Data Digitalization and Conformity Verification in Oil and Gas Industry Databooks Using Semantic Model Based on Ontology

Mario Ricardo Nascimento Marques Junior, Eder Mateus Nunes Gonçalves,
Silvia Silva da Costa Botelho and Emanuel da Silva Diaz Estrada
Center of Computational Sciences, Federal University of Rio Grande, Rio Grande, Brazil

Keywords: Data Digitalization, Oil and Gas Industry, Ontology, Semantic Model.

Abstract: Databooks are essential for monitoring and validating construction projects in the oil industry, containing crucial information like quality certificates and technical reports. However, manual analysis of these databooks is time-consuming, labor-intensive, and error-prone. This study proposes an intelligent system to streamline databook search and validation, enhancing efficiency and accuracy. Developing a valid conceptual model for databooks and their components presents a significant challenge. To overcome this, we focus on acquiring semantics for databooks and utilizing a semantic model for compliance checks. We introduce an ontology designed specifically for verifying completeness and compliance in Brazilian oil industry documents, encompassing domain knowledge and verification processes. Using the Methontology methodology, we create the ontology and integrate it with an annotation tool to validate its ability to incorporate semantic structures and facilitate compliance verification. Comparative analysis with manual verification by experts shows identical outcomes, confirming the effectiveness of the automated compliance checking process. The ontology-based approach offers advantages such as time savings, enhanced accuracy, and simplified work for specialists. This study contributes to oil industry document analysis by providing a semantic model that streamlines databook verification, with potential applications for compliance verification of complex documents in various domains.

1 INTRODUCTION

Databooks are pivotal documents in the oil and gas sector, housing essential equipment and work-related data. They compile an array of information, including engineering documents, modifications, purchase orders, tests, certificates, and inspection reports (Duarte, 2010). These records encapsulate an enterprise's history and components, potentially spanning thousands of pages for entities like ships or entire platforms.

In the context of Brazil's oil industry, the ABC of Inspection of Manufacture document (S.A, 2017) governs inspection guidelines for oil and gas equipment production. As per this document, manufacturing inspection is performed to verify the conformity of equipment or materials with contractual specifications at suppliers' or sub-suppliers' premises.

An Inspection and Testing Plan (ITP) is devised, based on inspected equipment, technical standards, and inspection levels. The ITP, part of the supplier's quality plan, outlines tests and certifications aligned with contract-defined quality and technical standards. All equipment documents are collated in a databook, serving as a customer's assurance certificate and containing crucial traceability information.

Post-manufacturing, a contracting party's inspector reviews documentation for completeness and compliance. Completeness entails ensuring all necessary information, certificates, reports, and tests are present in the databook. Compliance involves verifying that these documents adhere to contract-specified standards. If complete and compliant, material release occurs; otherwise, non-conforming product registration is initiated.

Professionals meticulously analyze and maintain these databooks, checking for anomalies such as missing reports or unsigned documents. The challenge lies in the time-intensive and costly nature of this manual analysis. Moreover, professionals verify databook completeness and compliance—tasks potentially optimized by an intelligent search system. Digitization and algorithms could streamline analysis, aiding in data retrieval, verifying completeness, and bolstering efficiency.

The majority of databook documents are scanned as PDF images, necessitating Optical Character

Nascimento Marques Junior, M., Gonçalves, E., Botelho, S. and Estrada, E.
Data Digitalization and Conformity Verification in Oil and Gas Industry Databooks Using Semantic Model Based on Ontology.
DOI: 10.5220/0012176700003543
In *Proceedings of the 20th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2023) - Volume 1*, pages 180-187
ISBN: 978-989-758-670-5; ISSN: 2184-2809

Figure 1: Proposal system.

Recognition (OCR) for conversion into structured text files. However, diverse companies contribute to databooks in varying formats, posing data modeling challenges. In this context, ontologies are more advantageous than data schemas, as they provide clearer semantics and mechanisms for content verification.

Ontologies model reality and logical representations in computer science, defining concepts, properties, and relationships. They enhance knowledge sharing within specific domains through a vocabulary. Works in oil and gas employ ontologies, such as IKBO for inspection knowledge (Rachman and Ratnayake, 2019) and ISO 15926-based high-level models (Batres et al., 2007). DoCO (Constantin et al., 2016) serves as a general-purpose vocabulary for academic texts.

The proposed system, as depicted in Figure 1, transforms image files into structured Q&A formats, leveraging an ontology for semantic inferences. This paper's focus is on imbuing a Q&A-structured document with semantics for conformity verification, using the Methontology methodology (Fernández-López et al., 1997). The study is part of a broader project automating verification using data science in a Big Data context.

The article follows this structure: Section 2 covers background and OntoToT development using Methontology. Section 3 presents ontology incorporation of actual documents. Section 5 concludes and discusses future work.

2 DATA DIGITALIZATION AND CONFORMITY VERIFICATION IN DATABOOKS

The process of digitizing data in a Databook involves several complex steps. Firstly, the PDF file of the Databook undergoes an OCR process. This Optical Character Recognition (OCR) process is responsible for converting the text embedded within images into a machine-readable text file. By doing so, it enables the computer to capture data from sources that were previously inaccessible or difficult to process.

However, not all the information present in a databook is relevant. To filter out only the pertinent information, an annotation tool has been developed. Through this tool, experts manually select the necessary information based on predefined labels.

Once the relevant data has been selected, it is annotated in a question and answer format, following the concepts and relationships defined by the ontology. Additionally, the annotation process generates a file that captures spatial and geometric aspects of the digitized document, which will be incorporated into a dataset. This dataset will be utilized for machine learning purposes, aiming to optimize the document understanding and form recognition process, thereby reducing or eliminating the need for manual annotations by the user.

The labeled data can then be inserted into a database, where it is stored and managed alongside the OCR-digitized data. This database facilitates the verification of data integrity and enables retrieval and consultation of the digitized information. The ontology is fed with this database, allowing for inferences related to completeness and compliance based on the representation of concepts and relationships. The process described above is illustrated in Figure 2.

Overall, this approach enhances the efficiency and accuracy of data digitization by combining manual selection, machine learning, and ontology-based inferences.

2.1 OntoToT - Ontology for Completeness and Compliance Verification in Petroleum Industry Documents

Developing an ontology is an iterative process that requires the use of a methodology, similar to software development (Aminu et al., 2020). However, despite the existence of different methodologies, there is no widely accepted standard for ontology development (Mendonça and Almeida, 2014).

In a comparative study conducted in (Aminu et al., 2020), the individual weaknesses of various methodologies are identified. Three methodologies, in particular, stand out: Methodology 101 (Noy and McGuinness, 2001), Methontology (Fernández-López et al., 1997), and the Methodology of Gruninger and Fox (Grüninger and Fox, 1995). These methodologies are

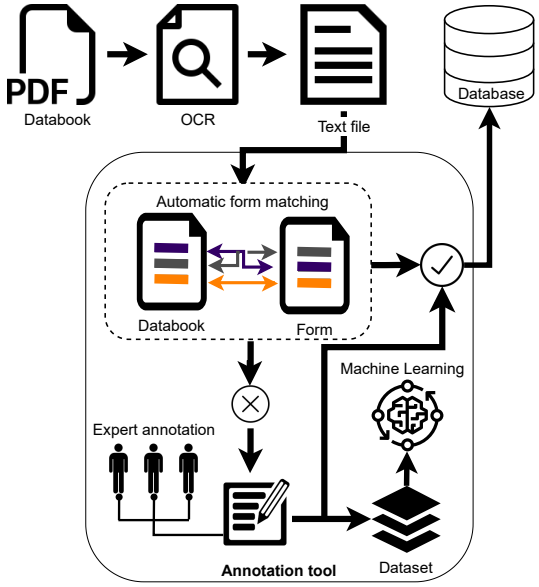

Figure 2: Databook digitalization process.

well-known and frequently used in the fields of software engineering and knowledge representation.

The analysis reveals that some approaches primarily focus on development activities, particularly ontology implementation (as seen in Methodology 101). These approaches tend to overlook crucial aspects such as project management, feasibility study, maintenance, and ontology evaluation, placing excessive emphasis on implementation details (Silva et al., 2008). It is worth noting that among these methodologies, Methontology is the only one that addresses project management and maintenance phases.

For the development of the OntoToT ontology, the Methontology methodology was chosen. Through research, it was determined that this methodology is the most suitable for the development of the ontology, as it provides a clear definition of the development stages and the corresponding activities. This ensures a proper and systematic development of the ontology. The following sections will provide a detailed explanation of the main stages involved in the development of the OntoToT ontology.

Specification

The ontology has the representation of documents found during the equipment inspection process as its primary objective. In addition, the ontology must include aspects related to the completeness verification process of databooks.

Acquisition of Knowledge

In the knowledge acquisition phase, it was necessary to read several databooks, ITPs, and purchase orders referring to different types of equipment. In addition, periodic meetings were held with employees of the partner company to clarify specific points and solve possible mistrust. It is also worth noting that its employees are specialists with several years of experience in the field studied here. Finally, the entire knowledge acquisition process was documented in text and slide shows.

Conceptualization

The conceptual modeling phase organizes and represents the knowledge acquired about the domain through intermediate representations. For example, in the ontology developed here, we used document analysis and the creation of trees to represent the concepts and their relationships.

Formalization

The formalization activity transforms the conceptual model into a formal or semi-computable model through a formal language(Gómez-Pérez et al., 2004). When tools like the ontology editor are used, the conceptualization model can be implemented directly in several ontology languages and, consequently, formalization is not a mandatory step (Gómez-Pérez et al., 2004). Thus, as the Protégé environment will be used to implement the ontology, this step will not be performed.

Implementation

To develop the ontology, the Protégé environment (Protégé, 2020), developed by Stanford University, was utilized. This is an open-source tool that employs a language based on descriptive logic (OWL-DL - Ontology Web Language Description Logics). Protégé consists of an ontology editor and a library of plugins with various functionalities.

For making inferences, we utilize HermiT (HermiT, 2020), Protégé's inherent inference engine tailored for handling OWL-based ontologies. HermiT adeptly checks the ontology's coherence, identifies subsumption connections among classes, and assesses other relevant attributes sourced from an OWL file. Remarkably, HermiT completes these assessments within mere seconds.

Documentation

We also used LODE (*Live OWL Documentation Environment*) (Peroni et al., 2012), an online service that automatically generates a human-readable description of any OWL ontology, taking into account both ontological axioms and annotations and sorting them with the appearance and functionality of a W3C recommendations document. This documentation is presented to the user as an HTML page with embedded links for easy navigation.

Ontology Description

Through the information obtained through the conceptual modeling, where the knowledge about the studied domain was organized and represented, artifices were generated that facilitated the ontology implementation. The definition of class and subclasses that make up the ontology is described below. The class hierarchy is illustrated in Figure 3.

Classes Description

- ***PDF File:*** Class connected directly to *Thing* class. This class represents the PDF file, the original source of the data present in the documents of the studied domain;
- ***PDF ID:*** Subclass of *PDF File*. Represents a unique identifier for each PDF file required for file retrieval from the data storage system. Thus, *ID PDF* represents an integer;
- ***PDF File Name:*** Subclass of *PDF File*. It is a set of characters that aims to identify a file, and can also be used to recover the file;
- ***Page:*** Subclass of *PDF File*. As a PDF can contain several pages and each page can contain a different quality certificate, it is necessary to know which page of the PDF this certificate is located on. Therefore, *Page* represents an integer;
- ***Documents:*** Class linked directly to *Thing* class. This class is responsible for grouping the types of documents and all the attributes they can contain in the documents of the studied domain;
- ***Documents Type:*** Subclass of *Documents*. It groups the types of documents that the ontology must include, namely: Quality certificate, Purchase order and ITP.
- ***Quality Certificate:*** A quality certificate is a document that gathers information that attests to the quality of a given product/service according to pre-defined standards;
- ***Purchase Order:*** It is a legal contractual instrument where the customer specifies the equipment to the supplier. This document has a detailed description of the item, such as quantity, standards that must be met and type of inspection performed;
- ***Inspection and Testing Plan - ITP:*** It is a document prepared by the supplier contained in its quality plan, following the standards established by the quality management norms and technical norms defined in the contract;
- ***Document Attributes:*** Subclass of *Documents*. It is responsible for grouping all the information that must be extracted from the documents. Therefore, its subclasses are *Signature, Stamp, Customer, NCM Code, Race, Date, Description, Equipment, Inspection Phases, Supplier, Item, Reference Standard, Certificate Number, Purchase Order Number, ITP Number, Quantity, Record, Test Result*;
- ***Signature:*** A signature or signature is a mark or writing in some document that aims to give it validity or identify its authorship;
- ***Stamp:*** Seal used for the recognition, proof or attestation of authenticity of a document;
- ***Client:*** Potential buyer or user of the supplier's products.
- ***Heat Number:*** It is a unique identification code that a technician stamps on a piece of metal to provide information about its origins;
- ***NCM Code:*** Any and all merchandise that circulates in Brazil must have the NCM code (Mercosur Common Nomenclature) and this code must be informed when filling in the invoice and other foreign trade documents;
- ***Date:*** Informs the time in the calendar when a certain document was generated;
- ***Description:*** Description of the equipment or component being purchased through a purchase order or evaluated through a quality certificate;
- ***Equipment:*** Equipment that is being acquired or evaluated before an essay or test;
- ***Inspection Steps:*** Phases of the inspection process of a certain equipment;
- ***Provider:*** Individual or legal entity that produces, assembles, creates, builds, transforms, imports; exports, distributes or markets products or services;
- ***Item:*** Code used to identify a specific piece of equipment or component in a document;

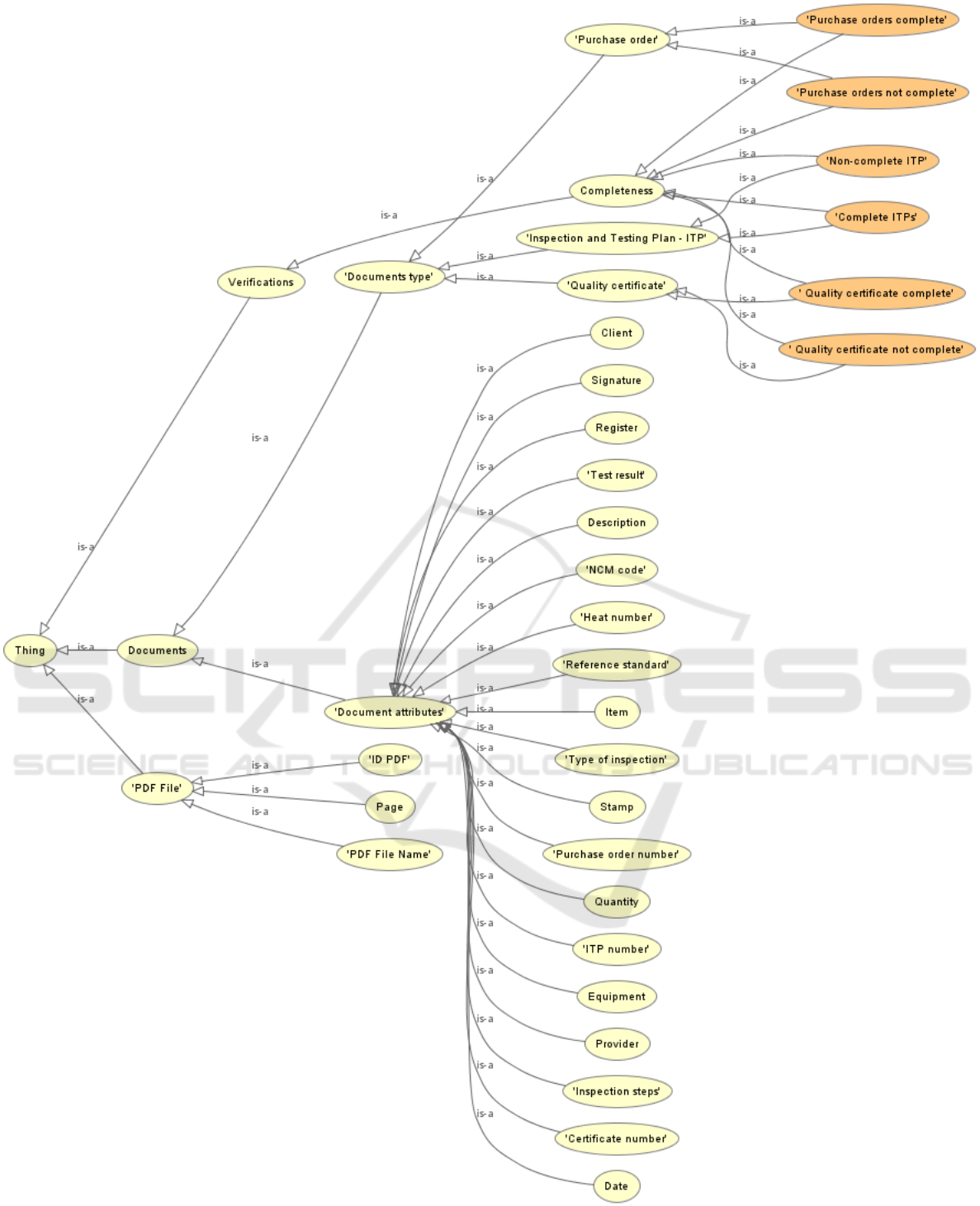

Figure 3: OntoToT class hierarchy.

- ***Reference Standard:*** It is a document, produced by an official body accredited for this purpose, which establishes rules, guidelines, or characteristics about a material, product, process or service;
- ***Certificate Number:*** Identifying number of a given quality certificate;
- ***Purchase Order Number:*** Identifying number of a given purchase order;

- ***ITP Number:*** Identifier number of a given ITP;
- ***Quantity:*** Number of items in a given purchase order;
- ***Register:*** Describes which document is the record that a certain inspection phase was performed;
- ***Test Result:*** Result of a given test or test for a physical or chemical property;
- ***Type of Inspection:*** Indicates the type of manufacturing inspection, which varies according to the criticality of the material, operational complexity of the material, complexity or novelty of the manufacturing process and quality control, and complexity or uniqueness of the project;

Object Properties

Object properties are basically divided into two groups, *must have* and *may have*. During the elaboration of the ontology, it was argued that, in the case of the quality certificate, some attributes must always appear, i.e., they are mandatory for the verification of completeness. However, other attributes have a certain frequency in the documents, but they are not mandatory for verifying completeness, i.e., the quality certificate can have these attributes. In the table 1 the properties, their domains, and their ranges are displayed.

Axioms

The axioms developed have the function of imposing restrictions on the classes used for completeness checking. That is, these axioms represent the rules that a certain type of document must respect in order to be considered complete or incomplete. To implement these axioms, the descriptive logic provided by Protégé was used. For each of the axioms a subclass of *Completeness* was created. These classes are described below.

- Class ***Quality Certificate Complete*****:** A quality certificate is said to be complete if all relevant information is correctly extracted. This information is defined by the object property *Quality certificate must have* and are: *Signature, Stamp, Customer, Date, Certificate Number* and *Test Result*;
- Class ***Quality Certificate Not Complete*****:** A quality certificate is not complete if at least one of the classes defined in the property *Quality certificate must have* is not extracted;
- Class ***Purchase Orders Complete:*** A purchase order is considered complete if it has all relevant information present. That is, a purchase order is considered complete if the *Signature, Stamp, NCM Code, Date, Description, Supplier, Item, Reference Norm, Purchase Order Number* and *Quantity* information is extracted successfully;
- Class ***Purchase Orders Not Complete:*** A purchase order is considered incomplete if at least one of the relevant information is not extracted;
- Class ***Complete ITPs:*** A ITP is said to be complete if it has all relevant information present. That is, it must contain *Signature, Stamp, Customer, Date, Equipment, Inspection Phases, Supplier, ITP Number, Registration*;
- Class ***Non-complete ITPs:*** A ITP is said to be non-complete if at least one of the relevant information is not extracted.
- ***ASTM A193 B7 Compliant Chemical Analysis Certificates:*** For a chemical analysis certificate to comply with the ASTM A193 B7 standard, it is necessary that the reference standard is the one referred to and that the values of the chemical elements correspond to those shown in the Table 2;
- ***Non-ASTM A193 B7 Chemical Analysis Certificates*****:** For a chemical analysis certificate not to comply with the ASTM A193 B7 standard, it is necessary that the reference standard is the one referred to and that at least one of the values presented in the Table 2 is different from that found in the certificate.

2.2 Pipeline for Data Conformity

As described in this section, the data extracted through OCR is stored in a database. Subsequently, a Python script is utilized to query the database and convert the data into instances within the ontology, establishing the necessary relationships for accurate representation.

In the Protégé environment, the generated file, containing only the classes and instances, is opened along with the OntoToT ontology file, which encompasses all the developed classes and rules. Following this, the inference mechanism is executed, utilizing the axioms to perform analysis on the completeness and conformity of the documents. The results of this analysis are displayed on the screen. The entire process is visually depicted in Figure 4.

3 RESULTS

To evaluate the proposed system, a set of databooks that align with the scope of the study was chosen.

Table 1: OntoToT object properties.

Object property	Domain	Range
Quality certificate must have	Quality certificate	Signature, Stamp, Customer, Date, Description, Certificate Number, Test Result
Quality certificate may have	Quality certificate	Heat number, Reference Standard, Purchase Order Number, Quantity
Purchase order must have	Purchase order	Signature, Stamp, NCM Code, Date, Description, Supplier, Item, Reference Standard, Purchase Order Number, Inspection Type
ITP must have	Inspection and Test Plan - ITP	Signature, Stamp, Customer, Date, Equipment, Inspection Phases, Supplier, ITP Number, Registration

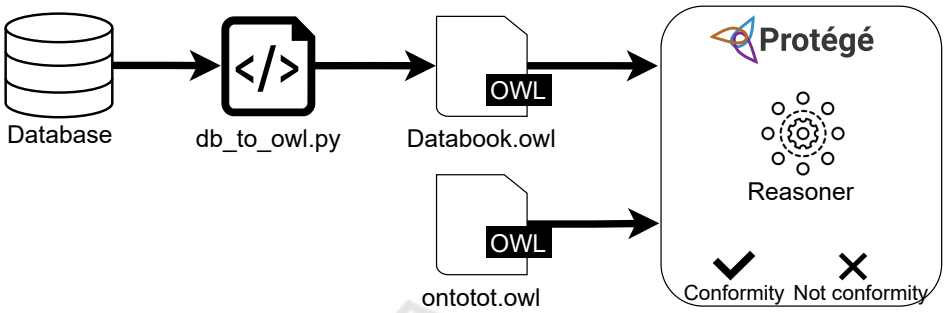

Figure 4: Ontology inference process.

Table 2: Chemical requirements requested by ASTM A193 Gr. B7.

Chemical element	Range (%)
Carbon	0.37 - 0.49
Chrome	0.75 - 1.20
Sulfur	max 0.040
Phosphor	max 0.035
Manganese	0.65 - 1.10
Molybdenum	0.15 - 0.25
Silicon	0.15 - 0.35

As the machine learning stage is still in development, each of these documents was manually annotated by an expert using the annotation tool, as depicted in Figure 2. The extracted data from these annotated documents were then stored in a database.

Table 3: Verification of compliance in chemical analysis certificates ASTM A193 Gr. B7 manually and through the ontology.

Certificate number	Compliance check by human	Compliance check by ontology
105501	Conforming	Conforming
2447/15	Conforming	Conforming
1236/17	Conforming	Conforming
105502	Conforming	Conforming
0390/18	Nonconforming	Nonconforming

Using the script illustrated in Figure 4, the data stored in the database was used to generate instances in the ontologies. Each generated file was subsequently inserted into Protégé along with the OntoToT file.

The inference mechanism in Protégé was then executed, utilizing the axioms to provide information about the completeness and conformity of the documents. The results of this analysis are presented in Tables 4 and 3. For comparison, the same set of documents was analyzed by humans using the same criteria, and the results obtained from this manual analysis are also included in the tables.

4 CONCLUSION AND FUTURE WORK

This work presents a system proposal for digitizing and ensuring compliance in oil and gas industry databooks. The first step involves converting image files into text files to enable computer understanding. To provide a semantic structure and enable inferences, an ontology called OntoToT was developed using the Methontology methodology, ensuring proper development, documentation, and incorporating project management and maintenance stages.

The annotation process confirmed that the ontology offers sufficient semantic structure for representing the analyzed databooks and demonstrated successful integration with the annotation tool. The proposed approach enables the digitization of documents, adds semantics to the data, and facilitates automatic

Table 4: Verification of completeness in quality certificates performed by human and by ontology.

Certificate Number	Completeness check by human	Completeness check by ontology
3418/17	Complete	Complete
1049/18	Complete	Complete
0700/17	Complete	Complete
2650/2018	Complete	Complete
111845	Complete	Complete
1078/2018	Complete	Complete
106210	Complete	Complete
3503/2017	Complete	Complete
1338739/2007	Not complete	Not complete
108277	Not complete	Not complete

compliance verification, thus simplifying the work of specialists.

Future work includes expanding the ontology to handle more complex documentation and verifications, such as assessing the completeness of an entire databook for specific equipment. The development of a dictionary to unify different labels for the same attribute could also be considered. It is essential to address ontology maintenance to accommodate new documents and verifications effectively.

Additionally, enhancements to the annotation tool using machine learning techniques could optimize expert annotation by requesting it only when necessary. Improvements in the inference process could involve developing a script for automated inference without manual intervention, storing the results in a database.

REFERENCES

Aminu, E. F., Oyefolahan, I. O., Abdullahi, M. B., and Salaudeen, M. T. (2020). A review on ontology development methodologies for developing ontological knowledge representation systems for various domains. *International Journal of Information Engineering & Electronic Business*, 12(2).

Batres, R., West, M., Leal, D., Price, D., Masaki, K., Shimada, Y., Fuchino, T., and Naka, Y. (2007). An upper ontology based on iso 15926. *Computers & Chemical Engineering*, 31(5-6):519–534.

Constantin, A., Peroni, S., Pettifer, S., Shotton, D., and Vitali, F. (2016). The document components ontology (doco). *Semantic web*, 7(2):167–181.

Duarte, G. D. (2010). *O Controle da Qualidade em Processos de Produção Mecânica Não-Seriada*. Monografia (Graduação em Engenharia Mecânica), Escola Politécnica da Universidade de São Paulo, São Paulo, Brasil.

Fernández-López, M., Gómez-Pérez, A., and Juristo, N. (1997). Methontology: from ontological art towards ontological engineering. In *AAAI 1997*. American Asociation for Artificial Intelligence.

Gómez-Pérez, A., Fernández-López, M., and Corcho, O. (2004). Methodologies and methods for building ontologies. *Ontological Engineering: with examples from the áreas of Knowledge Management, e-Commerce and the Semantic Web*, pages 107–197.

Grüninger, M. and Fox, M. S. (1995). Methodology for the design and evaluation of ontologies. In *IJCAI 1995*. Citeseer.

HermiT (2020). Hermit owl reasoner. http://www.hermit-reasoner.com/. Acess em 03/08/2020.

Mendonça, F. M. and Almeida, M. B. (2014). Princípios metodológicos para desenvolvimento de ontologias: análise das práticas correntes e proposição de melhorias. *XV Encontro Nacional de Pesquisa em Ciência da Informação - ENANCIB*.

Noy, N. F. and McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. Technical report, Stanford Knowledge Systems Laboratory.

Peroni, S., Shotton, D., and Vitali, F. (2012). The live owl documentation environment: a tool for the automatic generation of ontology documentation. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 398–412. Springer.

Protégé (2020). Protégé. https://protege.stanford.edu/. Acess 20/05/2020.

Rachman, A. and Ratnayake, R. C. (2019). An ontology-based approach for developing offshore and onshore process equipment inspection knowledge base. In *International Conference on Offshore Mechanics and Arctic Engineering*, volume 58783, page V003T02A084. American Society of Mechanical Engineers.

S.A, P. B. (2017). ABC da inspeção de fabricação. Instrumento informativo, Petróleo Brasileiro S.A, Brasil.

Silva, D. L. d., Souza, R. R., and Almeida, M. B. (2008). Ontologias e vocabulários controlados: comparação de metodologias para construção. *Ciência da informação*, 37:60–75.