALR" are at the top due to their priority.

Let All_C_1 be this list. New_C_1 is defined as a sub-list of C_1 in which the document age is less than or equal to *y*, and Old_C_1 contains documents older than *y*.

Let $A = \frac{N}{100} \times x$ be the length of New_C_1 and $B = \frac{N}{100} \times (100 - x)$ be the length of Old_C_1. To take into account the three selections made in sub-step D.

Note that, when the number of documents in All_C_1 is less than N, all the documents are considered affinity matches for the ALR; in that case, the MLTC selection is ignored.

However, when there are not enough documents whose age is less than or equal to *y* to satisfy the

MLTC selection, a new MLTC is provided in order to reach the number *A*. But if the researcher requires the MLTC selection to be met, some documents are removed from New_C$_1$ in order to meet the selected MLTC(x, y).

If an "OR" has been placed between the researcher selections, the LR corpus will be defined as the union of the C2 subsets provided by the MLTC process, the Number of references process and the "To be included in the ALR" tags.

4.3. ALR Machine Learning

ALR Machine Learning (Step 2 of Fig. 5) for semantic ALR selection is the main process of STELLAR. It is a supervised MLM that makes use of a training set in order to provide the learning model.

For the rest of this sub-section, cited document denotes the paper cited by another paper while the citing document denotes the paper citing another paper.

4.3.1. Section recognition learning model

The section recognition learning model in STELLAR allows to identify each section of a paper in order to know the section of each sentence. Indeed, knowing the section in which a sentence appears may change its context. For example, citations in the 'Related Work' section do not carry the same weight as those in the 'Discussion' section in terms of identifying existing papers in a specific domain. To perform automatic section detection, manual training model is used.

4.3.2. Citations-based learning model

A citations-based learning model has been designed to identify and extract citations in documents. This learning model is divided as follows (see Table 4):

A. Citation style learning model based on citation style

B. Citation classification learning model based on rhetorical categories, cue phrases.

A cue phrase is the phrase that often occurs in a certain rhetorical category. In the case of citation classification, the verb plays the main role. Researchers are asked to read and detect the cue phrases associated with each citation polarity and category; this makes it possible to build a training model of cue phrases and their classifications, which is integrated into the "Training Model".

Table 4: Citations-based learning model

A. Citation style learning model	
Style marker	**Description**
Numerical	The syntax of this citation style is the number between brackets.
Textual	This citation style: (<names of authors>, year) or < names of authors > (year).
Personalization	This style is based on the set of texts that refer to cited papers.
B. Citation classification model	
Citation category	**Description**
Relevant	According to the citing document, the cited document is relevant.
Problem	The cited document presents the issues that led to the research.
Uses	The cited document proposes a solution that is used in the citing document.
Extension	The cited document proposes a solution that is extended by the citing document.
Comparison	The cited document proposes a solution that is compared with the citing document solution in terms of performance.

Next, based on semantic similarities, any rhetorical category that was not detected manually is detected automatically and added to the model. The polarity model is proposed in order to indicate whether the citation is positive or negative.

4.3.3. Text-based learning model

To define the text-based learning model, text categories have been predefined as follows: problem, solution and results. As in the citation-based learning model, rhetorical expressions are detected by means of cue phrases. The text-based learning model is organized as follows:

1. The cue phrase learning model containing a list of cue phrases (CPs): problem CP, solution CP and result CP.
2. The thematic learning model (TRs):

a. Problem learning model: list of problem rhetorical expressions (P_TR)

b. Solution learning model: list of solution rhetorical expressions (S_TR)

c. Result learning model: list of result rhetorical expressions (R_TR).

4.4. ALR Refine & Recommendation MLM

Making use of the relevant and enriched papers identified automatically by STELLAR and contained into ALR Corpus according to the RS and RAs, the recommended selections parameters are provided to a researcher. This MLM engine recommends three different aspects of the ALR selection as shown in Fig. 7.

In other word, this engine suggests new RS parameters to the researchers in order to maximum the relevant papers for his ALR.

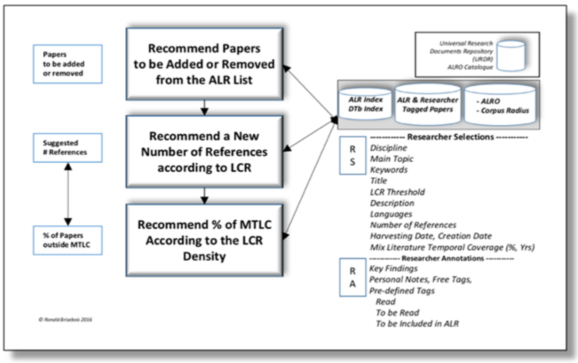


Fig. 7: Refinement & Recommendation MLM[4]

5. STELLAR PERFORMANCE EVALUATION THROUGH SIMULATIONS

This section presents an evaluation of the performance of the STELLAR prototype through a number of simulations to the identification and ranking of relevant papers.

5.1. Datasets

Two datasets were used for the simulations:

i. A dataset harvested from databases

ii. A baseline dataset.

5.1.1. Dataset harvested from databases

For the simulations, 2,000 scientific papers were collected from databases such as Science Direct and Scopus. The papers dealt with various research topics in Computer Science. Two sub-domains were chosen, each with 1,000 papers: (1) Artificial Intelligence, and (2) Information Systems. For these simulations, the sub-domains are treated as domains. The other metadata were collected as bibliographic references.

For each paper, the downloaded bibliographic files were parsed to extract the metadata and were input into the SMESE V3 platform with the paper itself. Here, a scenario was defined as a set of two simulator runs, one on each domain dataset. For the simulator run parameters, the metadata of one paper in the dataset (discipline, language, title, topic, keywords and abstract) were used as the RS and RA parameters.

5.1.2. Baseline dataset

For the present study, we had already produced a manual ALR that is listed in the References section. The baseline dataset consisted of 58 papers dealing with both general and specific topics within the domain. Here, a scenario was defined as one simulator run where the 58 papers constituted the dataset. For the simulator run parameters, the metadata of the present study (discipline, language, title, topic, keywords and abstract) were used as the RS and RA parameters.

[4] See Appendix D for a more readable version of Fig. 7

5.2. Performance criteria

As in (Rúbio & Gulo, 2016), two performance criteria were used to assess the relevancy of the papers for the researchers:

i. Accuracy: the percentage of true classifications

ii. Precision: the percentage of the classified items that are relevant

Considering the sets of relevant papers (REL) and non-relevant papers, (NREL), true relevant (TR) denotes the papers classified as REL when they really are, while false relevant (FR) denote the papers classified as REL when they are not. Thus, with the same logic, the papers classified as NREL can be true non-relevant (TN) or false non-relevant (FN).

Accuracy, denoted by a, was computed as follows for each scenario:

$$a = \frac{TR + FR}{TR + FR + TN + FN}$$

Precision, denoted by p, was computed as follows for each scenario:

$$p = \frac{TR}{TR + FR}$$

To identify TR, FR, TN and FN for each scenario, a target paper was chosen for the domain; next, the metadata of this target paper were used as the selection parameters and the references papers from the output set were compared to the cited papers of the target paper. Through this comparison, TR, FR, TN and FN were defined. Let $a_{i,j}$ be the accuracy of the scenario i^{th} of the dataset j; the average accuracy is defined as follows:

$$Avg\_a_i = \frac{\sum_{j=1}^{D} a_{i,j}}{D}$$

Similarly, the precision of the scenario i^{th} of the dataset j is defined as:

$$Avg\_p_i = \frac{\sum_{j=1}^{D} p_{i,j}}{D}$$

where D denotes the number of datasets.

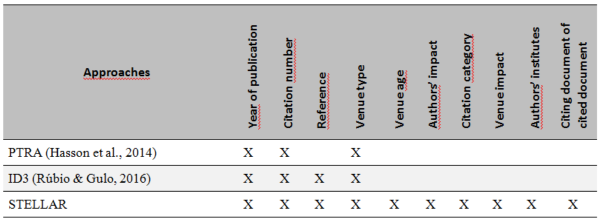
5.3. Related ranking approaches for comparison purposes

There are two other works on scientific paper ranking:

- PTRA (Hasson et al., 2014)
- ID3 (Rúbio & Gulo, 2016).

PTRA and ID3 are described in section 2.1. Table 5 presents a summary of the criteria taken into account by each ranking approach: the bottom line of Table 5 lists all the criteria used in the STELLAR ranking approach.

Table 5. Criteria taken into account in three paper ranking approaches



Approaches	Year of publication	Citation number	Reference	Venue type	Venue age	Authors' impact	Citation category	Venue impact	Authors' institutes	Citing document of cited document
PTRA (Hasson et al., 2014)	X	X		X						
ID3 (Rúbio & Gulo, 2016)	X	X	X	X						
STELLAR	X	X	X	X	X	X	X	X	X	X

The performance of the STELLAR approach was compared against the performance of PTRA (Hasson et al., 2014) and ID3 (Rúbio & Gulo, 2016) on the same datasets and scenarios. In the previous Table 7, it is observed that for ranking a cited paper as relevant, STELLAR considers more criteria, such as venue age, citation, authors' impact, etc.

5.4. Analysis of the simulation results

This section presents the analysis of the simulation results in terms of papers' relevancy for the two datasets.

5.4.1. Simulation using the dataset harvested from databases

Fig. 8 shows the average accuracy for the three different simulations (STELLAR, ID3 and PTRA). The horizontal axis represents the sequence number of the simulation scenarios and the vertical axis represents the average accuracy of the associated scenario.

It is observed that STELLAR performs better than ID3 (in green) and PTRA (in blue): STELLAR has an average accuracy of 0.91 per scenario while ID3 has an average of 0.60 per scenario. The average relative improvement in accuracy (defined as [Avg_a of STELLAR ¯ Avg_a of ID3]) of STELLAR in comparison to ID3 is 0.32 (32%).

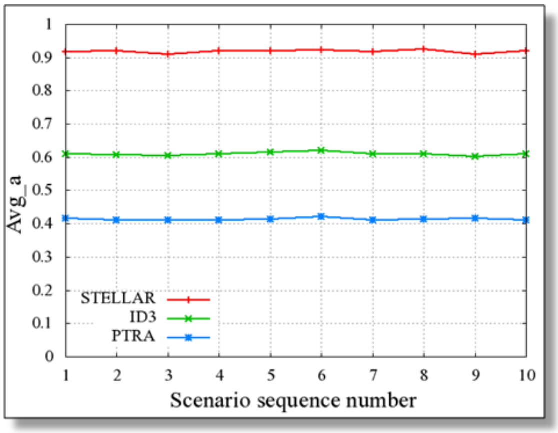


Fig. 8: Average accuracy vs Scenario sequence number – Harvested from databases

Fig. 9 shows the average precision for the same scenarios of Fig. 8. The x-axis represents the simulations scenario sequence number while the y-axis represents the average precision of the associated scenario. STELLAR performed better than ID3 and PTRA: it produced an average precision of 0.96 per scenario while ID3, the better of the two approaches used for comparison, had an average of 0.65 per scenario. The average relative improvement (defined as [Avg_p of STELLAR ¯ Avg_p of ID3]) of STELLAR in comparison to ID3 is 0.31 (31%) per scenario.

In both simulations and criteria, STELLAR outperformed ID3 and PTRA. This performance might be attributable to the use of additional bibliometric metadata.

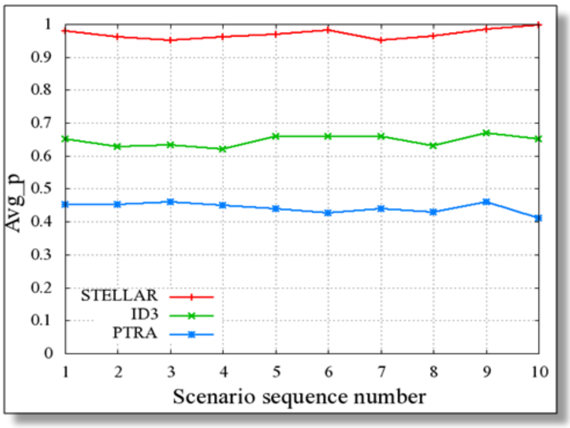


Fig. 9: Average precision vs Scenario sequence number – Harvested from databases

5.4.2. Simulation using the baseline dataset

Table 6 presents the accuracy and precision when the list of papers in the baseline dataset (i.e., the references cited in this paper) is used as the dataset for simulations with the three ranking approaches.

Table 6. Summary of performance criteria (accuracy and precision) using the baseline dataset

Approaches	Avg_a (%)	Avg_p (%)
PTRA (Hasson et al., 2014)	39.19	27.16
ID3 (Rúbio & Gulo, 2016)	53.98	41.97
STELLAR	76.09	68.73

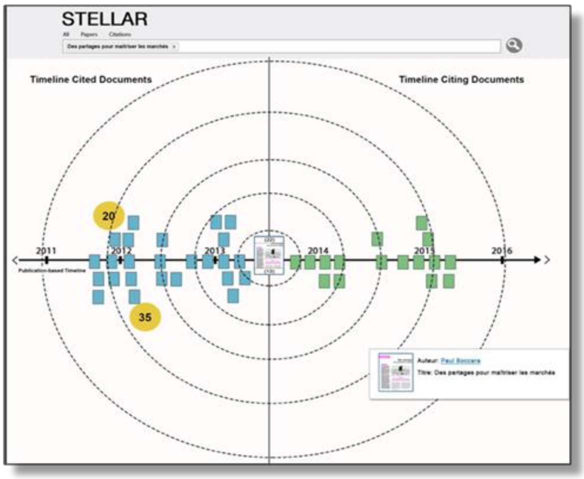
i. STELLAR produced an average accuracy (Avg_a) of 76.09% while ID3 produced an accuracy of 53.98%. The relative improvement in accuracy of STELLAR as compared to ID3 is 22.11%.

ii. STELLAR produced an average precision (Avg_p) of 68.73% while ID3 produced a precision of 41.97%. The relative improvement in precision of STELLAR as compared to ID3 is 26.76%.

Note that all the simulations are based on limited datasets, and should be extended later to larger datasets.

5.5. STELLAR prototype

This section presents a number of STELLAR's input screens. It can be seen that the radius of the paper at the top of the list is 0.0: indeed, this is the target paper. Fig. 10 represents the timeline of a document-based literature corpus radius, with the horizontal axis indicating the year of publication (here, from 2011 to 2016).


STELLAR
All
Papers
Citations
Des partages pour maîtriser les marchés
Timeline Cited Documents
Timeline Citing Documents
2011
2012
2013
2014
2015
2016
Publication-based Timeline
20
35

Auteur: Paul Boccara
Titre: Des partages pour maîtriser les marchés

Fig. 10: Timeline of a Document-based Literature Corpus Radius (LCR)

The radius denotes the temporal distance from the document at center to the cited documents and to the citing documents. The yellow circles on the left side represent multiple documents—here, 20 to 35 documents.

6. CONCLUSION AND FUTURE WORK

This paper has proposed an assisted literature review (ALR) prototype, called STELLAR (Semantic Topics Ecosystem Learning-based Literature Assistant Review). STELLAR is based on machine learning model (MLM) and a semantic metadata ecosystem (SMESE) to identify, rank and recommend relevant papers for an ALR according to researchers' selection parameters and annotations. Using text and data mining (TDM) techniques, MLM and a classification model, STELLAR assists the researcher to search relevant papers that meet his selection of parameters. The learning models applied by STELLAR use researchers' annotated (RA) data and semantic enriched metadata as training data. STELLAR also recommends selection parameters to researcher in order to refine the search.

The STELLAR prototype is based on SMESE V3, described in (Brisebois, Abran, Nadembega, et al., Unpublished results). The contributions of STELLAR include:

i. MLM designed to semantically harvest a Universal Research Documents Repository;
ii. Enhancement of Literature Corpus Radius, which compute the distance from each paper to the center of the Literature Corpus;
iii. MLM that help the researcher discover, find and refine the list of papers recommended for inclusion.

The performance of the STELLAR prototype has been evaluated through a comparison against a baseline manual LR using a number of simulations. In terms of accuracy, the STELLAR ALR provided an average accuracy of 0.91 per scenario while ID3 provided an average of 0.60 per scenario. In terms of precision, STELLAR produced an average of 0.96 per scenario while ID3 had an average of 0.65 per scenario. In comparison to ID3, STELLAR yielded an average relative improvement in accuracy of 32% per scenario and an average relative improvement in precision of 31%.

As STELLAR future work (i.e., STELLAR V2), the next contribution will focus on "Abstract of Abstracts summarization (AoA)" in order to extend STELLAR. More specifically, papers' abstracts will be used as input for our scientific paper summarization technique to generate the AoA. STELLAR V2 will allow enhancing the SMESE V3 prototype to harvest semantic metadata from more different sources as TV guides, radio channel schedule, books, music and other events calendar and create triplets to enriching metadata.

REFERENCES

[1] Beel, J., Langer, S., Genzmehr, M., Gipp, B., Breitinger, C., & Nurnberger, A. (2013). *Research paper recommender system evaluation: a quantitative literature survey*. Paper presented at the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, Hong Kong, China.

[2] Bertin, M., Atanassova, I., Sugimoto, C. R., & Lariviere, V. (2016). The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. *Scientometrics, 109*(3), 1417-1434. doi:10.1007/s11192-016-2134-8

[3] Boote, D. N., & Beile, P. (2005). Scholars Before Researchers: On the Centrality of the Dissertation Literature Review in Research Preparation. *Educational Researcher, 34*(6), 3-15. doi:http://dx.doi.org/10.3102/0013189x034006003

[4] Bornmann, L., Stefaner, M., Anegón, F. d. M., & Mutz, R. (2014). Ranking and mapping of universities and research-focused institutions worldwide based on highly-cited papers: A visualisation of results from multi-level models. *Online Information Review, 38*(1), 43-58. doi:http://dx.doi.org/doi:10.1108/OIR-12-2012-0214

[5] Bornmann, L., Stefaner, M., Anegón, F. d. M., & Mutz, R. (2015). Ranking and mappping of universities and research-focused institutions worldwide: The third release of excellencemapping.net. *COLLNET Journal of Scientometrics and Information Management, 9*(1), 65-72. doi:http://dx.doi.org/10.1080/09737766.2015.1027090

[6] Brisebois, R., Abran, A., & Nadembega, A. (Unpublished results). *A Semantic Metadata Enrichment Software Ecosystem (SMESE) based on a Multi-platform Metadata Model for Digital Libraries*. International Journal for Digital Libraries.

[7] Brisebois, R., Abran, A., Nadembega, A., & N'techobo, P. (Unpublished results). *A Semantic Metadata Enrichment Software Ecosystem based on Sentiment/Emotion Analysis Enrichment (SMESE V3)*. Information Systems.

[8] Carlos, A. S. J. G., & Thiago, R. P. M. R. (2015). *Text Mining Scientific Articles using the R Language*. Paper presented at the 10th Doctoral Symposium in Informatics Engineering, Porto, Portugal.

[9] Cataldi, M., Di Caro, L., & Schifanella, C. (2016). *Ranking Researchers Through Collaboration Pattern Analysis*. Paper presented at the European Conference on Machine Learning and Knowledge Discovery in Databases, Riva del Garda, Italy. http://dx.doi.org/10.1007/978-3-319-46131-1_11

[10] Dong, Y., Johnson, R. A., & Chawla, N. V. (2016). Can Scientific Impact Be Predicted? *IEEE Transactions on Big Data, 2*(1), 18-30. doi:http://dx.doi.org/10.1109/TBDATA.2016.2521657

[11] Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2015). Influence of omitted citations on the bibliometric statistics of the major Manufacturing journals. *Scientometrics, 103*(3), 1083-1122. doi:http://dx.doi.org/10.1007/s11192-015-1583-9

[12] Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (Vol. 6th

ed.(1996). xxii 788 pp). White Plains, NY: England: Longman.

[13] Gulo, C. A. S. J., Rubio, T. R. P. M., Tabassum, S., & Prado, S. G. D. (2015). Mining Scientific Articles Powered by Machine Learning Techniques. *OASIcs-OpenAccess Series in Informatics, 49*, 21-28. doi:http://dx.doi.org/10.4230/OASIcs.ICCSW.2015.21

[14] Hasson, M. A., Lu, S. F., & Hassoon, B. A. (2014). Scientific Research Paper Ranking Algorithm PTRA: A Tradeoff between Time and Citation Network. *Applied Mechanics and Materials, 551*, 603-611. doi:http://dx.doi.org/10.4028/www.scientific.net/AMM.551.603

[15] He, Z., Chen, C., Bu, J., Wang, C., Zhang, L., Cai, D., & He, X. (2015). Unsupervised document summarization from data reconstruction perspective. *Neurocomputing, 157*, 356-366. doi:http://dx.doi.org/10.1016/j.neucom.2014.07.046

[16] Madani, F., & Weber, C. (2016). The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis. *World Patent Information, 46*, 32-48. doi:http://dx.doi.org/10.1016/j.wpi.2016.05.008

[17] Marx, W., & Bornmann, L. (2016). Change of perspective: bibliometrics from the point of view of cited references—a literature overview on approaches to the evaluation of cited references in bibliometrics. *Scientometrics, 109*(2), 1397-1415. doi:http://dx.doi.org/10.1007/s11192-016-2111-2

[18] MASIC, I., & BEGIC, E. (2016). Evaluation of Scientific Journal Validity, It's Articles and Their Authors. *Stud Health Technol Inform., 226*, 9-14. doi:http://dx.doi.org/10.3233/978-161499-664-4-93-5

[19] Mayr, P., Scharnhorst, A., Larsen, B., Schaer, P., & Mutschke, P. (2014). *Bibliometric-Enhanced Information Retrieval*. Paper presented at the 36th European Conference on IR Research (ECIR), Amsterdam, The Netherlands. http://dx.doi.org/10.1007/978-3-319-06028-6_99

[20] Packalen, M., & Bhattacharya, J. (2015). Neophilia Ranking of Scientific Journals. *National Bureau of Economic Research Working Paper Series, 21579*. doi:http://dx.doi.org/10.3386/w21579

[21] Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). *WordNet::Similarity: measuring the relatedness of concepts*. Paper presented at the Demonstration Papers at Human Language Technology conference/North American chapter of the Association for Computational Linguistics (HLT-NAACL), Boston, Massachusetts, USA.

[22] Rúbio, T. R. P. M., & Gulo, C. A. S. J. (2016). *Enhancing Academic Literature Review through Relevance Recommendation*. Paper presented at the 11th Iberian Conference on Information Systems and Technologies, Gran Canaria, Canary Islands, Spain.

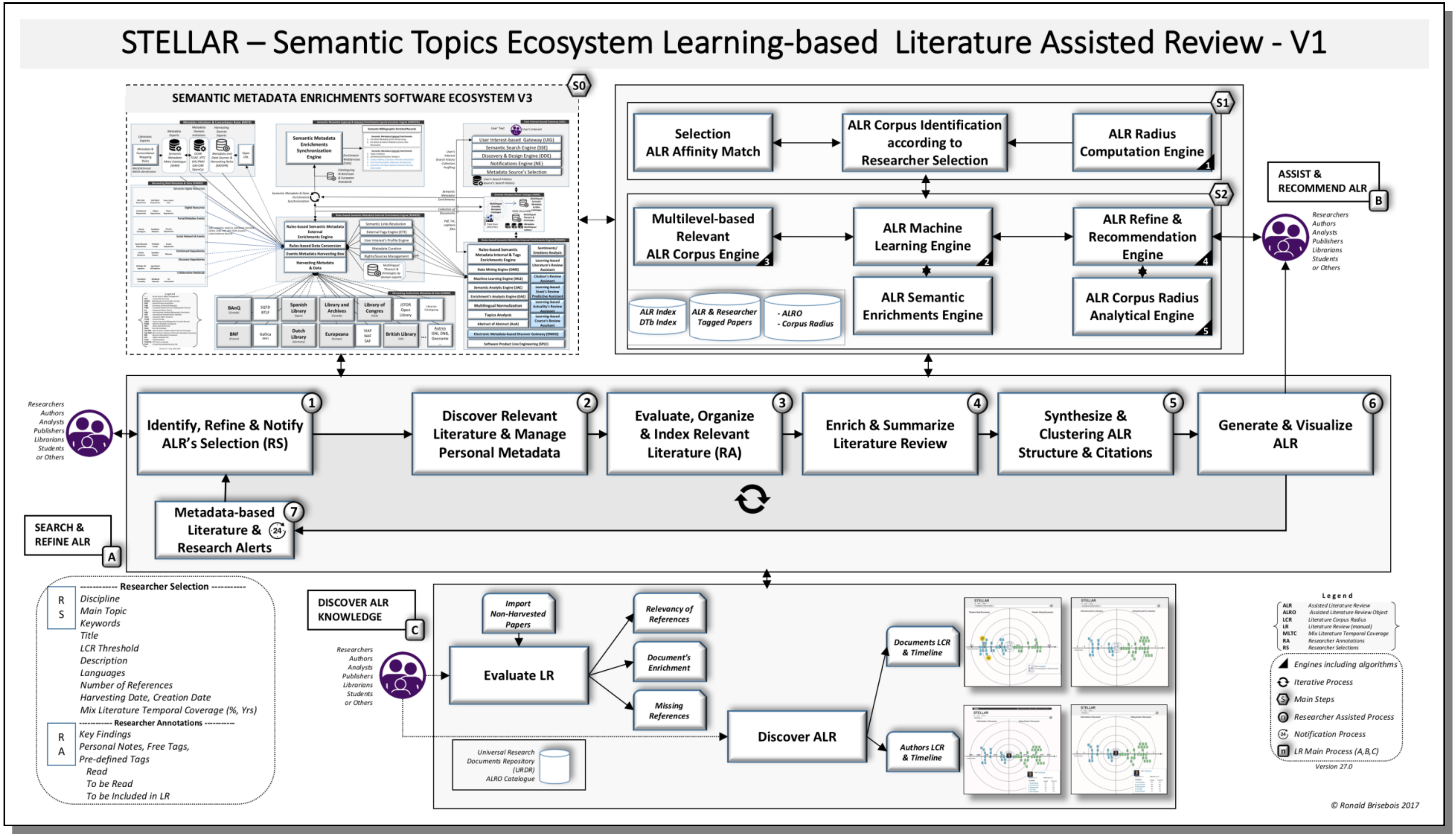
[23] Wan, X., & Liu, F. (2014). WL-index: Leveraging citation mention number to quantify an individual's scientific impact. *Journal of the Association for Information Science and Technology, 65*(12), 2330-1643. doi:http://dx.doi.org/10.1002/asi.23151

[24] Wang, S., Xie, S., Zhang, X., Li, Z., Yu, P. S., & Shu, X. (2014). *Future Influence Ranking of*

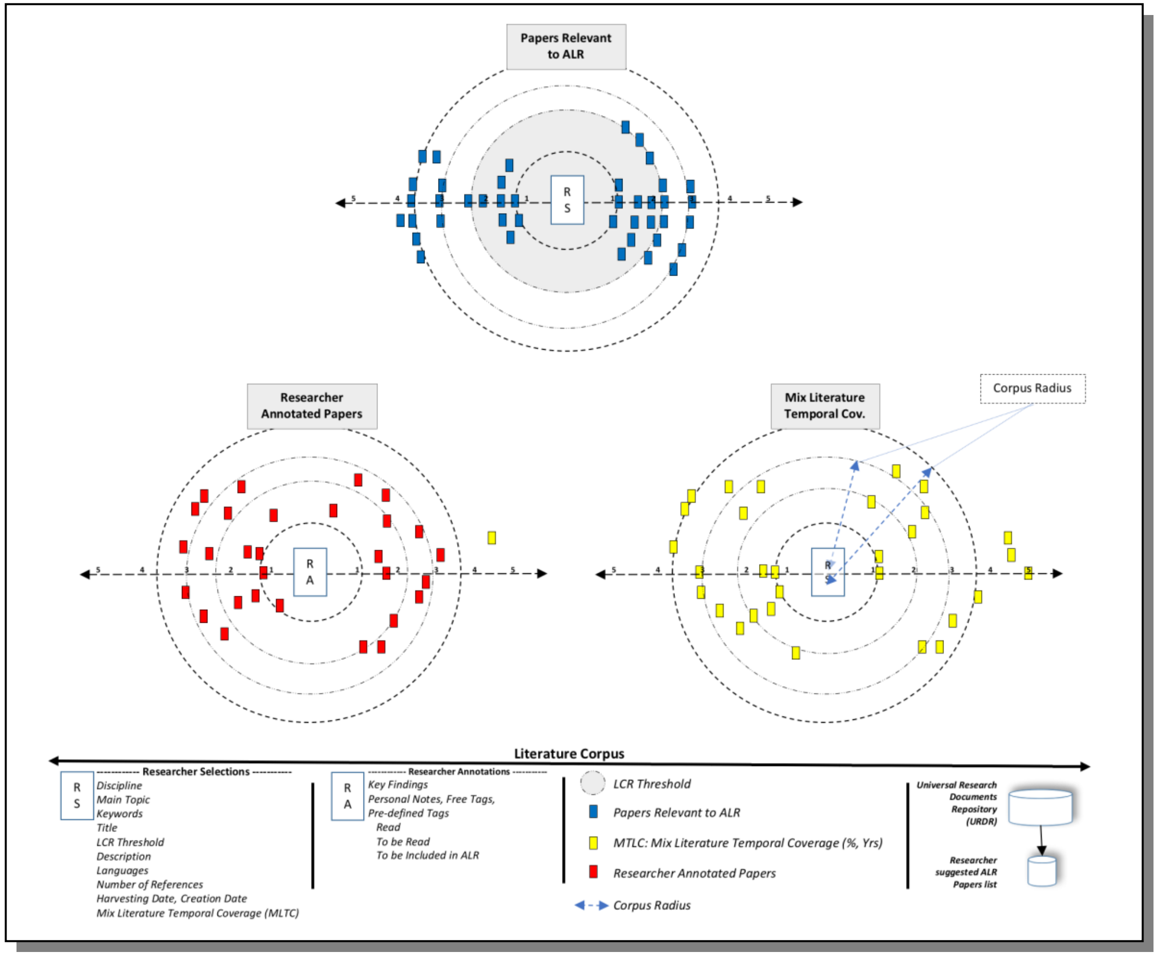
Scientific Literature. Paper presented at the Society for Industrial and Applied Mathematics (SIAM) International Conference on Data Mining, Philadelphia, Pennsylvania, USA. http://epubs.siam.org/doi/abs/10.1137/1.9781611973440.86

[25] Zhang, M., Zhang, X., & Hu, Y. (2015). *Ranking of Collaborative Research Teams Based on Social Network Analysis and Bibliometrics*. Paper presented at the 12th International Conference on Cooperative Design, Visualization, and Engineering (CDVE), Mallorca, Spain. http://dx.doi.org/10.1007/978-3-319-24132-6_30

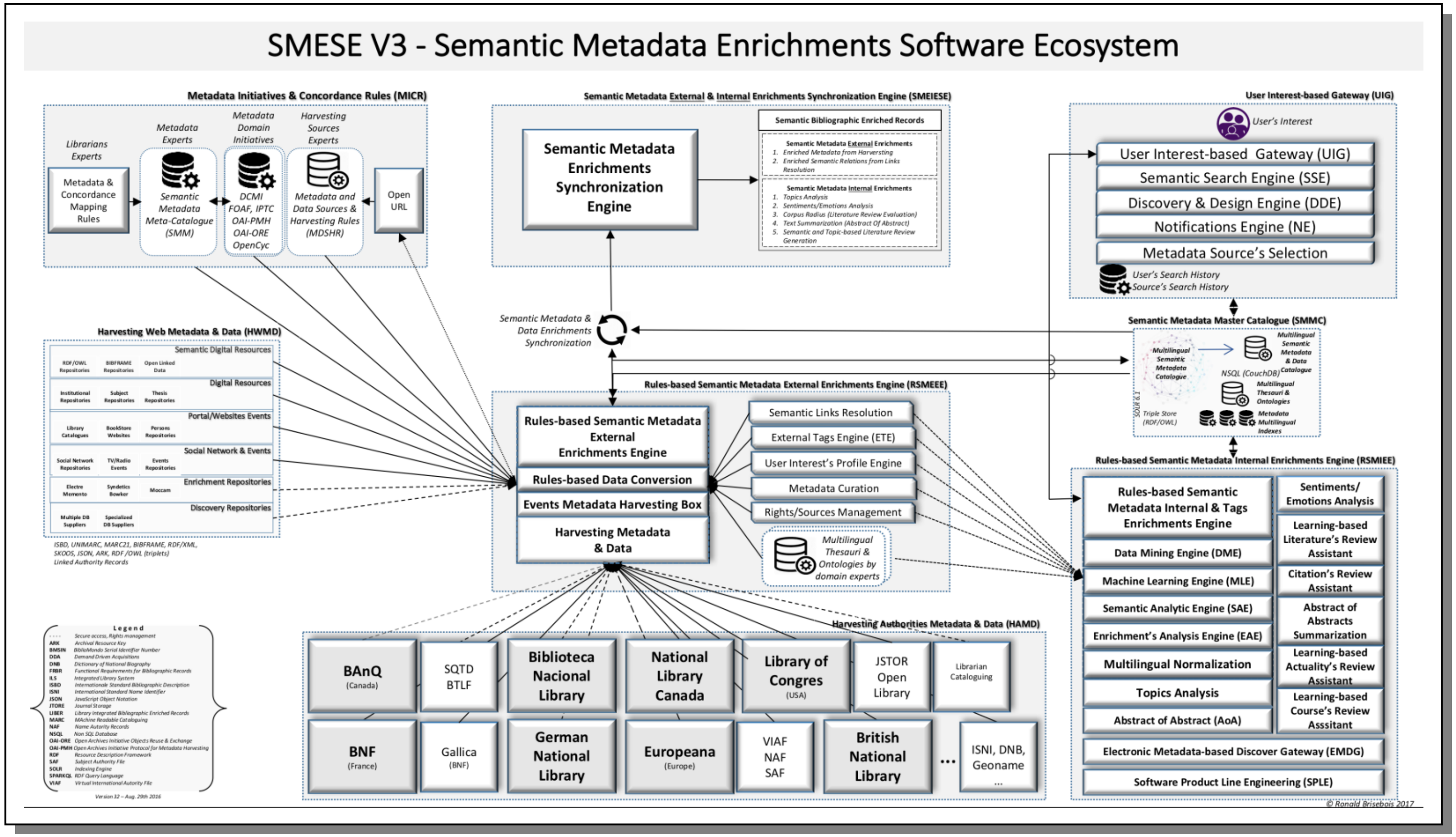
Appendix A: STELLAR – Semantic Topics Ecosystem Learning-based Literature Assisted Review - V1


STELLAR – Semantic Topics Ecosystem Learning-based Literature Assisted Review - V1
SEMANTIC METADATA ENRICHMENTS SOFTWARE ECOSYSTEM V3
S0
S1
Selection
ALR Affinity Match
ALR Corpus Identification according to Researcher Selection
ALR Radius Computation Engine
1
S2
Multilevel-based Relevant ALR Corpus Engine
3
ALR Machine Learning Engine
2
ALR Refine & Recommendation Engine
4
ALR Semantic Enrichments Engine
ALR Corpus Radius Analytical Engine
5
ALR Index
DTb Index
ALR & Researcher Tagged Papers
- ALRO
- Corpus Radius
ASSIST & RECOMMEND ALR
B
Researchers
Authors
Analysts
Publishers
Librarians
Students
or Others
Identify, Refine & Notify ALR's Selection (RS)
1
Discover Relevant Literature & Manage Personal Metadata
2
Evaluate, Organize & Index Relevant Literature (RA)
3
Enrich & Summarize Literature Review
4
Synthesize & Clustering ALR Structure & Citations
5
Generate & Visualize ALR
6
Metadata-based Literature & Research Alerts
7
24
SEARCH & REFINE ALR
A
---------- Researcher Selection ----------
R
S
Discipline
Main Topic
Keywords
Title
LCR Threshold
Description
Languages
Number of References
Harvesting Date, Creation Date
Mix Literature Temporal Coverage (%, Yrs)
---------- Researcher Annotations ----------
R
A
Key Findings
Personal Notes, Free Tags,
Pre-defined Tags
Read
To be Read
To be Included in LR
DISCOVER ALR KNOWLEDGE
C
Import Non-Harvested Papers
Evaluate LR
Relevancy of References
Document's Enrichment
Missing References
Discover ALR
Universal Research Documents Repository (URDR)
ALRO Catalogue
Documents LCR & Timeline
Authors LCR & Timeline
Legend
ALR Assisted Literature Review
ALRO Assisted Literature Review Object
LCR Literature Corpus Radius
LR Literature Review (manual)
MLTC Mix Literature Temporal Coverage
RA Researcher Annotations
RS Researcher Selections
Engines including algorithms
Iterative Process
Main Steps
Researcher Assisted Process
Notification Process
LR Main Process (A,B,C)
Version 27.0
© Ronald Brisebois 2017

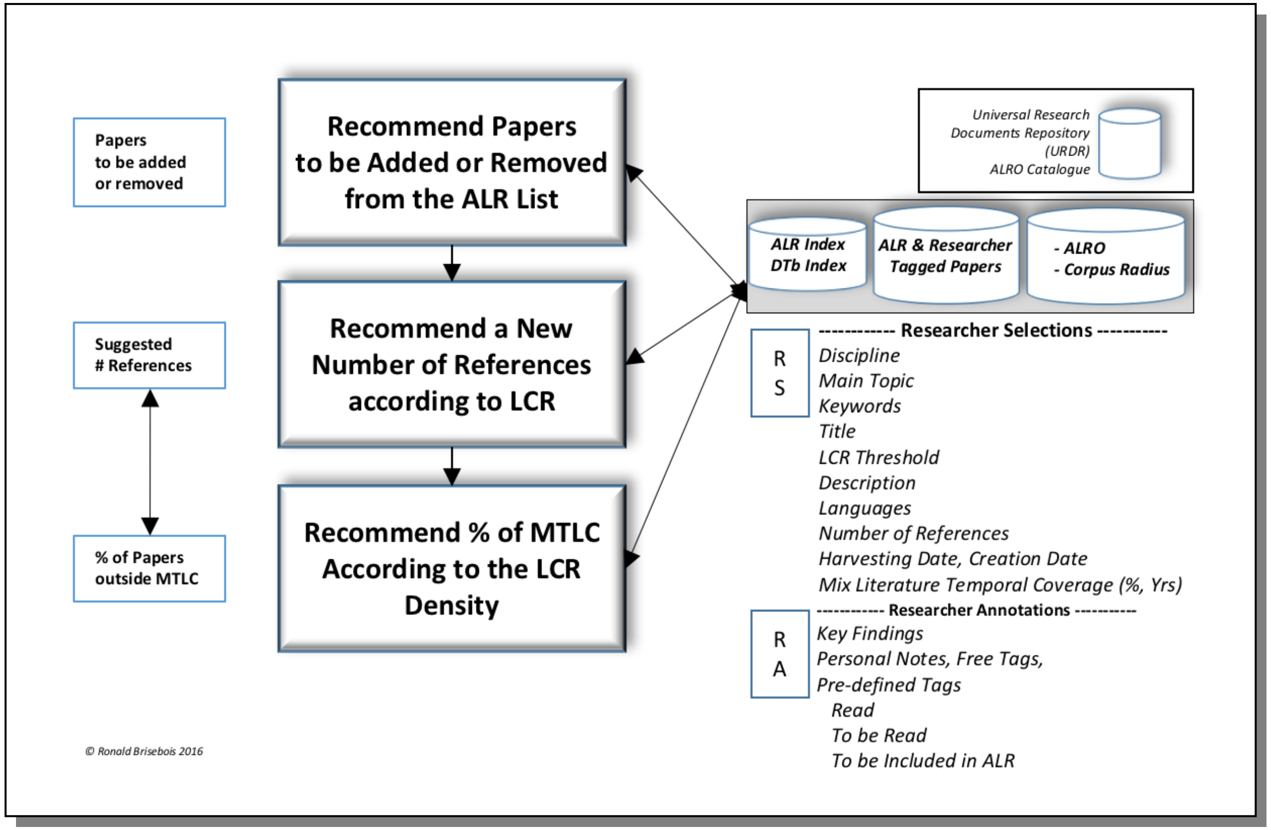
Appendix B: Fig. 3 - Sources used to build the suggested list of ALR papers


Papers Relevant to ALR
R S
Researcher Annotated Papers
R A
Mix Literature Temporal Cov.
R S
Corpus Radius
1
2
3
4
5
Literature Corpus
------------ Researcher Selections ------------
R S
Discipline
Main Topic
Keywords
Title
LCR Threshold
Description
Languages
Number of References
Harvesting Date, Creation Date
Mix Literature Temporal Coverage (MLTC)
------------ Researcher Annotations ------------
R A
Key Findings
Personal Notes, Free Tags,
Pre-defined Tags
Read
To be Read
To be Included in ALR
LCR Threshold
Papers Relevant to ALR
MTLC: Mix Literature Temporal Coverage (%, Yrs)
Researcher Annotated Papers
Corpus Radius
Universal Research Documents Repository (URDR)
Researcher suggested ALR Papers list

Appendix C: Fig. 4 - SMESE V3 - Semantic Metadata Enrichments Software Ecosystem


SMESE V3 - Semantic Metadata Enrichments Software Ecosystem
Metadata Initiatives & Concordance Rules (MICR)
Librarians Experts
Metadata & Concordance Mapping Rules
Metadata Experts
Semantic Metadata Meta-Catalogue (SMM)
Metadata Domain Initiatives
DCMI FOAF, IPTC OAI-PMH OAI-ORE OpenCyc
Harvesting Sources Experts
Metadata and Data Sources & Harvesting Rules (MDSHR)
Open URL
Semantic Metadata External & Internal Enrichments Synchronization Engine (SMEIESE)
Semantic Metadata Enrichments Synchronization Engine
Semantic Bibliographic Enriched Records
Semantic Metadata External Enrichments
1. Enriched Metadata from Harversting
2. Enriched Semantic Relations from Links Resolution
Semantic Metadata Internal Enrichments
1. Topics Analysis
2. Sentiments/Emotions Analysis
3. Corpus Radius (Literature Review Evaluation)
4. Text Summarization (Abstract Of Abstract)
5. Semantic and Topic-based Literature Review Generation
User Interest-based Gateway (UIG)
User's Interest
User Interest-based Gateway (UIG)
Semantic Search Engine (SSE)
Discovery & Design Engine (DDE)
Notifications Engine (NE)
Metadata Source's Selection
User's Search History
Source's Search History
Semantic Metadata & Data Enrichments Synchronization
Semantic Metadata Master Catalogue (SMMC)
Multilingual Semantic Metadata Catalogue
Multilingual Semantic Metadata & Data Catalogue
NSQL (CouchDB)
Multilingual Thesauri & Ontologies
SOLR 6.1
Triple Store (RDF/OWL)
Metadata Multilingual Indexes
Harvesting Web Metadata & Data (HWMD)
Semantic Digital Resources
RDF/OWL Repositories
BIBFRAME Repositories
Open Linked Data
Digital Resources
Institutional Repositories
Subject Repositories
Thesis Repositories
Portal/Websites Events
Library Catalogues
BookStore Websites
Persons Repositories
Social Network & Events
Social Network Repositories
TV/Radio Events
Events Repositories
Enrichment Repositories
Electre Memento
Syndetics Bowker
Moccam
Discovery Repositories
Multiple DB Suppliers
Specialized DB Suppliers
ISBD, UNIMARC, MARC21, BIBFRAME, RDF/XML, SKOOS, JSON, ARK, RDF /OWL (triplets)
Linked Authority Records
Rules-based Semantic Metadata External Enrichments Engine (RSMEEE)
Rules-based Semantic Metadata External Enrichments Engine
Rules-based Data Conversion
Events Metadata Harvesting Box
Harvesting Metadata & Data
Semantic Links Resolution
External Tags Engine (ETE)
User Interest's Profile Engine
Metadata Curation
Rights/Sources Management
Multilingual Thesauri & Ontologies by domain experts
Rules-based Semantic Metadata Internal Enrichments Engine (RSMIEE)
Rules-based Semantic Metadata Internal & Tags Enrichments Engine
Data Mining Engine (DME)
Machine Learning Engine (MLE)
Semantic Analytic Engine (SAE)
Enrichment's Analysis Engine (EAE)
Multilingual Normalization
Topics Analysis
Abstract of Abstract (AoA)
Sentiments/ Emotions Analysis
Learning-based Literature's Review Assistant
Citation's Review Assistant
Abstract of Abstracts Summarization
Learning-based Actuality's Review Assistant
Learning-based Course's Review Asssitant
Electronic Metadata-based Discover Gateway (EMDG)
Software Product Line Engineering (SPLE)
Harvesting Authorities Metadata & Data (HAMD)
BAnQ (Canada)
SQTD BTLF
Biblioteca Nacional Library
National Library Canada
Library of Congres (USA)
JSTOR Open Library
Librarian Cataloguing
BNF (France)
Gallica (BNF)
German National Library
Europeana (Europe)
VIAF NAF SAF
British National Library
...
ISNI, DNB, Geoname ...
Legend
Secure access, Rights management
ARK Archival Resource Key
BMSIN BiblioMondo Serial Identifier Number
DDA Demand Driven Acquisitions
DNB Dictionary of National Biography
FRBR Functional Requirements for Bibliographic Records
ILS Integrated Library System
ISBD Internationale Standard Bibliographic Description
ISNI International Standard Name Identifier
JSON JavaScript Object Notation
JTORE Journal Storage
LIBER Library Integrated Bibliographic Enriched Records
MARC MAchine Readable Cataloguing
NAF Name Autority Records
NSQL Non SQL Database
OAI-ORE Open Archives Initiative Objects Reuse & Exchange
OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting
RDF Resource Description Framework
SAF Subject Authority File
SOLR Indexing Engine
SPARKQL RDF Query Language
VIAF Virtual International Autority File
Version 32 – Aug. 29th 2016
© Ronald Brisebois 2017

Appendix D: Fig. 7 - Refinement & Recommendation MLM


Papers to be added or removed
Recommend Papers to be Added or Removed from the ALR List
Suggested # References
Recommend a New Number of References according to LCR
% of Papers outside MTLC
Recommend % of MTLC According to the LCR Density
Universal Research Documents Repository (URDR) ALRO Catalogue
ALR Index DTb Index
ALR & Researcher Tagged Papers
- ALRO
- Corpus Radius
----------- Researcher Selections ----------
R S
Discipline
Main Topic
Keywords
Title
LCR Threshold
Description
Languages
Number of References
Harvesting Date, Creation Date
Mix Literature Temporal Coverage (%, Yrs)
----------- Researcher Annotations ----------
R A
Key Findings
Personal Notes, Free Tags,
Pre-defined Tags
Read
To be Read
To be Included in ALR
© Ronald Brisebois 2016