KNOWLEDGE-BASED INFORMATION EXTRACTION FROM DATASHEETS OF SPACE PARTS

F. Murdaca [1], A. Berquand [1], K. Kumar[2], A. Riccardi[1], T. Soares[3], S. Gerené[4], N. Brauer[5]

[1] University of Strathclyde, Glasgow, UK
[2] spacejunkies V.O.F. (satsearch), Noordwijk, the Netherlands
[3] ESA, Noordwijk, the Netherlands
[4] RHEA group , Leiden , the Netherlands
[5] AIRBUS, Bremen, Germany

ABSTRACT

Selection of the right space parts is an essential step during the design of complex engineering systems and requires information that is typically embedded in unstructured documents like datasheets, Interface Control Documents (ICDs) and technical manuals. Satsearch (https://satsearch.co) aims to consolidate global space supply chain information within a single platform, by converting unstructured datasheets into machine-readable, human-readable, electronic datasheets (EDS). After satsearch's initial, manual efforts at generating EDS from source datasheets for space parts, they realized that the process is not scalable. A possible solution is to employ knowledge base information extraction systems. The Design Engineering Assistant (DEA) team from the University of Strathclyde is currently working on the automation of the extraction of information from unstructured documents (e.g. textbooks, reports, datasheets, research papers, etc.) through the development of an expert system. This paper summarizes the approach and outcomes of a feasibility study for the DEA, assessing benefits and obstacles for the implementation of a fully-automated information extraction process, focusing at this stage only on datasheets for space parts for preliminary mission design.

1. INTRODUCTION

In the era of big data, new, cutting-edge solutions have to be developed to be able to harness and extract information for advanced applications in many fields. In particular, in the space field, data is generated across the lifecycle of a mission. Not only is a vast amount of data generated during satellite operations, but also during the design phases, as mission complexity increases rapidly. There is a huge source of knowledge collected and stored during the mission design process that is underutilized. These data can be structured or semi-structured, but are mainly unstructured. The manual parsing of unstructured documents is time-consuming, subjective, difficult, error-prone, and can be tremendously costly. During the preliminary phases of a design study, some experts within the team might not be aware of existing similar studies or types of design, for example due to lack of time to conduct extensive research or lack of knowledge about old missions. The importance of having quick and easy access to lot of relevant information could benefit the experts' daily work, relieving their workload. It is important to explore the use of this knowledge, especially in the preliminary phases of the design, as it could benefit concept generation, analysis, and trade-off studies; hence saving time for more critical tasks.

One possible solution for the use of these data is their conversion into machine-readable format that can be easily retrieved and exploited by algorithms. Information Extraction (IE) is a process used to convert unstructured information embedded in text into structured data, for example used to populate a knowledge base to enable semantic search. Knowledge-based IE instead is a subfield of IE where a knowledge base is used to guide the information extraction process from text [8]. It contains the model of the domain of interest which is used to support the extraction of the data.

The paper will focus on the description of the data adopted in this preliminary study of a knowledge base information extraction system, in the frame of an expert system (section 2), the description of the manual extraction approach (section 3) and of the automatic extraction approach (section 4), both used to parse the source documents, and the results of the comparisons between the two approaches (section 5). The last part is dedicated to the main issues that were discovered for automatic extraction and the conclusions of this project (section 6).

2. SOURCE DATA USED FOR COMPARISON

Since the goal of this study is to investigate the use of automatic algorithms to extract information from unstructured documents, to support the space mission design process, this section is devoted to describing the source dataset. Our main interest is to compare the results achieved using computer algorithms against a manual, human-centered approach to convert unstructured documents into structured information. For a fair comparison, to enable assessment of the effectiveness of the automatic extraction process, a benchmark is needed. The benchmark for our study was generated by the satsearch team, by manually extracting data and populating EDS [9] from a common source dataset. Section 3 details this manual approach. Subsequently, section 4 provides insights into EDS generation using the automatic approach from the same source dataset. In this section, we describe the common source dataset that enables us to compare the results using both approaches.

The source dataset we chose for this study consists of a collection of datasheets. A “datasheet” is a means of communication between suppliers and potential buyers [4]. In particular, datasheets provide technical and non-technical details about specific products, services or technologies offered by suppliers. They are also often called specification sheets or specsheets, as they provide high-level technical specifications like mass, power modes, operating temperature, interfaces, etc, that are used in preliminary design. In a nutshell, the objective of a datasheet is to provide clear and unambiguous information, to foster trade and communication. The importance of a datasheet therefore lies in the accurate specifications and the detailed information it contains.

Developed in a paper format for decades, datasheets are characterized by different layouts from different organizations, potentially incorporating dissimilar levels of information. For spacecraft design, a paper, document-centric approach requires manual translation of documents like datasheets to simulators and mission control system databases [5], which is both time-consuming and error prone. Datasheets in the space industry are often found on supplier websites or are handed out during conferences & exhibitions. They are typically a few pages long, and include text, figures, and tables. Since there are no standards at present for the presentation of datasheets in the space industry, the task of combing through these documents to find the information necessary for space mission design can be laborious. Fig. 1 illustrates the severity of the lack of standardization of datasheets through a mosaic of thrusters available on the market, from a variety of suppliers spread across the world. Not only is the presentation of information dependent on the supplier; in addition, the choice of data to present and the terminology or language used to describe properties of a product or service also vary between suppliers.

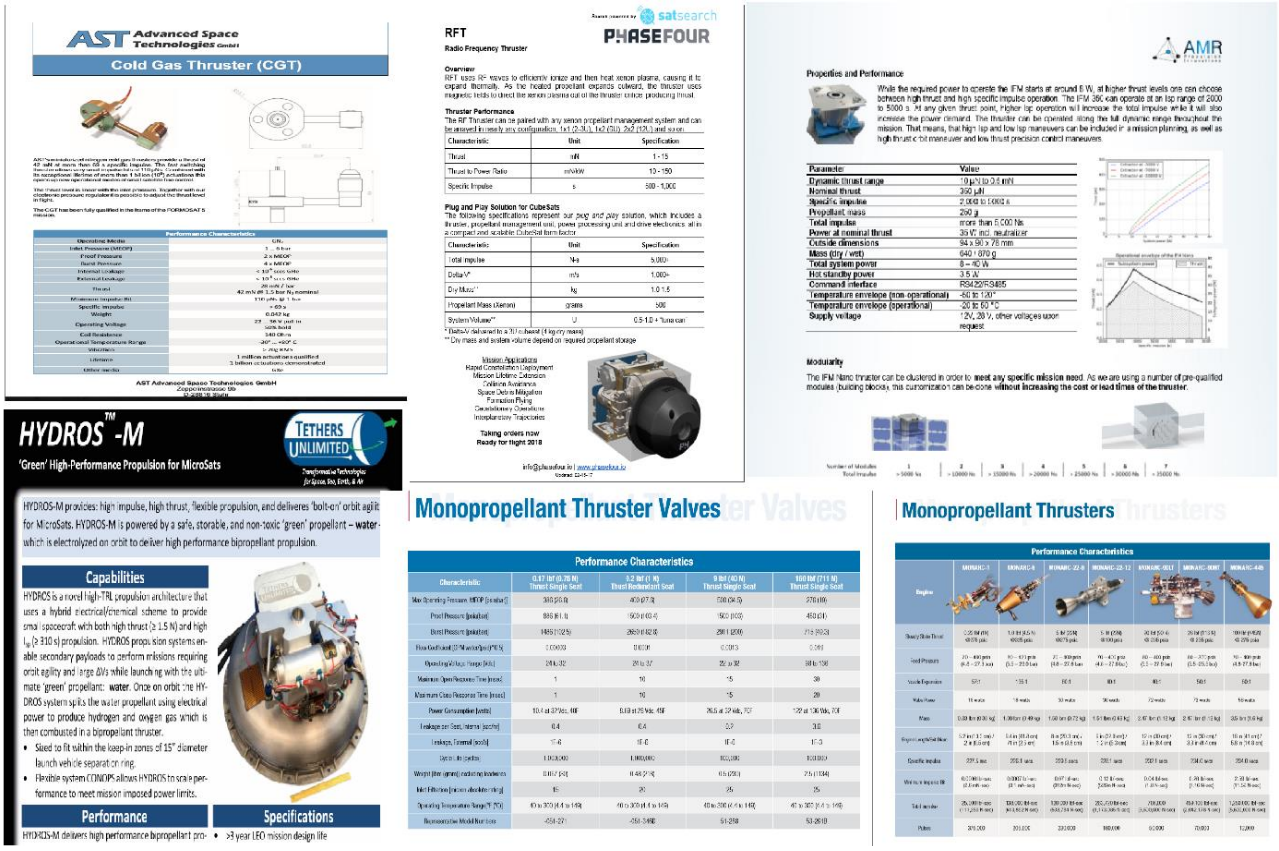


Figure 1. Mosaic of datasheets created by suppliers, highlighting the lack of standardization

Although the ultimate goal of this study is to develop robust, automatic, generic extraction algorithms to convert arbitrary datasheets in the space industry into EDS, this is not a problem that can be solved in one fell swoop. Instead, in this paper, we considered datasheets for a set of reaction wheels, as the basis for our comparison between the manual and automatic approaches. Reaction wheels, which are actuators used for attitude control, are employed in many spacecraft design concepts that require a degree of pointing control. Given the diversity of reaction wheels on the market, we chose this as a representative product to investigate the effectiveness of the automatic extraction approach compared to the manual approach.

For this study, we selected 23 reaction wheel datasheets collected by satsearch, containing a total of 33 products (some datasheets contain multiple products). These were obtained from various supplier websites and were made available to the authors by satsearch, as the original PDF documents created and disseminated publicly by suppliers.

The following sections will describe the approach towards converting these source datasheets into EDS using the two approaches, followed by a comparison of the results.

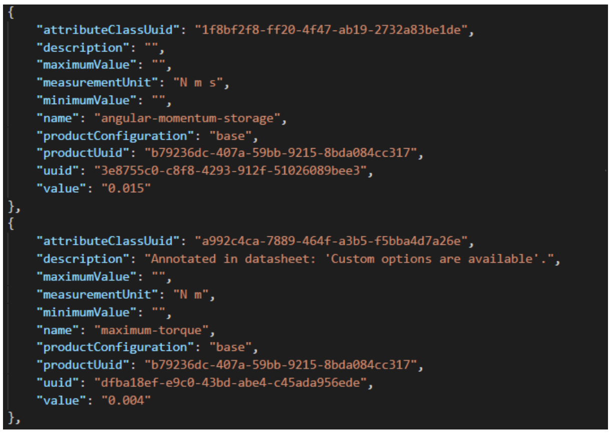
3. MANUAL APPROACH

Supply chain knowledge sits at the core of the space mission design process and is currently highly scattered, incomplete and unstructured. The growing challenges faced by the space engineering community in handling complex design data underpins satsearch's mission to curate, harmonize, and structure supply chain knowledge. Satsearch is developing a richer, structured, standardized format of representation for supply chain knowledge that will enable deep, complex querying to support optimization, sensitivity analysis, and risk mitigation during the design process [7].

To achieve this goal, satsearch started by converting original PDF datasheets into EDS, which is a cumbersome and lengthy process. For this study, instead of taking a "top-down" approach, and developing a formal ontology, or language, to describe reaction wheels, satsearch made use of a "bottom-up" approach. The idea of this bottom-up approach was to discover the language used by suppliers to communicate information about their reaction wheels to engineers. In this manner, the end result would provide more insight into the actual language used to transact across the supply chain, rather than a formal approach based on ideal, theoretical assumptions. In effect, satsearch engineers studied all 23 datasheets in the common dataset (section 2) and developed a strategy to extract the high-level specifications captured within each document.

By fully describing all the attributes of a datasheet in a model-based manner, the structured format underpinning the satsearch database allows engineers to evaluate space products much more effectively. One of the key features of the EDS developed by satsearch is that every product attribute uniquely addressable. Fig. 2 provides a snapshot of a EDS, illustrating the idea of unique attribute IDs.

The satsearch team generated EDS for all 33 reaction wheels. The EDS were made available for this study through the satsearch Application Programming Interface (API) (https://api.satsearch.co), which enables engineers to query against supplier and product attributes. The EDS generated in this manner manually was treated as the benchmark to compare against the results of the automatic extraction approach described in section 4.



```
{
    "attributeClassUuid": "1f8bf2f8-ff20-4f47-ab19-2732a83be1de",
    "description": "",
    "maximumValue": "",
    "measurementUnit": "N m s",
    "minimumValue": "",
    "name": "angular-momentum-storage",
    "productConfiguration": "base",
    "productUuid": "b79236dc-407a-59bb-9215-8bda084cc317",
    "uuid": "3e8755c0-c8f8-4293-912f-51026089bee3",
    "value": "0.015"
},
{
    "attributeClassUuid": "a992c4ca-7889-464f-a3b5-f5bba4d7a26e",
    "description": "Annotated in datasheet: 'Custom options are available'.",
    "maximumValue": "",
    "measurementUnit": "N m",
    "minimumValue": "",
    "name": "maximum-torque",
    "productConfiguration": "base",
    "productUuid": "b79236dc-407a-59bb-9215-8bda084cc317",
    "uuid": "dfba18ef-e9c0-43bd-abe4-c45ada956ede",
    "value": "0.004"
},
```

Figure 2. A section of a satsearch Electronic Data Sheet, illustrating uniquely identifiable attribute IDs

4. AUTOMATIC APPROACH

The Design Engineering Assistant (DEA) project aims at developing an expert system that will interact with the experts working in a Concurrent Design Facility (CDF) to enhance and ease their work, by providing them with new insights derived from large amount of knowledge accumulated within their field. DEA needs to be able to learn from past and future knowledge from structured, semi-structured and mainly unstructured data sources. Unstructured data sources are the most challenging to be dealt because algorithms have to be careful developed to enable automatic conversion into a machine-readable format.

One strategy foresees two phases: the generation of an ontology that captures all relevant concepts, relationships, hierarchies and rules extracted from the data for the specific domain of interest; the population of the ontology with data useful for space mission design.

The selection of the type of ontology, the ontology language and the reasoner are not trivial. The final decision will be taken through a deep analysis of the different configurations and at the same time taking into consideration the requirements for the tool. These requirements will be collected from two sources: the first source are the data, therefore the type of documents and the information inside them; the second source are the users because depending on the answers they want to obtain from the expert system they will also foster the use of a specific solution.

DEA architecture is composed by two main tools, smart-dog and smart-squid. In this paper we focus on the former one. In the frame of DEA, it is used for knowledge base generation and population, enrichment, consistency check, merging and evaluation with the reasoner. It contains a modular architecture which allows flexibility for the different tasks that shall be performed. [6] In the frame of this paper, we focused on the domain ontology population task using a simple ontology for the IE process.

5. COMPARISON OF EXTRACTED DATA BETWEEN SATSEARCH AND DEA

One of the purpose of this paper is to show the potential of the use of Artificial Intelligence (AI) methods to perform the extraction of data from unstructured documents. Two strategies have been identified to conduct the comparison between manual extracted data (satsearch) and automatic extracted data (smart-dog):

1) Compare the two ontologies, the one built by satsearch and the one built through the automatic tool (smart-dog) used in the DEA project;
2) Compare the extracted data (manual vs automatic) using a simple ontology (using some part of the one provided by satsearch but adding other concepts), mainly a terminological ontology.

We opted for the second strategy, therefore the data from satsearch API have been extracted to be part of a simple ontology for smart-dog.
In order to prove the feasibility and potential of this preliminary work, two constraints have been considered:

- ❑ Consider datasheets of one specific space part: reaction wheels
- ❑ Consider a subset of all the possible attributes available for that specific space part. The choice of this sublist has been made because it represents a typical spectrum of information used during the preliminary design stage. This subset is shown in Tab. 1.

Table 1. Selected subset of attributes for reaction wheels

Attributes
mass
length
width
height
maximum torque
lifetime
power
data interface

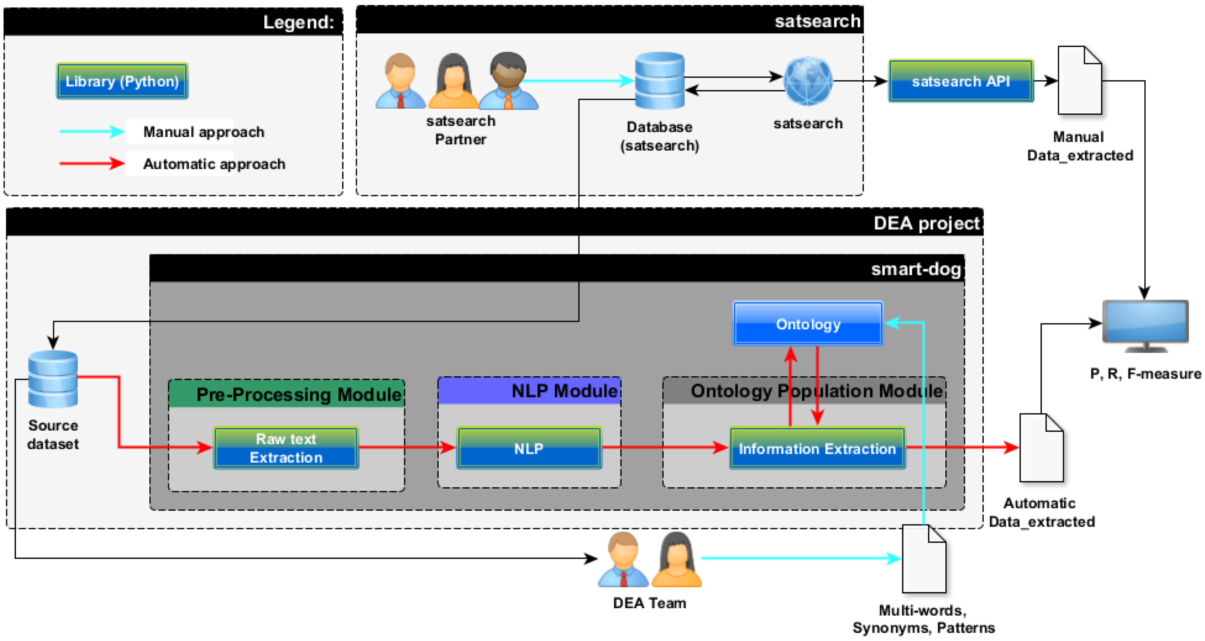


Figure 3. smart-dog modules for Knowledge Base Information Extraction

The performance evaluation of an IE system are related to the concept of relevance. In particular the main parameters used are Precision (P), Recall (R) and F-Measure [10]:

Precision (P) is defined as “the ratio of relevant items retrieved with respect to all retrieved items, or the probability, given that an item is relevant, that it will retrieved”:

$$\boldsymbol{Precision} = \frac{\#\,(\boldsymbol{relevant\ items\ retrieved})}{\#\,(\boldsymbol{retrieved\ items})} \quad (1)$$

Recall (R) is defined as “the ratio of relevant items retrieved with respect to all relevant items in a file (a collection), or the probability, given that an item is relevant, that it will be retrieved”:

$$\boldsymbol{Recall} = \frac{\#\,(\boldsymbol{relevant\ items\ retrieved})}{\#\,(\boldsymbol{relevant\ items})} \quad (2)$$

A single measure that trades off precision versus recall is the F-measure:

$$F_{measure} = \frac{2\,P\,R}{P+R} \qquad (3)$$

Two versions of the knowledge base IE process have been performed. The first version used a simple algorithm able to extract specific information according to general expected patterns (e.g., [attribute, number, measurement unit]) provided by the User. This process relied on three modules of smart-dog, shown in Fig. 3: Pre-Processing Module which rely on TIKA java library to convert pdf into raw text. [11], [12]; Natural Language Processing (NLP) Module used to perform sentence tokenization, work tokenization, POS tagging [13], [14], [15]; simple Ontology Population Module to perform Information Extraction using the simple ontology. The outputs of the manual and automatic approach are then compared using specific evaluation metrics.

Considering the issues identified in the first version, a deep analysis of the attributes in the datasheet was performed to extract near-synonyms of attributes, multi-words terms for normalization and more specific patterns (in the frame of the DEA, smart-dog will automatically extract near-synonyms, multi-words and patterns). The results are shown in Fig. 4 and Fig. 5. In Tab. 2 the F-measure average values for the two versions are provided.

Table 2: F-measure average for the two versions

Version	**F-measure$_{\text{AVERAGE}}$**
1	0.376
2	0.454

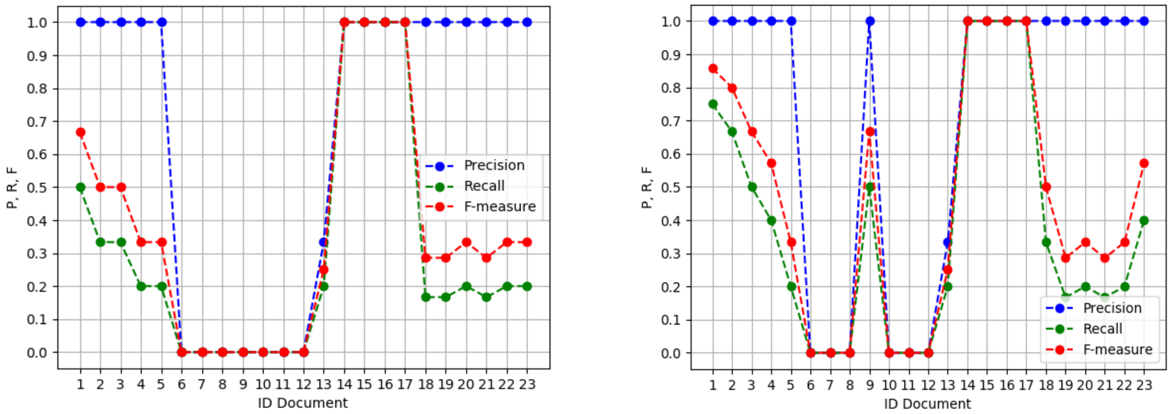


Figure 4. Results for v1 (left) and v2 (right) of the algorithm for Precision, Recall and F-measure

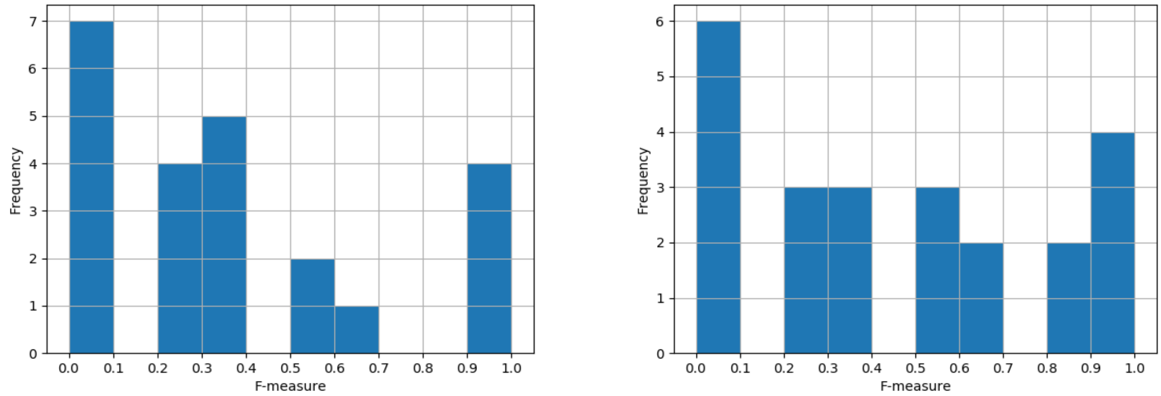


Figure 5. Histograms for v1 (left) and v2 (right) of the algorithm used for knowledge base information extraction

6. ANALYSIS

In this section the issues that emerged in the knowledge base IE process have been analysed. Some issues can be generalized to all kind of data, while some issues are specific to the type of documents used, in this case reaction wheel datasheets. It highlights the importance of the analysis of the corpus of documents (or source dataset) to understand where the algorithm lacks in succeeding, always with the goal of converging on a general algorithm that works for different documents within the same domain of expertise.

General issues:

- **Conversion error:** Error due to the conversion of the documents in raw text because of the format and layout used for the datasheets, which is not standardized;
- **Typo**: Typographical errors found inside the documents;
- **Synonyms recognition:** Limited list of synonyms and near-synonyms because they have to be identified a priori through the analysis of a bigger corpus. In the frame of DEA project, smart-dog will automatically extract them;
- **Multi-word recognition**: Limited recognition of the multi-words available in space mission design because of the short list added to the simple ontology. In the frame of DEA project, smart-dog will automatically extract them;
- **Patterns recognition**: Limited pattern list and not specific for each attribute. Due to the lack of standardization in the datasheet, several attributes can have several ways to be presented and accordingly several patterns. These patterns need to be identified automatically, because the manual effort to check thousands of documents is too time-consuming. In the frame of DEA project, smart-dog will automatically extract them introducing a new layer in the ontology learning layer cake [1], [2], [3];
- **Semantic need**: The lack of semantics in the pipeline and in the ontology limits the possibilities of improvements in the information extraction task and therefore population of the knowledge base;
- **Knowledge base model**: The lack of hierarchies in the ontology makes it impossible to extract some data that are not directly defined with a technical term inside the datasheets;
- **Uncertainty**: the uncertainty of the data is necessary in order to be able to provide consistent and reliable data during the design. Sometimes, the technical terms in the datasheets do not refer to the same definition of the attribute itself (e.g., mass could be related to a specific part of the space component or to the whole component and this information is fundamental for an appropriate selection of the space parts for a space mission) or the data are hidden in the text;

Some of the identified issues are specific for the datasheets:

- **Multiple space parts**: Some datasheets have several space parts. The initial assumption foresees one space part in each datasheet;
- **Multiple configurations**: Some datasheets can have multiple configurations of the same space part;
- **Different modes**: Some attributes are not uniquely provided, but sometimes it is possible to find them in several modes of the mission;
- **Multiple types of data**: Lot of data are not explicitly given in text format, but through images and equations, which requires further analysis in specific fields

The introduction of automation in the generation and population of a knowledge graph used for semantic search is a difficult task. This paper shows the high potential of knowledge base IE but also the limitations that arose, for the type of corpus used and for the type of ontology selected. This also highlights the importance of the choice of the ontology type and language that strongly depends on the corpus and also on the application for which it is addressed to.

7. CONCLUSIONS

This paper highlights the difficulties in the process of knowledge base information extraction on a specific type of document. In the frame of the DEA project, this is important in order to understand the actions that shall be taken to deal with several types of documents, not only datasheets. The issues analysed show two fundamental outcomes: the IE process requires a formal ontology because the documents are highly non-standard and containing mainly unstructured data. Hence, it’s necessary to give a corpus of documents to the algorithms that can capture the information about space mission design. In this way it is possible to expect accurate, consistent, and complete results. The other outcome regards the almost complete lack of standardization in the datasheets, which makes the extraction complicated and emphasizes the need of rapid solutions to support the generation of machine-readable documents. More generally the shift towards the implementation and use of the new technologies in the field of AI could be slowed down due to these obstacles. Therefore, an action is required from users generating the data, especially the unstructured source documents, to make them machine-readable and compliant with the AI technologies they want to rely on.

8. REFERENCES

1. Baclawski K., Bennett M., Berg-Cross G., Fritzsche D., Schneider T., Sharma R., Sriram R. D. and Westerinen A. (2017). Ontology Summit 2017 Communiqué – AI, Learning, Reasoning and Ontologies.
2. Staab S. and Studer (2007). Handbooks on Ontologies.
3. Petasis G., Karkaletsis V., Paliouras G., Krithara A. and Zavitsanos E. (2011). Ontology Population and Enrichment: State of the Arts.
4. Dewey, F.R. (1998). A Complete Guide to Data Sheets, Sensors Magazine.
5. Fowell, S. (2013). SOIS Electronic Data Sheets for Onboard Devices – Current Status, 7th ESA Workshop on Avionics Data, Control and Software Systems (ADCSS-2013).
6. Murdaca F., Berquand A. Riccardi A., Soares T., Gerene S., Brauer N., Kumar K. (2018). ARTIFICIAL INTELLIGENCE FOR EARLY DESIGN OF SPACE MISSIONS IN SUPPORT OF CONCURRENT ENGINEERING SESSIONS
7. Kumar K., Vaccarella A., Nagendra N.P., Gerené S., Lindblad L. (2018). Integrated Mission Design using satsearch, SECESA 2018.
8. Jurafsky D., Martin J. H., (2017). Speech and Language Processing
9. Prochazka, M. (2017). Electronic Data Sheets at ESA: Current Status and Roadmap, Workshop on Spacecraft Flight Software 2017, Applied Physics Laboratory, Johns Hopkins University, Laurel, Maryland, USA.
10. Manning C. D., Raghavan P., Schütze H., (2009). An Introduction to Information Retrieval.
11. TIKA library (2018): https://tika.apache.org/
12. Python tika library (2018): https://github.com/chrismattmann/tika-python
13. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python.
14. Hardeniya, N. (2016). Natural Language Processing : Python and NLTK Table of Contents.
15. Indurkhya, N., & Damerau, F. J. (2010). Handbook of NATURAL LANGUAGE PROCESSING.