Terminological Saturation in Retrospective Text Document Collections

The Influence of the Order of Adding Documents to Datasets on Terminological Saturation

Authors:
Victoria Kosa (ZNU),
David Chaves-Fraga (UPM),
Dmitriy Naumenko (BWT),
Eugene Yuschenko (BWT),
Hennadiy Dobrovolskyi (ZNU),
Svitlana Moiseenko (ZNU),
Alexander Vasileyko (ZNU),
Carlos Badenes-Olmedo (UPM)
Vadim Ermolayev (ZNU),
Oscar Corcho (UPM),
Aliaksandr Birukou (SPRINGER)

21.11.2018

Document Information

Project Information:					
Project ID:		**Acronym:**	TS-RTDC		
Full Title:	Terminological Saturation in Retrospective Text Document Collections				
Project URL:	N/A				
Project Manager:	Vadim Ermolayev	**Partner:**	ZNU	**E-mail:**	vadim@ermolayev.com

Deliverable Information:				
Type:	Technical Report		Planned	Actual
Status:	Version 2 Revision 1.0 Final	**Date of Delivery:**	31.10.2018	21.11.2018
Dissemination: (to check)	Public: [V] Consortium: [ ]	**Document ID:**	TS-RTDC-TR-2018-2-v2	
URL:	https://github.com/OntoElect/Docs/tree/master/Reports/TS-RTDS-TR-2018-2-v2.pdf			

Author Information:			
Authors (Partners):	Victoria Kosa (ZNU), David Chaves Fraga (UPM), Dmitriy Naumenko (BWT), Eugene Yuschenko (BWT), Hennadiy Dobrovolskyi (ZNU), Svitlana Moiseenko (ZNU), Alexander Vasileyko (ZNU), Carlos Badenes (UPM), Vadim Ermolayev (ZNU), Oscar Corcho (UPM) and Aliaksandr Birukou (SPRINGER)		
Responsible Author:	Victoria Kosa	**Partner:**	ZNU
		E-mail:	v.kosa@znu.edu.ua

Document Information:	
Abstract: (for dissemination)	This document reports the results of our experimental study aimed to find out the impact of different orders of adding documents to datasets for measuring terminological saturation. The motivation for this research activity lies in the fact that real world document collections are characterized by terminological drift in time. We empirically investigated the proper ways to cope with this temporal drift and its influence on terminological saturation. Our premise was that there could be several different orders of adding documents to the processed datasets, dealing with the time of publication: (i) **chronological**; (ii) **reversed-chronological**; (iii) **bi-directional**; (iv) **random**; (v) **descending citation frequency**. Experiments were performed using three different real world document collections coming from different domains, where the collections of high-quality documents were available as scientific papers. In the presence of different levels of noise, it has also been checked if different orders are differently sensitive in detecting excessive noise. Based on the comparison of experimental results, we recommended that the **descending citation frequency** order of adding documents to datasets is preferable as it demonstrated the most balanced performance.
Keywords:	OntoElect, Automated Term Extraction, Terminological Saturation, Order of Adding Documents
Citation:	Kosa, V., Chaves-Fraga, D., Naumenko, D., Yuschenko, E., Moiseenko, S., Dobrovolskyi, H., Vasileyko, A., Badenes-Olmedo, C., Ermolayev, V., Corcho, O., and Birukou, A.: The Influence of the Order of Adding Documents to Datasets on Terminological Saturation. Technical Report TS-RTDC-TR-2018-2, 21.11.2018, Dept. of Computer Science, Zaporizhzhia National University, Ukraine, 72 p.

Revision Log:				
Version/Revision	**Date**	**Change(s)**	**Submitted by**	**Comment**
0.1, Draft	01.12.2017	**Initial** draft containing experimental set-up and partial results (DMKD collection)	Victoria Kosa	
0.2, Draft	22. 12.2017	**Intermediate** draft. Added the results on the TIME collection.	Victoria Kosa, Vadim Ermolayev	
0.3, Draft	29.12.2017	**Intermediate** draft. Added the results on the DAC naturelle datasets.	Victoria Kosa, Vadim Ermolayev	
0.4, Draft	01.01.2018	**Intermediate** draft. Added the results on the DAC cleaned datasets. Added the final comparison of results.	Victoria Kosa, Vadim Ermolayev	
0.5, Pre-final	02.01.2018	**Pre-final** draft. Finished: Introduction, Motivation, and Conclusions sections	Victoria Kosa, Vadim Ermolayev	
0.6, Final	20.01.2018	**Final**: cross-read, several typos and mistakes corrected	Victoria Kosa, Vadim Ermolayev	Made public on ResearchGate
0.7, Draft	20.08.2018	Rewritten the dataset generator and catalogue generator modules; extended State-of-the-Art, added the order of descending citation frequency to experiments.	Hennadiy Dobrovolskyi, Svitlana Moiseenko, Alexander Vasileyko, Victoria Kosa, Vadim Ermolayev	The OntoElect GitHub repository has been set up, supporting files and docs deployed there
0.9, Pre-final	30.10.2018	DCF order experiments done, added the analysis of the DCF order experiments	David Chaves Fraga, Victoria Kosa, Vadim Ermolayev	Supporting files and docs deployed to the OntoElect GitHub repository
1.0, Final	05.11.2018	Added missing bits and references, text scanned for typos.	Hennadiy Dobrovolskyi, Victoria Kosa, Vadim Ermolayev	Data and experimental results published @Mendeley Data

Authors and Affiliations

Victoria Kosa, Hennadiy Dobrovolskyi,
Svitlana Moiseenko, **Alexander Vasileyko**,
and **Vadim Ermolayev**

Department of Computer Science

Zaporizhzhya National University (**ZNU**)

Zhukovskogo st. 66,

69600, Zaporizhzhia

Ukraine

David Chaves-Fraga,
Carlos Badenes-Olmedo,
and **Oscar Corcho**

Ontology Engineering Group,

Universidad Politécnica de Madrid (**UPM**)

Madrid,

Spain

Dmitriy Naumenko and **Eugene Yuschenko**

BWT Group (**BWT**)

Mayakovskogo st. 11,

69035, Zaporizhzhia,

Ukraine

Aliaksandr Birukou

Springer-Verlag GmbH (**SPRINGER**)

Tiergartenstrasse 17,

69121, Heidelberg,

Germany

Table of Contents

Executive Summary

This report presents our results in the development of the methodological components for extracting representative (complete) sets, having minimal possible size, of significant terms extracted from the representative sub-collections of textual documents with time stamps. The approach to assess the representativeness does so by evaluating terminological saturation in a document (sub-)collection.

One of the important aspects in this work is that the constituent documents in a collection have been published at different times. Hence, the temporal drift in terminology has to be appropriately taken into account. We focus on empirically investigating the proper ways to cope with this temporal drift and its influence on terminological saturation. Our premise is that there could be several different orders of adding documents to the processed datasets, dealing with the time of publication: (i) **chronological**; (ii) **reversed-chronological**; (iii) **bi-directional**; (iv) **random**; and (v) **descending citation frequency**.

We perform our experiments on three different real world document collections coming from different domains, where the collections of high-quality documents are available as scientific papers. The collections and the respective datasets are presented in Section 6. These collections also have different proportions of noise. Therefore, we are able to assess the impact of different orders of adding documents in the presence of different levels of noise and check if different orders are differently sensitive for excessive noise.

For each collection, we:

- Extract the bags of terms from the prepared datasets using the UPM Term Extractor software
- Measure terminological saturation in the pairs of the extracted bags of terms using our THD module
- Measure the two additional characteristics that further help us analyse the influence of an order on terminological saturation. These are: (i) the **proportion** of the **retained** terms to **all extracted** terms in percent; (ii) the **volatility** of terminological difference (*thd*) values – a discrete analogue to the 1st derivative to these values – computed as the difference between the current and previous values.
- Build the diagrams and analyse the results

In addition to these activities, we also look at the effect of removing stop terms after doing term extraction. By removing these stop terms, which represent the noise in the pre-processed documents, we de-noise the output. This is why, for the DAC collection, we also compare noisy (DAC naturelle) and cleaned (DAC cleaned) bags of terms for all the orders.

In assessing the impact of an order of adding documents, we analyse the following measurable aspects:

- Which of the orders result in the earlier or later entry into the saturation zone (*thd* lower than *eps*)
- Which of the orders result in, integrally, the smallest No of retained terms
- Which of the orders result in, integrally, the highest individual term significance thresholds *eps*
- Which of the orders result in, integrally, the highest proportions of all extracted to retained terms
- Which of the orders result in, integrally, the least volatile *thd* values

In assessing the sensitivity of an order of adding documents to excessive noise, we base the comparison on the following measurable aspects:

- Which of the orders result in the earlier or later *eps* peak
- Which of the orders result in the higher or lower *eps* peak
- Which of the orders result in the earlier or later *thd* peak
- How many measurement steps are taken to confirm excessive noise after the *eps* peak
- Which of the orders result in the earlier or later no-of-retained-terms peak
- How many measurement steps are taken to confirm excessive noise after no-of-retained-terms peak

Based on the results of the comparison, we recommend that the **descending citation frequency** order of adding documents to datasets might be used as the one demonstrating the best and most balanced performance.

This result sounds reasonable as the choice of the **descending citation frequency** order allows processing the documents in the order of their descending impact on the domain, measured in citation frequency. Hence, taking documents with more impact first, results in getting all the significant terms introduced in these documents earlier. This order, due to the use of citation frequency as a measure for impact, also balances those significant terms that are new to the terms that survived as significant through time. Due to that, it remedies terminological drift in time better that the other orders.

Our future experimental work will focus on answering our research question about finding an optimal size of a dataset increment for having quicker and more stable terminological saturation. For that, the **descending citation frequency** will be used as proven optimal.

List of Figures

List of Tables

1 Introduction

Many research activities are undertaken currently to improve the quality of automated term extraction (ATE) results. These activities focus on different aspects, including: new or improved extraction algorithms; combining linguistic and statistical approaches to extraction; developing new or refined metrics which allow higher quality extraction; developing new extraction tools which yield better results and scale to fit current dataset size requirements. The mainstream criteria used to assess the quality of extracted results are adopted from information retrieval and based on recall and precision metrics. However, to the best of our knowledge, there were **no reports on approaches to assess the completeness of the document collection from which term extraction is done** [1].

The research presented in this report[1] is the part of the development of the methodological and instrumental components for extracting representative (complete) sets of significant terms from the representative sub-collections of textual documents having minimal possible size. It is assumed that the documents in a collection cover a single and well circumscribed domain and have a timestamp associated with them. The main hypothesis, put forward in this work, is that a sub-collection can be considered as representative to describe the domain, in terms of its terminological footprint, if any additions of extra documents from the entire collection to this sub-collection do not noticeably change this footprint. Such a sub-collection is further considered as a complete terminological core. Hence, a representative bag of significant terms can be extracted from this sub-collection. These bags of terms can further be used e.g. for learning an ontology [3] in the domain of the document collection. The approach to assess the representativeness does so by evaluating terminological saturation in a sequence of incrementally extended sub-collections of the entire collection.

One of the important aspects in minimizing the size of the terminological core of the collection is that the constituent documents were published at different times. Hence, the temporal drift in terminology has to be appropriately taken into account. Secondly, different documents may have different terminological contributions to the collection. Therefore, the differences in the terminological impacts of individual documents have to be taken into account. Our research hypothesis in this respect is that both aspects could be accounted for if a proper order of adding documents to sub-collections is used.

In this report, we focus on empirically finding out which of the possible orders of adding documents to the datasets for detecting terminological saturation is the most efficient and effective. For being efficient, it has to allow detecting saturation faster and using smaller sub-collections. For being effective, it has to be stable in detecting saturation. Our premise is that there could be several different ways, to add

[1] This research is performed as the PhD project by the first author. Its exposé has been presented in [2].

documents for the processed datasets, dealing with terminological temporal drift and the differences in the terminological impacts of individual documents: (i) **chronological**; (ii) **reversed-chronological**; (iii) **bi-directional**; (iv) **random**; and (v) **descending citation frequency**.

Section 3 presents the review of the related work in the relevant fields of automated term extraction from English texts and text similarity (distance) measurement for the bags of terms.

The approach to measure terminological saturation is based on the use of the THD algorithm developed in our OntoElect project [3]. This part of OntoElect methodology is briefly presented in Section 4.

Section 5 of the report provides the information on our experimental settings. The experimental workflow is presented in Section 5.1. Section 5.2 describes the instrumental software, which has been developed [1] to facilitate the routine and laborious activities in our experiments. Section 5.3 describes the objectives and set-up of our experimental series. Section 5.4 focuses on presenting the document collections used in our experiments. It also describes the characteristics of the datasets generated from the document collections. All the software modules and experimental datasets are available publicly to ensure the reproducibility of our results. The links to these resources are provided in Sections 5.2 and 5.4.

Section 6 presents our experimental results and discusses these comparatively using several important features that help assess the impact of different orders on: (i) yielding quicker and more stable terminological saturation, hence, more compact outputs; but also (ii) detecting excessive noise in the input datasets that may result in yielding saturated noise. As an outcome of this analysis, a comparative table with a numeric assessment of the performance of all the orders is given in Section 5.4. Based on this comparison, the best performing order of adding documents to datasets is recommended.

Finally, we summarize our results and outline our plans for the future work in Section 7, which concludes the report.

2 Motivation

Extracting terminology from texts is a complicated and laborious process, which requires a substantial part of highly qualified human effort. Despite that, it is more and more often used in many important applications, e.g. for engineering ontologies [3], [4]. Hence, knowing the smallest possible representative document collection for a domain is very important to develop ontologies efficiently and with satisfactory domain coverage. Therefore, laying out a method to determine a terminologically saturated subset of documents of the minimal size within a collection is topical. It is also important to make this method as efficient and automated as possible to lower the overhead on the core knowledge engineering workflow.

One of the important aspects in such a kind of document collections is that the constituent documents are published at different times. Therefore, the terminology used in these documents reflects the understanding of the domain by the authors at these different times of publication. These times may vary significantly, so as the terms used or introduced by the authors. Therefore, temporal drift in terminology has to be appropriately taken into account. Secondly, different documents may have different terminological contributions to the collection. Hence, the differences in the terminological impacts of individual documents have to be appropriately dealt with.

In the reported research, we investigate the possible ways to cope with these two important differences in individual documents and their influence on terminological saturation. Our premise is that there could be several different ways, to add documents to the processed datasets:

(i) The documents may be added in the **chronological order** of their publication dates. This way reflects a natural order of the formation of a collection.
(ii) The **reversed-chronological order** allows processing the most recent documents first. This order first considers the most recent documents going back to the oldest ones. Using this way allows focusing on the most recently introduced terminology and account for the terms that passed the test of time.
(iii) The **bi-directional** order offers a more temporally balanced way to mix the newest and the oldest documents while adding these from a collection to datasets for processing. The documents are picked in turns, the newest, the oldest, again the oldest left, and so on.
(iv) The **random order** ignores the time of publication and picks the documents randomly for their inclusion in the dataset increments. This way may be reasonable if a collection is big, as it may allow not losing some significant collection members in the middle.
(v) The order of **decreasing citation frequency** takes the most frequently cited documents first as, presumably, these bring the highest terminological contribution

Our **objective** in the reported work is to find out if any of these orders of adding documents to datasets have a measurable impact on terminological saturation. More specifically, is there an order, which, if used in the processing pipeline, results in:

- A more stable[2] terminological saturation
- Detecting the smallest yet statistically representative document sub-collection which is a terminological core

[2] Terminological saturation is meant to be more stable if it is less likely that adding a new increment of documents to a dataset makes it not saturated.

3 Related Work

The work presented in this report aims at improving the efficiency of the measures of terminological difference between the bags of terms extracted from the consecutive pairs of incrementally extended textual datasets. The datasets are formed by adding textual documents taken from a document collection as described in Section 4. The documents for these additions could be picked in different orders.

The measurement of terminological saturation is done with the help of our THD algorithm [3], which uses character strings equivalence measures for comparing terms. The THD algorithm returns the values of the individual term significance threshold (*eps*), absolute terminological difference (*thd*) and relative terminological difference (*thdr*). The latter two are in fact the similarity measures for the pair of the bags of significant (retained) terms extracted from the respective datasets in a pair. It is also the premise in our work that the terms are multi-word, extracted from plain text files, and accompanied by numeric significance (rank) values. The terms also have to be English. Therefore, the work related to the presented research is in:

- Automated term extraction (ATE) from English texts
- Text similarity (distance) measurement for the bags of terms, containing the lists of: (i) terms; and (ii) corresponding numeric term significance values

3.1 Automated Term Extraction

In the majority of approaches to ATE, e.g. [6] or [7], processing is done in two consecutive phases: linguistic processing and statistical processing. Linguistic processors, like POS taggers or phrase chunkers, filter out stop words and restrict candidate terms to *n*-gram sequences: nouns or noun phrases, adjective-noun and noun-preposition-noun combinations. Statistical processing is then applied to measure the ranks of the candidate terms. These measures are [8]: either the measures of unithood, which focus on the collocation strength of units that comprise a single term; or the measures of termhood, which point to the association strength of a term to domain concepts.

For unithood, the measures are used such as mutual information [9], log likelihood [9], t-test [6], [7], modifiability and its variants [10], [7]. The measures for termhood are either term frequency-based (unsupervised approaches) or reference corpora-based (semi-supervised approaches). The most used frequency-based metrics are TF/IDF (e.g. [11], [12]), weirdness [13], and domain pertinence [14]. More recently, hybrid approaches were proposed, that combine unithood and termhood measurements in a single value. A representative metric is c/nc-value [15]. C/nc-value-based approaches to ATE have received their further evolution in many works, e.g. [6], [14], [16] to mention a few.

Linguistic processing is organized and implemented in a very similar fashion in all ATE methods, except some of them that also include filtering out stop words. Stop words could be filtered out also at a cut-off step after statistical processing. Statistical processing is sometimes further split in two consecutive sub-phases of term

candidate scoring, and ranking. For term candidates scoring, reflecting its likelihood of being a term, known methods could be distinguished by being based on (c.f. [11]) measuring occurrences frequencies (including word association), assessing occurrences contexts, using reference corpora, e.g. Wikipedia [17], topic modelling [18], [27].

The cut-off procedure, takes the top candidates, based on scores, and thus distinguishes significant terms from insignificant (or non-) terms. Many cut-off methods rely upon the scores, coming from one scoring algorithm, and establish a threshold in one or another way. Some others that collect the scores from several scoring algorithms use (weighted) linear combinations [38], voting [8], [3], or (semi-)supervised learning [39]. In our set-up, we do cut-offs after term extraction based on retaining a simple majority vote, as explained in Section 4. Therefore, the ATE solutions, which perform cut-offs together with scoring are not relevant for our approach.

Based on the evaluations in [8], [11], [40], the most widely used ATE algorithms, for which their performance assessments are published, are listed in Table 1. The table also provides the assessments based on the aspects we use for selection.

Table 1. The comparison of the most widely used ATE measures and algorithms

Method [Source]	**Domain-independence (+/-)**	**Supervizion (U/SS)**	**Metrics**	**Term Significance**	**Cut-off (+/-)**	**Precision (GENIA; average)**	**Run Time (%/c-value)**
TTF [41]	+	U	Term (Total) Frequency	+	-	0.70; 0.35	0.34
ATF [40]	+	U	Average Term Frequency	+	-	0.71; 0.33	0.37
						0.75; 0.32	0.35
TTF-IDF [42]	+	U	TTF+Inverse Document Frequency	+	-	0.82; 0.51	0.35
RIDF [43]	+	U	Residual IDF	-		0.71; 0.32	0.53
						0.80; 0.49	0.37
C-value [15]	**+**	**U**	c-value, nc-value	**+**	**-**	0.73; **0.53**	1.00
						0.77; **0.56**	1.00
Weirdness [13]	+/-	SS	Weirdness	-		0.77; 0.47	0.41
						0.82; 0.48	1.67
GlossEx [38]	+	SS	Lexical (Term) Cohesion, Domain Specificity	-		0.70; 0.41	0.42
TermEx [14]	+	SS	Domain Pertinence, Domain Consensus, Lexical Cohesion, Structural Relevance	-	+	0.87; 0.46	0.52
PU-ATR [17]	-	SS	nc-value, Domain Specificity	-	+	0.78; 0.57	809.21

Comments:
Domain Independence: "+" stands for a domain-independent method; "-" marks that the method is either claimed to be domain-specific by its authors, or is evaluated only on one particular domain. We are looking for a domain-independent method.
Supervision: "U" – unsupervised; "SS" – semi-supervised. We are looking for an unsupervised method.
Term Significance: "+" – the method returns a value for each retained term, which could further be used as a measure of its significance compared to the other terms; "-" marks that such a measure is not returned or the method does the cut-off itself. We are looking for receiving a measure to do cut-offs later.
Cut-off: "+" – the method does cut-offs itself and returns only significant terms; "-" – the method does not do cut-offs. We are looking for "-".
Precision and Run Time: The values are based on the comparison of the 2 cross-evaluation experiments reported in [11] / [40]. Empty cells in the table mean that there was no data for this particular method in this particular experiment. Survey [11] used ATR4S – open-source software written in Scala. It evaluated 13 different methods, implemented in ATR4S, on 5 different datasets, including GENIA. Survey [40] used JATE 2.0, free software written in Java. It evaluated 9 different methods, implemented in JATE, on 2 different datasets, including GENIA. Hence, the results on GENIA are the baseline for comparing the Precision. Two values are given for each reference experiment: precision on GENIA; average precision. Both [11] and [40] experimented with *c-value* method, which was the slowest on average for [40]. So, the execution times for *c-value* were used as a baseline to normalize the rest in the **Run Time** column.

After looking at Table 1, we support the conclusion of [40] stating that *c-value* is the most reliable method. The *c-value* method obtains consistently good results, in terms of precision, both on the 2 different mixes of datasets – [40] and [11]. We also note that *c-value* is one of the slowest methods among the group of unsupervised and domain-independent, though its performance is comparable with the fastest ones. Still, *c-value* outperforms the domain-specific methods, sometimes significantly – as it is in the case with PU-ATR.

Hence, we have chosen *c-value* as the method for our experimental framework. Therefore, we looked at the software tools, which implement *c-value* and are publicly freely available. We checked the implementations of term extraction tools at several popular web resources like at http://inmyownterms.com/terminology-extraction-tools/ or https://en.wikipedia.org/wiki/Terminology_extraction. In addition to the reference implementations mentioned before: ATR4S [11] and JATE 2.0 [40], we have identified the following freely available ATE software tools as indicated in Table 2.

Table 2. Free ATE software tools (listed alphabetically)

Name / Owner	Website	Short description	Algorithm / Metric	Domain	Constraints
BioTex / LIRMM	http://tubo.lirmm.fr/biotex/	extracts biomedical terms from free text		Bio-medical	Domain-specific
FiveFilters / Medialab-Prado	http://fivefilters.org/term-extraction/	extracts terms through a web service; relies on a PHP port of Topia's Term Extraction; a	Occurrence (TTF) and word count in a term	independent	Web service, size of text constrained

Name / Owner	Website	Short description	Algorithm / Metric	Domain	Constraints
		simple alternative to Yahoo Term Extraction service			
TaaS (Terminology as a Service EU Project)	https://term.tilde.com/	Identify term candidates in your documents and extract them automatically. Uses CollTerm (linguistic) or Kilgray (statistical) services	Frequency-based	independent	Does not provide term significance scores
TerMine / NaCTeM	http://www.nactem.ac.uk/software/termine/	Extracts terms from plain English texts, provides the Batch mode (access to be requested for non-UK academic users)	*c-value*	independent	The service requests to avoid heavy bulk processing
TermFinder / Translated.net	https://labs.translated.net/terminology-extraction/	A Web application; extracts terms from the inserted plain text; compares the frequency of words in the given text with their frequency in the generic language corpus	Poisson statistics, Maximum Likelihood Estimation and IDF	requires language corpus	Returns the score of a term as a numeric value (%)
TBXTools [44] / Universitat Oberta de Catalunya	https://sourceforge.net/projects/tbxtools/	A Python toolset using NLTK (Natural Language Toolkit)	TTF	Independent, multilingual, requires language corpus	Deletes n-grams with stop words
UPM Term Extractor [28] / Dr Inventor EU project	https://github.com/ontologylearning-oeg/epnoi-legacy	A Java software for extracting terms and relations from scientific papers	*c-value*	Independent	Takes text input data of at most 15 Mb

After the analysis of the tools listed in Table 2, we found out that the most appropriate candidates were NaCTeM TerMine and UPM Term Extractor. Those two ATE tools were cross-evaluated [19]. Based on this cross-evaluation, UPM Term Extractor has been chosen as it outperformed NaCTeM TerMine in several aspects.

3.2 Text Similarity Measurement

In the recent surveys on text similarity measurement approaches, e.g. [30], [31], methods (or measures)[3] are grouped based on analysing: (i) characters and their sequences; (ii) tokens; (iii) terms; (iv) text corpora; or (v) synsets. In [31] hybrid

[3] In this context, we do not distinguish a method and a measure. A method is understood as a way to implement the corresponding measure function.

measures are also mentioned that allow fuzzy matching between tokens. Brief characteristics of these groups are given immediately below. The individual methods belonging to the groups are detailed in Table 3.

Character- and character sequence-based measures compare characters and their sequences in strings, taking into account also the order of characters. These include the measures of common character sequences, e.g. substrings; edit distance; the number and order of the common characters between two strings.

Token-based methods model a string as a set of tokens: individual characters, character *n*-grams, or separate words. Quantification is done by computing the size of the overlap normalized by a measure of string length.

Term-based measures are similar to token-based measures but the tokens are different. Those are not character *n*-grams but terms, which are word *n*-grams with possibly varying *n*. Furthermore, the weights of the terms, e.g. their frequencies of occurrence, are taken into account. These measures apply more to long character strings, or documents, hence are better suited to measure document or text dataset similarity. The *thd* and *thdr* measures used in our work fall into this category. Therefore, comparing *thd* and *thdr* to the other term-based measures is relevant.

Corpus-based and **synset-based** (or knowledge-based) **methods** are marginally relevant to our purposes in this report. Corpus-based approaches determine the similarity between words based on (statistical) information gained from large text corpora. Synset-based approaches rely on semantic networks, like WordNet [36], to derive semantic similarity between words. Both approaches are therefore too bulky computationally, though may be applied to ATE – e.g. for deciding about cut-offs. In OntoElect, we decide about cutting off insignificant terms after ATE, based on the notion of a simple majority vote, as described in Section 4. Hence, we omit looking closer at corpus- and synset-based measures.

The overview of the most popular text / string similarity measures, grouped by method types, is provided in Table 3. This overview is by far not complete as many other variants of SSM are available in the literature. Those we omit are however based on the same principles compared to the listed in Table 3, to the best of our knowledge.

The approach followed in our work is finding the terminological core of a document collection by measuring terminological saturation [3],[29]. This measurement is done using our terminological difference measure (*thd*, [3]) which is a variant of a Manhattan distance measure (see e.g. [30]) or Minkovski's distance with p=1 [37].

Table 3. The overview of text similarity / distance measures

Name, source	Description	Specifics	Relevance	
			Term Similarity	*thd* [4]
Character- and character sequence-based measures				
Longest Common Substring [32]	common character sequence based measure	returns the integer length of the longest common substring; could be normalized by the total length	moderate	irrelevant
Levenshtein distance [20]	edit distance based measure	returns an integer number of required edits	marginal	irrelevant
Hamming distance [25]	edit distance based measure	strings have to be of equal length	marginal	irrelevant
Monger-Elkan distance [21]	edit distance based measure	returns an integer number of required edits	marginal	irrelevant
Jaro distance [22]	counts the minimal number of one character transforms in one string for arriving at the other string	returns a normalized real value from [0, 1]	good	irrelevant
Jaro-Winkler distance [23]	refines Jaro measure by using a prefix scale value – prioritizes the stings that match at the beginning	returns a normalized real value from [0, 1]	good	irrelevant
Token-based measures				
Sørensen-Dice coefficient [26],[34]	counts the ratio of identical character *bi*-grams to the overall number of bi-grams in both strings	returns a normalized real value from [0, 1]	good	irrelevant
Jaccard similarity [24]	counts the ratio between the cardinalities of the intersection and union of the character sets (*uni*-grams) in the strings	returns a normalized real value from [0, 1]	good	irrelevant
Cosine similarity [31]	size of the overlap in character *uni*-grams divided by the square root of the sum of the squared total numbers of *uni*-grams	returns a normalized positive real value	marginal (computa-tionally hard)	irrelevant

[4] *thd* is the measure for terminological difference developed in OntoElect [3] and used in our approach – see also Section 5. Hence, "relevant" in this column means being appropriate for measuring terminological difference between documents of text datasets.

Name, source	Description	Specifics	Relevance	
			Term Similarity	*thd* [4]
	in both strings			
Term-based measures				
Euclidian distance [33]	Measures traditional Euclidian distance in an *n*-dimensional metric space (of positive reals)	works for documents; returns a real positive value	irrelevant	relevant
Cosine similarity [35]	Computes a cosine between two vectors in the term space; vectors are specified by term weights (e.g. TF of C-value)	works for documents; returns a normalized positive real value	irrelevant	marginal
Pearson correlation [33]	Computes Pearson correlation for a pair of vectors in the term vector space	works for documents; returns a normalized real value that ranges from +1 to −1; it is 1 when vectors are fully identical	irrelevant	marginal
Manhattan (block) distance [30]	the distance to be traveled to get from one data point to the other if a grid-like path is followed	works for documents; resembles the *thd* measure [2]	irrelevant	relevant
thd (*thdr*) [3]	a vector space-based measure that returns an absolute (normalized) difference between the sets of terms in the pair of compared bags of retained significant terms	works for a pair of bags of terms; every term is supplemented with its significance value (*c-value* [19])	irrelevant	the same

3.3 Contributions and Background Work

In this work, we do not contribute any novel method for ATE. The *c-value* method [15] implemented in the UPM Term Extractor is used as this combination of the method and implementation has been experimentally proven to be the best appropriate for detecting terminological saturation [19].

In difference and hopefully complementary to the abovementioned relevant work, we contribute the experimental analysis and comparison of the influence of different orders of adding documents to datasets on terminological saturation. Terminological saturation is measured using our THD algorithm. Based on this comparative analysis we recommend the descending citation frequency order of adding documents as the best performing.

4 OntoElect Saturation Measure and Measurement Pipeline

OntoElect, as a methodology, seeks for maximizing the fitness of the developed ontology to what the domain knowledge stakeholders think about the domain. Fitness is measured as the stakeholders' "votes" – a measure that allows assessing the stakeholders' commitment to the ontology under development, reflecting how well their sentiment about the requirements is met. The more votes are collected, the higher the commitment is expected to be. If a critical mass of votes is acquired (say 50%+1, which is a simple majority vote), it is considered that the ontology meets the requirements satisfactorily.

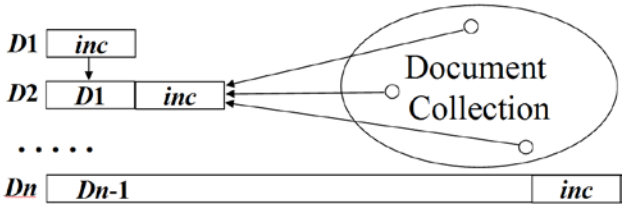
Unfortunately, direct acquisition of requirements from domain experts is not very realistic. The experts are expensive and not willing to do the work, which falls out of their core activity. That is why the OntoElect approach focuses on the indirect collection of the stakeholders' votes by extracting these from high quality and reasonably high impact documents authored by the stakeholders.

An important feature to be ensured for knowledge extraction from text collections is that the dataset needs to be representative to cover the opinions of the domain knowledge stakeholders satisfactorily fully. OntoElect suggests a method to measure the terminological completeness of the document collection by analysing the *saturation* of terminological footprints of the incremental slices of the document collection [3].

The approach followed in our work is finding the terminological core of a document collection by measuring terminological saturation [3], [29]. This measurement is done using our terminological difference measure (*thd*, [3]) which is a variant of a Manhattan distance measure.

The full texts of the documents from a collection are grouped in datasets in the order of their timestamps. As pictured in Fig. 1 (a), the first dataset D_1 contains the first portion of (*inc*) documents. The second dataset D_2 contains the first dataset D_1 plus the second incremental slice of (*inc*) documents. Finally, the last dataset D_n contains all the documents from the collection.

At the next step of the OntoElect workflow, the bags of multi-word terms B_1, B_2, …, B_n are extracted from the datasets D_1, D_2, …, D_n together with their significance (*c-value*) scores, using UPM Term Extractor software [28]. An example of an extracted bag of terms is shown in Fig. 1 (b).



(a) Incrementally enlarged datasets in OntoElect

Term	*c-value*
temporal	6428.5
time	3950
temporal logic	2558
logic	2186.5
interval	2011
model	1957
formula	1750
relations	1613
temporal representation	1604
relation	1574.5
constraints	1557
reasoning	1549
....	

(b) an example of an extracted bag of terms

Fig. 1. OntoElect: the approach to form datasets (a) and an example of a bag of extracted terms

At the subsequent step, every extracted bag of terms B_i, i = 1, …, n is processed as follows. Firstly, an individual term significance threshold (*eps*) is computed to cut off those terms that are not within the majority vote. The sum of *c-value*s with individual values above *eps* form the majority vote if this sum is higher than ½ of the sum of all *c-value*s. Secondly, insignificant term candidates are cut-off at *c-value* < *eps*. Thirdly, the normalized scores are computed for each individual term: *n-score* = *c-value* / max(*c-value*). Finally, the result is saved in the bag of retained significant terms T_i.

After this step only significant terms, that represent the majority vote, are retained in the bags of terms. T_i are then evaluated for saturation by measuring pair-wise terminological difference between the subsequent bags T_i and T_{i+1}, i = 0, …, n-1. It is done by applying the baseline THD algorithm [3] pictured in Fig. 2.

```
Algorithm THD. Compute Terminological Difference between Bags of Terms
Input:
 Ti, Ti+1 – the bags of terms with grouped similar terms.
      Each term Ti.term is accompanied with its T.n-score.
      Ti, Ti+1 are sorted in the descending order of T.n-score.
 M –  the name of the string similarity measure function to compare terms
 th – the value of the term similarity threshold from within [0,1]
Output: thd(Ti+1, Ti), thdr(Ti+1, Ti)
1.  sum := 0
2.  thd := 0
3.  for k := 1, |Ti+1|
4.    sum :=  sum + Ti+1.n-score[k]
5.    found : = .F.
6.    for m := 1, |Ti|
7.       if (Ti+1.term[k] = Ti.term[m])  if (M(Ti+1.term[k], Ti.term[m], th))
8.           then
9.             thd += |Ti+1.n-score[k] - Ti.n-score[m]|
10.            found := .T.
11.   end for
12.   if (found = .F.) then thd += Ti+1.n-score[k]
13. end for
14. thdr := thd  / sum
```

Fig. 2. THD algorithm [3] for comparing a pair of the bags of extracted terms. It has been modified, compared to [3], for computing also the *thdr* value.

In fact, THD accumulates, in the *thd* value for the bag T_{i+1}, the *n-score* differences if there were linguistically the same terms in *Ti* and *Ti*+1. If there was no the same term in *Ti*, it adds the *n-score* of the orphan to the *thd* value of *Ti*+1. After *thd* has been computed, the relative terminological difference *thdr* receives its value as *thd* divided by the sum of *n-scores* in *Ti*+1.

Absolute (*thd*) and relative (*thdr*) terminological differences are computed for further assessing if T_{i+1} differs from T_i more than the individual term significance threshold *eps*. If not, it implies that adding an increment of documents to D_i for producing D_{i+1} did not contribute any noticeable amount of new terminology. So, the subset D_{i+1} of the overall document collection may have become terminologically saturated. However, to obtain more confidence about the saturation of D_{i+1}, OntoElect suggests that more subsequent pairs of T_i and T_{i+1} are evaluated. The rationale for that is that, sometimes, a terminological peak may occur after saturation has been observed in the previous pairs of *T*. Normally this peak indicates that a highly innovative document with a substantial number of new terms has been added in the increment. If stable saturation is observed, then the process of looking for a minimal saturated sub-collection (terminological core) could be stopped. An example of saturation evaluation for the TIME document collection [5] using OntoElect is pictured in Fig. 3.

Datasets Pair	Bag of Terms	C-value > 1	eps	Retained Terms (C-value > eps)	thd, value	thdr,%
empty-D1	8546	3439	3.0000	838	54.4448	100.0000
D1-D2	14597	5654	3.1699	1179	35.9807	62.3806
D2-D3	23992	9223	3.7549	1548	36.0855	59.6366
D3-D4	31427	12127	4.0000	2104	23.7044	35.4153
D4-D5	38122	14923	4.7549	2183	22.4341	30.7901
D5-D6	42788	17256	5.0000	2400	14.9911	18.7218
D6-D7	49986	20127	5.0000	2821	17.4853	20.7287
D7-D8	59294	23950	5.0000	3430	23.1877	26.9035
D8-D9	65627	26418	5.0000	3767	13.1819	15.3747
D9-D10	75171	30473	5.6147	3584	25.0810	36.7663
D10-D11	81617	33296	6.0000	3893	9.6005	13.8278
D11-D12	91692	37637	6.0000	4410	13.3894	19.7595
D12-S13	101190	41803	6.0000	4903	9.0502	12.6376
D13-D14	108203	44655	6.0000	5255	7.3260	9.8946
D14-D15	115493	47792	6.0000	5658	8.5976	11.7790
D15-D16	121832	50475	6.0000	6007	6.6174	9.0302
D16-D17	128171	53180	6.3043	5564	6.3422	9.0829
D17-D18	137918	57368	6.3399	6043	13.0734	20.2061
D18-D19	145173	60346	6.3549	6109	5.1033	8.0395
D19-D20	151075	62839	6.6667	6259	5.4895	8.7677

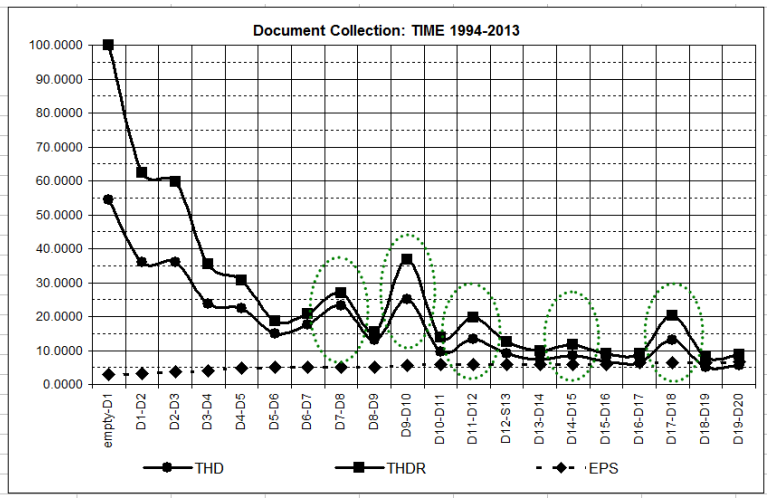


Fig. 3. The results of evaluating the saturation of the TIME document collection (adapted from [5]). Terminological peaks are observed at D_7-D_8, D_9-D_{10}, D_{14}-D_{15}, and D_{17}-D_{18}. As explained in [5], the peaks are correlated with the added frequently cited papers. It is also worth noticing that the number of retained terms in *T*-s is significantly lower that the number of extracted terms in *B*-s.

5 Experimental Workflow and Set-up

In our prior work [2], we have formulated several research questions about the important aspects that may influence terminological saturation in document collections. One of these aspects is the order of adding documents to the increments in the datasets, as described in Section 3. Our research hypothesis is that using different orders of adding documents to datasets may result in better (quicker, smoother) or worse (slower, more volatile) saturated sequences of datasets. The objective of the reported part of our research is to find out, experimentally, which of the orders outperforms the rest.

For conducting experiments in the reported work, the instrumental toolset has been developed to assist in an experimental workflow. The workflow is presented in Section 5.1, the instrumental toolset – in Section 5.2, the objectives and set-up of our experiments – in Section 4.3, and the document collections and datasets – in Section 5.4.

5.1 Experimental Workflow

The experimental workflow, outlined in Fig. 4, is based on the OntoElect processing pipeline described in Section 3. This workflow could be generically applied (using Configure Experiment step) to perform all the experimental series planned in the project, as described in [2].

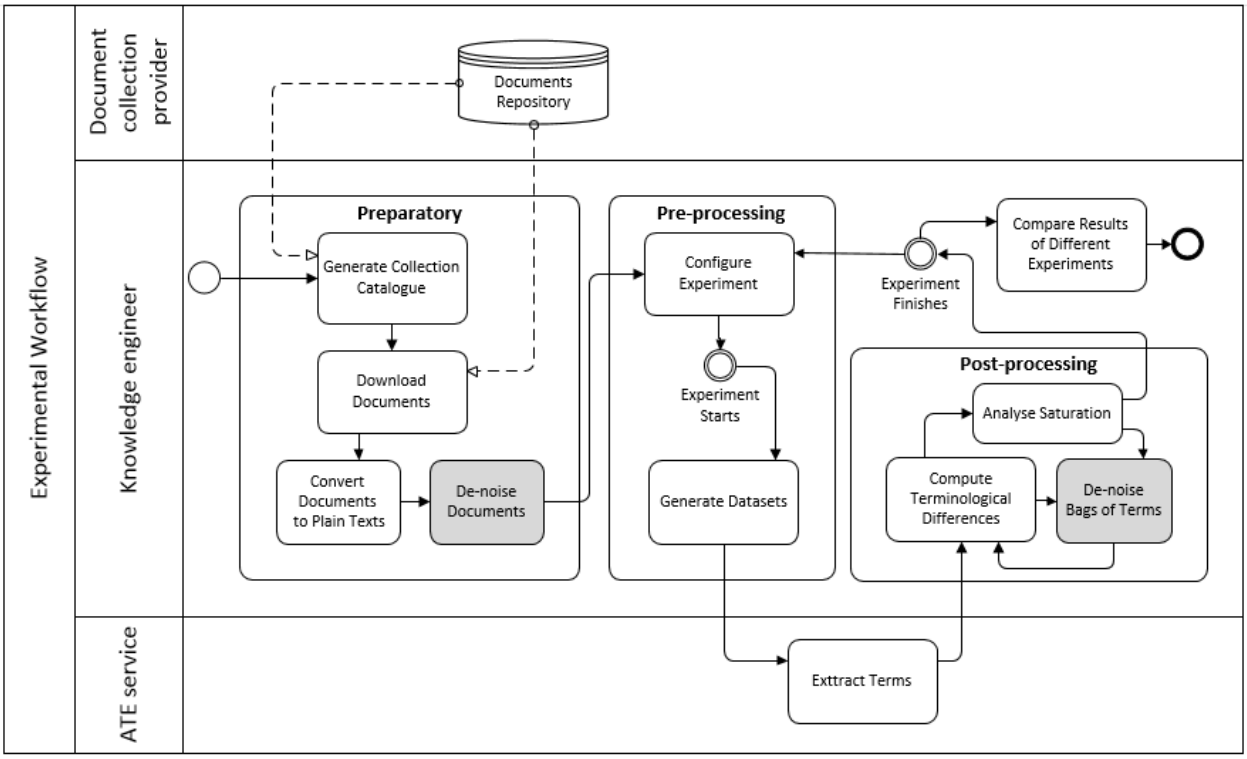


Fig. 4. Experimental workflow. The Knowledge Engineer pipeline includes the Preparatory, Pre-processing, Term Extraction, and Post-processing phases. The De-noise Documents and De-noise Bags of Terms tasks (greyed) are optional.

The workflow covers the **preparatory** activities for documents acquisition, **pre-**

processing, including the **generation** of experimental datasets, term **extraction**, **saturation measurement**, and the **analysis** of the results. Some of the steps can only be performed manually, like Configure Experiment, Analyse Saturation, and Compare Results of Different Experiments. These steps are not too laborious, however, and the effort does not noticeably grow with the number of documents. The rest of the steps require instrumental software support, especially for large document collections. The instrumental software that has been developed to support this workflow is presented in Section 5.2.

The **Preparatory** phase includes:

- **The Generation** of the catalogue of the chosen document collection using the information publicly available at the document repository. This catalogue includes all the metadata for the documents, including their abstracts, and the numbers of their citations acquired from Google Scholar[5].
- **The Download** of the full texts of the papers, usually in PDF format, based on the information in the catalogue. This step may require the permission granted by the owner of the collection to bulk-download their full texts.
- **The Conversion** of the full texts of the downloaded documents to the plain text format for further term extraction.

The De-noising of the plain text documents. This step is optional and may require the use of the techniques the use of a text cleansing technique. Relevant approaches range from simple regular expression search and replacement to complex long short-term memory (LSTM) language models [45]. In OntoElect De-noising of the Bags of Terms is performed at Post-processing and only for regular noise (stop terms) that, terminologically, have high *c-value*s, as described in [1].

The **Pre-processing** phase comprises **Experiment Configuration** and **Dataset Generation** steps.

Configuring an experiment makes the workflow generic as it allows choosing among different kinds of experiments and the data to be processed. For this, several choices have to be made: the document collection (in plain texts); the order of adding documents to datasets; the size of the increment (*inc*) for the datasets.

The **Dataset Generation** task takes these configuration parameters in and generates the sequence of datasets using the texts of the document collection as the input. The texts are added to the increments in the chosen order.

The phase of **term extraction** consequently applies the chosen software tool to the generated datasets: D_1, D_2, …, D_n. It outputs the bags of extracted terms B_1, B_2, …, B_n.

The **Post-processing** phase comprises the task for **computing terminological differences** using the THD algorithm, optional task for **De-noising the Bags of**

[5] http://scholar.google.com/

Terms, and **Saturation Analysis** step.

The **computation of terminological differences** is done for the pairs of the bags of terms (B_1, empty), (B_2, B_1), (B_3, B_2), …, (B_n, B_{n-1}) as explained in Section 4. It outputs the results in the tabular form as, for example, pictured in Fig. 3.

The saturation **analysis** step and further **comparison** with the other experimental results are done manually, using any appropriate software tool. MS Excel is used in our experiments at this step.

It might be detected by examining the bags of terms manually or analysing the output of the THD algorithm, that the texts of the documents are too noisy. In the case of the presence of regular noise, its detection is quite easy at the Post-processing phase, as described in Section 6.3. In such cases there is a good chance that regular noise dominates real terms in the sets of retained significant terms T_1, T_2, …, T_n. For de-noising, the removal of the regular noise from the bags of terms is performed in a semi-automated way. After removing excessive noise, the computation of terminological differences and saturation analysis are repeated.

5.2 Instrumental Software Toolset

Our experimental workflow, presented in Section 4.1, is appropriately supported by the developed and used instrumental software. The toolset is concisely presented in Table 4 and further explained in more detail.

Table 4: The modules of the instrumental software toolset

Phase / Task	Tool	Input	Output	Implementation Language, Link	Constraints
Preparatory Phase					
Generate Collection Catalogue	Catalogue Generator – Springer Journals	Document repository	Catalogue in CSV/XLS; contains documents metadata and the No of citations	PHP, https://github.com/bwtgroup/SSRTDC-Springer-article-parser	A parser tailored to a specific document repository; implemented for Springer Link journal pages
	Catalogue Generator – document files	the folder containing PDF files		Java, https://github.com/OntoElect/Code/tree/master/CatGen-PDF	A parser for the downloaded collection of PDF documents
Download Documents	Full Text Downloader	Catalogue (CSV)	the folder with PDF documents	PHP, https://github.com/bwtgroup/SSRTDC-Collections-Springer-PDF-Downloader	
Convert Documents to Plain Texts	PDF to Plain Text Convertor	the folder with PDF documents	the folder with Plain Text documents	Python, https://github.com/bwtgroup/SSRTDC-PDF2TXT	
Pre-Processing Phase					
Configure Experiment - done manually					

Phase / Task	Tool	Input	Output	Implementation Language, Link	Constraints
Generate Datasets	Dataset Generator	the folder with Plain Text documents	the folder with Plain Text datasets	Python, https://github.com/OntoElect/Code/tree/master/DataSetGen	
Terms Extraction Phase					
Extract Terms	UPM Term Extractor	the folder with Plain Text datasets	the folder with the bags of terms	Java, https://github.com/ontologylearning-oeg/epnoi-legacy	English texts only, *c-value* method, takes in text input data of at most 15 Mb
Post-processing Phase					
Compute Terminological Differences	Baseline THD	the folder with the bags of terms	The table containing *eps*, *thd*, *thdr* values for the consecutive pairs (B_1, empty), (B_2, B_1), (B_3, B_2), …, (B_n, B_{n-1})	Python, https://github.com/bwtgroup/SSRTDC-modules/tree/master/THD	uses the baseline THD algorithm (Section 3)
De-noise Bags of Terms	Stop Terms Remover	the list of stop terms (regular noise); the folder with the bags of terms	the folder with the de-noised bags of terms	Pascal, https://github.com/bwtgroup/SSRTDC-modules/tree/master/STR	the list of stop terms has to be compiled manually after examining the last bag of terms in the sequence
Analyse Saturation - done manually					
Compare Results of Different Experiments - done manually					

The **Preparatory** phase of our experimental workflow is supported by the following three software modules.

Catalogue Generator. We found out in the pre-implementation phase that developing a generic parser for creating the catalogue of the papers is not feasible due to the layout differences at different publisher resources / repositories. For example, the journal and proceedings pages, from which the information about the papers needs to be parsed, look very differently for Springer, IEEE, or ACM. Besides that, some document collections could only be available as the sets of PDF or Plain Text files. Therefore, we opted to develop tailored parsers for different cases. The parser for Springer journal pages has been developed. The parser takes a Springer Link journal web page URL as its input and stores the list of all the papers of this journal in the specified .csv file[6]. The information about a paper contains all its reference (metadata) information, the abstract, and the no of citations acquired from Google Scholar. We have also developed the parser for a collection of PDF

[6] The catalogues of the acquired journal papers for the KM collection in .XLSX format are available at: https://github.com/bwtgroup/SSRTDC-PaperCatalogues/. The data has been collected on December 3-4, 2016.

document files, which extracts some metadata from the documents and acquires the numbers of citations from Google Scholar.

Full Text Downloader. For downloading the full texts of the papers, another software module has been developed. It receives a .csv list of the documents to be downloaded and downloads the full texts of these documents based on their DOI information taken from the catalogue. The papers in PDF are stored in a folder specified as a parameter. The PDF files are named, using the information from the catalogue, as follows: <journal_ID>+"-"+<year>+<vol>+"("+<issue>+")-("+<pages>+")-"+<DOI>+".pdf"

PDF to Plain Text Convertor. This software module performs batch conversions of the full texts in PDF to plain text format. It gets a path to the directory where PDF documents are stored, as a parameter. It produces the outputs for each input file in plain text (ANSI) format in which hyphenations are removed; ligatures are substituted by appropriate letter combinations, and each sentence occupies a separate line for better term extraction.

The **Pre-processing** phase of our experimental workflow is supported by the **Dataset Generator** module. This module takes the following inputs:

1. **in_txt_dir** – the name of the folder containing the TXT documents for forming the datasets. It assumes that the text files in this folder are named using the following convention: <journal_ID>+"-"+<year>+<vol>+"("+<issue>+")-("+<pages>+")-"+<DOI>+".txt". Hence, the information about the time of publication (timestamp) is encoded in the name of a file: <year>+<issue>.
2. **out_dataset_dir** – the name of the folder to store the generated datasets
3. **strategy** – the order in which the documents are picked for adding to datasets. Four different values are possible: (i) "**time-asc**" for the chronological order; (ii) "**time-desc**" for reversed-chronological order; (iii) "**time-bidir**" for bi-directional order; (iv) "**random**" for picking the documents randomly; and (v) "**citation-desc**" for picking the documents in the descending order of citations.
4. **increment_size** – the number of papers to be included in a dataset increment
5. [**citations**] – the name of the XLS file containing the information about the numbers of citations

The datasets are formed following the OntoElect procedure described in Section 3. The documents are added following the specified order.

The **Terms Extraction** phase is supported by the **UPM Term Extractor**.

UPM Term Extractor has been developed in the Dr Inventor project. The tool takes an English (PDF or plain text) corpus of documents and returns the bag of extracted terms as a CSV file. Each term is provided in a separate line with its *c-value*.

The **Post-processing** phase is supported by the **Baseline THD** and **Stop Terms Remover** modules.

The **Baseline THD module** have been developed in Python to implement the THD

algorithm for the input bags of terms taken from the UPM Term Extractor. The module processes the sequence of the pairs of the bags of terms as presented in Section 3. The bags of terms have to be stored in separate CSV files. The list of the files to be processed is taken from the "list.txt" configuration file.

Finally, the **Stop Term Remover** module has been developed to lower the effort needed to remove the set of manually selected stop terms from the datasets. It takes the list of the manually selected terms in a plain text input file and deletes all these terms from the bags of terms extracted by UPM Term Extractor.

Only two modules in the toolset are constrained by some specifics in data. The **Catalogue Generator – Springer Journals** is tailored to **Springer Link journal pages** – so it is document collection dependent. **UPM Term Extractor** takes in only **English texts**. The rest of the software modules can be used to process any document collection, coming from an arbitrary domain, and in any language.

5.3 Experiments: Objectives and Set-up

This report covers our study of the influence of the order of adding documents to datasets on terminological saturation. This influence has been evaluated experimentally using three different document collections, DMKD, TIME, and DAC, described in detail in Section 4.3. Our objective was to find out if there is a particular order that results in detecting terminological saturation quicker, therefore on a more compact sub-collection of the entire document collection.

The experimental workflow, described in Section 5.1, and software toolset, presented in Section 5.2, supporting the workflow have been used.

Overall:

- The **Preparatory** phase has been executed three times – one per document collection
- The **Pre-processing** and **Terms Extraction** phases have been executed fifteen times – one per an order of adding documents within a collection
- The **Post-processing** phase has been executed twenty times – fifteen times for different orders and, additionally, five times after removing stop terms in the DAC collection

In assessing the impact of an order of adding documents, we analysed the following measurable aspects:

- Which of the orders resulted in the earlier or later entry into the saturation zone (*thd* lower than *eps*)
- Which of the orders resulted in, integrally, the smallest No of retained terms
- Which of the orders resulted in, integrally, the highest individual term significance thresholds *eps*
- Which of the orders resulted in, integrally, the highest proportions of all extracted to retained terms
- Which of the orders resulted in, integrally, the least volatile *thd* values

In assessing the sensitivity of an order of adding documents to excessive noise, we based the comparison on the following measurable aspects:

- Which of the orders resulted in the earlier or later *eps* peak
- Which of the orders resulted in the higher or lower *eps* peak
- Which of the orders resulted in the earlier or later *thd* peak
- How many measurement steps were taken to confirm excessive noise after *eps* peak
- Which of the orders resulted in the earlier or later no-of-retained-terms peak
- How many measurement steps were taken to confirm excessive noise after no-of-retained-terms peak

All the computations, except terms extraction, have been run using a Windows 7 64-bit PC with: Intel® Core™ i5 CPU, M520 @ 2.40 GHz; 8.0 Gb on-board memory; NVIDIA Geforce GT330M GPU. For terms extraction we used an Ubuntu 16.04LTS PC with: Intel® Xeon® CPU E5-2603 v3 @ 1.60GHz 12 cores, 128G on-board memory and 3.5T hard disk.

5.4 Document Collections and Datasets

The experiments have been performed on the three high quality real documents collections coming from different domains. In this section, we describe the data used in our experiments. These data come from three real document collections – TIME, DMKD, and DAC.

These document collections are all composed of the high quality papers published at the peer-reviewed international venues in three different domains. The TIME collection contains the full text papers of the proceedings of the TIME Symposia series[7]. The DMKD collection contains the subset of full text articles from the Springer journal on Data Mining and Knowledge Discovery[8]. The DAC collection contains the subset of full text papers of the Design Automation Conference[9].

The domain of the TIME collection is Time Representation and Reasoning. The publisher of these papers is IEEE. This collection has been acquired in our previous research reported in [5]. The complete TIME collection contains all the papers published in the TIME symposia proceedings between 1994 and 2013, which are 437 full text documents in total. The papers of the TIME collection have been processed manually, including their conversion to plain texts and cleaning of these texts. Therefore, the resulting datasets were not very noisy. We have chosen the increment for generating the datasets to be 20 papers. So, based on the available texts, we have generated 22 incrementally enlarged datasets D_1, D_2, …, D_{22}[10].

---

[7] http://time.di.unimi.it/TIME_Home.html
[8] https://link.springer.com/journal/10618
[9] http://dac.com/
[10] **TIME** collection in plain texts: http://dx.doi.org/10.17632/knb8fgyr8n.1#folder-d1e5f2b6-c51e-4572-b10d-0e2ebccead02; incrementally enlarged datasets generated of these texts: http://dx.doi.org/10.17632/knb8fgyr8n.1#folder-06cc93de-de58-43b9-a378-743b61ef9093

The domain of DMKD collection is Data Mining and Knowledge Discovery, which falls into our broader target domain of Knowledge Management as its essential part. This collection is provided by Springer based on their policy on full text provision for data mining purposes[11]. To the DMKD document collection, we have included 300 papers published in the journal on Data Mining and Knowledge Discovery between 1997 and 2010. All the papers in their full texts were automatically processed using our instrumental pipeline presented in Section 5. In difference to the TIME collection, no manual cleaning of document texts was applied. For generating the datasets, the increment has been chosen to be 20 papers. Therefore, based on the available documents we have generated 15 incrementally enlarged datasets D_1, D_2, …, D_{15}[12].

The domain of the DAC collection is Engineering Design Automation. The publisher of these papers is IEEE. For the collection, we have chosen 506 papers published between 2004 and 2010. The papers of the DAC collection have been automatically converted to plain text using our instrumental software. Similarly to TIME, we have chosen the increment for generating the datasets to be 20 papers. So, based on the available texts, we have generated 26 incrementally enlarged datasets D_1, D_2, …, D_{26}[13]. We deliberately skipped manual cleaning of the plain texts to be able to compare the results between very noisy bags of terms (without stop term removal) and not very noisy bags of terms (after stop term removal), similarly to what has been done for the Random Articles from Wikipedia (RAW) collection in [1].

The characteristics of all the three document collections and datasets are summarized in Table 5 below.

Table 5: The features of the used document collections and datasets

Collec-tion	Type	Paper Type and Layout	No Doc	Noise	Processing	Inc	No Datasets
TIME	real	conference, IEEE 2-column	437	manually cleaned	manual conversion to plain text, automated dataset generation	20 papers	22
DMKD	real	journal, Springer 1-column	300	not cleaned, moderately noisy	automated	20 papers	15
DAC	real	conference, IEEE 2-column	506	not cleaned, quite noisy	automated	20 papers	26

[11] https://www.springer.com/gp/rights-permissions/springer-s-text-and-data-mining-policy/29056

[12] **DMKD** collection in plain texts: http://dx.doi.org/10.17632/knb8fgyr8n.1#folder-637dc34c-fa29-4587-9f63-df0e602d6e86; incrementally enlarged datasets generated of these texts: http://dx.doi.org/10.17632/knb8fgyr8n.1#folder-9043e879-5f77-486f-bc56-cb6af3cdd306

[13] **DAC** collection in plain texts: http://dx.doi.org/10.17632/knb8fgyr8n.1#folder-010b1add-cd5c-4b33-b6ce-8c93301d880b; incrementally enlarged datasets generated of these texts: http://dx.doi.org/10.17632/knb8fgyr8n.1#folder-e07d55cc-6830-4d9e-9b90-d69010050508

6 Experimental Results and Discussion

In this section, we report the results of our experiments on the datasets generated from all the three data collections and using the five different orders of adding documents. We also discuss these results. The experiments have been set and performed using the workflow, instrumental tools, and data presented in Section 5.

In the experiment with each collection we:

- Extracted the bags of terms from the prepared datasets using the UPM Term Extractor software
- Measured terminological saturation in the pairs of the extracted bags of terms using the Baseline THD module
- Built the diagrams and analysed the results

In addition to the above activities, for the DAC collection we also looked at the effect of removing stop terms after doing term extraction. By removing these stop terms, which represented the noise in the pre-processed documents, we de-noised the output. The lists of the stop terms were prepared manually based on the extractions from the last (largest) dataset D_{26}. These stop terms were further automatically removed from all the datasets using our Stop Term Remover module. Hence, for the DAC collection we also compared noisy and cleaned bags of terms for all the orders.

In all the experiments, we also computed two additional characteristics that further helped us analyse the influence of order aspect on terminological saturation. These were:

- The **proportion** of the **retained** terms to **all extracted** terms in percent (**% Retained Terms** column)
- The **volatility** of *thd* – a discrete analogue to the 1st derivative to *thd* values – computed as the different between the current and previous *thd* values (**thd volatility** column)

All the results of measuring individual term significant thresholds (*eps*), retained terms numbers and ratios, terminological differences (*thd*, *thdr*), and *thd* volatility in these experiments are available as a apart of our public OntoElect Dataset[14].

6.1 Experiments on DMKD

For the datasets extracted from the DMKD document collection, the results look as follows.

We processed at the bags of terms extracted from the dataset sequences generated for all five orders: chronological, reverse-chronological, bi-directional, random, and descending citation frequency. The results of measuring individual term significant thresholds (*eps*), retained terms numbers and their ratios to all extracted terms, terminological differences (*thd*, *thdr*), and *thd* volatility are presented in the

[14] http://dx.doi.org/10.17632/knb8fgyr8n.1#folder-348f6201-7a55-4c96-a67a-47ec54c9d558

saturation measurement analysis Tables 6 – 10. The measurements indicating the entries into the saturation areas are bolded.

Table 6: Terminological saturation measurements on DMKD datasets – **chronological** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	23957	9.5098	1480	6.1777	35.8837	35.8837	100.0000
*D*2-*D*1	49334	14.0000	2212	4.4837	27.6977	-8.1860	57.5551
*D*3-*D*2	67913	15.5000	2770	4.0787	17.0595	-10.6382	31.6816
*D*4-*D*3	89617	**17.0000**	**3242**	**3.6176**	**16.2008**	-0.8587	27.8074
*D*5-*D*4	112286	19.0196	3770	3.3575	13.2253	-2.9755	21.3724
*D*6-*D*5	130147	20.0000	4113	3.1603	10.3880	-2.8373	15.7243
*D*7-*D*6	147162	22.0000	4448	3.0225	9.9962	-0.3918	14.0580
*D*8-*D*7	164007	23.7744	4666	2.8450	8.8620	-1.1342	12.0373
*D*9-*D*8	182192	24.0000	5190	2.8486	9.4518	0.5898	12.1387
*D*10-*D*9	200840	26.5000	4986	2.4826	9.4208	-0.0310	12.3912
*D*11-*D*10	230283	28.5293	5709	2.4791	12.0487	2.6279	15.2067
*D*12-*D*11	250739	36.0000	4825	1.9243	15.5742	3.5255	17.5651
*D*13-*D*12	275270	36.0000	5285	1.9199	8.5915	-6.9827	9.5324
*D*14-*D*13	298786	37.0000	5503	1.8418	6.5399	-2.0516	6.9874
*D*15-*D*14	320025	38.0391	5800	1.8124	7.5363	0.9964	7.7269

Table 7: Terminological saturation measurements on DMKD datasets – **reversed-chronological** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	33272	12.0000	1701	5.1124	126.2163	126.2163	100.0000
*D*2-*D*1	63567	15.5098	2562	4.0304	83.9106	-42.3056	69.2365
*D*3-*D*2	93654	16.5000	3344	3.5706	53.5087	-30.4019	46.1484
*D*4-*D*3	115284	29.5000	2367	2.0532	56.9859	3.4772	41.5804
*D*5-*D*4	145109	31.0196	3109	2.1425	38.7499	-18.2359	28.1125
*D*6-*D*5	164943	33.2193	3400	2.0613	37.0169	-1.7330	32.5304
*D*7-*D*6	182768	**33.2842**	**3774**	**2.0649**	**18.9340**	-18.0829	17.6584
*D*8-*D*7	200575	34.0000	4070	2.0292	13.6271	-5.3069	13.0017
*D*9-*D*8	215651	34.0000	4407	2.0436	9.9870	-3.6402	9.3644
*D*10-*D*9	232629	34.8289	4617	1.9847	12.0437	2.0567	11.8396
*D*11-*D*10	253574	35.5000	4919	1.9399	9.7006	-2.3430	9.5723
*D*12-*D*11	271979	36.0000	5257	1.9329	9.2925	-0.4081	9.2286
*D*13-*D*12	287495	36.0000	5552	1.9312	7.6252	-1.6673	7.6833
*D*14-*D*13	305907	36.5000	5723	1.8708	6.8389	-0.7863	6.9056
*D*15-*D*14	319449	38.0000	5917	1.8523	6.9907	0.1518	7.2115

Table 8: Terminological saturation measurements on DMKD datasets – **bi-directional** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	28345	9.5098	1838	6.4844	63.7814	63.7814	100.0000
*D*2-*D*1	55707	14.0000	2483	4.4572	47.1538	-16.6275	62.4848
*D*3-*D*2	77313	15.5098	3128	4.0459	29.1141	-18.0397	33.4804
*D*4-*D*3	104312	17.0000	3614	3.4646	28.6699	-0.4442	34.1047
*D*5-*D*4	128426	19.0196	4228	3.2922	19.4916	-9.1783	22.7516
*D*6-*D*5	148353	19.6515	4726	3.1856	15.1394	-4.3522	17.3759
*D*7-*D*6	169973	21.0000	5021	2.9540	12.7482	-2.3911	14.2898
*D*8-*D*7	189926	28.5293	4188	2.2051	24.7463	11.9980	24.6445

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*9-*D*8	213503	30.0000	4532	2.1227	12.6361	-12.1102	12.3758
*D*10-*D*9	305956	22.0000	8099	2.6471	45.7049	33.0688	43.9035
*D*11-*D*10	319441	**24.0000**	**8054**	**2.5213**	**13.9341**	-31.7709	13.8163
*D*12-*D*11	319450	28.0000	7122	2.2295	5.7703	-8.1638	5.7409
*D*13-*D*12	319463	31.0196	6622	2.0729	14.4390	8.6687	14.1818
*D*14-*D*13	319463	34.0000	6191	1.9379	5.2553	-9.1837	5.2549
*D*15-*D*14	319463	38.0000	5917	1.8522	5.9141	0.6589	6.1024

Table 9: Terminological saturation measurements on DMKD datasets – **random** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	29189	11.6096	1576	5.3993	62.7723	62.7723	100.0000
*D*2-*D*1	51950	14.0000	2365	4.5525	35.0536	-27.7187	49.7339
*D*3-*D*2	75163	16.0000	3004	3.9966	28.0942	-6.9595	36.8698
*D*4-*D*3	103367	19.0196	3478	3.3647	23.8669	-4.2273	27.8512
*D*5-*D*4	124331	**20.5000**	**3739**	**3.0073**	**17.9221**	-5.9448	20.8018
*D*6-*D*5	143485	22.0000	4289	2.9892	13.3707	-4.5513	14.8096
*D*7-*D*6	165875	33.2193	3267	1.9696	28.8455	15.4747	29.5475
*D*8-*D*7	189043	33.2193	3846	2.0345	13.4110	-15.4344	13.2340
*D*9-*D*8	205640	34.0000	4092	1.9899	13.9783	0.5673	14.6923
*D*10-*D*9	223054	34.5000	4366	1.9574	9.6418	-4.3365	10.2040
*D*11-*D*10	242435	35.5000	4683	1.9317	8.1446	-1.4971	8.3633
*D*12-*D*11	257940	36.0000	4974	1.9284	7.3427	-0.8020	7.5166
*D*13-*D*12	279642	36.0000	5355	1.9149	7.6851	0.3424	7.8558
*D*14-*D*13	298073	36.0000	5668	1.9015	6.7205	-0.9646	6.8720
*D*15-*D*14	319428	38.0000	5916	1.8521	7.9479	1.2274	8.2001

Table 10: Terminological saturation measurements on DMKD datasets – **descending citation frequency** order of adding documents

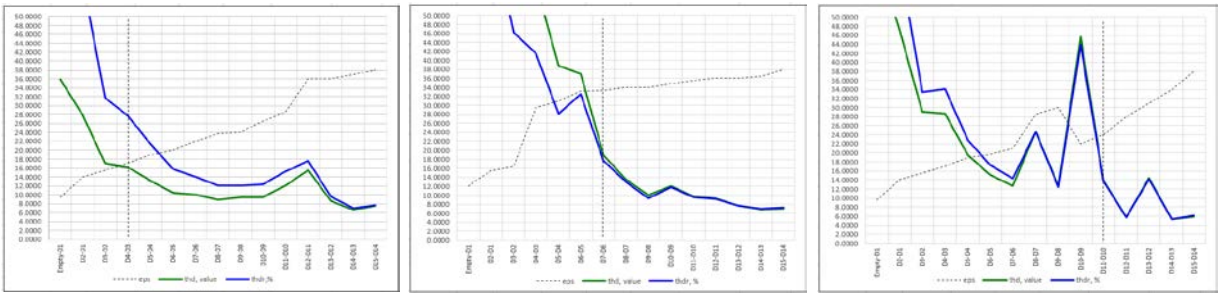
Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	39132	12.0000	1864	4.7634	62.8927	62.8927	100.0000
*D*2-*D*1	63676	15.5000	2537	3.9842	31.5074	-31.3853	43.6818
*D*3-*D*2	85596	18.0000	3022	3.5305	23.3825	-8.1249	32.9367
*D*4-*D*3	109076	**19.6515**	**3399**	**3.1162**	**18.0366**	-5.3458	27.2375
*D*5-*D*4	135006	23.2193	3983	2.9502	15.8003	-2.2363	21.4395
*D*6-*D*5	152777	24.0000	4279	2.8008	10.0685	-5.7318	13.7836
*D*7-*D*6	173655	24.0000	4954	2.8528	10.6702	0.6018	13.5940
*D*8-*D*7	193998	26.0000	5020	2.5877	9.1376	-1.5326	11.3216
*D*9-*D*8	215213	28.0000	5345	2.4836	8.6805	-0.4571	10.3356
*D*10-*D*9	237382	28.5293	5656	2.3827	7.5625	-1.1180	8.6988
*D*11-*D*10	258522	28.5293	6177	2.3894	7.3048	-0.2578	8.1332
*D*12-*D*11	275178	28.5293	6625	2.4075	6.5320	-0.7728	6.9789
*D*13-*D*12	291482	30.0000	6661	2.2852	6.4337	-0.0983	6.8160
*D*14-*D*13	306387	38.0000	5685	1.8555	15.4687	9.0351	15.8708
*D*15-*D*14	313506	38.0000	5815	1.8548	3.5291	-11.9396	3.6601

These measurements are visualized in Fig. 5 – 8. The dashed vertical lines point at the *thd* measurement in which saturation has been observed as *thd* went below *eps* and did not go up above it afterwards. The corresponding values of *eps*, no of retained terms, ratios of retained terms to the total numbers of terms in the bags of terms, and *thd* for these bags of terms are bolded in Tables 6 – 10. As it may be

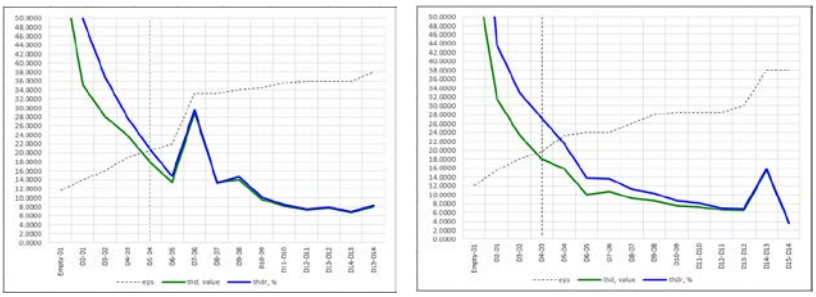
noticed in Tables 6 – 10 and Fig. 5 – 6, all orders yielded stable terminological saturation on DMKD, but at different measurement points. The first to reach saturation were the chronological and descending citation frequency orders for which the absolute terminological difference measures (*thd*) went below the individual terms significance threshold curve (eps) already at *D*4-*D*3 – see Tables 6 and 10, and Fig. 5(a) and (e). The next order was random at *D*5-*D*4 – see Table 9 and Fig. 5(d). It has been followed by reversed-chronological at *D*7-*D*6 – see Table 7 and Fig. 5(b). The last one was bi-directional at *D*11-*D*10 – see Table 8 and Fig. 5(c). The measures at saturation points for all the five evaluated orders are provided in Table 11 for comparison.

Table 11: DMKD. The comparison of saturation measurements for all orders at their saturation points.

Order	Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
Chrono	*D*4-*D*3	89617	**17.0000**	**3242**	**3.6176**	**16.2008**	-0.8587	27.8074
Random	*D*5-*D*4	124331	**20.5000**	**3739**	**3.0073**	**17.9221**	-5.9448	20.8018
Rev-chrono	*D*7-*D*6	182768	**33.2842**	**3774**	**2.0649**	**18.9340**	-18.0829	17.6584
Bi-dir	*D*11-*D*10	319441	**24.0000**	**8054**	**2.5213**	**13.9341**	-31.7709	13.8163
DCF	*D*4-*D*3	109076	**19.6515**	**3399**	**3.1162**	**18.0366**	-5.3458	27.2375



(a) Chronological (b) Reversed-chronological (c) Bi-directional



(d) Random (e) Descending citations frequency

Fig. 5: DMKD. Terminological difference measures for different individual orders (a–e) of adding documents to datasets.

As it could be noticed in Tables 6 – 10, and in Fig. 6, the individual term significance thresholds grow monotonically for the reversed-chronological order and reach much higher values both earlier in the measurement process and at the respective saturation point. This indicates that the reversed-chronological order might lead to more stable

saturation and cuts-off relatively more insignificant terms than the other orders. The curve for the descending citation frequency order is however integrally the least volatile in values and fastest growing up to its saturation point at *D*4-*D*3, which also might indicate saturation stability. This expectation for the reversed chronological order is also supported by its ratio of retained to all terms, which is the lowest at saturation points – see Table 11.

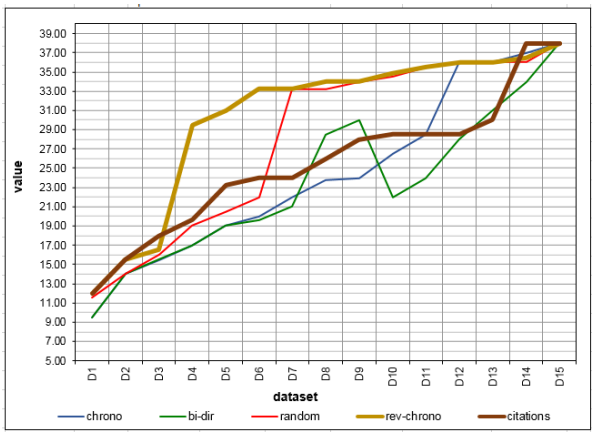


Fig. 6: DMKD. Individual term significance thresholds (*eps*) for different orders of adding documents to datasets (a–e in Fig. 5). The curve for **reversed-chronological** order is rendered **thicker** as it has the highest values. The curve for **descending citation frequency** order is rendered **thicker** as it has the least volatile values.

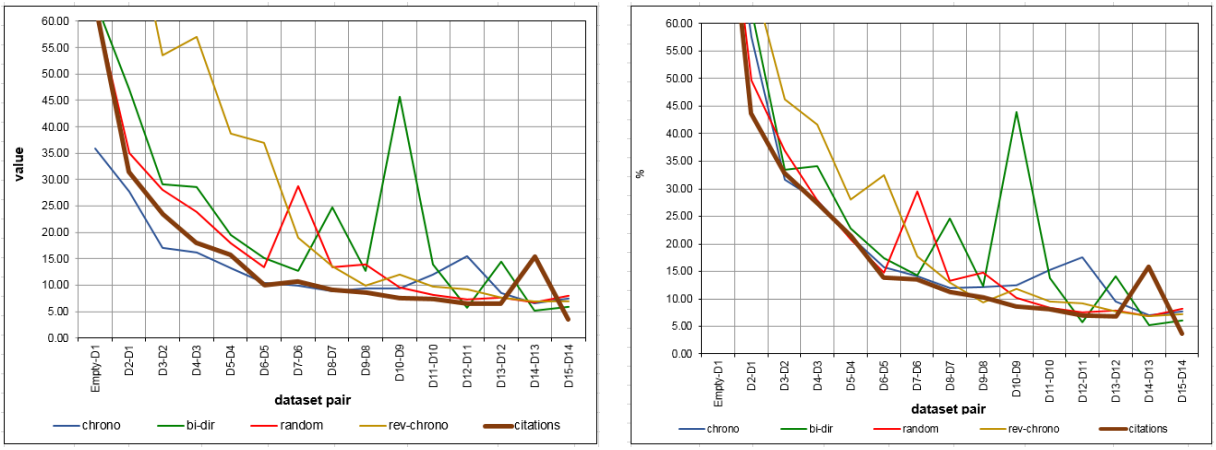
Figure 7 pictures the *thd* (a) and *thdr* (b) curves for different orders, grouped together for comparison. It could be seen in Fig. 7(a) that the *thd* curve for the descending citation frequency enters the saturation area together with the chronological order, but has a higher *thd* value at this point. In the later measurements, however, for example in the D_{11}-D_{10} to D_{13}-D_{12} area, the *thd* values for the chronological order are substantially more volatile. Therefore, the stability of the descending citation frequency order is higher in terms of terminological difference. The *thd* value for DCF peaks only at *D*14-*D*13, 10 points after entering the saturation zone, and does not go above *eps*. This peak might be an indicator of excessive accumulated noise, to which DCF appears to be the least sensitive.

The third fastest order to reach its saturation point is reversed chronological. The *thd* curves for the bi-directional and random orders reach the area of saturation later than the reversed-chronological curve. Their values also oscillate much more than the values in the reversed-chronological order curve.

Fig. 8 pictures how different orders relate to each other by the numbers of retained terms and the proportions of retained to all extracted terms. As it could be seen in in Table 11 and Fig. 8(a), the descending citation frequency curve retains the second lowest number of terms[15] at saturation point (*D*3), but yields the lowest ratio of

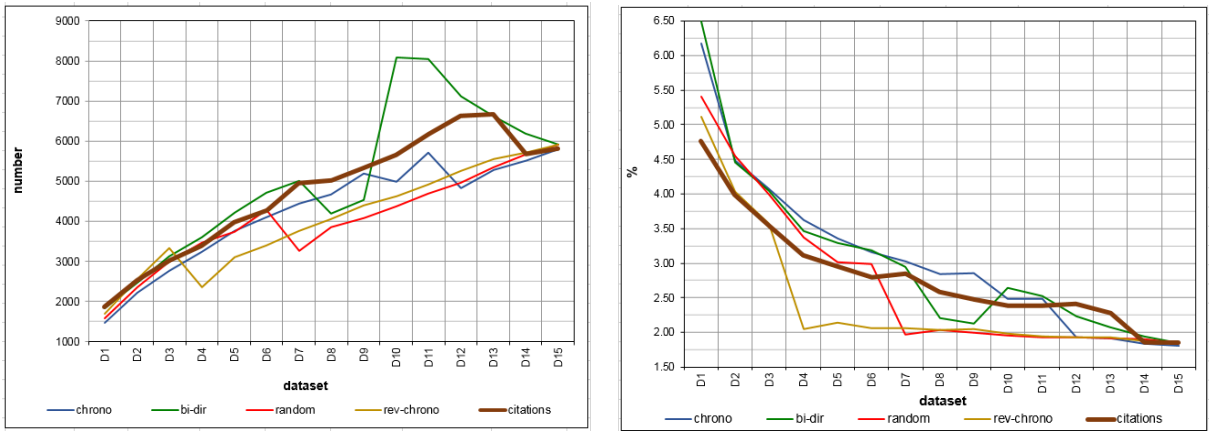
[15] In the context of our research objectives, the lower the number of retained terms is the better. Indeed, a smaller quantity of retained terms, if saturated, results in a more compact, though representative ranked list of terms. The more compact it is, the less laborious is its further processing.

retained to all extracted terms (Fig. 8(b)) due to the value of eps which is higher than the one reached by the chronological order. However, Fig. 8(b) clearly pictures that the reversed-chronological order is the first to reach the stable ratio of retained to all extracted terms (of ~2 percent) already at *D*4, which is the saturation point for the chronological order. The lowest ratio at saturation point is reached by the reversed chronological order. Random at and bi-directional orders are too much volatile in their behaviour, so – unstable. at *D*13. Judging by the shapes of the curves in Fig. 8, DCF is the most stable order.



(a) *thd* measures (b) *thdr* measures

Fig. 7: DMKD. The comparison of absolute (*thd*) and relative (*thdr*) terminological difference measures for different orders of adding documents to datasets (a–e in Fig. 5). The curve for **descending citation frequency** order is rendered **thicker** as the smoothest, indicating the most stable terminological saturation, and the one having, integrally, the lowest *thd* values.



(a) Numbers of retained terms (b) Ratios of retained to all terms

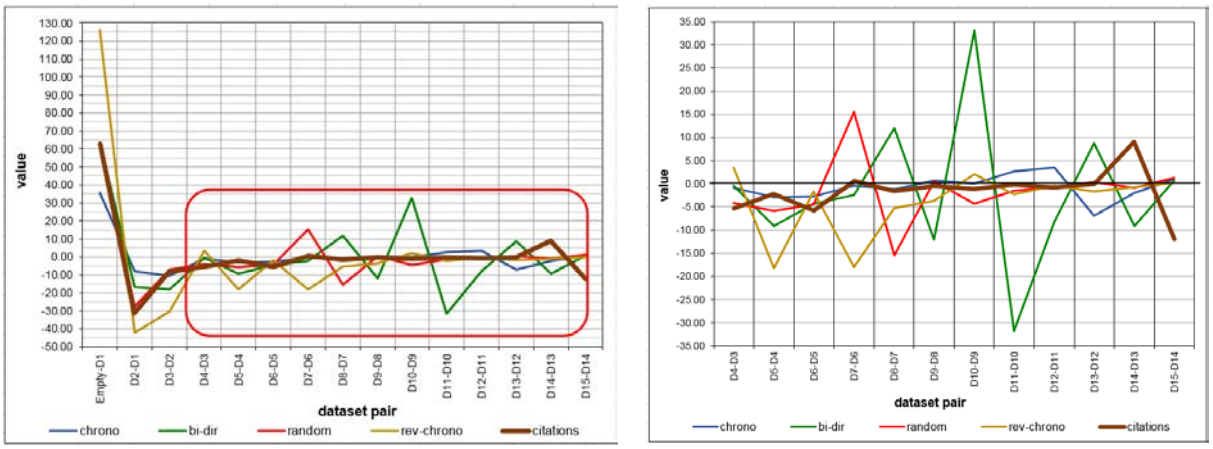
Fig. 8: DMKD. The numbers and ratios of retained terms for different orders of adding documents to datasets (a–e in Fig. 5). The curve for **descending citation frequency** order is rendered **thicker** because this order is the most stable in terms of retained significant terms and their ratios to all extracted terms.

Perhaps, the integral view on the stability of different orders with respect to terminological saturation is best given in the *thd* volatility diagram of Fig. 9. Fig. 9(a) covers the whole set of the measured volatility values, while Fig. 9(b) presents in finer detail the area within the rounded rectangular in Fig. 9(a). It can be

noticed in the figure that the oscillations of volatility values for the random and bi-directional orders are too big. Therefore, these orders cannot be considered as stable regarding terminological saturation.

Integrally, the least *thd*-volatile order is DCF. The comparison of the volatility curves of the chronological versus reversed-chronological order reveals one interesting observation. The chronological order curve demonstrates relatively low volatility in initial *thd* measurements, including the saturation point. However, in later measurements, volatility increases quite noticeably. This can be interpreted as an indicator pointing to the insufficient stability of the chronological order to the accumulation of excessive regular noise.

On the contrary, the reversed chronological order is moderately *thd*-volatile in the initial measurements, including its saturation point (higher than chronological but lower than random and bi-directional) at the initial *thd* measurements. It then persistently converges to smaller and smaller volatilities in the saturation area. Therefore, in terms of the stability of terminological saturation the reversed-chronological method demonstrates better results compared to random and bi-directional. The volatility curve of the random order in resembles the reversed-chronological curve but with sharper oscillations in the beginning of the saturation area (*D*6-*D*5 – *D*10-*D*9). The bidirectional order is the highest *thd*-volatile.



(a) (b)

Fig. 9: DMKD. The comparison of *thd* volatility for different orders of adding documents to datasets (a–e in Fig. 5). Diagram (b) pictures the fragment within the rounded rectangle in diagram (a) in finer detail. The curve for **descending citation frequency** order is rendered **thicker** because this order has the least volatile *thd* and, therefore, results in the most stable saturation.

For summarizing the analysis of our results on the DMKD collection, it may be stated that:

- The **bi-directional** order of adding documents to datasets clearly demonstrated the **worst performance** regarding terminological saturation. This conclusion is supported by: the latest entrance into the saturation area (Fig. 7); the lowest individual term significance thresholds (Fig. 6); the highest numbers of retained terms and ratios of retained to all extracted terms (Fig. 8); the highest volatility of *thd* values (Fig. 9).

- The **random** order of adding documents to datasets demonstrated **balanced performance** and **noticeable instability** regarding terminological saturation. This conclusion is supported by: rather an early entrance into the saturation area (Fig. 7); the third highest individual term significance thresholds (Fig. 6), comparable with DCF; integrally the lowest numbers of retained terms, similar to reversed chronological, though not converging to the stable saturated ratio of retained to all terms (of ~2 percent) as quickly as the reversed chronological curve (Fig. 8); a sharp oscillating *thd* volatility curve (Fig. 9).
- The **chronological** order of adding documents to datasets demonstrated **balanced performance** for the entire range of the datasets and **noticeable instability** within the saturation area. This conclusion is supported by: the earliest entrance into the saturation area (Fig. 7); the lowest individual term significance thresholds (Fig. 6), at its saturation point and integrally; the third highest numbers of retained terms in the proximity of the saturation point and the slowest convergence to the stable saturated ratio of retained to all terms (Fig. 8); a substantially sharper oscillating *thd* volatility curve, compared to DCF (Fig. 9).
- The **reversed-chronological** order of adding documents to datasets demonstrated **integrally the second best** and **balanced performance** compared to all the other evaluated orders. This conclusion is supported by: the timely entrance into the saturation area (Fig. 7); integrally the highest individual term significance thresholds (Fig. 6); integrally the second lowest numbers of retained terms and a remarkably quick convergence to the stable saturated ratio of retained to all terms (Fig. 8); the second smoothest *thd* volatility curve with decreasing values on the whole range of thd measurements (Fig. 9).
- The **descending citation frequency** order of adding documents to datasets demonstrated **integrally the best** and **most balanced performance** compared to all the other evaluated orders. This conclusion is supported by: the earliest entrance into the saturation area (Fig. 7); the least volatile in values and fastest growing up to its saturation point regarding individual term significance thresholds (Fig. 6); integrally being the most stable in terms of retained significant terms and their ratios to all extracted terms (Fig. 8); the least *thd*-volatile on the whole range of measurements (Fig. 9).

Hence, it could be stated, based on the **DMKD** experiments, that the **descending citation frequency** order is **the best** to yield **quicker** and more **stable** terminological saturation.

6.2 Experiments on TIME

For the datasets extracted from the TIME document collection, the results look as follows.

We processed at the bags of terms extracted by UPM Extractor from the dataset sequences generated for all five orders: chronological; reverse-chronological; bi-

directional; random; and descending citation frequency. The results of measuring individual term significant thresholds (*eps*), retained terms numbers and ratios, terminological differences (*thd*, *thdr*), and *thd* volatility are presented in the saturation measurement analysis Tables 12 – 16. The measurements indicating the entries into the saturation areas are bolded.

These measurements are visualized in Fig. 10 – 13. The dashed vertical lines in Fig. 10 point to the *thd* measurement in which saturation has been observed as *thd* went below *eps* and did not go up above it. The corresponding values of *eps*, no of retained terms, ratios of retained terms to the total numbers of terms in the bags of terms, and *thd* for these bags of terms are bolded in Tables 12 – 16. As it may be noticed in Tables 12 – 17 and Fig. 10, all orders yielded stable terminological saturation on TIME, but at different measurement points. The first to reach saturation was the reversed-chronological order for which the absolute terminological difference measures (*thd*) went below the individual terms significance threshold curve (*eps*) at *D*8-*D*7 – see Table 13 and Fig. 10(b). The second best was the descending citation frequency order for which the absolute terminological difference measures (*thd*) went below the individual terms significance threshold curve (*eps*) already at *D*9-*D*8 – see Table 16 and Fig. 10(e). These two best performing orders were followed by the chronological and random at *D*12-*D*11 – see Tables 12, 15 and Fig. 10(a, d).

Table 12: Terminological saturation measurements on TIME datasets – **chronological** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	53478	28,0000	1379	2,5786	112,2408	112,2408	100,0000
*D*2-*D*1	91701	24,0000	2473	2,6968	72,4258	-39,8150	59,3896
*D*3-*D*2	114061	21,5000	3028	2,6547	24,2654	-48,1604	17,3121
*D*4-*D*3	129896	19,6515	3997	3,0771	32,8794	8,6139	20,2957
*D*5-*D*4	145796	19,6515	4466	3,0632	32,6222	-0,2571	17,8096
*D*6-*D*5	162746	20,0000	4587	2,8185	44,6462	12,0240	27,0271
*D*7-*D*6	190263	21,0000	5133	2,6978	38,0715	-6,5747	24,0767
*D*8-*D*7	200176	22,0000	5413	2,7041	26,8691	-11,2024	18,5984
*D*9-*D*8	217461	22,0000	5855	2,6924	18,7762	-8,0929	13,1105
*D*10-*D*9	245967	23,2193	6453	2,6235	26,9142	8,1381	18,2810
*D*11-*D*10	263034	24,0000	6428	2,4438	24,1645	-2,7497	16,6888
*D*12-*D*11	287887	**23,7744**	**7110**	**2,4697**	**18,1096**	-6,0550	12,7371
*D*13-*D*12	298367	23,7744	7383	2,4745	12,5737	-5,5358	9,1441
*D*14-*D*13	320500	24,0000	7723	2,4097	13,3350	0,7612	9,6244
*D*15-*D*14	333975	23,7744	8298	2,4846	14,4039	1,0690	10,6986
*D*16-*D*15	350741	24,0000	8426	2,4023	16,4281	2,0242	13,1356
*D*17-*D*16	369316	24,0000	8877	2,4036	9,6426	-6,7855	7,6385
*D*18-*D*17	389022	24,0000	9617	2,4721	11,4165	1,7739	8,7843
*D*19-*D*18	399553	24,0000	10005	2,5040	8,0421	-3,3744	6,1366
*D*20-*D*19	420464	24,0000	10574	2,5148	11,6557	3,6136	8,6524
*D*21-*D*20	435075	26,0000	9751	2,2412	9,7817	-1,8740	7,2973
*D*22-*D*21	449719	26,0000	10139	2,2545	6,9261	-2,8555	5,1092

Table 13: Terminological saturation measurements on TIME datasets – **reversed-chronological** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	25782	11,6096	1187	4,6040	72,8703	72,8703	100,0000
*D*2-*D*1	48330	16,0000	1920	3,9727	105,0817	32,2114	65,4299
*D*3-*D*2	69775	19,0000	2324	3,3307	107,4275	2,3457	122,8275
*D*4-*D*3	82981	19,6515	2759	3,3249	24,8418	-82,5856	22,3473
*D*5-*D*4	104720	23,0000	3252	3,1054	37,4318	12,5900	25,8496
*D*6-*D*5	123337	22,0000	3741	3,0332	22,9828	-14,4490	14,2555
*D*7-*D*6	143019	22,0000	4240	2,9646	50,0895	27,1067	38,1934
*D*8-*D*7	155715	**22,0000**	**4550**	**2,9220**	**20,1549**	-29,9347	16,4026
*D*9-*D*8	178556	22,0000	5108	2,8607	16,1536	-4,0013	12,5713
*D*10-*D*9	197655	21,0000	5568	2,8170	15,5612	-0,5924	12,4019
*D*11-*D*10	211023	22,0000	5815	2,7556	14,2711	-1,2900	12,0184
*D*12-*D*11	229671	23,7744	5890	2,5645	17,5596	3,2885	14,2635
*D*13-*D*12	258418	24,0000	6469	2,5033	17,6597	0,1001	13,7726
*D*14-*D*13	274916	24,0000	6832	2,4851	11,4580	-6,2018	9,0909
*D*15-*D*14	286870	24,0000	7131	2,4858	9,5748	-1,8832	7,7490
*D*16-*D*15	310664	24,0000	7885	2,5381	14,7069	5,1321	11,6869
*D*17-*D*16	329705	24,0000	8357	2,5347	11,3683	-3,3386	8,8473
*D*18-*D*17	338980	24,0000	8601	2,5373	10,1071	-1,2612	8,2525
*D*19-*D*18	358698	24,0000	9000	2,5091	7,7166	-2,3905	6,3152
*D*20-*D*19	371627	24,0000	9244	2,4874	6,9751	-0,7415	5,8371
*D*21-*D*20	409248	24,0000	10253	2,5053	10,6987	3,7236	8,3531
*D*22-*D*21	449718	26,0000	10137	2,2541	13,0843	2,3856	9,6525

Table 14: Terminological saturation measurements on TIME datasets – **bi-directional** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	36106	15,5098	1399	3,8747	98,3131	98,3131	100,0000
*D*2-*D*1	76462	24,0000	2154	2,8171	120,0323	21,7191	75,2510
*D*3-*D*2	103240	24,0000	2900	2,8090	73,4954	-46,5369	49,8514
*D*4-*D*3	135404	24,0000	3737	2,7599	54,4605	-19,0349	30,0058
*D*5-*D*4	152844	23,7744	4290	2,8068	32,4939	-21,9666	16,6469
*D*6-*D*5	176777	23,2193	4812	2,7221	36,8631	4,3693	18,3746
*D*7-*D*6	193005	23,2193	5323	2,7580	30,4235	-6,4396	14,7646
*D*8-*D*7	203900	23,2193	5627	2,7597	25,7468	-4,6767	13,1313
*D*9-*D*8	221621	23,2193	6089	2,7475	28,8725	3,1257	15,3694
*D*10-*D*9	239462	24,0000	6519	2,7224	40,1387	11,2662	22,6283
*D*11-*D*10	260401	24,0000	6978	2,6797	22,2248	-17,9140	12,7857
*D*12-*D*11	272824	24,0000	7364	2,6992	20,0006	-2,2242	12,0463
*D*13-*D*12	299125	24,0000	7821	2,6146	18,4918	-1,5088	11,6070
*D*14-*D*13	317468	24,0000	8417	2,6513	27,3097	8,8179	18,4483
*D*15-*D*14	329240	**24,0000**	**8723**	**2,6494**	**11,4488**	-15,8609	7,9220
*D*16-*D*15	338373	24,0000	8985	2,6554	12,1865	0,7377	8,7825
*D*17-*D*16	359213	24,0000	9402	2,6174	9,8107	-2,3758	7,0887
*D*18-*D*17	375384	24,0000	9935	2,6466	9,6494	-0,1613	6,8499
*D*19-*D*18	392956	24,0000	10217	2,6000	8,0828	-1,5666	5,8418
*D*20-*D*19	419273	24,0000	10990	2,6212	14,0746	5,9918	9,9795
*D*21-*D*20	428192	24,0000	11188	2,6128	6,4098	-7,6648	4,6255
*D*22-*D*21	449748	26,0000	10133	2,2530	11,2181	4,8083	8,2767

Table 15: Terminological saturation measurements on TIME datasets – **random** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	32965	9,5098	1566	4,7505	76,0049	76,0049	100,0000
*D*2-*D*1	60105	11,6096	2679	4,4572	66,2003	-9,8047	65,6513
*D*3-*D*2	77643	11,6096	3781	4,8697	49,3656	-16,8347	44,8633
*D*4-*D*3	103460	15,5098	3964	3,8314	59,9308	10,5652	48,6021
*D*5-*D*4	119993	15,5098	4744	3,9536	33,5940	-26,3368	28,2242
*D*6-*D*5	139986	16,0000	5062	3,6161	29,3439	-4,2501	23,6073
*D*7-*D*6	159249	16,0000	5717	3,5900	22,1217	-7,2222	18,2979
*D*8-*D*7	184785	17,5000	5988	3,2405	20,5853	-1,5363	16,9324
*D*9-*D*8	208303	19,0196	6339	3,0432	17,7630	-2,8223	13,9856
*D*10-*D*9	221770	19,0196	6810	3,0707	12,7057	-5,0573	9,9084
*D*11-*D*10	243994	20,0000	7139	2,9259	25,9328	13,2272	18,7526
*D*12-*D*11	252767	**22,0000**	**7129**	**2,8204**	**10,4441**	-15,4888	7,7871
*D*13-*D*12	277540	23,2193	7577	2,7301	18,6628	8,2187	14,2772
*D*14-*D*13	294762	22,0000	8160	2,7683	11,3405	-7,3223	8,7474
*D*15-*D*14	310047	23,2193	8352	2,6938	11,5158	0,1753	9,0984
*D*16-*D*15	336433	24,0000	8690	2,5830	13,1724	1,6566	9,9282
*D*17-*D*16	355677	24,0000	9121	2,5644	10,9545	-2,2178	8,0544
*D*18-*D*17	369876	24,0000	9506	2,5701	9,5206	-1,4340	7,0747
*D*19-*D*18	393997	24,0000	10080	2,5584	11,0838	1,5633	8,2072
*D*20-*D*19	407552	26,0000	9352	2,2947	14,4440	3,3601	10,6993
*D*21-*D*20	430390	26,0000	9746	2,2645	7,5803	-6,8637	5,5637
*D*22-*D*21	449685	26,0000	10137	2,2542	8,7302	1,1499	6,4404

Table 16: Terminological saturation measurements on TIME datasets – **descending citation frequency** order of adding documents

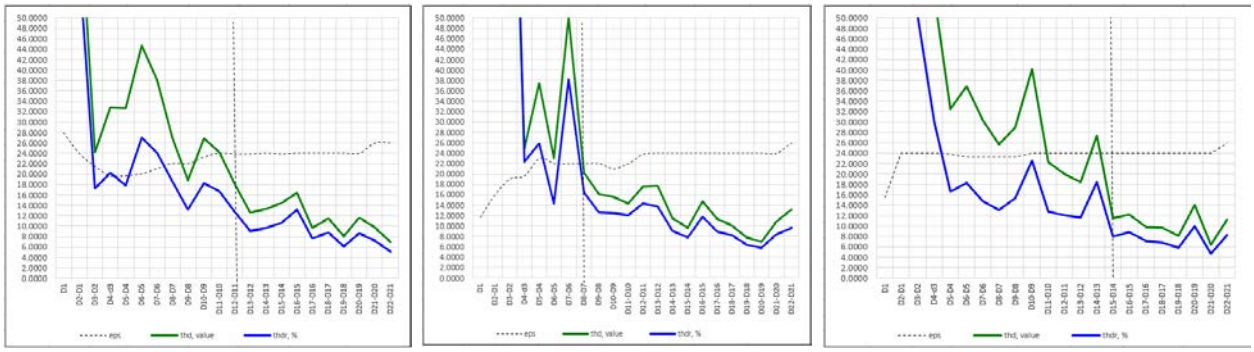
Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	45105	10.0000	2192	4.859772	169.2485	169.2485	100.0000
*D*2-*D*1	65281	11.6096	3210	4.917204	114.3297	-54.9188	76.2112
*D*3-*D*2	78974	14.2647	2817	3.566997	89.1144	-25.2153	87.4492
*D*4-*D*3	99178	15.5098	3549	3.578415	39.7380	-49.3764	29.0388
*D*5-*D*4	116311	16.0000	4142	3.561142	42.8644	3.1264	29.6232
*D*6-*D*5	133188	16.0000	4768	3.579902	31.9212	-10.9433	23.3479
*D*7-*D*6	150418	16.0000	5357	3.561409	23.1770	-8.7442	17.5910
*D*8-*D*7	162355	18.0000	5270	3.245973	19.6145	-3.5625	15.2524
*D*9-*D*8	174103	**19.0000**	**5493**	**3.155029**	**15.4377**	-4.1768	12.3733
*D*10-*D*9	185752	19.0196	5846	3.147207	16.9234	1.4857	14.3311
*D*11-*D*10	199958	19.6515	6004	3.002631	17.5711	0.6477	15.4221
*D*12-*D*11	213081	19.6515	6437	3.020917	10.4792	-7.0919	9.1463
*D*13-*D*12	227509	19.6515	6746	2.965157	8.7331	-1.7461	7.7909
*D*14-*D*13	236451	19.6515	7052	2.982436	9.7958	1.0627	9.0996
*D*15-*D*14	249585	19.6515	7440	2.980948	8.9930	-0.8029	8.4819
*D*16-*D*15	259850	19.6515	7703	2.964403	6.6584	-2.3346	6.4194
*D*17-*D*16	269191	19.6515	8033	2.984127	5.8903	-0.7681	5.6380
*D*18-*D*17	279687	19.6515	8329	2.977972	4.8594	-1.0308	4.6199
*D*19-*D*18	290545	19.6515	8609	2.963052	4.7192	-0.1402	4.4203
*D*20-*D*19	303238	20.0000	8431	2.780324	7.6192	2.8999	7.0529
*D*21-*D*20	308710	22.0000	8158	2.64261	4.3863	-3.2329	4.1531
*D*22-*D*21	315474	22.5000	8105	2.56915	5.1304	0.7441	4.8031

The worst performing was the bi-directional order at *D*15-*D*14 – see Table 14 and Fig. 10(c). The measures at saturation points for all the five evaluated orders are provided in Table 17 for comparison.

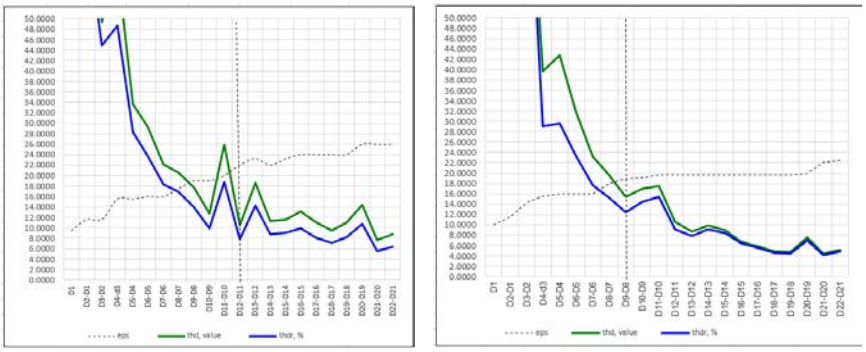
Table 17: TIME. The comparison of saturation measurements for all orders at their saturation points

Order	Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
Chrono	*D*12-*D*11	287887	**23,7744**	**7110**	**2,4697**	**18,1096**	-6,0550	12,7371
Random	*D*12-*D*11	252767	**22,0000**	**7129**	**2,8204**	**10,4441**	-15,4888	7,7871
Rev-chrono	*D*8-*D*7	155715	**22,0000**	**4550**	**2,9220**	**20,1549**	-29,9347	16,4026
Bi-dir	*D*15-*D*14	329240	**24,0000**	**8723**	**2,6494**	**11,4488**	-15,8609	7,9220
DCF	*D*9-*D*8	174103	**19.0000**	**5493**	**2.5692**	**15.4377**	-4.1768	12.3733

As it could be seen in Tables 12 – 17 and also in Fig. 11, and in difference to DMKD experiments, the individual term significance thresholds (*eps*) for bi-directional order reach the highest values both earlier in the measurement process and at the respective saturation point. The second highest eps values have been demonstrated by the reversed chronological order. This observation should have indicated that the bi-directional order leads to most stable saturation and cuts-off relatively more insignificant terms than the other orders, regarding TIME datasets. However, this is not true, as clearly demonstrated below. For example, the bi-directional order leads to retaining the biggest number of terms and reaching saturation later than the other orders (Table 17). The descending citation frequency order, though results in integrally the lowest *eps*, yields the best saturation indicators – as presented below. The second best in saturation indicators is the reversed chronological order.



(a) Chronological (b) Reversed chronological (c) Bi-directional



(d) Random (e) Descending citation frequency

Fig. 10: TIME. Terminological difference measures for different individual orders (a–e) of adding documents to datasets.

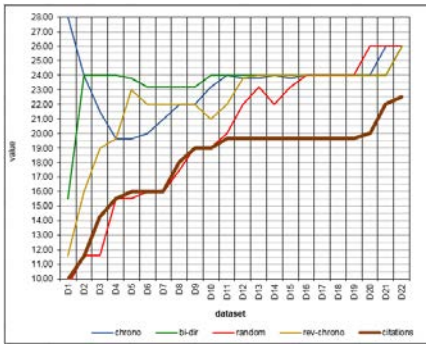
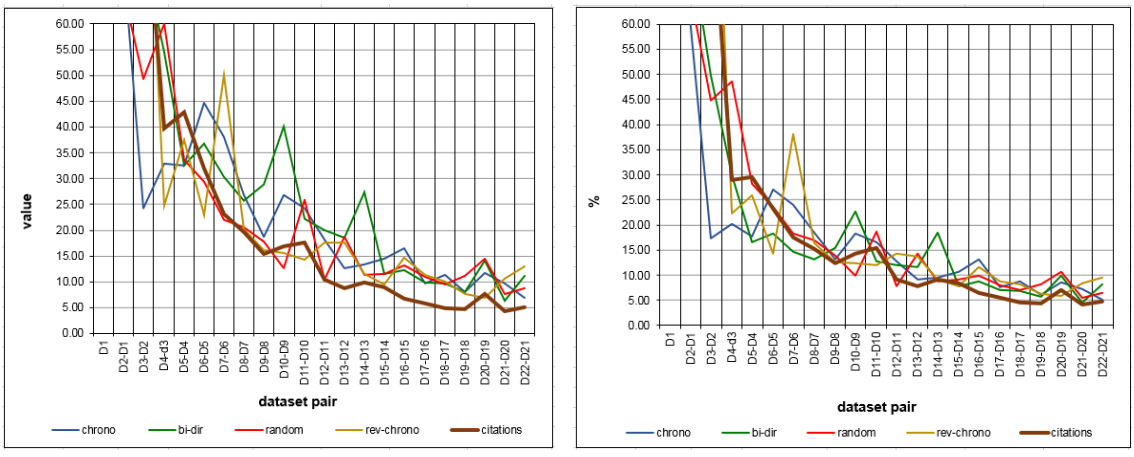


Fig. 11: TIME. Individual term significance thresholds (*eps*) for different orders of adding documents to datasets (a–e in Fig. 10). The curve for **DCF** order is rendered **thicker** as it renders the lowest values.

Fig. 12 pictures the *thd* (a) and *thdr* (b) grouped together for comparison. It could be seen in Fig. 12(a) that the *thd* curve for the reversed chronological order is the first to enter the area of saturation and clearly outperforms the rest of the orders regarding this factor. The reason, perhaps, is that eps values are the highest for the reversed chronological order, as pictured in Fig. 11. Notably, and similarly to the DMKD results, the descending citation frequency curve demonstrates the least volatile *thd* values also after reaching the saturation zone. The second least volatile was reversed chronological order curve. Therefore, the stability of the descending citation frequency order is the highest in terms of terminological difference. The second most stable order was reversed chronological.

The curves of the absolute terminological difference for the chronological, random, and bi-directional orders reach the area of saturation later than the reversed chronological curve and their values oscillate much more than the values in the reversed-chronological order curve. The curve of the descending citation frequency order enters the saturation zone just one point later than the reversed chronological curve, but is the least volatile in *thd* values among all the five compared orders.

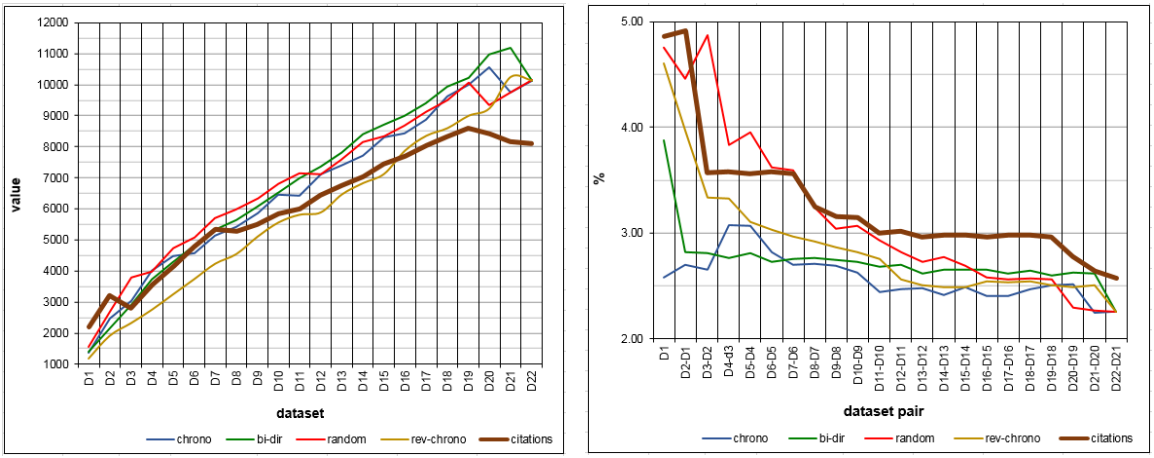


(a) *thd* measures (b) *thdr* measures

Fig. 12: TIME. The comparison of absolute (*thd*) and relative (*thdr*) terminological difference measures for different orders of adding documents to datasets (a–e in Fig. 10). The curve for **DCF** order is rendered **thicker** as the smoothest, and lowest indicating the most stable terminological saturation.

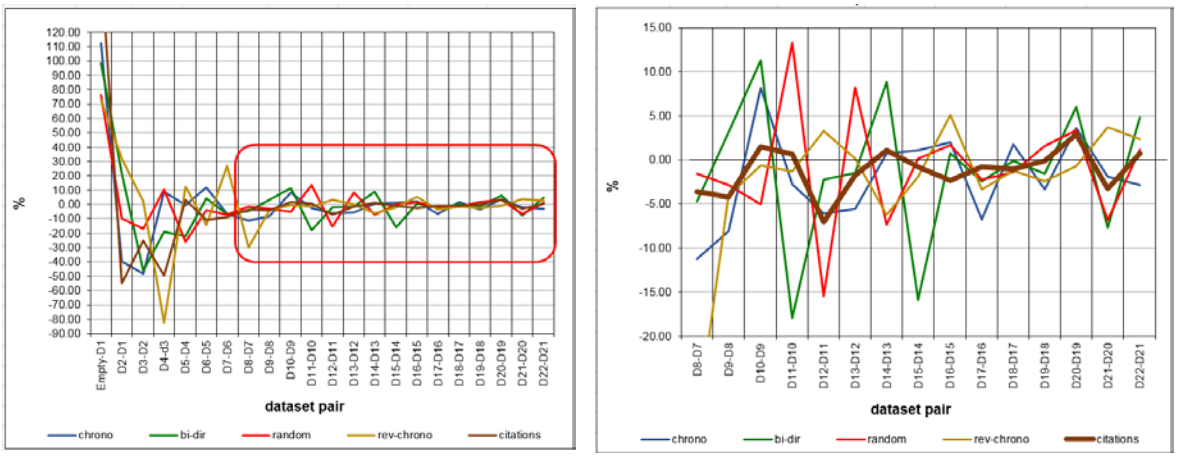
Fig. 13 pictures how different orders relate to each other by the numbers of retained terms and the proportions of retained to all extracted terms. As it could be seen in Fig. 13(a), the reversed chronological order retains the least number of terms at all measurement points. At saturation point, the descending citation frequency order is the second lowest. Fig. 13(b) pictures that all the orders allow reaching the stable ratio of retained to all extracted terms at *D*12. This ratio is circa 2.6-2.8 percent for all orders, except the descending citation frequency. For descending citation frequency, the ratio is higher – circa 3.0 percent. This fact, together with the lowest *eps* values for the descending citation frequency order, indicate that the DCF retains significant terms most completely compared to the other orders.

With respect to the *thd* volatility indicator, Fig. 14(a) covers the whole set of the measured volatility values while Fig. 14(b) presents in finer detail the area within the rounded rectangular in Fig. 14(a). It could be seen in Fig. 14 that the descending citation frequency order is the least volatile. It is followed by the reversed chronological and chronological orders, which both yield integrally similar volatility values and outperform random and bi-directional orders. Therefore, the descending citation frequency order yields the most stable saturation compared to the rest. In their turn, the chronological and reversed chronological orders yield more stable saturation compared to random or bi-directional.



(a) Numbers of retained terms (b) Ratios of retained to all terms

Fig. 13: TIME. The numbers and ratios of retained terms for different orders of adding documents to datasets (a–e in Fig. 10). The curve for **DCF** order is rendered **thicker** because this order yields the second lowest numbers of retained terms but the highest ratio of retained to all extracted terms.



(a) (b)

Fig. 14: TIME. The comparison of *thd* volatility for different orders of adding documents to datasets (a–e in Fig. 10). Diagram (b) pictures the fragment within the rounded rectangle in diagram (a) in finer detail. The curve for **DCF** order is rendered **thicker** as it is the least volatile.

For summarizing the analysis of our results on the TIME collection, it may be stated that:

- The **bi-directional** order of adding documents to datasets clearly demonstrated the **worst performance** regarding terminological saturation. This conclusion is supported by: the latest entrance into the saturation area (Fig. 11); the highest numbers of retained terms (Fig. 13); the highest volatility of *thd* values (Fig. 14).
- The **random** improves the bi-directional order only by entering the saturation area earlier (Fig. 11)
- The **chronological** order of adding documents to datasets demonstrated **balanced performance** for the entire range of the datasets and outperformed both random and bi-directional orders. However it lost to the reversed chronological order by: the later entrance to the saturation area (Fig. 12); the integrally lower individual term importance threshold values (Fig. 11); and, therefore, integrally higher numbers of retained terms (Fig. 13).
- The **reversed chronological** order of adding documents to datasets demonstrated the **second best** performance.
- The **DCF** order of adding documents to datasets demonstrated **integrally the best** and **most balanced performance** compared to all the other evaluated orders. This conclusion is supported by: the second earliest entrance into the saturation area (Fig. 12); the second lowest numbers of retained terms and the best convergence to the stable highest saturated ratio of retained to all terms (Fig. 13); the smoothest *thd* volatility curve (Fig. 14). This order yielded the lowest individual term significance thresholds (Fig. 11). This factor in the combination with the rest indicates that the most complete term set is retained.

Hence, it could be concluded that, based on the **TIME** experiments, the **DCF** order is again **the best** to yield **quicker** and more **stable** terminological saturation.

6.3 Experiments on DAC

As it has been reported in Section 5.4, the datasets generated of the DAC document collection are quite noisy. In the presence of such an amount of noise, any deliberation about terminological saturation would be not reasonable. As demonstrated in [1], any saturation in the noisy DAC datasets is in fact the saturation of its noise. Further, if the noise is (partially) eliminated by removing the stop terms, the collection becomes appropriate also for investigating terminological saturation. Therefore, we conducted two experiments on the DAC datasets:

(i) On the **noisy** datasets (identified as **DAC naturelle**). In this experiment, it has been verified if any order of adding documents is more helpful in detecting excessive noise.
(ii) On the **partially cleaned** datasets (identified ad **DAC cleaned**). In this experiment, we measured the influence of different orders of adding documents on terminological saturation – similarly to DMKD and TIME experiments.

6.3.1 DAC Naturelle

In this experiment, we did not expect any meaningful terminological saturation, but looked at how sensitive an order of adding documents was in indicating excessive noise. Therefore, instead of comparing *thd* measures and entry times into the area of saturation, we focused on looking how early and sharply the peaks occurred in our measurements.

For **DAC naturelle** datasets, the results look as follows.

We processed at the bags of terms extracted by UPM Extractor from the dataset sequences generated for all five orders: chronological; reverse-chronological; bi-directional; random; and DCF. The results of measuring individual term significant thresholds (*eps*), retained terms numbers and ratios, terminological differences (*thd*, *thdr*), and *thd* volatility are presented in the saturation measurement analysis Tables 18 – 22. The measurements pointing at a characteristic peak indicating noise are bolded.

Table 18: Terminological saturation measurements on DAC naturelle datasets – **chronological** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	27395	9,5098	1529	5,5813	144,2445	144,2445	100,0000
*D*2-*D*1	46574	11,6096	2344	5,0329	131,8707	-12,3738	52,5414
*D*3-*D*2	77317	11,6096	3698	4,7829	160,8954	29,0247	73,7342
*D*4-*D*3	96012	11,6096	4747	4,9442	71,4044	-89,4910	30,8001
*D*5-*D*4	112551	24,0000	2080	1,8481	**154,2663**	82,8619	**225,3759**
*D*6-*D*5	138766	21,0000	2848	2,0524	12,7574	-141,5089	15,7099
*D*7-*D*6	156527	36,0000	1661	1,0612	19,0872	6,3298	35,4148
*D*8-*D*7	169982	33,2193	2107	1,2395	4,9757	-14,1114	8,4518
*D*9-*D*8	184272	32,0000	2471	1,3410	4,5231	-0,4526	7,1349
*D*10-*D*9	212542	28,5293	3510	1,6514	8,9889	4,4658	12,4184
*D*11-*D*10	230726	**18294,0372**	**34**	**0,0147**	**1,6057**	-7,3833	5,6242
*D*12-*D*11	256595	16058,6816	37	0,0144	1,4631	-0,1425	4,8751

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*13-*D*12	281606	13940,0664	39	0,0138	0,8284	-0,6348	2,6860
*D*14-*D*13	301187	11712,0000	41	0,0136	1,2682	0,4399	4,0453
*D*15-*D*14	321240	17240,8068	39	0,0121	0,7296	-0,5386	2,3546
*D*16-*D*15	337402	15149,0716	41	0,0122	0,7574	0,0278	2,3858
*D*17-*D*16	357543	10985,1798	44	0,0123	0,9089	0,1515	2,7834
*D*18-*D*17	386999	6814,8590	48	0,0124	0,8308	-0,0781	2,4810
*D*19-*D*18	406035	2352,0000	56	0,0138	0,6644	-0,1664	1,9455
*D*20-*D*19	427894	1074,0000	78	0,0182	0,8952	0,2308	2,5545
*D*21-*D*20	453189	710,0000	128	0,0282	1,1009	0,2057	3,0457
*D*22-*D*21	470374	567,0000	175	0,0372	0,7905	-0,3104	2,1402
*D*23-*D*22	497532	464,0000	246	0,0494	1,0068	0,2163	2,6535
*D*24-*D*23	515285	454,0000	262	0,0508	0,8218	-0,1851	2,1684
*D*25-*D*24	543322	398,0000	319	0,0587	0,7312	-0,0906	1,8929
*D*26-*D*25	552077	376,0000	346	0,0627	0,2793	-0,4519	0,7178

Table 19: Terminological saturation measurements on DAC naturelle datasets – **reversed-chronological** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	29240	8,0000	2221	7,5958	242,0037	242,0037	100,0000
*D*2-*D*1	64356	19,6515	1085	1,6859	138,4368	-103,5669	294,8455
*D*3-*D*2	98358	15,5098	2693	2,7380	24,0498	-114,3870	33,8719
*D*4-*D*3	119501	15,5098	3497	2,9263	16,5831	-7,4666	18,9337
*D*5-*D*4	136311	15,5098	4300	3,1546	20,4135	3,8304	18,9016
*D*6-*D*5	169960	15,5098	5380	3,1655	24,8984	4,4849	18,7691
*D*7-*D*6	193968	15,5098	6240	3,2170	20,0061	-4,8923	13,1048
*D*8-*D*7	221199	15,5098	7257	3,2808	22,0490	2,0428	12,6202
*D*9-*D*8	240854	15,5098	7991	3,3178	17,3833	-4,6656	9,0493
*D*10-*D*9	262840	15,5098	8857	3,3697	21,5850	4,2016	10,1015
*D*11-*D*10	277775	32,0000	3191	1,1488	**107,9931**	86,4081	**141,1055**
*D*12-*D*11	304472	38,0391	2989	0,9817	12,7612	-95,2319	18,2791
*D*13-*D*12	320634	36,0000	3318	1,0348	3,3279	-9,4333	4,5500
*D*14-*D*13	349264	33,2193	4247	1,2160	6,3878	3,0599	8,0321
*D*15-*D*14	366384	33,2193	4541	1,2394	3,6445	-2,7433	4,3819
*D*16-*D*15	393997	**3107,6393**	**51**	**0,0129**	**3,3400**	-0,3046	9,8195
*D*17-*D*16	411646	1074,5000	73	0,0177	1,0073	-2,3327	2,8763
*D*18-*D*17	423208	708,0000	106	0,0250	0,8095	-0,1978	2,2592
*D*19-*D*18	439374	543,5000	155	0,0353	0,9173	0,1079	2,4963
*D*20-*D*19	460291	660,0000	129	0,0280	0,2997	-0,6177	0,8306
*D*21-*D*20	475015	803,0000	108	0,0227	0,2452	-0,0545	0,6906
*D*22-*D*21	487543	641,5000	141	0,0289	0,6442	0,3990	1,7822
*D*23-*D*22	511686	542,0572	202	0,0395	0,9349	0,2907	2,5212
*D*24-*D*23	526837	466,6588	253	0,0480	0,7221	-0,2128	1,9102
*D*25-*D*24	546806	390,0000	327	0,0598	0,9010	0,1789	2,3280
*D*26-*D*25	552117	376,0000	346	0,0627	0,2027	-0,6983	0,5210

Table 20: Terminological saturation measurements on DAC naturelle datasets – **bi-directional** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	32018	8,0000	2467	7,7050	170,5324	170,5324	100,0000
*D*2-*D*1	53552	10,0000	2685	5,0138	138,0222	-32,5101	50,8087
*D*3-*D*2	87933	19,0196	1740	1,9788	194,4128	56,3906	311,5034
*D*4-*D*3	105881	16,0000	2797	2,6416	19,0218	-175,3911	23,3588
*D*5-*D*4	135605	16,0000	3685	2,7175	20,3778	1,3561	20,0154

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*6-*D*5	163835	15,5098	5116	3,1227	23,2783	2,9005	18,6094
*D*7-*D*6	186012	15,5098	5901	3,1724	17,4823	-5,7960	12,2621
*D*8-*D*7	203199	15,5098	6611	3,2535	16,6696	-0,8127	10,4682
*D*9-*D*8	219983	15,5098	7266	3,3030	18,7327	2,0632	10,5256
*D*10-*D*9	234634	23,7744	4378	1,8659	**81,0028**	62,2701	**97,4883**
*D*11-*D*10	265068	23,0000	5106	1,9263	8,2838	-72,7190	9,0659
*D*12-*D*11	291022	22,0000	5818	1,9992	9,1891	0,9053	9,1377
*D*13-*D*12	307284	28,5293	4483	1,4589	25,3236	16,1345	34,6434
*D*14-*D*13	329797	28,5293	4859	1,4733	4,0895	-21,2342	5,2981
*D*15-*D*14	345955	28,5293	5185	1,4987	3,9755	-0,1139	4,8982
*D*16-*D*15	364067	28,0000	5686	1,5618	4,2112	0,2357	4,9326
*D*17-*D*16	385492	26,0000	6242	1,6192	4,3408	0,1296	4,8384
*D*18-*D*17	399124	26,0000	6609	1,6559	3,8294	-0,5114	4,0937
*D*19-*D*18	421694	26,0000	7170	1,7003	5,7900	1,9606	5,8288
*D*20-*D*19	445817	24,0000	8880	1,9918	8,4431	2,6532	7,8339
*D*21-*D*20	463660	**352,0000**	**291**	**0,0628**	**18,4497**	10,0066	49,1533
*D*22-*D*21	476233	707,5000	132	0,0277	1,7616	-16,6881	4,8579
*D*23-*D*22	504080	536,0000	205	0,0407	1,1565	-0,6052	3,0905
*D*24-*D*23	525513	466,6588	255	0,0485	1,2348	0,0783	3,2653
*D*25-*D*24	525513	466,6588	255	0,0485	0,0000	-1,2348	0,0000
*D*26-*D*25	552095	376,0000	346	0,0627	1,0912	1,0912	2,8047

Table 21: Terminological saturation measurements on DAC naturelle datasets – **random** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	27877	8,0000	2327	8,3474	167,1250	167,1250	100,0000
*D*2-*D*1	52013	48,0000	270	0,5191	66,9223	-100,2027	179,5947
*D*3-*D*2	80883	28,0000	1023	1,2648	12,3577	-54,5646	24,9043
*D*4-*D*3	102745	23,2193	1764	1,7169	10,9360	-1,4217	18,0591
*D*5-*D*4	135407	20,0000	2496	1,8433	11,6943	0,7583	16,1857
*D*6-*D*5	159658	19,6515	3312	2,0744	11,4883	-0,2061	13,7191
*D*7-*D*6	176775	19,0196	3982	2,2526	10,5266	-0,9617	11,1669
*D*8-*D*7	201587	19,0196	4704	2,3335	13,0321	2,5055	12,1457
*D*9-*D*8	226218	19,6515	5694	2,5170	22,3558	9,3237	17,4172
*D*10-*D*9	241292	19,6515	6128	2,5397	9,4156	-12,9402	6,8342
*D*11-*D*10	265540	48,0000	1934	0,7283	**51,5208**	42,1052	**87,8267**
*D*12-*D*11	286161	**19616,9148**	**36**	**0,0126**	**0,5045**	-51,0163	1,6982
*D*13-*D*12	304375	17693,1612	38	0,0125	0,9338	0,4292	3,0471
*D*14-*D*13	324691	15878,8163	39	0,0120	0,4181	-0,5157	1,3460
*D*15-*D*14	354206	15406,7638	40	0,0113	0,9792	0,5611	3,1248
*D*16-*D*15	369237	16404,3619	40	0,0108	0,4805	-0,4987	1,5366
*D*17-*D*16	391458	13263,0620	43	0,0110	1,0132	0,5328	3,1386
*D*18-*D*17	410289	9662,5898	46	0,0112	0,7471	-0,2661	2,2620
*D*19-*D*18	432587	3190,5295	51	0,0118	0,6844	-0,0628	2,0300
*D*20-*D*19	453547	1660,0000	63	0,0139	0,6370	-0,0474	1,8544
*D*21-*D*20	470513	954,0000	96	0,0204	0,9312	0,2942	2,6394
*D*22-*D*21	482735	730,0000	127	0,0263	0,6442	-0,2870	1,7932
*D*23-*D*22	501325	574,0000	185	0,0369	0,9418	0,2976	2,5545
*D*24-*D*23	528036	455,0000	259	0,0490	1,0209	0,0791	2,6945
*D*25-*D*24	545756	389,0000	323	0,0592	0,7607	-0,2602	1,9682
*D*26-*D*25	552089	376,0000	346	0,0627	0,2566	-0,5041	0,6596

Table 22: Terminological saturation measurements on DAC naturelle datasets – **DCF** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	24618	9.5098	1358	5.5163	125.5089	125.5089	100.0000
*D*2-*D*1	43821	11.6096	2193	5.0044	134.5471	9.0382	53.8154
*D*3-*D*2	73775	11.6096	3358	4.5517	158.6158	24.0687	77.9195
*D*4-*D*3	88737	11.6096	4261	4.8018	65.4120	-93.2038	29.7694
*D*5-*D*4	106814	24.0000	1906	1.7844	**149.3096**	83.8976	**230.4051**
*D*6-*D*5	131635	22.0000	2665	2.0245	12.6100	-136.6996	16.2893
*D*7-*D*6	146308	38.0537	1417	0.9685	16.9707	4.3607	32.8790
*D*8-*D*7	161997	34.0000	1886	1.1642	5.1084	-11.8623	9.0058
*D*9-*D*8	175249	33.2193	2224	1.2691	4.1533	-0.9551	6.8225
*D*10-*D*9	198801	31.0196	2897	1.4572	7.1593	3.0060	10.5228
*D*11-*D*10	218764	**19338.4980**	**33**	**0.0151**	**1.4475**	-5.7119	5.1671
*D*12-*D*11	247663	17724.0546	35	0.0141	1.0546	-0.3928	3.6281
*D*13-*D*12	266810	16058.6816	37	0.0139	0.9442	-0.1105	3.1459
*D*14-*D*13	291484	13940.0664	39	0.0134	1.3531	0.4089	4.4150
*D*15-*D*14	307814	18497.3958	38	0.0123	0.7204	-0.6327	2.3566
*D*16-*D*15	327248	15957.0054	40	0.0122	0.8083	0.0880	2.5762
*D*17-*D*16	347251	15149.0716	41	0.0118	0.3689	-0.4395	1.1619
*D*18-*D*17	367472	10985.1798	44	0.0120	0.9089	0.5400	2.7834
*D*19-*D*18	381385	8003.4174	47	0.0123	0.6648	-0.2440	1.9954
*D*20-*D*19	412279	1855.0000	59	0.0143	0.9847	0.3199	2.8705
*D*21-*D*20	427243	1041.5000	80	0.0187	0.8011	-0.1836	2.2819
*D*22-*D*21	444702	786.0000	115	0.0259	0.8136	0.0125	2.2652
*D*23-*D*22	465132	585.0000	167	0.0359	0.9211	0.1075	2.5002
*D*24-*D*23	497157	592.8457	174	0.0350	0.7380	-0.1831	2.0088
*D*25-*D*24	514541	520.5000	226	0.0439	0.7814	0.0434	2.0827
*D*26-*D*25	519686	490.0000	241	0.0464	0.2124	-0.5689	0.5630

The measures at the points of appearance of characteristic accumulated noise peaks are provided in Table 23 for comparison.

Table 23: DAC naturelle. The comparison of noise indications for all the orders.

Order	Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
(a) *eps* peaks								
Chrono	*D*11-*D*10	230726	**18294,0372**	**34**	**0,0147**	1,6057	-7,3833	5,6242
Random	*D*12-*D*11	286161	**19616,9148**	**36**	**0,0126**	0,5045	-51,0163	1,6982
Rev-chrono	*D*16-*D*15	393997	**3107,6393**	**51**	**0,0129**	3,3400	-0,3046	9,8195
Bi-dir	*D*21-*D*20	463660	**352,0000**	**291**	**0,0628**	18,4497	10,0066	49,1533
DCF	*D*11-*D*10	218764	**19338.4980**	**33**	**0.0151**	1.4475	-5.7119	5.1671
(b) *thd*/*thdr* peaks								
Chrono	*D*11-*D*10	277775	32,0000	3191	1,1488	**107,9931**	86,4081	**141,1055**
Random	*D*11-*D*10	265540	48,0000	1934	0,7283	**51,5208**	42,1052	**87,8267**
Rev-chrono	*D*11-*D*10	277775	32,0000	3191	1,1488	**107,9931**	86,4081	**141,1055**
Bi-dir	*D*10-*D*9	234634	23,7744	4378	1,8659	**81,0028**	62,2701	**97,4883**
DCF	*D*5-*D*4	106814	24.0000	1906	1.7844	**149.3096**	83.8976	**230.4051**

As it could be noticed in Tables 18 – 23 and in Fig. 15, the individual term significance thresholds (*eps*) peaked significantly, indicating excessive noise, at different times, and having different *eps* scales. A clear negative outlier was the bi-directional order, which was not sufficiently sensitive in terms of collecting noise terms with high *c-values*. Therefore, the *eps* values for the bi-directional order were

quite low. It produced a small peak quite lately and with the *eps* value been lower by the order of magnitude compared to the rest four orders. On the contrary, DCF, random, and chronological orders produced the highest *eps* peaks at almost the same time and very similar values, while chronological and DCF been slightly faster than random. The reversed-chronological order though lost the competition to the chronological and random in being quick to detect noise and the height of the *eps* peak, still gave the indication of a noise with the *eps* value of the similar scale in the percentage of retained terms.

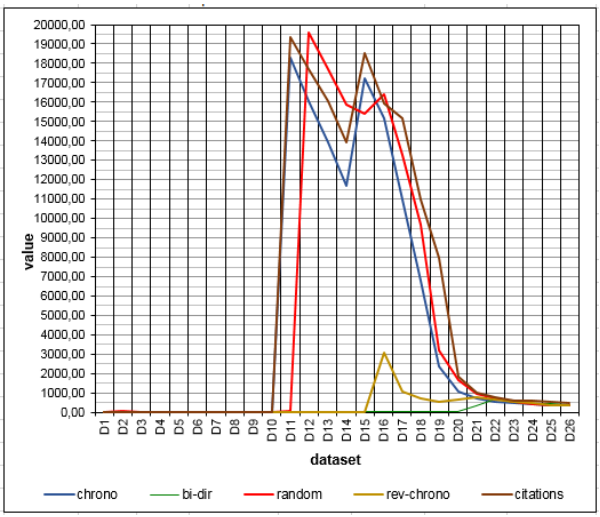
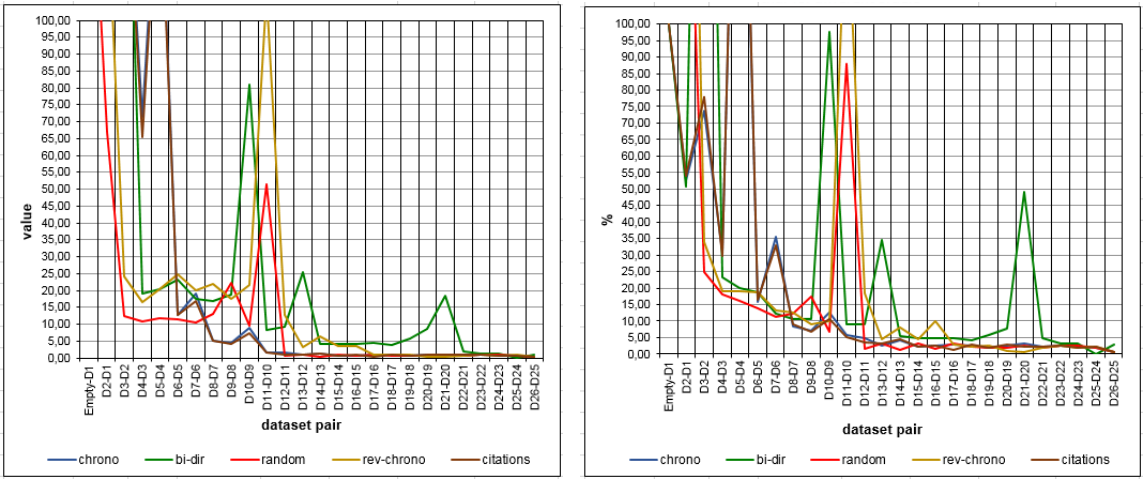


Fig. 15: DAC naturelle. Individual term significance thresholds (*eps*) for different orders of adding documents to datasets (Tables 18 – 22).

The curves in Fig. 16, picturing the *thd* and *thdr* measurements for all the orders, support the conclusion regarding the *eps* indicator. Please see also Table 23(b) for the values. Indeed, *thd* and *thdr* peaked much higher and earlier for the DCF order. The bi-directional order was the second earliest, though not the second highest in the amplitude of the peak, after DCF. The chronological and reversed-cronological orders were the latest, though second highest in the amplitudes. Finally, the random order was the clear negative outlier.

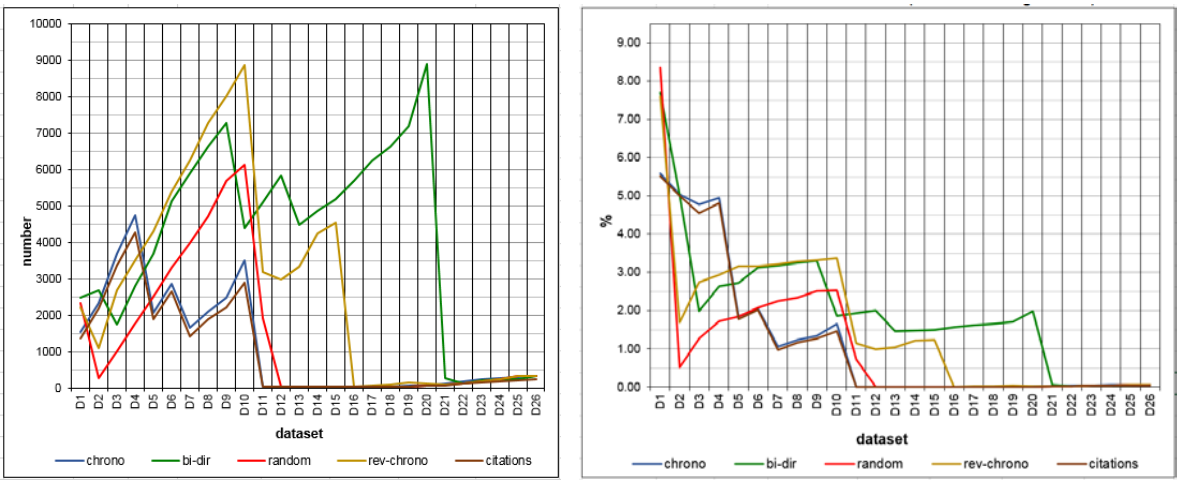


(a) *thd* measures (b) *thdr* measures

Fig. 16: DAC naturelle. The comparison of absolute (*thd*) and relative (*thdr*) terminological difference measures for different orders of adding documents to datasets.

Fig. 17 pictures how different orders relate to each other by the numbers of retained terms and the proportions of retained to all extracted terms. As it could be seen in Fig. 17(a), the chronological and DCF order curves were the first to peak at their highest numbers of retained terms (at *D*4, ~4 700 for chronological and ~4 300 for DCF) among the other curves. Both chronological and DCF order curves than dropped down to the very few retained terms (indicating excessive noise at *D*11) in 7 measurement steps. The curves of the reversed-chronological and random orders peaked at their highest numbers of retained terms later (at *D*10) but with substantially higher values (~8 900 for rev-chrono and ~6 100 for random). The curve for the random order dropped down to the very few retained terms immediately at *D*12 and the curve for the reversed-chronological order did it less quickly at *D*16 – in 5 measurement steps. Finally, the curve for the bi-directional order peaked the latest (at *D*20) with the value similar to the reversed-chronological order (~8 900). It went down also immediately at *D*21.

The curves of the ratios of retained to all extracted terms for the orders presented in Fig. 17(b) show a very similar picture. The chronological and DCF orders were the quickest to peak and drop down to very small values. These were followed by random, reversed-chronological, and bi-directional.



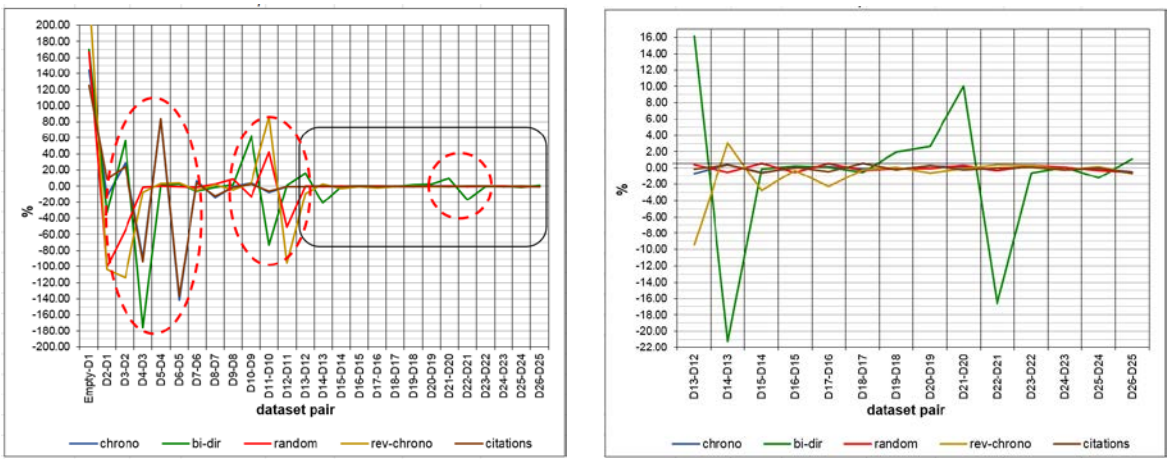
(a) Numbers of retained terms (b) Ratios of retained to all terms

Fig. 17: DAC naturelle. The numbers of retained terms for different orders of adding documents to datasets.

Hence, judging by the peaks and drops in the numbers of retained terms:

- The two most sensitive orders were chronological and DCF. These orders peaked and went down the earliest. The peaks were also for the highest ratios of retained to all extracted terms.
- The trhird sensitive order was random as it peaked quite highly and sufficiently early and dropped down immediately (in 2 measurement points)
- The fourth sensitive order was reversed-chronological as it peaked the highest and dropped down sufficiently quickly (in 5 measurement points)
- The least sensitive order was bi-directional as it was the latest in indicating excessive noise

Perhaps, the overall picture of the sensitivity of different orders to excessive noise is provided by the analysis of the thd volatility, shown in Fig. 18. In this figure, one can see three regions of oscillations outlined by dashed ovals. Those represent the indications of excessive noise. The earliest oscilation region (1) with the highest volatility amplitudes corresponds to the bi-directional, chronological, and DCF orders. The second region corresponds to bi-directional, random, and reversed-chronological orders. Finally, the third region corresponds to the bi-directional order. Hence, the chronological and DCF orders were the most densitive in detecting excessive noise, both the earliest and with the highest *thd* volatilities (the first oscillation region). The random and reversed chronological orders were less sensitive – refer to the oscillation region 2. Finally, the bi-directional continued oscillating in the region 3. So it was the least sensitive.



(a) All measurement points (b) Points after D12-D11 in finer detail

Fig. 18: DAC naturelle: *thd* volatility for different orders of adding documents to datasets.

As a summary, it could be stated based on our **DAC naturelle** experiments, that the **chronological** and **DCF** orders are **the best** to detect excessive noise in the datasets used to measure terminological saturation.

6.3.2 DAC Cleaned

In difference to **DAC naturelle**, the datasets of **DAC cleaned** were processed, using our Stop-Term Remover module, to remove the top-scored stop-terms in the bags of terms extracted from DAC down to *c-values* equal to 40. This allowed us to de-noise the datasets to the noise levels comparable to the TIME and DMKD collections.

Therefore, we expected to observe terminological saturation in DAC cleaned and processed it similarly to DMKD (Section 6.1) and TIME (Section 6.2) using the same features for results analysis.

We processed at the bags of terms extracted by UPM Extractor from the dataset sequences generated for all four orders: chronological; reverse-chronological; bi-directional; and random. The results of measuring individual term significant thresholds (*eps*), retained terms numbers and ratios, terminological differences (*thd*, *thdr*), and *thd* volatility are presented in the saturation measurement analysis

Tables 24 – 28. The measurements indicating the entries into the saturation areas are bolded.

Table 24: Terminological saturation measurements on DAC cleaned datasets – **chronological** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	27167	8,0000	2230	8,2085	296,0600	296,0600	100,0000
*D*2-*D*1	46308	10,0000	2467	5,3274	196,5636	-99,4964	83,6827
*D*3-*D*2	77006	10,0000	3866	5,0204	148,0610	-48,5025	73,5442
*D*4-*D*3	95687	10,0000	5015	5,2410	68,4898	-79,5712	31,7024
*D*5-*D*4	112158	11,6096	5641	5,0295	75,9056	7,4158	40,4425
*D*6-*D*5	138360	11,6096	7000	5,0593	49,1559	-26,7497	23,9032
*D*7-*D*6	156098	12,0000	6856	4,3921	44,2846	-4,8713	21,8258
*D*8-*D*7	169544	12,0000	7646	4,5097	43,2259	-1,0587	23,1400
*D*9-*D*8	183827	12,0000	8409	4,5744	24,5146	-18,7113	12,7884
*D*10-*D*9	212050	14,0000	9039	4,2627	32,7411	8,2264	16,2564
*D*11-*D*10	230131	14,2647	9457	4,1094	25,7116	-7,0295	12,7382
*D*12-*D*11	255956	14,2647	10813	4,2246	30,9879	5,2763	14,7150
*D*13-*D*12	280958	14,5000	11012	3,9194	23,3313	-7,6566	11,3419
*D*14-*D*13	300438	15,5098	11917	3,9665	30,2086	6,8773	14,9713
*D*15-*D*14	320473	15,5098	12655	3,9489	18,4493	-11,7593	9,2021
*D*16-*D*15	336624	15,5098	13298	3,9504	19,8337	1,3843	10,0905
*D*17-*D*16	356749	15,5098	14204	3,9815	16,6898	-3,1439	8,3662
*D*18-*D*17	386197	15,5098	15322	3,9674	16,0698	-0,6200	7,7320
*D*19-*D*18	405224	15,5098	16139	3,9827	17,1996	1,1297	8,2627
*D*20-*D*19	427074	15,5098	16904	3,9581	13,2064	-3,9932	6,2536
*D*21-*D*20	452306	15,5098	17980	3,9752	15,1762	1,9698	6,9597
*D*22-*D*21	469478	15,5098	18659	3,9744	11,0274	-4,1488	4,9663
*D*23-*D*22	496635	15,5098	19635	3,9536	15,8308	4,8034	7,1513
*D*24-*D*23	514364	**15,5098**	**20558**	**3,9968**	**13,6261**	-2,2046	6,0687
*D*25-*D*24	542392	15,5098	21166	3,9023	13,1869	-0,4392	5,9699
*D*26-*D*25	551147	15,5098	21467	3,8950	3,9058	-9,2811	1,7635

Table 25: Terminological saturation measurements on DAC cleaned datasets – **reversed-chronological** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	29021	8,0000	2145	7,3912	229,3454	229,3454	100,0000
*D*2-*D*1	64041	9,5098	2864	4,4721	173,8704	-55,4750	101,8027
*D*3-*D*2	98005	9,5098	4765	4,8620	112,1951	-61,6754	60,8375
*D*4-*D*3	119127	10,0000	5606	4,7059	64,0111	-48,1840	32,0186
*D*5-*D*4	135864	11,6096	6125	4,5082	52,0473	-11,9638	24,9215
*D*6-*D*5	169498	11,6096	7745	4,5694	59,6194	7,5722	25,2380
*D*7-*D*6	193496	11,6096	9045	4,6745	58,4578	-1,1616	25,2964
*D*8-*D*7	220722	11,6096	10588	4,7970	48,6135	-9,8443	20,0947
*D*9-*D*8	240356	11,6096	11654	4,8486	36,4494	-12,1640	14,8903
*D*10-*D*9	262330	11,6096	12926	4,9274	43,8389	7,3895	18,1554
*D*11-*D*10	277228	12,0000	11131	4,0151	32,8241	-11,0149	15,6494
*D*12-*D*11	303824	13,5000	11758	3,8700	33,1084	0,2843	15,4197
*D*13-*D*12	319984	14,0000	12401	3,8755	25,9261	-7,1823	12,6584
*D*14-*D*13	348562	14,2647	13365	3,8343	27,8678	1,9416	13,2665
*D*15-*D*14	365669	14,2647	14234	3,8926	19,9225	-7,9452	9,3379
*D*16-*D*15	393190	14,2647	15327	3,8981	19,3706	-0,5519	8,8459
*D*17-*D*16	410794	15,5098	15097	3,6751	18,6246	-0,7460	8,5401
*D*18-*D*17	422348	15,5098	15763	3,7322	18,2870	-0,3376	8,5447
*D*19-*D*18	438507	15,5098	16530	3,7696	16,3788	-1,9082	7,6506

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*20-*D*19	459405	15,5098	17430	3,7940	15,1482	-1,2306	6,9951
*D*21-*D*20	474125	15,5098	18287	3,8570	18,6145	3,4663	8,6940
*D*22-*D*21	486649	**15,5098**	**18865**	**3,8765**	**10,9433**	-7,6712	5,0772
*D*23-*D*22	510781	15,5098	19763	3,8692	13,2852	2,3418	6,0915
*D*24-*D*23	525928	15,5098	20360	3,8713	12,9281	-0,3571	6,0111
*D*25-*D*24	545879	15,5098	21269	3,8963	11,6063	-1,3218	5,2641
*D*26-*D*25	551187	15,5098	21467	3,8947	2,6804	-8,9259	1,2102

Table 26: Terminological saturation measurements on DAC cleaned datasets – **bi-directional** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	31789	8,0000	2362	7,4302	285,8752	285,8752	100,0000
*D*2-*D*1	53283	9,5098	2846	5,3413	191,7138	-94,1614	75,1298
*D*3-*D*2	87582	10,0000	3818	4,3593	155,2341	-36,4797	75,6281
*D*4-*D*3	105516	11,6096	4622	4,3804	80,7515	-74,4826	40,1074
*D*5-*D*4	135200	11,6096	6182	4,5725	74,6582	-6,0933	35,0775
*D*6-*D*5	163415	11,6096	7343	4,4935	65,4651	-9,1931	33,7102
*D*7-*D*6	185586	11,6096	8567	4,6162	38,4185	-27,0467	18,7340
*D*8-*D*7	202758	11,6096	9628	4,7485	31,5727	-6,8458	14,5570
*D*9-*D*8	219517	11,6096	10576	4,8179	33,2532	1,6805	15,0153
*D*10-*D*9	234106	12,0000	9598	4,0999	37,3321	4,0788	18,6980
*D*11-*D*10	264535	12,0000	10872	4,1099	28,6374	-8,6946	13,2045
*D*12-*D*11	290472	12,0000	11888	4,0926	26,7571	-1,8803	12,0993
*D*13-*D*12	306718	14,0000	11922	3,8870	25,1814	-1,5757	11,5849
*D*14-*D*13	329221	14,0000	12949	3,9332	24,6805	-0,5009	11,2244
*D*15-*D*14	345371	14,2647	13120	3,7988	22,0608	-2,6197	10,3344
*D*16-*D*15	363478	14,2647	14005	3,8531	21,2047	-0,8561	10,0478
*D*17-*D*16	384894	14,2647	14774	3,8385	14,4548	-6,7499	6,6919
*D*18-*D*17	398513	14,5000	14523	3,6443	16,1654	1,7106	7,6401
*D*19-*D*18	421035	15,5098	15434	3,6657	19,0313	2,8658	8,7484
*D*20-*D*19	445152	15,5098	16441	3,6933	18,7163	-0,3150	8,5498
*D*21-*D*20	462897	15,5098	17094	3,6928	15,1104	-3,6059	6,9907
*D*22-*D*21	475448	15,5098	17782	3,7401	15,9848	0,8744	7,4599
*D*23-*D*22	503245	15,5098	18809	3,7375	13,9552	-2,0296	6,3775
*D*24-*D*23	524591	15,5098	20381	3,8851	20,1050	6,1499	8,9189
*D*25-*D*24	542936	15,5098	21176	3,9003	16,5913	-3,5137	7,5284
*D*26-*D*25	551165	**15,5098**	**21468**	**3,8950**	**3,7072**	-12,8841	1,6738

Table 27: Terminological saturation measurements on DAC cleaned datasets – **random** order of adding documents

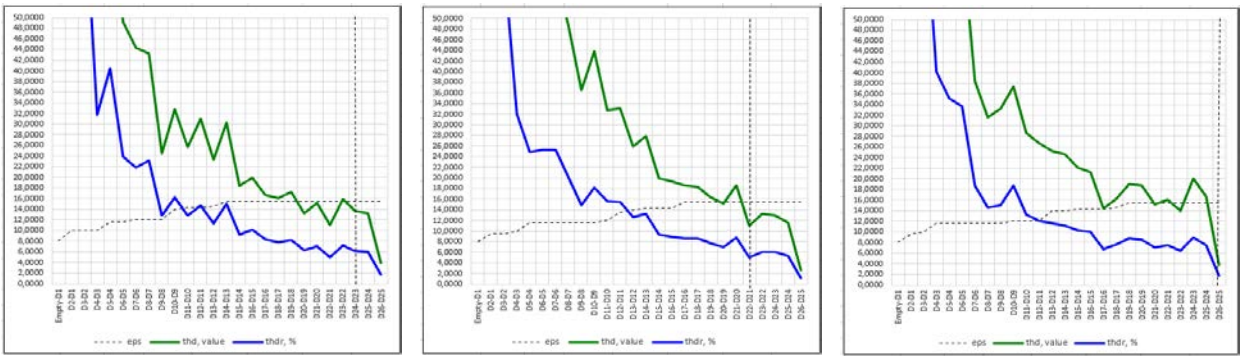
Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	27644	8,0000	2246	8,1247	156,7791	156,7791	100,0000
*D*2-*D*1	51691	9,5098	2982	5,7689	108,2269	-48,5522	68,6316
*D*3-*D*2	80538	9,5098	4623	5,7401	87,9187	-20,3082	45,1867
*D*4-*D*3	102380	10,0000	5351	5,2266	69,9225	-17,9962	37,0801
*D*5-*D*4	135032	11,6096	6037	4,4708	52,1233	-17,7992	28,2881
*D*6-*D*5	159270	11,6096	7256	4,5558	42,9927	-9,1307	22,4643
*D*7-*D*6	176370	11,6096	8286	4,6981	36,1830	-6,8097	18,5510
*D*8-*D*7	201130	11,6096	9789	4,8670	39,4528	3,2697	19,1924
*D*9-*D*8	225656	12,0000	9673	4,2866	46,5972	7,1444	21,7680
*D*10-*D*9	240724	13,0000	9651	4,0092	23,6224	-22,9748	11,3965
*D*11-*D*10	264944	14,0000	10774	4,0665	31,2482	7,6258	14,7723
*D*12-*D*11	285463	14,2647	11190	3,9199	26,7024	-4,5459	12,9893

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*13-*D*12	303672	14,2647	11991	3,9487	19,7136	-6,9887	9,5264
*D*14-*D*13	323981	14,2647	12843	3,9641	18,6721	-1,0415	8,9015
*D*15-*D*14	353486	14,2647	13834	3,9136	21,5275	2,8554	10,2828
*D*16-*D*15	368496	15,0000	13686	3,7140	21,0973	-0,4302	10,5161
*D*17-*D*16	390670	15,5098	14594	3,7356	17,6403	-3,4570	8,4942
*D*18-*D*17	409485	15,5098	15440	3,7706	15,2754	-2,3650	7,1337
*D*19-*D*18	431778	15,5098	16274	3,7691	15,8182	0,5428	7,3475
*D*20-*D*19	452726	15,5098	17051	3,7663	16,4516	0,6334	7,7723
*D*21-*D*20	469665	15,5098	17936	3,8189	17,7837	1,3322	8,4191
*D*22-*D*21	481885	**15,5098**	**18516**	**3,8424**	**10,8970**	-6,8867	5,1331
*D*23-*D*22	500415	15,5098	19435	3,8838	13,9125	3,0155	6,4501
*D*24-*D*23	527115	15,5098	20487	3,8866	14,8776	0,9651	6,8483
*D*25-*D*24	544829	15,5098	21251	3,9005	9,8297	-5,0479	4,4465
*D*26-*D*25	551159	15,5098	21467	3,8949	3,6884	-6,1413	1,6653

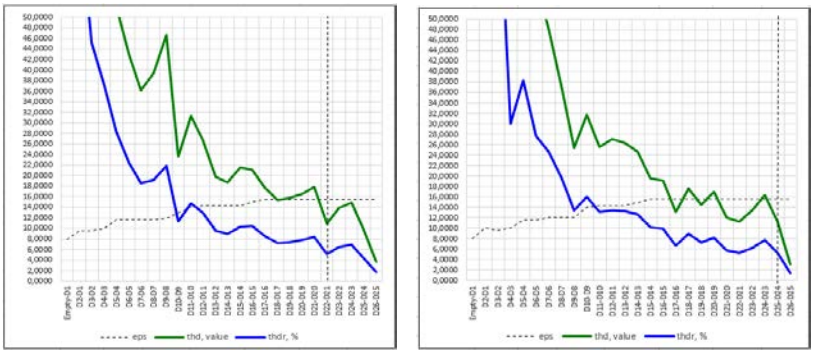
Table 28: Terminological saturation measurements on DAC cleaned datasets – **DCF** order of adding documents

Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
*D*1-empty	24401	8.0000	1976	8.0980	294.4309	294.4309	100.0000
*D*2-*D*1	43563	10.0000	2296	5.2705	210.7306	-83.7003	90.3284
*D*3-*D*2	73480	9.5098	4002	5.4464	158.0617	-52.6689	78.9850
*D*4-*D*3	88428	10.0000	4513	5.1036	62.0073	-96.0544	30.0686
*D*5-*D*4	106440	11.6096	5181	4.8675	71.9988	9.9914	38.3356
*D*6-*D*5	131251	11.6096	6616	5.0407	57.2432	-14.7556	27.7304
*D*7-*D*6	145891	12.0000	6501	4.4561	48.1318	-9.1114	24.6035
*D*8-*D*7	161572	12.0000	7286	4.5094	37.2009	-10.9309	19.7249
*D*9-*D*8	174815	12.0000	7975	4.5620	25.3817	-11.8192	13.4853
*D*10-*D*9	198322	14.0000	8521	4.2965	31.7604	6.3786	16.0161
*D*11-*D*10	218188	14.2647	8906	4.0818	25.5103	-6.2500	13.1271
*D*12-*D*11	247035	14.2647	10106	4.0909	27.1576	1.6473	13.4638
*D*13-*D*12	266125	14.2647	10950	4.1146	26.3493	-0.8084	13.2537
*D*14-*D*13	290784	15.0000	11212	3.8558	24.6727	-1.6766	12.7239
*D*15-*D*14	307048	15.5098	11828	3.8522	19.6402	-5.0325	10.1810
*D*16-*D*15	326475	15.5098	12631	3.8689	19.0322	-0.6080	9.8547
*D*17-*D*16	346470	15.5098	13259	3.8269	13.0470	-5.9853	6.6743
*D*18-*D*17	366679	15.5098	14171	3.8647	17.6378	4.5908	8.9017
*D*19-*D*18	380583	15.5098	14830	3.8967	14.4754	-3.1624	7.3363
*D*20-*D*19	411426	15.5098	16094	3.9118	16.9466	2.4713	8.2255
*D*21-*D*20	426370	15.5098	16703	3.9175	12.0883	-4.8584	5.7770
*D*22-*D*21	443817	15.5098	17380	3.9160	11.3268	-0.7615	5.3211
*D*23-*D*22	464226	15.5098	18284	3.9386	13.3909	2.0642	6.2007
*D*24-*D*23	496241	15.5098	18976	3.8239	16.3654	2.9744	7.7950
*D*25-*D*24	513621	**15.5098**	**19796**	**3.8542**	**11.2104**	-5.1550	5.2838
*D*26-*D*25	518765	15.5098	19991	3.8536	3.1238	-8.0866	1.4674

These measurements are visualized in Fig. 19 – 22. The dashed vertical lines point to the *thd* measurement in which saturation indicator has been observed as *thd* went below *eps* and did not go up above it. The corresponding values of *eps*, no of retained terms, ratios of retained terms to the total numbers of terms in the bags of terms, and *thd* for these bags of terms are bolded in Tables 24 – 28.



(a) Chronological (b) Reversed-chronological (c) Bi-directional



(d) Random (e) Descending citation frequency

Fig. 19: DAC cleaned. Terminological difference measures for different individual orders (a–e) of adding documents to datasets.

As it may be noticed in Tables 24 – 28 and Fig. 19 – 20, all orders, but bi-directional, yielded stable terminological saturation on DAC, but at different measurement points. The *thd* curve for the bi-directional order went below the *eps* curve only at the last measurement point, which only hints about possible saturation. It is also observable that, in difference to DMKD and TIME measurements, *thdr* values in DAC cleaned are noticeably lower than the *thd* values. This may hint[16] that the datasets were not sufficiently de-noised.

The first to reach saturation were the reversed-chronological and random orders for which the absolute terminological difference measures (*thd*) went below the individual term significance threshold curve (*eps*) at *D*22-*D*21 – see Fig. 19(b and d). The next order was chronological, at *D*24-*D*23 – see Fig. 19(a). The fourth order was DCF, at *D*25-*D*24 – see Fig. 19(e). The last one was bi-directional, at *D*26-*D*25 – see Fig. 19(c). The measures at saturation points for all the five evaluated orders are provided in Table 21 for comparison.

One of the early indicators of the existence of potential saturation is reaching the maximum value of the individual term significance threshold. Therefore, the quicker the order reaches this maximum, the more sensitive it is to detecting saturation. As it could be noticed in Tables 24 – 28 and in Fig. 20, the individual term significance thresholds (*eps*) for the chronological order was the earliest to reach the maximum

[16] This hypothesis has not been checked, as it did not influence the result evaluated using *thd*.

– at *D*14. The values for DCF reached the same maximum one step later – at *D*15. The random and reversed-chronological orders were the third at *D*17. The bi-directional was the last at *D*19.

Table 29: DAC cleaned. The comparison of saturation measurements for all orders at their saturation points.

Order	Dataset Pair	No Terms (in the Bag of Terms)	*eps*	Retained Terms (*c-value* > *eps*)	% Retained Terms	*thd*, value	*thd* volatility	*thdr*,%
Chrono	*D*24-*D*23	514364	**15,5098**	**20558**	**3,9968**	**13,6261**	-2,2046	6,0687
Random	*D*22-*D*21	481885	**15,5098**	**18516**	**3,8424**	**10,8970**	-6,8867	5,1331
Rev-chrono	*D*22-*D*21	486649	**15,5098**	**18865**	**3,8765**	**10,9433**	-7,6712	5,0772
Bi-dir	*D*26-*D*25	551165	**15,5098**	**21468**	**3,8950**	**3,7072**	-12,8841	1,6738
DCF	*D*25-*D*24	518765	**15.5098**	**19991**	**3.8536**	**3.1238**	-8.0866	1.4674

Fig. 21 pictures the *thd* (a) and *thdr* (d) curves for different orders grouped together for comparison. Fig. 21(a) presents the full curves measured for all dataset pairs in the diagram at the left. The diagram in Fig. 21(b) visualizes the measurements in finer detail, for the area in the rounded rectangle of Fig. 21(a). The diagram in Fig. 21(c) visualizes the upper envelope curves for the measurements in Fig. 21(b). The dashed lines in Fig. 21(a-d) represent the measurements of *eps* for the bi-directional order. Bi-directional has been selected as the one having the lowest *eps* values compared to the other orders.

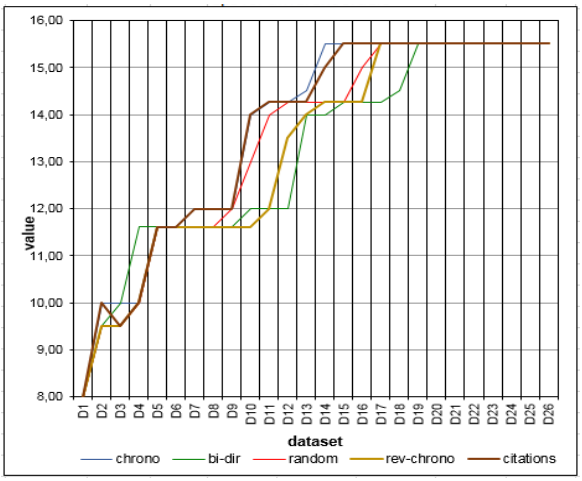
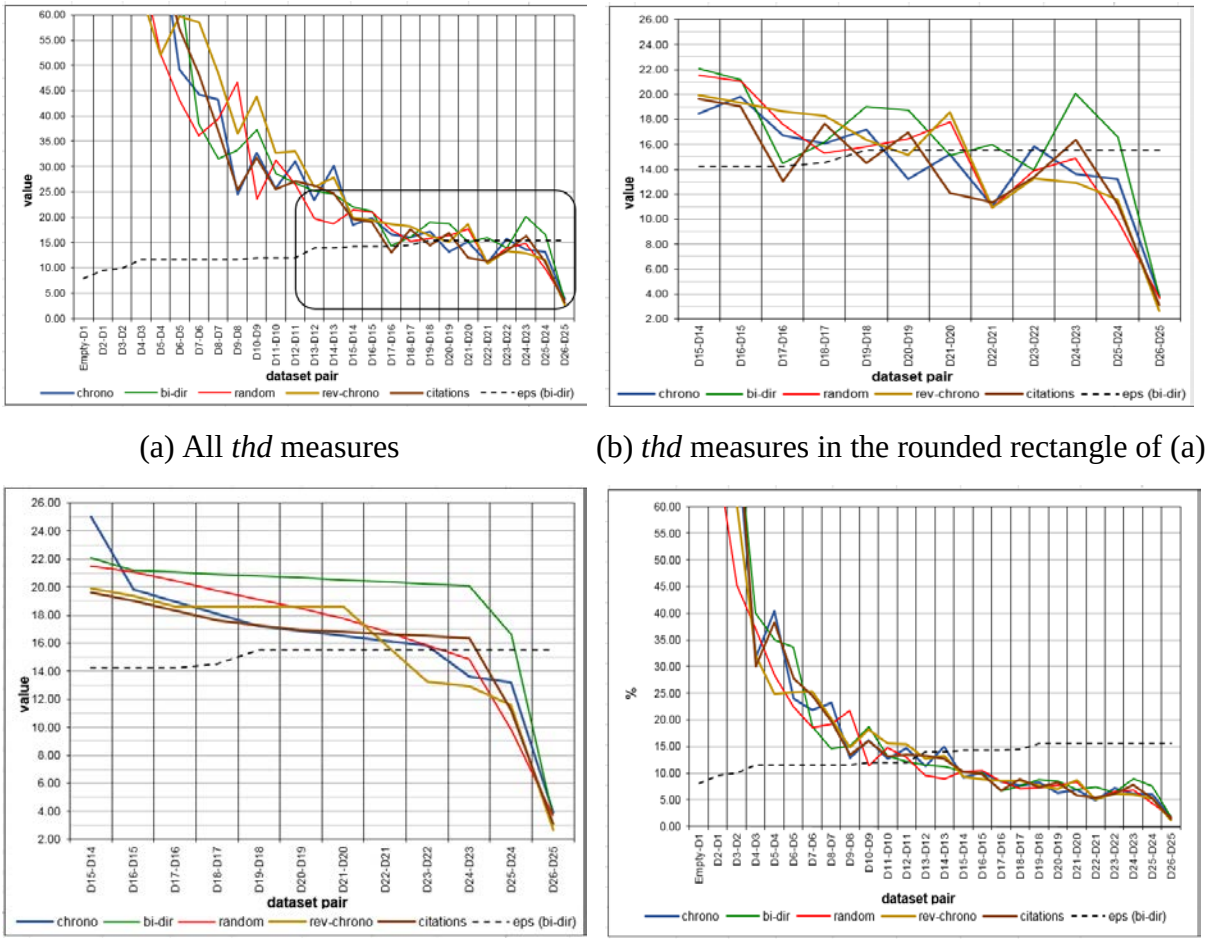


Fig. 20: DAC cleaned. Individual term significance thresholds (*eps*) for different orders of adding documents to datasets (a–e in Fig. 19).

It could be seen in Fig. 21(a-c) that the *thd* curve for the reversed-chronological order is the first to enter the area of saturation. It is followed by random, chronological, DCF, and finally bi-directional curves.



(a) All *thd* measures

(b) *thd* measures in the rounded rectangle of (a)

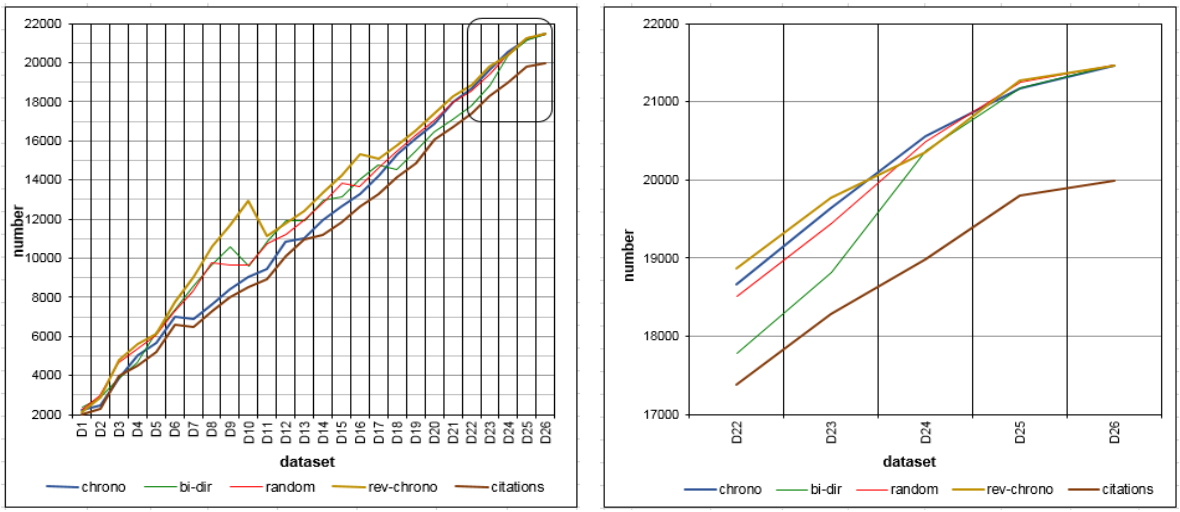
(c) Upper envelopes for *thd* measures in the rounded rectangle of (a)

(d) *thdr* measures

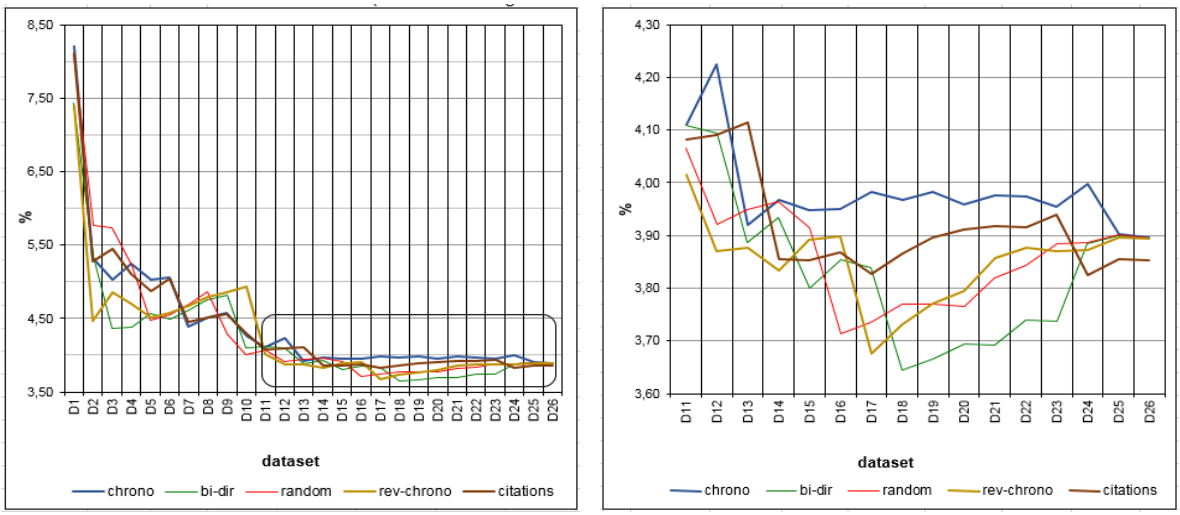
Fig. 21: DAC cleaned. The comparison of absolute (*thd*) and relative (*thdr*) terminological difference measures for different orders of adding documents to datasets (a–e in Fig. 19).

The picture of *thdr* curves in Fig. 21(d) shows very similar shapes compared to Fig. 21(a). The difference is that *thdr* curves go below *eps* substantially earlier than *thd* curves.

Fig. 22 pictures how different orders relate to each other by the numbers of retained terms and the proportions (ratios) of retained to all extracted terms. The diagrams at the right, like in Fig. 21(a), give finer-grained views of the areas in rounded rectangles of the diagrams at the left. As it could be seen in Fig. 22(a), the DCF order curve retains integrally the least number of terms, which results in the most compact saturated set of significant terms. The second integrally best order is chronological, though it is outperformed by the bi-directional order in the area, which is close to saturation (*D*22 – *D*26).



(a) Numbers of retained terms

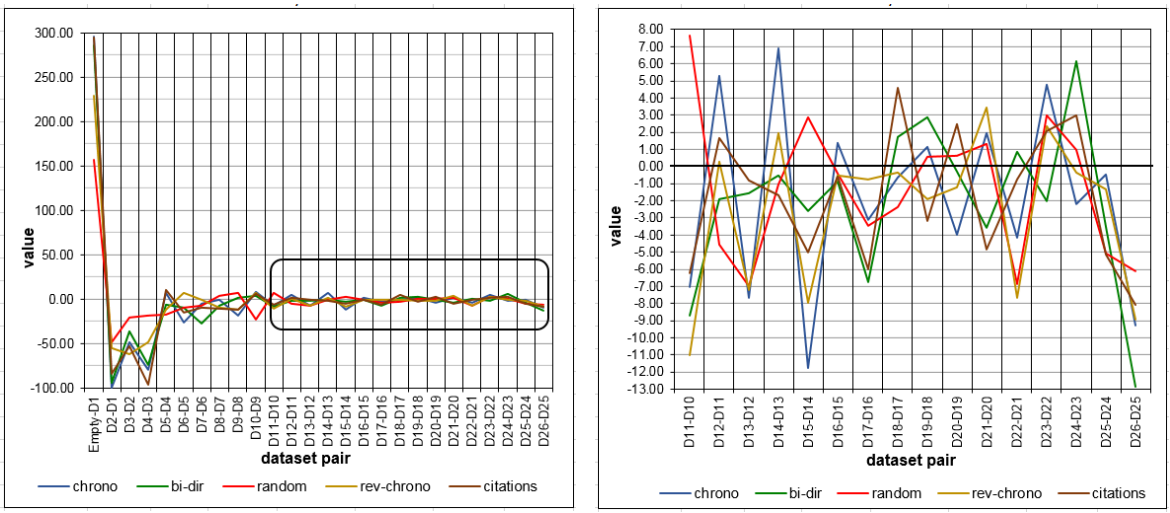


(b) Ratios of retained to all terms

Fig. 22: DAC cleaned. The numbers and ratios of retained terms for different orders of adding documents to datasets (a–e in Fig. 19).

Overall, in the saturation area, the best performance in retaining significant terms is demonstrated by the DCF order. It is followed by bi-directional, then – random, then – reversed-chronological, and, finally – chronological. Fig. 22(b) clearly pictures, however, that the chronological order yields the maximal ratio of retained to all extracted terms, which may be regarded as positive in terms of completeness. The second best ratio is yielded by DSF, the third is shared by random and reversed chronological. Bi-directional is a clear negative outlier in this aspect.

Finally, an integral view on the stability of different orders with respect to terminological saturation is given in the *thd* volatility diagram of Fig. 23. Fig. 23(a) embraces the whole set of the measured volatility values, while Fig. 23(b) presents in finer detail the area within the rounded rectangular in Fig. 23(a).



(a) (b)

Fig. 23: DAC cleaned. The comparison of *thd* volatility for different orders of adding documents to datasets (a–e in Fig. 19).

It can be noticed in the figure that the oscillations of volatility values for all the orders are quite substantial, perhaps due to insufficient dataset cleaning. To find out how different orders stand in terms of their volatility, the summation of the absolute volatility values has been done as given in Table 30.

Table 30: DAC cleaned. Absolute *thd* volatility values for all orders.

	Order				
Measurement Point	**Random**	**Bi-Directional**	**Chronological**	**Reversed-Chronological**	**DCF**
11	7.625783	8.69464	7.02945345	11.01485	6.250041
12	4.545857	1.88035	5.2762859	0.284332	1.647281
13	6.988732	1.57566	7.65656636	7.182291	0.808353
14	1.041535	0.5009	6.87731116	1.941638	1.676596
15	2.855447	2.61969	11.7593041	7.945231	5.032455
16	0.430211	0.85609	1.38433503	0.551942	0.607958
17	3.456982	6.7499	3.14389333	0.746046	5.985253
18	2.364968	1.71062	0.61996061	0.337561	4.590808
19	0.542809	2.86585	1.12974804	1.908199	3.162422
20	0.63338	0.31497	3.99316958	1.230595	2.471265
21	1.332161	3.60587	1.96980845	3.466348	4.858355
22	6.886722	0.87436	4.14880034	7.671209	0.76153
23	3.015471	2.02965	4.80335987	2.341824	2.064194
24	0.965149	6.14985	2.20463236	0.357095	2.974442
45	5.047942	3.51371	0.43922206	1.321804	5.154998
26	6.141314	12.8841	9.28111903	8.925865	8.086551
Sum:	**53.87446**	**56.8262**	**71.7169697**	**57.22683**	**56.1325**
Rank:	**1**	**3**	**5**	**4**	**2**

The calculations clearly show that the chronological order is the most volatile order. The rest of the orders demonstrate very similar integral volatility values with a very small advantage of the random order.

For summarizing the results of our experiments on **DAC cleaned**, it may be stated that all the orders, overall, demonstrated comparable results, been individually better in some aspects and worse regarding the other. Overall, for DAC cleaned, the best and most balanced performance was demonstrated by the **random** order. DCF was the second best, chronological and reversed-chronological were the third best. Bi-

directional was the negative outlier.

It could be stated, based on the **DAC cleaned** experiments, that **random** was **slightly better** than **DCF**, but not sufficiently for a definitive choice.

6.4 Comparison and Recommendations

This section summarizes and compares our findings and judgements after analysing the results of the experiments on cross-evaluating different orders of adding documents to datasets for measuring terminological saturation (using DMKD, TIME, DAC cleaned collections) or detecting excessive noise (using DAC naturelle collection). The summary is structured along the document collections and features we analysed in our experiments as presented in Table 31. In the order columns, the scores per feature are given for the performance of individual orders. The scores are summed per collection and the ranks of the orders are assigned – the higher the score, the lower the performance and the higher the rank. Finally, the scores and ranks for different collections are summed at the bottom of the table. Hence, the sum of the scores indicates how good or bad was an order, overall, regarding the chosen evaluation criteria. The sum of the ranks, however, outlines how stable was the performance of an order across all the collections and experiments.

It could be noticed in Table 31 that, overall, the order that demonstrated the **best performance** was **DCF**. This order appeared to be remarkably **stable and balanced** in its performance for all the experiments on all collections. This stability is proven by the facts that:

- It was the **second best** in all three cases of measuring saturation
- It was the **best** for detecting excessive noise

It is also clear from Table 31 that the **bi-directional** order is a **negative outlier**, as it both scored and ranked the highest for all the collections.

The situation is a bit unclear for chronological, reversed-chronological, and random orders (tied group). For instance, the reversed-chronological is the best in scores, but the worst in ranks among those three orders. The reason for that is that the number of the evaluation criteria in the case of DAC naturelle (the detection of excessive noise) is six but not eight as it is in the cases for measuring saturation. Therefore, the sums of the scores in the case of DAC naturelle are lower than in the other three cases. Reversed-chronological is the worst in the tied group in detecting excessive noise, which is reflected by the ranks. It was however less penaltized by the scores in this case, compared to the score penalties in the cases for measuring saturation. Hence, a fair way to do final ranking in the tied group is to use the sum of the ranks but not the sum of the scores – as reflected in the last line of Table 31.

Table 31: The comparison of the performance of different orders of adding documents to datasets in measuring terminological saturation or detecting excessive noise.

Collection	Order				
Features	**Chrono-logical**	**Reversed-Chrono-logical**	**Random**	**Bi-Direc-tional**	**DCF**
DMKD (saturation)					
Earliest in the saturation zone (*thd* lower than *eps*)	1	4	3	5	1
Integrally, the smallest No of retained terms	3	1	2	5	4
The smallest No of retained significant terms at saturation point	1	4	3	5	2
Integrally, the highest individual term significance thresholds *eps*	4	1	2	5	3
The highest individual term significance threshold *eps* at saturation point	5	1	3	2	4
Integrally, the highest proportions of all extracted to retained terms	1	5	4	3	2
The highest proportion of all extracted to retained terms at saturation point	1	5	3	4	2
Integrally, the least volatile in *thd* values	2	3	4	5	1
Score:	**18**	**24**	**24**	**34**	**19**
Rank:	**1**	**3**	**3**	**5**	**2**
TIME (saturation)					
Earliest in the saturation zone (*thd* lower than *eps*)	3	1	3	5	2
Integrally, the smallest No of retained terms	3	1	4	5	2
The smallest No of retained significant terms at saturation point	3	1	4	5	2
Integrally, the highest individual term significance thresholds *eps*	3	2	5	1	4
The highest individual term significance threshold *eps* at saturation point	2	3	3	1	5
Integrally, the highest proportions of all extracted to retained terms	5	3	2	4	1
The highest proportion of all extracted to retained terms at saturation point	5	1	2	3	4
Integrally, the least volatile in *thd* values	3	2	4	5	1
Score:	**27**	**14**	**27**	**29**	**21**
Rank:	**3**	**1**	**3**	**5**	**2**
DAC naturelle (excessive noise)					
Earliest *eps* peak	1	4	3	5	1
Height of *eps* peak	3	4	1	5	2
Earliest *thd* peak	5	5	3	2	1
Earliest no-of-retained-terms peak	1	3	3	5	1
Height of no-of-retained-terms peak	4	1	3	1	5
Steps to confirm excessive noise after no-of-retained-terms peak	3	2	1	5	3
Score:	**17**	**19**	**14**	**23**	**13**
Rank:	**3**	**4**	**2**	**5**	**1**
DAC cleaned (saturation)					
Earliest in the saturation zone (*thd* lower than *eps*)	3	1	1	5	4
Integrally, the smallest No of retained terms	5	4	3	2	1
The smallest No of retained significant terms at saturation point	4	2	1	5	3
Integrally, the highest individual term significance thresholds *eps*	1	3	3	5	2
The highest individual term significance threshold *eps* at saturation point	1	1	1	1	1
Integrally, the highest proportions of all extracted to retained terms	1	3	3	5	2
The highest proportion of all extracted to retained terms at saturation point	1	3	5	2	4
Integrally, the least volatile in *thd* values	5	4	1	3	2
Score:	**21**	**21**	**18**	**28**	**19**
Rank:	**3**	**3**	**1**	**5**	**2**
Sum of Scores:	**83**	**78**	**83**	**114**	**72**
Sum of Ranks:	**10**	**11**	**9**	**20**	**7**
Final Rank:	**3**	**4**	**2**	**5**	**1**

7 Conclusions and Future Work

This report presented our results in the development of the methodological components for extracting representative (complete) sets, having minimal possible size, of significant terms extracted from the representative sub-collections of textual documents having time stamps. The approach to assess the representativeness does so by evaluating terminological saturation in a document (sub-)collection.

One of the important aspects in the reported part of research work was that the constituent documents in a collection have been published at different times. Therefore, the temporal drift in terminology had to be appropriately taken into account. We focused on empirically investigating the proper ways to cope with this temporal drift and its influence on terminological saturation. Our premise was that there could be several different orders of adding documents to the processed datasets, dealing with the time or impact of a publication: (i) **chronological**; (ii) **reversed-chronological**; (iii) **random**; (iv) **bi-directional**; and (v) **descending citation frequency**.

We performed our experiments on three different real world document collections coming from different domains, where the collections of high-quality documents were available as scientific papers. The collections and the respective datasets were presented in Section 5. These collections also have different proportions of noise. So, we were able to assess the impact of different orders of adding documents in the presence of different levels of noise and also check if different orders are differently sensitive in detecting excessive noise.

For each collection, DMKD, TIME, and DAC we:

- Extracted the bags of terms from the prepared datasets using the UPM Term Extractor software
- Measured terminological saturation in the pairs of the extracted bags of terms using the THD module
- Measured the two additional characteristics that further helped us analyse the influence of order on terminological saturation. These were: (i) the **proportion** of the **retained** terms to **all extracted** terms in percent; (ii) the **volatility** of *thd* – a discrete analogue to the 1st derivative to *thd* values – computed as the different between the current and previous *thd* values.
- Built the diagrams and analysed the results

In addition to the above activities, for the DAC collection we also looked at the effect of removing stop terms after doing term extraction. By removing these stop terms, which represented the noise in the pre-processed documents, we de-noised the output. So, for the DAC collection we also compared noisy (DAC naturelle) and cleaned (DAC cleaned) bags of terms for all the orders.

In assessing the impact of an order of adding documents, we analysed the following measurable aspects:

- Which of the orders resulted in the earlier or later entry into the saturation zone (*thd* lower than *eps*)
- Which of the orders resulted in, integrally, the smallest No of retained terms
- Which of the orders resulted in, integrally, the highest individual term significance thresholds *eps*
- Which of the orders resulted in, integrally, the highest proportions of all extracted to retained terms
- Which of the orders resulted in, integrally, the least volatile *thd* values

In assessing the sensitivity of an order of adding documents to excessive noise, we based the comparison on the following measurable aspects:

- Which of the orders resulted in the earlier or later *eps* peak
- Which of the orders resulted in the higher or lower *eps* peak
- Which of the orders resulted in the earlier or later *thd* peak
- How many measurement steps were taken to confirm excessive noise after *eps* peak
- Which of the orders resulted in the earlier or later no-of-retained-terms peak
- How many measurement steps were taken to confirm excessive noise after no-of-retained-terms peak

Based on the results of the comparison, we recommended that the **descending citation frequency** order of adding documents to datasets might be used as the one demonstrating the best and most balanced performance.

This result sounds reasonable as the **descending citation frequency** order processes the documents in the datasets in the order of their descending impact on the domain, measured in citation frequency. Hence, taking documents with more impact first results in getting all the significant terms introduced in these documents earlier. This order, due to the use of citation frequency as a measure for impact, also balances those significant terms that are new to the terms that survived as significant through time. Due to that, it remedies terminological drift in time better that the other orders.

Our future research work will focus on answering our research question about finding an optimal size of a dataset increment for having quicker and more stable terminological saturation. For that, the **descending citation frequency** will be used as proven optimal.

Acknowledgements

The first author is funded by PhD grants provided by Zaporizhzhia National University, the Ministry of Education and Science of Ukraine, and the Cabinet of Ministers of Ukraine.

The research leading to this report has been done in part in cooperation with the Ontology Engineering Group of the Universidad Politécnica de Madrid in frame of FP7 Marie Curie IRSES SemData project (http://www.semdata-project.eu/), grant agreement No PIRSES-GA-2013-612551.

A substantial part of the instrumental software used in the reported experiments has been developed in cooperation with BWT Group.

The data collection of Springer journal papers dealing with Knowledge Management, including DMKD, has been provided by Springer-Verlag.

References

1. Kosa, V., Chaves Fraga, D., Naumenko, D., Yuschenko, E., Badenes, C., Ermolayev, V., and Birukou, A.: Cross-evaluation of automated term extraction Tools. Technical Report TS-RTDC-TR-2017-1, 30.09.2017, Dept of Computer Science, Zaporizhzhia National University, Ukraine, 60 p. (2017) online: http://ermolayev.com/TS-RTDS-TR-2017-1.pdf DOI: 10.13140/RG.2.2.31187.07207
2. Kosa, V., Chugunenko, A., Yuschenko, E., Badenes, C., Ermolayev, V., Birukou, A.: Semantic saturation in retrospective text document collections. In: Mallet, F., Zholtkevych, G. (eds.) Proc. ICTERI 2017 PhD Symposium, CEUR-WS, vol. 1851, pp. 1--8, Kyiv, Ukraine, May 16-17 (2017) online
3. Tatarintseva, O., Ermolayev, V., Keller, B., Matzke, W.-E.: Quantifying ontology fitness in OntoElect using saturation- and vote-based metrics. In: Ermolayev, V., et al. (eds.) Revised Selected Papers of ICTERI 2013, CCIS, vol. 412, pp. 136--162 (2013)
4. Osborne, F., Motta, E.: Klink-2: Integrating multiple web sources to generate semantic topic networks. In: Arenas, M. et al. (eds.): ISWC 2015, Part I, LNCS, vol. 9366, pp. 408--424 (2015) doi : 10.1007/978-3-319-25007-6_24
5. Ermolayev, V., Batsakis, S., Keberle, N., Tatarintseva, O., Antoniou, G.: Ontologies of time: review and trends. International Journal of Computer Science and Applications 11(3), 57--115 (2014)
6. Fahmi, I., Bouma, G., van der Plas, L.: Improving statistical method using known terms for automatic term extraction. In: Computational Linguistics in the Netherlands, CLIN 17 (2007)
7. Wermter, J., Hahn, U.: Finding new terminology in very large corpora. In: Clark, P., Schreiber, G. (eds.) Proc.3rd Int Conf on Knowledge Capture, K-CAP 2005, pp. 137--144, Banff, Alberta, Canada, ACM (2005) doi: 10.1145/1088622.1088648
8. Zhang, Z., Iria, J., Brewster, C., Ciravegna, F.: A comparative evaluation of term recognition algorithms. In: Proc. 6th Int Conf on Language Resources and Evaluation, LREC 2008, Marrakech, Morocco (2008)
9. Daille, B.: Study and implementation of combined techniques for automatic extraction of terminology. In: Klavans, J., Resnik, P. (eds.) The Balancing Act: Combining Symbolic and Statistical Approaches to Language, pp. 49--66. The MIT Press. Cambridge, Massachusetts (1996)
10. Caraballo, S. A., Charniak, E.: Determining the specificity of nouns from text. In: Proc. 1999 Joint SIGDAT Conf on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 63--70 (1999)
11. Astrakhantsev, N.: ATR4S: toolkit with state-of-the-art automatic terms recognition methods in scala. arXiv preprint arXiv:1611.07804 (2016)

12. Medelyan, O., Witten, I. H.: Thesaurus based automatic keyphrase indexing. In: Marchionini, G., Nelson, M. L., Marshall, C. C. (eds.) Proc. ACM/IEEE Joint Conf on Digital Libraries, JCDL 2006, pp. 296--297, Chapel Hill, NC, USA, ACM (2006) doi: 10.1145/1141753.1141819
13. Ahmad, K., Gillam, L., Tostevin, L.: University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In: Proc. 8th Text REtrieval Conf, TREC-8 (1999)
14. Sclano, F., Velardi, P.: TermExtractor: A Web application to learn the common terminology of interest groups and research communities. In: Proc. 9th Conf on Terminology and Artificial Intelligence, TIA 2007, Sophia Antipolis, France (2007)
15. Frantzi, K. T., Ananiadou, S.: The c/nc value domain independent method for multi-word term extraction. J. Nat. Lang. Proc. 6(3), 145--180 (1999) doi : 10.5715/jnlp.6.3_145
16. Kozakov, L., Park, Y., Fin, T., Drissi, Y., Doganata, Y., Cofino, T.: Glossary extraction and utilization in the information search and delivery system for IBM Technical Support. IBM System Journal 43(3), 546--563 (2004) doi: 10.1147/sj.433.0546
17. Astrakhantsev, N.: Methods and software for terminology extraction from domain-specific text collection. PhD thesis, Institute for System Programming of Russian Academy of Sciences (2015)
18. Bordea, G., Buitelaar, P., Polajnar, T.: Domain-independent term extraction through domain modelling. In: Proc. 10th Int Conf on Terminology and Artificial Intelligence, TIA 2013, Paris, France (2013)
19. Kosa, V., Chaves-Fraga, D., Naumenko, D., Yuschenko, E., Badenes-Olmedo, C., Ermolayev, V., Birukou, A.: Cross-evaluation of automated term extraction tools by measuring terminological saturation. In: Bassiliades, N., et al. (eds.) ICTERI 2017. Revised Selected Papers. CCIS, vol. 826, pp. 135--163 (2018)
20. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10 (8), pp. 707--710 (1966)
21. Monger, A., Elkan, C.: The field-matching problem: algorithm and applications. In: Proc. 2nd Int Conf on Knowledge Discovery and Data Mining, pp. 267--270, AAAI Press (1996)
22. Jaro, M. A.: Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. J Amer Stat Assoc 84(406), 414--420 (1989)
23. Winkler, W. E.: String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In: Proc. Section on Survey Research Methods. ASA, pp. 354--359 (1990)
24. Jaccard, P.: The distribution of the flora in the alpine zone. New Phytologist 11, 37--50 (1912)

25. Hamming, R. W.: Error detecting and error correcting codes. Bell System Technical Journal 29(2), 147--160 (1950)
26. Dice, L. R.: Measures of the amount of ecologic association between species. Ecology 26(3), 297--302 (1945)
27. Badenes-Olmedo, C., Redondo-García, J. L., Corcho, O.: Efficient clustering from distributions over topics. In: Proc. K-CAP 2017, ACM, New York, NY, USA, Article 17, 8 p. (2017)
28. Corcho, O., Gonzalez, R., Badenes, C., Dong, F.: Repository of indexed ROs. Deliverable No. 5.4. Dr Inventor project (2015)
29. Ermolayev, V.: OntoElecting requirements for domain ontologies. The case of time domain. EMISA Int J of Conceptual Modeling 13(Sp. Issue), 86--109 (2018)
30. Gomaa, W. H., Fahmy. A. A.: A Survey of Text Similarity Approaches. Int J Comp Appl 68(13), 13--18 (2013)
31. Yu, M., Li, G., Deng, D., Feng, J.: String similarity search and join: a survey. Front. Comput. Sci. 10(3), pp. 399--417 (2016)
32. Arnold, M., Ohlebusch, E.: Linear Time Algorithms for Generalizations of the Longest Common Substring Problem. Algorithmica 60(4), pp. 806--818 (2011)
33. Huang, A.: Similarity Measures for Text Document Clustering. In: Proc. 6th New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand, pp. 49--56 (2008)
34. Sørensen, T.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Kongelige Danske Videnskabernes Selskab 5 (4), pp. 1--34 (1948)
35. Singhal, A.: Modern information retrieval: a brief overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4), pp. 35--43 (2001)
36. Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K.: WordNet: An online lexical database. Int. J. Lexicograph. 3(4), 235--244 (1990)
37. Minkowski, H.: Geometrie der Zahlen. Bibliotheca Mathematica Teubneriana, Band 40 Johnson Reprint Corp., New York-London, 256 pp. (1968) – in German
38. Park, Y., Byrd, R. J., Boguraev, B.: Automatic glossary extraction: beyond terminology identification. In: Proc. 19th Int Conf on Computational linguistics, pp. 1--7. Taipei, Taiwan (2002) doi : 10.3115/1072228.1072370
39. Nokel, M., Loukachevitch, N.: An experimental study of term extraction for real information-retrieval thesauri. In: Proc 10th Int Conf on Terminology and Artificial Intelligence, pp. 69--76 (2013)
40. Zhang, Z., Gao, J., Ciravegna, F.: Jate 2.0: Java automatic term extraction with Apache Solr. In: Proc.LREC 2016, pp. 2262--2269, Slovenia (2016)

41. Justeson, J., Katz, S. M.: Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering 1(1), 9--27 (1995). doi: 10.1017/S1351324900000048
42. Evans, D. A., Lefferts, R. G.: Clarit-trec experiments. Information processing & management 31(3), 385--395 (1995) doi: 10.1016/0306-4573(94)00054-7
43. Church, K. W., Gale, W. A.: Inverse document frequency (idf): a measure of deviations from Poisson. In: Proc. ACL 3rd Workshop on Very Large Corpora, pp. 121--130, Association for Computational Linguistics, Stroudsburg, PA, USA (1995) doi: 10.1007/978-94-017-2390-9_18
44. Oliver, A., V`azquez, M.: TBXTools: a Free, Fast and Flexible Tool for Automatic Terminology Extraction. In: Angelova, G/, Bontcheva, K., Mitkov, R. (eds.): Proc. Recent Advances in Natural Language Processing, pp. 473--479, Hissar, Bulgaria, Sep. 7-9 (2015)
45. De Boom, C., Demeester, T., Dhoedt, B.: Character-level recurrent neural networks in practice: comparing training and sampling schemes. *Neural Comput & Applic* (2018). https://doi.org/10.1007/s00521-017-3322-z