Documentary analysis of STAB and ISSW proceedings

Hélène Dartois, *Stagiaire de l'École nationale des Chartes*, helene.dartois@chartes.psl.eu
Paul Creismeas, *DGA Hydrodynamics 27105 Val de Reuil, France*, paul.creismeas@intradef.gouv.fr
Jean-François Leguen, *DGA Hydrodynamics 27105 Val de Reuil, France*, jean.françois.leguen@intradef.gouv.fr

ABSTRACT

From the beginning, the Stability R&D Committee (SRDC) made a huge effort to collect and dematerialize all the proceedings of the STAB and ISSW congresses since the first editions. All these dematerialized documents have been uploaded on the website www.shipstab.org. Search is now possible directly on the website. It is proposed to present work on the use of these data. The documentary tools presented will allow better search and an optimized classification of articles. Roadmap is suggested, such as the harmonization of keywords, providing recommendations to future conference STAB and ISSW authors and organisers.

Keywords: *stability, congress, documentation*

1. INTRODUCTION

The work presented in this paper was initiate by SRDC (Stability Research & Development Committee), with its help and under its control, in order to promote and disseminate the work presented during STAB and ISSW congresses which is a part of the SRDC mandate.

2. PREVIOUS WORK DONE

SRDC efforts

Since the beginning of SRDC (first meeting in Washington DC in 2011 during ISSW 2011), efforts were done to work on papers from STAB and ISSW congresses. Because it seemed that nobody has the complete set of proceedings, the first step was to centralise proceedings of all conferences from the first STAB conference in 1975 and the first ISSW in 1995, both in Glasgow. With help of all, it was possible with reasonable efforts to find a version, at least a paper version for oldest years, of all congresses and workshops. The second steps was the OCR process of every proceedings. This task was shared in many places. It was chosen to separate papers in several independent "pdf" format files. In parallel of this process, tentative to fill metadata of the files to facilitate research was performed but this task was never fully ended because it is a very time consuming task if it is handmade.

At the end of those first steps, it was possible to upload about a thousand independent "pdf" files of all STAB and ISSW papers on STAB/ISSW website, www.shipstab.org. Many files contains metadata (as date and place of the conference, session name, authors and including keywords) and have standardized names (made from author names to facilitate researches). An index was also written in an MS Excel file with usual information in order to facilitate the searching process outside of internet and to propose a standalone CD-ROM, which is useful or mandatory in some situations.

Now the website is up-to-date at the end of every congress and contain a search engine. In order to improve searching process and output information from this huge database it was decided at DGA Hydrodynamics to try some new tools and new methodologies.

The initial corpus

At DGA Hydrodynamics a documentation about ship stability expertise field is needed and expected in order to be able to produce very significant researches and tests. The SRDC helps the creation of a specific corpus for this test center since its creation in 2011. One of its objectives was to centralize every article about ship stability for tests centers around the world. The corpus presented at DGA Hydrodynamics is made of more than 1400 PDF files of nearly all the STAB conferences and ISSW workshops. This corpus is increased every

year with new documents from ISSW or STAB which proves the daily effectiveness and usefulness of those documents in this test center.

As previously mentioned, every congress and workshop books were split into smaller chunks in order to obtain a single PDF file for each article. This way, the indexation in an MS Excel file was easier and more relevant. Plus it helps the creation of more complete metadata for each article.

Through the year, DGA Hydrodynamics has obtained a nearly complete set of articles which is an advantage when it comes to make precise analysis on vocabulary or on the different point of interests about ship stability through the year.

3. NEW OBJECTIVES

The website: www.shipstab.org gives access to every STAB and ISSW articles plus thesis but there is some limits that need to be improved. The search engine doesn't allow the researcher to do a precise research on metadata, it only allows us to search in a full-text mode. The result page is also quite limited: every result for a research gives an access to a more specific result page where we can download the full proceeding book and not the article that we are interested in. The results do not specify on which article the keyword was found, which means we have to search within the PDF of all articles before actually finding the result announced on the search engine.

The work beginning at DGA Hydrodynamics intend to overcome those limits. The tools we propose will not only allow advanced research on the content of the article but also on the metadata fields of each article in order to study the evolution of stability work over the year. This paper also submits recommendations to authors and organisers to obtain a standardization of the outputs of congresses, a bibliographic structure easy to extract and a more precise research engine.

4. DOCUMENTATION TOOLS

The ISSW and STAB corpus collected by SRDC are made of PDF, some of them are quite old which means the OCR is not perfect. As most PDF, they are readable with any PDF viewer but they are not easily modified which is a problem when it comes to make those corpus more usable and searchable for the engineers. For instance, it's nearly impossible to add rich metadata to a PDF file and this can be a challenging point when it comes to give access to technical and scientific information. A corpus of technical documents is interesting only if we can search precisely on it and structured metadata are key to search and to use effectively any kind of corpus.

In order to help engineers getting a privileged access to this scientific and specific ship stability documentation, different tools were created to search onto those documents. Since the beginning of the project an emphasis was made on the use of freeware and license free software. The aim is to propose a re-usable and easy system to access the different corpus and to search them at DGA or in other institutions. The accessibility can only be guaranteed by freeware and license free programs. Most of the tools were coded in python language and required a python 3 version to run. At the moment, all the tools are available only on one computer, but if the solutions created are good enough, the different codes could be combined into the shipstab website and thus becoming accessible for everyone without the obligation to install python on a computer.

The OCR process

For older articles, a primary step is necessary before using the GROBID API to get the TEI : the articles need to be pass on a OCR software.

Sometimes, the OCR is quite difficult to produce because of the low quality of the original document (archives can be nearly unreadable or an old printed version of an article can be of low quality which induce many errors during the OCR process). To reduce the number of errors or to increase the effectiveness of the OCR process an simple image treating process can be made. At DGA-Hydrodynamics, M. Paul CREISMEAS uses the Omnipage Ultimate software, which is not a license free software but it allows us to treat the quality of the image by adjusting the contrast and luminosity, the orientation of the page, selecting the content zone, etc.

Those image treatments are essentials to produce a good quality TEI document and to use it effectively without spending hours on corrections over the original document. This all process is just an

example of what is possible to do with old article that still have a scientific interest.

GROBID

First and foremost, it was necessary to convert the PDF files into a format that allows us to add metadata and/or to specify the existing metadata.

To do so, the GROBID API[1] was used: "GROBID" is a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured TEI[2] -encoded documents"[3]. This new format allows us to create more accurate metadata or to increase the already existing metadata. To do so, we use the XML Copy Editor program which is, like GROBID, a license free program.

XML Copy Editor

This software is design to write XML type document. It provides a validation tool to verify the validity of the document towards TEI guidelines. This way, we make sure the document can be used by another XML program and can be exchange without damages. We are only modifying or creating metadata for each document, we will never modify the content of the document in itself.

Before adding anything to the document we create a list of essential metadata. This list contains every field that will be searchable on the application:

- author (forename, surname),
- affiliation (organization name and type (research, certification, industrial, test center), address (country, settlement),
- title,
- keywords,
- abstract.

We tried to create the smallest list possible because adding too much metadata will be time consuming and it will be in contradiction to our main objective: creating a simple application to search the collection of articles.

The TEI format is made to structure a document and add some metadata in order to explore a text document in different aspects such as metadata, specific formulas or bibliographic references. But like every markup language it is quite a heavy format to use to exchange data or documents. We choose to convert the new TEI document, once the metadata were added, to another web oriented format called JSON[4] (pronounced « Jason »). This format preserved all the information added with XML modifications but it simplifies the document structure and it work as an array in JavaScript. The conversion was made with a python scripts which parsed XML files and convert them into JSON files. All those files contain the necessary metadata to search precisely onto the documents.

Elasticsearch

Once the JSON files are ready we add them to a search engine program. Searching in full-text was one of the main requirement for the application besides the possibility to use structured metadata and to do so, we decide to use Elasticsearch[5], a JSON document oriented search engine compatible with Python, JavaScript, PHP etc. The JSON documents are indexed into Elasticsearch clusters according to their structure. This way, it will be possible, with a web application, to do full-text search and very specific and precise researches in the corpus such as a research by organization type or by bibliographic reference. The application will be coded to search on multiple fields of metadata in the documents such as authors, institution, date, keywords, abstract, and to search on plain-text, bibliographic references etc. On the frontend, the user deals with a search page and the application will return the PDF file for each response. That way, we made sure that none of the document could be modified or deleted by the user, the application only gave access to the non-changeable document, the PDF file.

5. PROCESS

The mind-map (Figure 1), represents the process to create the research application with all the tools

[1] API is an application programming interface.
[2] Text Encoding Initiative P5, the last version was launched in 2007 by Lou Burnard and Syd Bauman with the TEI Consortium.
[3] Grobid documentation can be found at : https//grobid.readthedocs.io/en/latest/

[4] Java Script Object Notation is a web exchange format created by JavaScript, during the 2000's, to simplify the data communication on the web.
[5] Elasticsearch is an API that allows to index document in a JSON format and to search upon the documents with a REST API compatible with many other langages such as PHP or Python.

that will be deployed at DGA Hydrodynamics. This process is already an ambitious one because it requires multiple actions to obtain the structured document in a good format before the indexation on a search engine. But this is also a starting point for a more precise analysis of this important documentation and it will give a privilege access to the content in a very simple way.

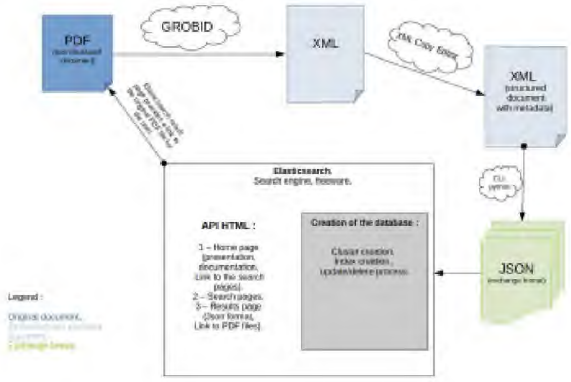

Figure 1: STAB and ISSW treatment process

In this mind map we assume the document is already in a good quality and in a PDF format. If the document needs to be dematerialized, an OCR process can be useful. Once this step is done, we first transform our PDF file into an XML document with the web API GROBID. Once we have obtained this new document we enrich the metadata with the XML Copy Editor software. Then, we transform the XML into a JSON format that simplify data exchange on the web and on the application. This conversion is made with a python script. Finally we add the JSON file to our Elasticsearch cluster so that we can interrogate it with all our other articles.

After all those steps, the documents and their metadata and content are searchable and ready for a more complete analysis.

6. EXAMPLE

In this part, we introduce examples to demonstrate what it is possible to obtain through the analysis of documents metadata. To do so, we extract from the whole initial corpus a subcorpus on which the work can be realizable by hand, the tools described upper in the text are under construction. This subcorpus is comprised of the communications from ISSW 2013, 2014, 2016 and 2017 and from STAB 2015 and 2018, totalling approximately 260 documents to be further analysed.

The set of metadata connected to each of these documents are keywords in relation with the activities in stability domain. To identify telling keywords, we are guided by a map depicting the main activities of the domain and the relations among them, Figure 2.

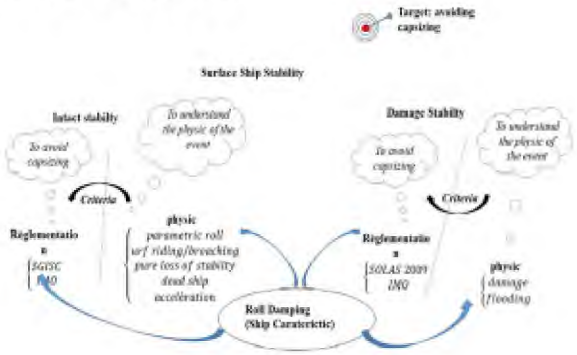

Figure 2: Map of the activities connected to the stability domain and the relations among them.

The list of the chosen keywords is given below:

- Roll
- Seakeeping
- Manoeuvrability
- Offshore
- Experimental technics
- Probability
- Intact Stability
- Damaged Stability

At each keyword, we associate a measure of the importance of the keyword in the documentation related to an element of the subcorpus. Such a measure is performed by computing the *density*, also called:

Term Frequency-Inverse Document Frequency, "Tf-Idf" (Leydesdor, 2011), (Thijs, 2011), which is a weighted counting of the keyword among the documentation. For example, we assess the importance of the activity *Roll* in the congress STAB 2015, $d_{Roll,STAB2015}$ through the following formulae:

$$d_{Roll,STAB2015} = \frac{N_{Roll,STAB2015}}{N_{STAB2015}} . log \frac{N_{subcorpus}}{N_{STAB2015}} \quad (1)$$

With:

- $N_{Roll,STAB2015}$ is the number of communications from the congress STAB 2015 concerned with rolling phenomena,

- $N_{STAB2015}$ is the total number of the documents from the congress STAB 2015,
- $N_{subcorpus}$ is the number of the considered documents in the subcorpus,
- *log* is the decimal logarithm function

In Figure 3 and Figure 4, the evolution of all the keywords is depicted versus ISSW workshops and STAB congress, in a chronological order. We must be careful before drawing out definitive conclusions, but it is very interesting to notice the importance of both *Roll* and *Seakeeping* activities during the period over six years from 2013, figure 3. But the other figure, figure, show us straightforwardly that the main point of concern over this period of six years is the *intact stability*.

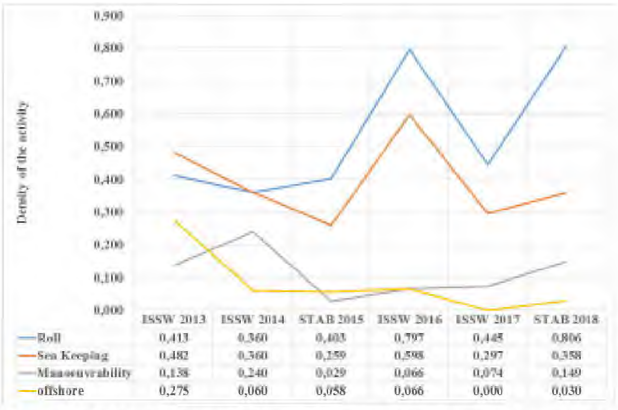

Figure 3: Evolution on the importance of some keywords related to intact stability versus the workshops ISSW and congress STAB, chronological order over a range of six years

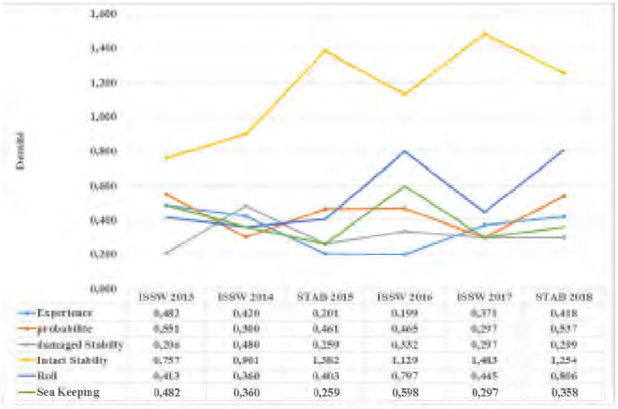

Figure 4: Evolution on the importance of some keywords versus the workshops ISSW and congress STAB, over a range of six years.

7. RECOMMENDATIONS

Because it was an objective of this study, preliminary version of recommendations are suggested below.

Recommendations to authors

Those analysis and uses of the corpus are only possible if a few recommendations are filled. Those recommendations do not requires a great amount of time for the authors but it determines the success of the future researches and investigations.

First, we must underline the importance of creating a structured article. This can be easily achieved with the article guide given by the organisers of each congress. The article guide is conform to GROBID training data so the result are accurate and quick. Moreover, this guide presents example for bibliographic references and figure model.

Another point of interest is on formulas. GROBID was not trained to achieve good results while recognizing mathematical formulas. But one of the possible evolution of the project could be to make research over specific formulas. In order to try to achieve this objective, authors could, when possible, name the formulas they used with a specific sentence such as: "GM calculation formula:" followed by the mathematical formula. This would help creating a model to search on all the articles.

Lastly, we recommend that the session name appeared as a keyword in the article whenever it is possible or added by organisers to increase the research possibility. This way, it would be possible to find every article belonging to the same session and to the same theme for a specific congress or workshop.

Recommendations for references

The use of the article guide with a bibliographic standard will give the possibility to generate a stability bibliography from all articles, and thus to have access to a large bibliography about this scientific field that can be increased regularly. We recommend the use of the APA[6] (2009) standard for bibliographic reference because the GROBID API

[6]The American Psychological Association created a bibliographic standard for students and psychologist in 1929.

model data use the APA standard to recognize any bibliographic references present in a document

This standard is well known among researchers and it also a standard that can be used when we want to produce a bibliography with Latex for example.

You can find all the information and examples to use this standard on the official website: http://www.apastyle.org/. This site also contains a tutorial presenting the basic rules for the standard.

Recommendations to organisers

To help the use of the STAB and ISSW corpus, it is important to maintain the habit to create an update of the database for every STAB and ISSW session. This update should present, at least, a few metadata field such as author(s), title, organization name and type, keywords and abstract. This will help adding metadata and update the Elasticsearch cluster regularly.

In the tool presentation section we made a list of the fields we want to be able to search within the python application. But all those fields are not required in an article. The only fields that we recommend to be put in every article are the following:

- title,
- author(s),
- keyword(s), preferably separated with commas,
- name of session,
- abstract.

Those fields represent the core metadata of each article. They will be used to search the document database.

For future congresses, we suggest that every participant has an easy access to a small documentation. This documentation have to work as a reminder of good practise while writing an article. This way it will be easier to have a homogeneous corpus of document that share the same structure.

This can be easily achieved by adding the documentation directly on the www.shipstab.org. This documentation page, on the website, should at least contain a guide for article redaction with the excepted structure. This will encourage author to write in a re-usable format for GROBID for instance. A link to the APA standard for bibliography would also help the redaction of article. Plus, the use of the standard will help for a better indexation on Google Scholar or on a university website. Finally a short documentation about the expected metadata for each article would help reducing the time to treat every article after the conferences.

8. REFERENCES

APA, Publication Manual for American Psychological Association, (6th edition), Washington DC, 2009.

Leydesdor, L. and Welbers, K., "The semantic mapping of words and co-words in context". Journal of Infometrics, 2011.

Thijs, B., "Mapping of science". In www.scientometrics-school.eu, editor, ESSS European Summer School for Scientometrics. 2011. Leuven, Belgium.