**Knowledge-Based Information Extraction from Datasheets of Space Parts**

F. Murdaca*1\**, A. Berquand*1*, K. Kumar2, A. Riccardi*1*, T. Soares*3,**S. Gerené 4, N. Brauer5*

*1University of Strathclyde, Glasgow, UK,*

*2Satsearch, Delft, Netherlands*

*3ESA, Noordwijk, Netherlands*

*4RHEA group, Leiden, Netherlands*

*5AIRBUS, Bremen, Germany*

*\*Primary author contact details: francesco.murdaca@strath.ac.uk*

1. **Introduction**

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 datasheets. Searching for and through component datasheets is an arduous task, requiring significant manual effort. This often results in an incomplete overview of the components available on the market and suboptimal, if not erroneous, design choices.

To tackle the problem of finding the right space parts, the founders of [satsearch](https://satsearch.co) have embarked on a mission to consolidate global space supply chain information within a single platform.  Satsearch has undertaken manual effort to collect, curate, and structure supply chain information by extracting attributes from unstructured datasheets and generating equivalent machine-readable, human-readable, electronic datasheets (EDS). Their vision is to ultimately change the way supply chain information is exchanged across the supply chain by “reinventing the datasheet”. Currently however, the company faces a large hurdle of parsing unstructured PDF datasheets, in order to insert consistent and reliable data into EDS format; a task that has been largely manual to date. Due to rapid growth of the datasheet collection, this workload has been rendered unfeasible using the manual technique. Hence, a new approach is necessary to speed up this process.

Datasheets are provided by suppliers to communicate information about products that they wish to sell. These datasheets typically contain information that do not follow any standardization, especially in the space field. Datasheets for comparable space parts often provide different attributes, with no standardized schema in place. In some cases, attributes with the same name can refer to different properties of the system. (e.g., the “mass” for an on-board computer might refer to the motherboard only or to motherboard and structure that encapsulate it). This characteristic of datasheets makes the parsing process more complex. Moreover, some information is not provided in the form of numbers and text, but using graphs and equations.

The solution presented in this paper to speed up the parsing pipeline is a knowledge-based information extraction tool to extract reliable data from unstructured datasheets. A formal ontology, also known as knowledge base, is able to capture the different concepts referring to the same entity, it allows to quickly and accurately identify similarities and relationships between concepts, it can be dynamically enriched, it allows automation of the process and reasoning thanks to the rules inside it. Manual generation of such ontology is a long process that requires a lot of time and domain knowledge, especially when the amount of data is continuously increasing. Therefore, the solution would be to rely on ontology learning techniques that can automate the process.

This solution is currently under development in the frame of a Design Engineering Assistant for Early Space Mission Design. In the frame of this project, the corpus of documents cannot be generated only by datasheets because the semantic knowledge relative to a concept is not present. Therefore, the corpus of documents is enhanced with material (e.g., books) that includes the required knowledge. The comparison of output between manual data extraction by satsearch and the automatic, kowledge-based approach will also serve as a validation step for the DEA project knowledge base generation.

The proposed procedure foresees the creation of a knowledge base for the AOCS subsystem, as starting point, that will be extended to the other subsystems of the satellite. The second step, once the knowledge base is ready, is to use it to extract the structured data needed for satsearch. These data will be compared with the manually extracted ones and assess the potential of automatic parsing and extraction of data.