Final Report

Development of a graphical interface for corrections and additions by users when expanding ontologies – OntoHuman

Applicant	Diana Peters, Kobkaew Opasjumruskit
Institution	German Aerospace Center (DLR), Institute of Data Science
Project Title	Development of a graphical interface for corrections and additions by users when expanding ontologies – OntoHuman
Funding Period	01.07.2021 – 30.06.2022
List of published work	• Short paper accepted to CDVE (Cooperative Design, Visualization and Engineering) 2022 • Source code is available on 10.5281/zenodo.6783007 • NFDI4Ing Tool Talk: https://nfdi4ing.de/tooltalk-dsat

Executive Summary

In this project we address and support two purposes: to enrich ontologies, which contain semantic information describing objects or concepts, and to extract information from technical documents (e.g. to enrich ontologies with this information).

Manual development and maintenance of ontologies are tedious tasks and require extra training to use ontology modelling tools. Therefore, a semi-automatic process to enrich ontologies can assist domain experts, who are not necessarily ontology experts, to map knowledge into ontologies.

Major sources of information for enriching ontologies are often documents, especially data sheets in engineering domain. These data sheets are used during the planning and designing of products and they are crucial for choosing components that fulfill project requirements. However, the information in data sheets are mostly not accessible in machine-readable formats. In current practice, they are manually extracted and converted to be compatible to computer applications.

We already developed prototypical solutions for automatic information extraction from technical documents with the support from ontologies in previous work [4]. To verify the results of an automatic process, we pursued a Human-in-the-Loop (HiL) approach, which requires humans to provide feedback to the system.

In this NFDI4Ing SeedFund project, we combined the HiL component with DSAT to generalize the automatic information extraction process, which is running on the backend. Prior to this project, our solution could only be used for space engineering related documents. We now can apply and customize ontologies used for extracting data from documents of other domains. Feedback from users can also be collected via a web-based user interface and used for updating ontologies further, which could, in return, ultimately improve the automatic process.

During the project period, we completed all of the initially proposed features: correction of automatically extracted data, resolution of word ambiguities, adding new annotations, and export function for annotations. We arranged two workshops participated by interested users from NFDI4Ing community to gather a list of improvement and evaluate the intermediate result. The list of additional implemented features is elaborated on in the results section. Finally, the outcome from this project are: one published publication (at CDVE), one NFDI tool talk, and the published source code [5].

Results Report

This project is based on previous work, which consists of a set of tools assisting users to annotate technical documents. We offered an automatic annotation using ontology and natural language processing (NLP) techniques. When the automatically detected information was displayed to users, we saw the opportunity that the users can correct or improve the results via an intuitive user interface, as known as HiL component. Then, the correction can be collected for updating the underlying ontology further.

Project Goal

In this project, we aimed to improve the HiL component in our previous work to support more functionality and better usability. Previously, we applied the tools within the space-system engineering domain. Therefore, to broaden the domain of usage, the feedback from NFDI4Ing community was crucial. We also focused on the interaction between users, system backends, and ontologies.

Development

The proposed functions to be implemented in this project are summarized in Table 1:

Functions/Features	Importance	Area	Status/Remark
Correction of results from automatic information extraction.	High	Functionality	Done
Resolving words ambiguities, e.g. detecting domain of knowledge	High	Functionality	Partially achieved. The disambiguation depends on the quality and amount of text.
Annotating information on the document that the automatic process missed to detect.	High	Functionality/ UI	Done
Export of annotated information.	High	UI	Done

Table 1 Initial requirements to be implemented in the project.

Furthermore, we conducted two workshops as planned in the work packages and milestones as following.

Work package 1: Workshop and Definition of Requirements
In the beginning of the project, we conducted the first workshop to refine the initial requirements and collect additional features. The actual workshop was held 3 weeks later than planned due to technical difficulties and availability of participants. The additional requirements collected from the first workshop are listed in Table 2.

Functions/Features	Importance	Area	Remark
To use other ontologies for annotating documents.	High	Functionality	Done
Handle semantic uncertainty, e.g. interpret the uncertainty of values like maximum, approximately, etc.	Med	Functionality	Not implemented. This feature requires further research and study and cannot be done within the project period.
To extract non-text information, e.g. values from graphs.	Low	Functionality	Not fully completed within the project. We developed this function independently and planned to integrate it into DSAT in the near future.
Multiple language support	Low	Functionality	Tested with documents in German, French and English. This relies on the ontology.
To upload a spreadsheet as a document	High	UI	Done
Specify page range in pdf file for processing.	Med	UI	Done
Enable the tools via a webpage	High	Deployment	There is no running web interface due to DLR's export control regulation. However, the published source code can be compiled and executed as a web-based service.

Table 2 Requirements collected from the first workshop

Work package 2: Implementation phase of the requirements from work package 1

We focused on the initial requirements and planned to complete most of the functions listed in Table 1 and Table 2 before the seventh month of the project. The request for resolving semantic uncertainty is a deep topic and should be studied specifically, so this function is not resolved. We implemented the function to extract values from graphs as a standalone application. It is planned to be integrated into DSAT in the near future.

Work package 3: Implementation of the export interface
Since there was no request for extra export format than JSON in the workshop, we keep the export of annotations to two options: storing in the database, or save as a JSON file.

Work package 4: Testing on specific use cases or set of documents

According to the work package 1, we planned to collect use cases and example documents from the workshop participants from NFDI4Ing community. However, there was no concrete set of documents suggested, so we fell back to our original use cases, which is space-engineering technical documents.

The second workshop was planned to be held at the middle of the project to collect further feedbacks and evaluate the implemented functions. The workshop was held one month late due to the holidays. The functions and features implemented in the work package 2 and 3 were presented. The workshop participants tried DSAT via their browser and provided feedback as summarized in Table 3. All the suggested features this time were UI-related topics.

Functions/Features	Importance	Area	Status/Remark
To display metadata / definition on an ontology, e.g. title, abstract, creator, license.	High	UI	Done
To add first time user guide introducing the tool and simplify the interface.	High	UI	Done
To add an annotation with a parameter name that doesn’t appear on a document.	Med	UI	Done
To highlight incomplete annotations.	Med	UI	Done
To use collections of ontologies for automatic detection of items, to get higher recall within one data sheet.	Low	UI	Done
To add “Delete all” annotations function.	Low	UI	Done

Table 3 Requirements collected from the second workshop

Results

After finishing all work packages and implementation, we published DSAT and related tools with source code. There is not only the user interface (DSAT) but there are also the ontology enricher (ConTrOn), database (DSAT DB), and information extraction package (PLIX) to facilitate the document annotation with ontologies. The tools are connected and work together as shown in Figure 1.

DSAT was improved to have a simplified interface (shown in Figure 2) for annotating documents. It also provides a feature to preview and upload ontologies for automatic annotation. DSAT DB is also bundled in the installation package for storing annotations and custom ontologies.

To process text in the document, the information extraction module is needed. However, this module is independently developed as a standalone python package called PLIX [6], which is out of OntoHuman's universe of discourse. Nevertheless, we bundled PLIX to our installation package for convenience in installation.

ConTrOn is used to process ontologies for the automatic annotation process. It works together with PLIX and finally returns the detected annotation to be displayed on DSAT. Users can then give feedback about the result via DSAT or update ontologies to get better automatic extraction result.

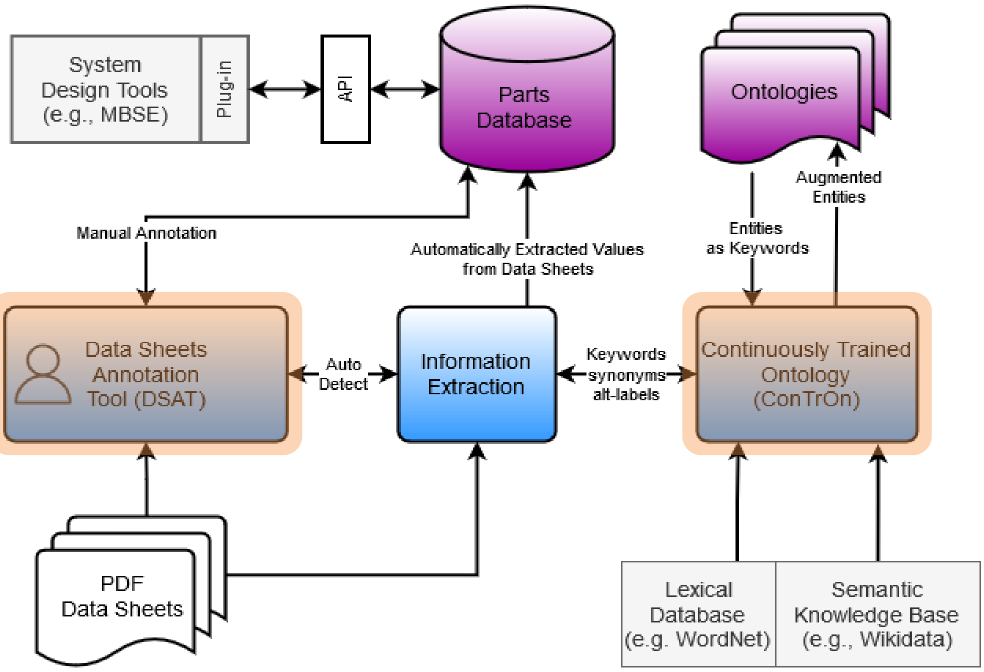
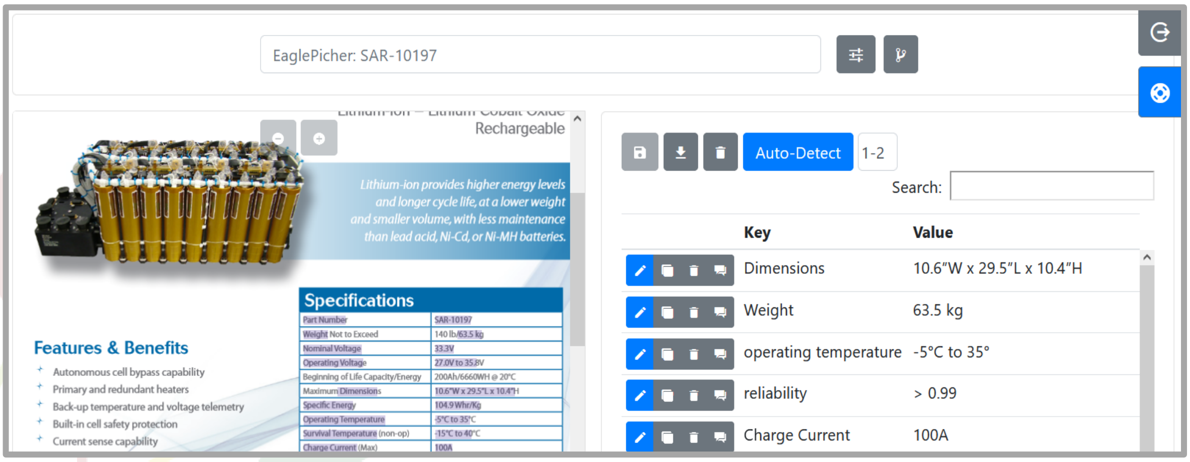

System Design Tools (e.g., MBSE)
Plug-in
API
Parts Database
Ontologies
Augmented Entities
Entities as Keywords
Manual Annotation
Automatically Extracted Values from Data Sheets
Data Sheets Annotation Tool (DSAT)
Auto Detect
Information Extraction
Keywords synonyms alt-labels
Continuously Trained Ontology (ConTrOn)
PDF Data Sheets
Lexical Database (e.g. WordNet)
Semantic Knowledge Base (e.g., Wikidata)

Figure 1 An overview of tools integrated for OntoHuman.



EaglePicher: SAR-10197

Lithium-ion – Lithium Cobalt Oxide
Rechargeable

Lithium-ion provides higher energy levels and longer cycle life, at a lower weight and smaller volume, with less maintenance than lead acid, Ni-Cd, or Ni-MH batteries.

Features & Benefits

- Autonomous cell bypass capability
- Primary and redundant heaters
- Back-up temperature and voltage telemetry
- Built-in cell safety protection
- Current sense capability

Specifications

Part Number	SAR-10197
Weight Not to Exceed	140 lb/63.5 kg
Nominal Voltage	33.3V
Operating Voltage	27.0V to 35.8V
Beginning of Life Capacity/Energy	200Ah/6660WH @ 20°C
Maximum Dimensions	10.6"W x 29.5"L x 10.4"H
Specific Energy	104.9 Whr/Kg
Operating Temperature	-5°C to 35°C
Survival Temperature (non-op)	-15°C to 40°C
Charge Current (Max)	100A

Auto-Detect 1-2

Search:

Key	Value
Dimensions	10.6"W x 29.5"L x 10.4"H
Weight	63.5 kg
operating temperature	-5°C to 35°
reliability	> 0.99
Charge Current	100A

Figure 2 A screenshot of DSAT used for annotating documents

Discussion

The automatic annotation of documents depends largely on the used ontologies. The ontologies offered by default are limited only to the domain of space-engineering hardware. In order to fully use the tools for other domains, users should know where to find relevant ontologies. An ontology search API could be used to assist the users to find the right ontologies in the future. Furthermore, the semantic disambiguation, multi-language support, and graph value extraction are currently being our research topics. They could be integrated to DSAT in the future.

Cooperation & Exchange

The project had cooperated and exchanged with NFDI4Ing community, from which we invited participants to our two workshops. In these workshops, we refined the requirements and improve functionality and usages of OntoHuman. As discussed in the results, we implemented all the initial requirements and followed most of the suggestions that could be managed in the project period. Additionally, we demonstrated OntoHuman in one of the NFDI Tool Talk. We also plan to present the project at the NFDI4Ing conference 2022.

Relevancy within the Community

We aimed to integrate the members from NFDI4Ing community throughout the course of the project, thus we hope that OntoHuman is relevant to the community. From the survey we conducted in one of our workshops, the participants rated the domain of usage of OntoHuman to generic purpose (rated 3.5 points out of 5), somewhat relevant to their colleagues' work (3 points out of 5), and not very relevant to their own work (2 points out of 5).

Usage within the Community

From the direct feedback from the workshops' participants, they expressed interest in using OntoHuman but the frequency of use might be once a year, and could be more often if all the important features are implemented. The tool is considered to be easy to used (6/7), supportive (5.5/7), efficient (6/7) and novel (5/7).

Publications

During the project, we conducted two workshops and gave an NFDI tool talk, which is open to the public [7]. As deliverables at the end of the project, we published the source code for OntoHuman including installation instruction on Zenodo [5]. A short paper describing the tools is accepted and will be presented at CDVE conference in September 2022.

Acknowledgement

We would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713.

References

[1] Al-Aswadi, F.N., Chan, H.Y. & Gan, K.H. Automatic ontology construction from text: a review from shallow to deep learning trend. Artif Intell Rev 53, 3901–3928 (2020). https://doi.org/10.1007/s10462-019-09782-9

[2] Rizvi, S. T. R., Mercier, D., Agne, S., Erkel, S., Dengel, A., & Ahmed, S. (2018). Ontology-based Information Extraction from Technical Documents. In ICAART (2) (pp. 493-500).

[3] H. Bast and C. Korzen, "A Benchmark and Evaluation for Text Extraction from PDF," 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, ON, Canada, 2017, pp. 1-10, doi: 10.1109/JCDL.2017.7991564.

[4] Opasjumruskit, K., Schindler, S., Schäfer, P. M., & Thiele, L. (2019). Towards Learning from User Feedback for Ontology-based Information Extraction. DI2KG

[5] Opasjumruskit, K. (2022). OntoHuman OpenSource Software (v1.0.1). Zenodo. https://doi.org/10.5281/zenodo.6783007

[6] Böning, S., Kiesewetter, C. PLIX (Information Extraction module) version 1.0, license Apache-2.0

[7] https://nfdi4ing.de/tooltalk-dsat, accessed 26.08.2022