

# Towards improving the information management of aircraft cabin retrofits by processing quantity-on-hand documents

Fabian Niklas Laukotka , , Markus Christian Berschik  and Dieter Krause 

Hamburg University of Technology (TUHH), Germany

 [fabian.laukotka@tuhh.de](mailto:fabian.laukotka@tuhh.de)

---

**ABSTRACT:** Retrofitting aircraft cabins is characterized by a large number of documents created and required, most of which are currently processed manually. Engineers need to identify which documents include the information that is required for a specific task. This paper proposes an approach that builds upon a digital knowledge base and moves towards automatically processing the quantity-on-hand documents to reduce the work required to identify the required documents without the labour-intensive creation of the knowledge base in beforehand. After describing the scenario this work faces, comparable approaches and promising techniques are discussed. A process-chain that builds upon these fundamentals is presented, including a selection of feasible techniques and algorithms. Finally, the steps towards an implementation as part of the transformation towards a data-driven value chain are presented.

**KEYWORDS:** document processing, digital/digitised engineering value chains, big data, artificial intelligence

---

## 1. Introduction

With the ongoing digitalization, the access to information is becoming ever easier. Conversely, the vast amount of created information, often lacking structure or standardization, likewise leads to difficulties regarding automated editing, searching or analyzing (Kolandaishamy et al., 2024). Additionally, such processing of unstructured data in general is always time-consuming and costly (Baviskar et al., 2021). New information is created and digitally documented consistently throughout the engineering design processes, while sophisticated systems allow for the storage and management of this data across teams. However, while this applies primarily to the initial design and manufacturing of products, long-lasting assets generate and require access to these documents and the information they contain throughout their lifecycle and across multiple stakeholders. For example, the planning and execution of maintenance or modification activities requires access to specific details of the individual asset as well as a general understanding of the affected systems.

A prime example is the aircraft cabin retrofit, where acquiring and accessing the required information on the specific aircraft to be modified is a major challenge. An innovative approach based on system modelling of generic information and combining this with specific information to create a knowledge base that provides engineers with a tailored and intuitive innovative access system has already been presented by (Laukotka et al., 2024). However, their approach faces the challenge that the necessary creation of the models and knowledge base requires manual work and manual processing of existing documents and specifications. At the same time, when working with a major player in the aircraft retrofit business, much of the required information is already available due to past performed retrofits and concomitant documentation. Yet, because of the vast number of systems and information that occur when considering aircraft, creating and extending this digital knowledge base is a tedious and labour-intensive task. This work tackles this situation by working out an approach that increases the interactively digitally available information by automatically processing available documentation and, thus, reducing the

manual work. The task at hand is an essential step in the transformation towards a data-driven value chain. The respective research question is formulated as follows.

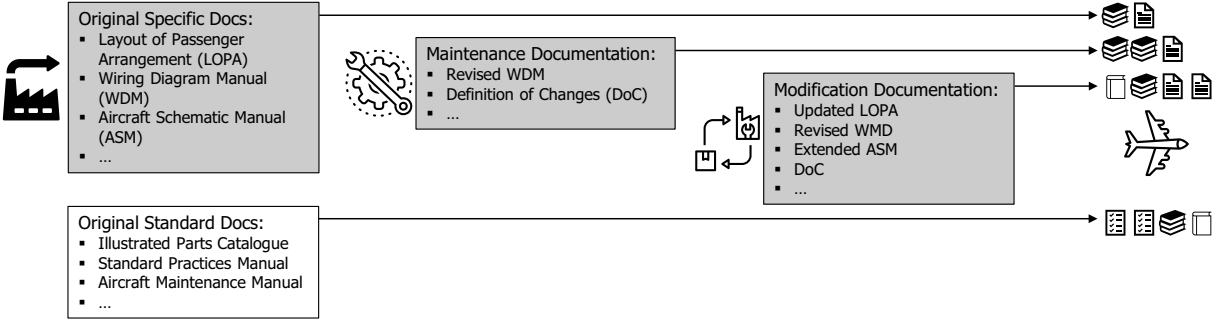
*How can processing quantity-on-hand documents for aircraft cabin retrofits be enhanced to reduce manual effort and support the integration of processed information into a digital knowledge base?*

## 2. Fundamentals

This section presents the fundamentals related to the task at hand, namely an introduction to aviation’s retrofit as well as possible recent approaches to document processing.

### 2.1. Aircraft cabin retrofits and its challenges of data management

Aircraft are composed of a comparatively high number of parts, typically amounting to approximately 400,000 (Stark, 2016). They also have a long lifespan, often over 20 years (Niță & Scholz, 2011). During their operational lifetime, aircraft are handled by a number of different stakeholders, including the original manufacturer, their suppliers, the operator, as well as a variety of maintenance organizations. To satisfy the high safety standards required of aircraft, they must be maintained thoroughly (Florio, 2016). Such maintenance is conducted regularly by specialized maintenance organizations and must be meticulously documented. To keep the cabin in a satisfactory condition or to adapt it to changing passenger requirements, the cabin is typically replaced (in part) every five to seven years (Niță & Scholz, 2009). During the process, known as a cabin retrofit, new components are installed into an existing aircraft, which requires a comprehensive understanding of the current state of the aircraft and the documentation of this information (Laukotka et al., 2024). Similarly, any changes made must be certified and, therefore, documented (Florio, 2016). However, this usually takes the form of documentation of the change made, resulting in a revision of certain aspects of the previous documentation. In addition to the separation into revisions, each domain involved in the modification or retrofit process creates its own set of documents: There might be one set for the electrical system, one for changes to the air conditioning or water/wastewater systems, and another for structural adaptations. As a result, there is a new dataset for each modification project that consists of individual documents for each domain or system and includes revisions as well as newly added documents. Sometimes, there are cross-references between the individual documents. However, these are mostly within the dataset of a particular domain, usually not between documents from different domains, and especially not between change projects. As a result, documentation becomes increasingly fragmented over the life of the asset, and the information required for future changes is spread across a variety of stakeholders, files, documents, and sometimes even analogue papers (Moenck, Laukotka, Krause, & Schüppstuhl, 2022). These documents include structured data (numerical values) and unstructured data (textual descriptions). Some have a semi-formal structure due to existing regulations and forms required for certification or simply due to established structures and systems in aviation. Figure 1 illustrates this accumulating set of documents resulting from the different adaptations throughout the asset’s lifespan.

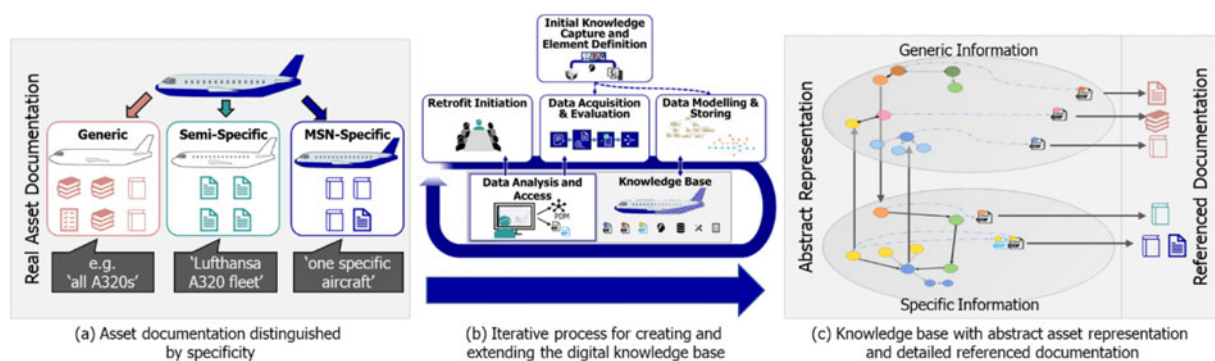


**Figure 1. Accumulating dataset of documents by maintenance and retrofits throughout the asset’s lifespan**

While this aggregated documentation of the performed changes is usually returned to the operator, it still does not result in an easily processable and single comprehensive representation of the individual physical asset. Due to the required certification of each stakeholder, including their standard processes, a

complete transformation to a purely digital representation without further use of the existing documents is not easily feasible. Even if a single stakeholder invests heavily in digital infrastructure, this usually does not directly involve the other stakeholders, leaving them with the established structures and typical exchanged documents. As a result, an easily usable virtual representation is not available for future modifications, at least not if they are carried out by a third-party organization, as is usually the case for cabin retrofits (Moenck, Laukotka, Deneke, et al., 2022). Instead, these organizations need to reconstruct the actual state of the asset based on these fragments, as the asset itself is still in operation and spatially unavailable at this point in the process. As described above, the documentation of the actual asset becomes crucial to creating these representations, which include numerous individual documents, each describing specific aspects of the asset, often as part of revisions and spread across different individual system descriptions.

Consequently, this requires extensive manual work to reconstruct the actual state, identify the relevant documents for the task, and then access the respective information, which depicts a significant challenge. An iterative approach that faces these challenges by combining different strategies that result in a digital knowledge base was presented by (Laukotka & Krause, 2024) (cf. Figure 2(b)). The base information of aircraft is separated by their specificity (cf. Figure 2(a)), and abstract representations of the aircraft are created using a sophisticated methodology. Their introduced data specificity describes how specific a certain document or set of information is, ranging from the applicability to all aircraft of the same type (e.g., all Airbus A320s) to only applicable to a specific aircraft, identifiable by its manufacturer serial number (MSN). The resulting representations depict a stripped-down definition of occurring elements and their relations. To these elements and relations, references to the documentation from which the information originated are added (cf. Figure 2(c)), which allows engineers to query the resulting knowledge base for relevant information, which results in a list of the appropriate documentation.



**Figure 2. The current process of creating a knowledge base of abstract aircraft representations with references to detailed documentation, as described by (Laukotka & Krause, 2024)**

Although this has improved access to the documents relevant to the currently planned modification, building this knowledge base still requires manual effort, as described in the introduction. Advancing the process using the semi-formal structure and standardized systematics of aviation could further reduce this manual effort. However, completely digitalizing all information and getting rid of the documents is currently impossible due to certification and intellectual property (IP) reasons; as described above, the documentation is required for certification and often is the only way the information is passed along during the lifespan and across stakeholders. Additionally, the processes and operations themselves need to be certified, which hinders the implementation of drastic changes, such as completely changing the basis for all information. Some documents are provided with IP restrictions and must be deleted after the modification is done and the resulting changes have been documented. Such records cannot be wholly digitalized or transferred into other storage means.

## 2.2. Document processing, semantic recognition and recent AI approaches

Independent from the challenges described for aviation, the processing of documents and its data is a major aspect of digitalization approaches. Data can be classified into structured, semi-structured and unstructured types (Meier & Kaufmann, 2019) based on the level of organization and ease of digital processing. Structured data adheres to a defined format, such as tables with columns and rows, allowing

for straightforward computational processing. Semi-structured data, such as XML, contains elements with a flexible but specific syntax that facilitates processing when content types are known. Unstructured data, including text documents, images and video, lacks a fixed structure and often requires interpretation to extract meaningful information. While structured and semi-structured data can quickly be directly processed and evaluated, this is not true for unstructured data per se (Feldman & Sanger, 2008). For example, while individual pixels of an image can be processed, understanding the image's content, such as identifying a number in a photograph, requires additional steps to interpret the semantics beyond mere computational analysis.

As a result, pre-processing approaches have evolved, which turn unstructured data into structured or semi-structured form by identifying the inherent information. Optical character recognition (OCR) was developed very early on to make PDF files of scanned documents editable. OCR represents a foundational technology for converting text from images into machine-readable format (Google Ireland Limited, 2024). While the “traditional methods for extracting information have their limitations, [...] artificial intelligence (AI) could provide a more fitting alternative” (Mahadevkar et al., 2024). Because the machine-readable format does not necessarily allow the underlying semantics to be processed, natural language processing (NLP) has recently been added to allow the computer to access the information inherent in the textual form (Giordano et al., 2024).

Similarly, large language models (LLMs) have become increasingly popular in recent years, such as OpenAI's GPT-4 or Meta's Llama 2. Building on these and accessing external sources such as internal company documents or email correspondence, retrieval augmentation generation (RAG) approaches enable LLMs to understand the context of queries better (Yadav, 2024). Besides these AI approaches focusing on text processing and recognition, there is also “a growing interest in understanding and extracting semantic relations from technical documents” (Han et al., 2021).

Automatic access to the vast number of digital documents created throughout the years “offers multiple benefits, facilitating enhanced reasoning, analysis, and manipulation of the knowledge embedded within design documents” (Giordano et al., 2024). The advances in AI have resulted in progress, which is mainly restricted to structured data (Ramakrishnan et al., 2024). Yet, nowadays, some innovative approaches and tools utilize AI to process engineering drawings, a typical example of unstructured data, like Werk24 (W24 Service GmbH, 2024). Some examples, such as NLP or other related aspects of AI, are described in Section 2.3.

### 2.3. Related works

While the context of this work is highly specific to the application to aircraft cabin retrofits, elements, especially the processing of documents, can also be found in other scientific work. Additionally, review papers like (Baviskar et al., 2021) describe available approaches and their potential in general. An excerpt of relevant works with related applications is described in this section. (Arnarsson et al., 2021) created a search engine that provides engineers with related and relevant documents by utilizing natural language processing and applying “clustering algorithms on top of traditional searches”. Hence, the user is provided with an overview of the different topics the documents relate to. Their approach mainly focussed on Engineering Change Requests, which contain the relevant information mostly in textual form. Thus, they utilize doc2vec algorithms that handle these texts and allow them to process the information of these documents further using NLP to create their search and clustering solution.

Also, natural language processing (Giordano et al., 2024) extracts the semantics of design documents to enhance “reasoning, analysis, and manipulation of the knowledge embedded within [them]”. Their approach focuses on technical documents like scientific papers, patents and literature. Furthermore, they describe a roadmap to guide and advance future studies. Similarly, to better exploit patent documents and improve the transfer and identification of knowledge across domains (Li et al., 2023), describe an approach that builds a knowledge space that processes the patent contents and citation relationships. They also utilize a doc2vec algorithm to process the textual descriptions within these documents. While these approaches and solutions show progress in processing unstructured data, such as texts originating from documents, the works briefly described above do not consider graphical elements. For this, other technologies and approaches are required.

Once again, focusing on information within patents (Hoque et al., 2022) presents a computer vision (CV) approach that improves the segmentation and consecutive processing of images from patent documents. They used an approach that builds upon a convolutional neural network (CNN).

To process large-scale technical drawings (Nguyen et al., 2021) propose a method that detects and recognizes selected visual objects and text patterns by combining object detection using a faster R-CNN neural network with a CNN-based binary sliding windows classifier for character recognition. Both techniques require the CNNs to be sufficiently trained, which requires respective annotated datasets. In this case, they used a set of “4630 technical drawings that were labelled manually by human operators” (Nguyen et al., 2021).

To help aerospace engineers quickly access relevant information, thus improving the efficiency of the information retrieval and communication processes (Yadav, 2024) presents a chatbot equipped with RAG. However, the proposed solution focuses on textual documents describing industry standards or communication protocols.

As not all applications have the resources available to perform such annotations, a review paper regarding computer vision applications (Paneru & Jeelani, 2021) describes the lack of annotated datasets as “one of the biggest challenges to implement deep-learning-based computer vision techniques”.

In conclusion, even this small insight into other research revealed that different approaches tackle somewhat similar challenges because the existing documentation has a vast inherent value. Yet none can be easily applied to the given scenario, as they are usually likewise tailored to their specific circumstances. Consequently, the following chapters develop and present a concept tailored to the information handling of aviation’s cabin retrofit.

### 3. Towards extending the knowledge base by automatically processing available quantity-on-hand documents

As described in Section 2.1, aviation is subject to strict regulation, and drastically changing processes or information storage is not easily done. While an ideal solution might include the complete transfer of all available information into a digital knowledge base, improved access to specific information and the respective documents is a feasible first step for now. Engineers need a better way to get to the information and documents required at a particular time. Building on the systematic principle of improved access to information as presented by (Laukotka & Krause, 2024), this work aims to identify further the correct documents and information for a given situation. A respective concept is presented below.

#### 3.1. Methodology

As already identified, some common generic references are related to the existing and pre-defined aircraft structure, like frames or seat-rails. These references are used to define the shown information within the airframe spatially. Besides these *structural references*, other typical references are ones related to the *asset ID* or *symbolic references*. A textual reference to a specific MSN usually does the former. The latter can take various forms, like symbols representing essential cabin elements like seats, doors, or lavatories. Figure 3

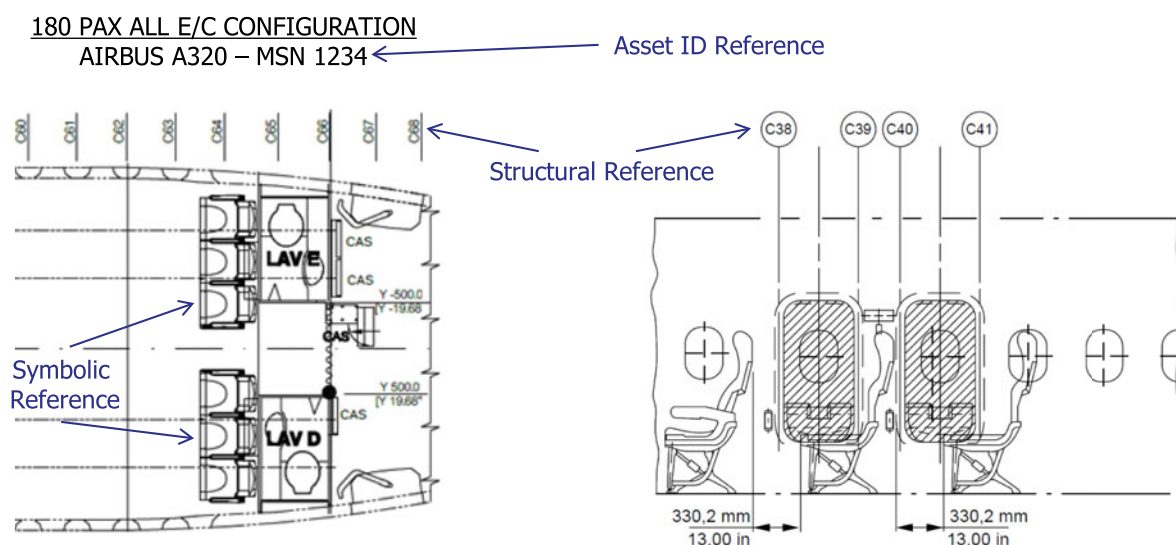


Figure 3. Excerpts of typical documents with structural, symbolic and asset ID references: left: planning position for a retrofit of new lavatories; right: emergency exit door arrangement

shows different excerpts of typical documents with structural, symbolic and asset ID references. On the left side, an excerpt of an *Interface Control Document* (Lufthansa Technik AG, 2024) is shown, including symbolic references to seats, an asset ID reference to the MSN 1234, and structural references to the aircraft's frames at the specific location. On the right side of the figure, an excerpt of the classic *Airbus A320 Cabin Configuration Guide* (AIRBUS S.A.S., 2005) is shown, including the positions of the emergency exit doors with another form of a structural reference to the surrounding frames of the aircraft.

In many documents, the definition of the relevant position within the fuselage is realized by references to the frames. These frames are uniformly defined within an aircraft family, such as the Airbus A320 with its additional A319, A318 and A321 variants, providing a reasonable generic basis for structural localization. Thus, it enables many cabin components to be referenced to the aircraft structure, as many are directly connected to frames. This can help narrow down relevant documents based on their proximity to a specific planned change. To some extent, this is already done today, albeit by manually reviewing the appropriate documents and identifying their relevance based on the displayed references. This is now to be extended by automatic processing of the information. Crucial additional factors that need to be considered in such document processing are

- Validity; the validity of automatically extracted information must be ensured.
- Traceability; in addition to the influence of the source material on the validity of the information, the specific files or documents from which the information originates need to be traced to ultimately make these files available to the engineers. In particular, the aim is to facilitate access to the files based on their content.
- Intellectual property restrictions should not include detailed information in the knowledge base, as this may be subject to restrictions. Instead, a general definition of what detailed information can be found in which document needs to be included, as well as the reference to the specific file or where it can be obtained.

### 3.2. Concept and process chain derivation

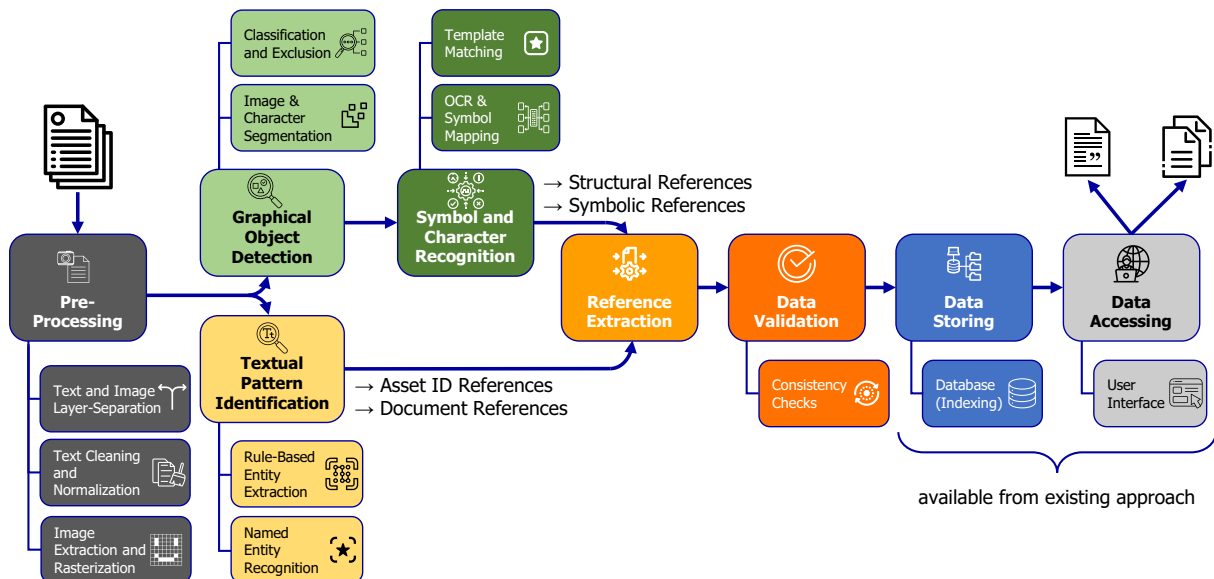
The concept includes the following key aspects in considering the circumstances described in [Section 2](#) and the additional factors listed above. As mentioned, the plan is not to store everything processed digitally (as in most LLM approaches) but to provide an innovative and easier way to access the already used and certified documents. As a result, the solution is more similar to an index than a comprehensive representation, which also means that IP-restricted information is not fully included in the knowledge base. For this reason, the following tasks are essential within the processing:

- Identification and differentiation between relevant textual and graphical elements
- Asset allocation based on textual references to MSN
- Spatial allocation through the identification of structural references by the processing of the graphical or textual references to frames or other structural elements
- Identification of installed components by processing the graphical representation or symbols
- Identification of textual references to other documents

The resulting extracted information is then validated and - in a reduced abstracted form - added to the knowledge base with links to the documents from which it was derived. Engineers are still expected to use these documents without the labour-intensive task of first going through all of them just to identify which ones are the right to use. Thus, the created knowledge base is structured around a hybrid data representation, combining an abstract virtual representation with the original detailed documentation. As a result, this approach allows easy indexing and, thus, identification of key information while respecting certification or IP restrictions, as the detailed information is still used in the required and current form. While this is highly specific to the given circumstances, it also represents the novelty of the approach presented. Based on common standard approaches and other comparable research (see [Section 2.3](#)), a process chain is derived that is visualized in [Figure 4](#) and described below.

#### ***Pre-processing***

To allow for the different occurring references, in text and graphical form to be processed individually and in parallel, texts and images within the document must be separated. This standard task can be accomplished by most PDF parsing libraries like PyMuPDF or pdflumber ([Adhikari & Agarwal, 2024](#)). Another pre-processing step lies in the first processing of the separated layer, e.g., pre-processing the texts to remove unwanted characters and spaces or normalize the font format. On the image side, the images or vector-based diagrams and symbols can be pre-processed and converted into a standard format



**Figure 4. Essential steps of the concept for document processing targeting aviation’s cabin retrofit**

for further processing. Depending on the requirements of the algorithms used later and the image types that occur, vector images might be rasterized, making them compatible with common computer vision techniques. Again, libraries like PyMuPDF can help with these tasks. The further processing is done separately and differently for texts and graphics.

**Textual pattern identification**

Textual patterns that hint at references might be identified or categorized using Named Entity Recognition Models (NER) (Baviskar et al., 2021). Pre-trained NER models, like spaCy, can be fine-tuned to the required dataset labels occurring within the asset documentation. A supplement or alternative to these NER models can be rule-based approaches that identify references with a clear, structured notation and repeatable pattern, e.g., “MSN: 1234” or “See Doc #T-1337”. Hence, in the case of aircraft modification documentation, the results of this step usually are references to the asset ID or other relevant documents, as shown in Figure 3.

**Graphical object detection**

After the pre-processing, the graphical elements are separated from the texts, which are further processed in this step. A key task is classifying the images and excluding unrelated imagery like logos to focus only on diagrams or schematics associated with components or references. A segmentation into distinct sub-elements is required for more complex visualizations, e.g., to identify areas of interest within diagrams. Such distinct elements, in this case, labels, components or regions of interest, can be detected using object detection models like YOLO or Faster R-CNN (Paneru & Jeelani, 2021).

**Symbol and character recognition**

With distinct graphical objects being identified, they are analyzed for the occurrence of references. Many documents use standard symbols (e.g., arrows, connectors, or certain aircraft components) representing functional relationships or elements. Template matching techniques or CNNs trained on respective symbol datasets are possible approaches to detect these icons (Baviskar et al., 2021) and, thus, inherit references. If these symbols include additional texts, another round of OCR, e.g., using Tesseract or Google Vision, can help extract text labels, part numbers, or other annotations within diagrams or symbols. Once symbols and texts are recognized, they are mapped to specific relationships or underlying meanings. The results of this step include visual references to the structure or symbols that represent specific installed components, as shown in Figure 3.

**Reference extraction**

With the respective references identified, in this phase, they are extracted, and the actual processing of the texts / visual references is conducted. This includes the creation of respective arrays, including the reference value and a reference back to the document being processed to allow for traceability and future linking to the specific document. This is probably done separately for each entity, in this case, object or value.

### ***Data validation***

The data is being validated before the entities are added to the final database and knowledge base. This includes a consistency check that verifies the format and validity of extracted references (e.g., ensuring part numbers meet expected patterns). This might be done using rule-based validation to filter out inconsistencies. A human review and feedback loop might be implemented to validate and correct extractions, especially in the early stages. This data can also improve the models' accuracy in future iterations through ongoing retraining. Validating the values by processing multiple identified entities and comparing them might also be a good approach.

### ***Data storing and accessing***

As presented by (Laukotka & Krause, 2024), the resulting information can be stored in a graph database (e.g., Neo4j) that depicts the growing knowledge base. With the increasing dataset size, their approach can be extended at this stage by integrating full-text indexing, e.g., using Elasticsearch, to increase the efficiency of the information retrieval based on keywords or reference patterns. A specifically tailored user interface is created to access the information. This UI can be made universally available across devices and to different stakeholders within the same network using a web application.

This generically described proposed concept and process chain described above shows how digitally processing the quantity-on-hand documents can leverage the planning of new retrofits by allowing for easier access to required information.

## **3.3. Roadmap to a future implementation**

At this stage, the process chain consists of generically described steps using different tools and approaches. While the individual steps can also be found in other approaches and have shown promise in their own right, they still need to be tailored to the application presented here. Therefore, the first step is to test and select valuable solutions and implementations for the different process steps presented, as many comparable but different algorithms and tools are often available. Together with an industry partner, the most promising solutions are currently being evaluated to allow progress in that research while likewise ensuring that it stays within intellectual property restrictions. Nevertheless, there are expected challenges that need to be overcome: Even when focusing on single, distinct aspects such as serial numbers or frame references, different visualization styles can occur within the dataset, as can be seen with the different styles of structural references in Figure 3. The processing algorithms must, therefore, cope with these different visualizations and correctly identify the respective elements. In theory, this can be achieved through extensive manual annotation and training. While huge datasets are available, at least for the companies currently facing these challenges, they are not always available for independent research purposes. In addition, the files of these datasets need to be manually annotated, as described in other approaches (see Section 2.3). A research project is currently being planned with leading aviation and retrofit industry players to address these challenges. The concept presented here is a first step towards this research and its future implementation. Because of the modular structure of the approach, independent from a project that targets the full implementation, work on individual modules is already done today. Focussing on smaller representative datasets enable the preliminary development of individual modules and their successive implementation, training, testing and evaluation of different approaches.

## **4. Discussion and conclusion**

This work introduced the specific challenges of data handling for information related to aircraft cabin retrofits. An existing approach to improve this situation was presented, but it still requires the challenging processing of quantity-on-hand documents. Various approaches and techniques that allow for the processing of documents and unstructured data in general were presented, and a concept including a process chain for using these techniques to move towards automated processing of retrofit-related documents was assembled. Directly referring back to the research question, the following answer can be derived:

*The processing of quantity-on-hand documents for aircraft cabin retrofits can be improved by developing a methodology that uses automated document processing techniques, such as optical character recognition (OCR) and natural language processing (NLP), to digitize and interpret documents effectively. By incorporating a hybrid data representation approach that combines an abstract virtual model with the*

*original documentation, key information can be easily indexed and identified while respecting certification and IP restrictions. In addition, machine learning algorithms can be used to improve the accuracy and efficiency of the methodology over time, reducing manual effort and ensuring seamless integration of processed information into a digital knowledge base, thereby improving data management for aircraft cabin retrofits.*

The current status of this highly specific research has been elaborated on, leading to next steps and future research, which includes the ongoing implementation, test and evaluation of individual modules of the presented approach as well as a future research project that incorporates all modules into a full implementation and finally a validation using comprehensive datasets.

## References

- Adhikari, N. S., & Agarwal, S. (2024, October 13). *A comparative study of PDF parsing tools across diverse document categories*. <http://arxiv.org/pdf/2410.09871>
- AIRBUS S.A.S. (2005). *A320 single aisle cabin configuration guide*.
- Arnarsson, I. Ö., Frost, O., Gustavsson, E., Jirstrand, M., & Malmqvist, J. (2021). Natural language processing methods for knowledge management—Applying document clustering for fast search and grouping of engineering documents. *Concurrent Engineering*, 29(2), 142-152. <https://doi.org/10.1177/1063293X20982973>
- Baviskar, D., Ahirrao, S., Potdar, V., & Kotecha, K. (2021). Efficient automated processing of unstructured documents using artificial intelligence: A systematic literature review and future directions. *IEEE Access*, 9, 72894-72936. <https://doi.org/10.1109/ACCESS.2021.3072900>
- Feldman, R., & Sanger, J. (2008). *The text mining handbook: Advanced approaches in analyzing unstructured data* (Reprinted ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511546914>
- Florio, F. de. (2016). *Airworthiness: An introduction to aircraft certification and operations*(3rd ed.). Elsevier/BH.
- Giordano, V., Consoloni, M., Chiarello, F., & Fantoni, G. (2024). Towards the extraction of semantic relations in design with natural language processing. *Proceedings of the Design Society*, 4, 2059-2068. <https://doi.org/10.1017/pds.2024.208>
- Google Ireland Limited. (2024). *OCR (Optical character recognition)*. <https://cloud.google.com/use-cases/ocr?hl=en>
- Han, J., Sarica, S., Shi, F., & Luo, J. (2021). Semantic networks for engineering design: State of the art and future directions. *Journal of Mechanical Design*, 1-45. <https://doi.org/10.1115/1.4052148>
- Hoque, M. R. U., Wei, X., Choudhury, M. H., Ajayi, K., Gryder, M., Wu, J., & Oyen, D. (2022). Segmenting technical drawing figures in US patents. *CEUR Workshop Proceedings* (3), 2022 Workshop on Scientific Document Understanding.
- Kolandaisamy, R., Rajagopal, H., Kolandaisamy, I., & Sinnappan, G. S. (2024). The smart document processing with artificial intelligence. In M. Sugisaka (Ed.), *Proceedings of the 2024 International Conference on Artificial Life and Robotics (ICAROB 2024)*(pp. 534-554). International Steering Committee of International Conference on Artificial Life and Robotics.
- Laukotka, F. N., & Krause, D. (2024). Improving knowledge capture, access, and reuse by a model-based and data-driven management using the example of aviation's retrofit. In *Proceedings of the ASME 2024 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Volume 2B: 44th Computers and Information in Engineering Conference (CIE)*. American Society of Mechanical Engineers. <https://doi.org/10.1115/DETC2024-138000>
- Laukotka, F. N., Moenck, K., Schüppstuhl, T., & Krause, D. (2024). Graph-based information management for retrofitting long-living assets. *Data-Centric Engineering*, 5, 46. <https://doi.org/10.1017/dce.2024.49>
- Li, M., Wang, Z., Yan, Z., & Liu, J. (2023). Exploiting patent documents for cross-domain knowledge transfer in innovative engineering design: A Doc2Vec-GAT-based approach. In Yi J. (Ed.), *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*(pp. 1-6). IEEE. <https://doi.org/10.1109/CASE56687.2023.10260662>
- Lufthansa Technik AG. (2024). *Exemplary interface control document for a planned retrofit of a lavatory in an A320's aft section*.
- Mahadevkar, S. V., Patil, S., Kotecha, K., Soong, L. W., & Choudhury, T. (2024). Exploring AI-driven approaches for unstructured document analysis and future horizons. *Journal of Big Data*, 11(1). <https://doi.org/10.1186/s40537-024-00948-z>
- Meier, A., & Kaufmann, M. (2019). *SQL & NoSQL databases* (1st ed.). Springer Fachmedien. <https://doi.org/10.1007/978-3-658-24549-8>

- Moenck, K. H. W., Laukotka, F. N., Deneke, C., Schüppstuhl, T., Krause, D., & Nagel, T. J. (2022). Towards an intelligent digital cabin twin to support an aircraft's retrofit and base maintenance. *SAE Technical Paper Series*. SAE International. <https://doi.org/10.4271/2022-01-0046>
- Moenck, K. H. W., Laukotka, F. N., Krause, D., & Schüppstuhl, T. (2022). Digital twins of existing long-living assets: Reverse instantiation of the mid-life twin. In *Proceedings of the 33rd Symposium Design for X (DfX 2022)* (p. 10). The Design Society. <https://doi.org/10.35199/dfx2022.20>
- Nguyen, T., van Pham, L., Nguyen, C., & van Nguyen, V. (2021). Object detection and text recognition in large-scale technical drawings. In *Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods* (pp. 612-619). SCITEPRESS - Science and Technology Publications. <https://doi.org/10.5220/0010314406120619>
- Niță, M. F., & Scholz, D. (2009). The process chain to a certified cabin design and conversion. *Deutscher Luft- Und Raumfahrtkongress 2009*.
- Niță, M. F., & Scholz, D. (2011). Business opportunities in aircraft cabin conversion and refurbishing. *Journal of Aerospace Operations*, 1(1-2), 129-153. <https://doi.org/10.3233/AOP-2011-0008>
- Paneru, S., & Jeelani, I. (2021). Computer vision applications in construction: Current state, opportunities & challenges. *Automation in Construction*, 132, 103940. <https://doi.org/10.1016/j.autcon.2021.103940>
- Ramakrishnan, S., Bhatt, H. S., Raja, S., & Jawahar, C. V. (2024). Unlocking the potential of unstructured data in business documents through document intelligence. In Natarajan S. *et al.* (Eds.), *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (CODS-COMAD 2024)* (pp. 505-509). ACM. <https://doi.org/10.1145/3632410.3633293>
- Stark, J. (2016). *Product lifecycle management: Volume 2 - The devil is in the details*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-24436-5>
- W24 Service GmbH. (2024). *Werk24: Feature extraction from engineering drawings with AI*. <https://werk24.io/blog/structured-data-vs-unstructured-data>
- Yadav, S. (2024). AeroQuery RAG and LLM for aerospace query in designs, development, standards, certifications. In *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)* (pp. 1-6). IEEE. <https://doi.org/10.1109/CONECCT62155.2024.10677028>