K-Hub: a modular ontology to support document retrieval and knowledge extraction in Industry 5.0

Anisa Rula[1 [0000−0002−8046−7502]], Gloria Re Calegari[2 [0000−0002−4558−229X]], Antonia Azzini[2 [0000−0002−9066−1229]], Davide Bucci[2 [0000−0001−6051−6701]], Alessio Carenini[2 [0000−0003−1948−807X]], Ilaria Baroni[2 [0000−0001−5791−8427]], and Irene Celino[2 [0000−0001−9962−7193]]

[1] University of Brescia, Italy
anisa.rula@unibs.it
[2] Cefriel – Politecnico di Milano, Italy
{name.surname}@cefriel.com

Abstract. Digitalization is entering the industrial sector and different needs are emerging to support shop floor operators; in particular, they need to retrieve information to support their operations (e.g., during maintenance activities), from structured and unstructured sources, as well as from other people's experience. Sharing knowledge and making it accessible to industrial workers is therefore a key challenge that Semantic Web technologies are able to address and solve. In this paper, we present a modular ontology that we engineered in order to support the collection, extraction and structuring of relevant information for industrial operators in a "knowledge hub" (K-Hub). In particular, our K-Hub ontology covers several aspects, from document annotation/retrieval to procedure support, from manufacturing domain concepts to company-specific information. We discuss its engineering process, extensibility and availability, as well as its current and future application scenarios to support industrial workers.

Keywords: Industry 5.0 · document retrieval · knowledge extraction.

Resource type	Ontology
Licencse	CC BY 4.0 International
DOI	https://doi.org/10.5281/zenodo.7443000
URL	https://knowledge.c-innovationhub.com/k-hub/

1 Introduction

The manufacturing industry is advanced by a technological revolution, often referred to as Industry 4.0 [13], where the future trend lies in the convergence of several technologies including artificial intelligence, smart manufacturing, Internet of Things and web-based knowledge management. Moreover, the advent of

the so-called Industry 5.0[3] is shedding light not only on the adoption of digital technologies, but also on their actual uptake by industry workers, thus making industry sustainable, human-centric and resilient. With specific reference to knowledge management, manufacturing companies face the challenge of managing, maintaining and transferring different kinds of knowledge between people and across company functions such as product design, process definition, production lines, system maintenance and customer service. This knowledge can be present in documents like user manuals, troubleshooting instructions, guidelines, internal processes and so on. Those documents should ensure optimal comprehensibility by the operator to safely and effectively install, operate, maintain and service the industrial systems. Given the high number and diversity of such documents, the operators often have to go through a laborious and time-consuming process of searching them and trying to filter their content to find the relevant information to answer their questions.

In this scenario, enterprises call for tools and methods for extracting knowledge from unstructured information encoded in documents (e.g., PDF or text files), using diverse state-of-the-art Natural Language Processing (NLP) [12] techniques that involve three main tasks: Named-Entity-Recognition (NER) [19], Entity-Linking (EL) [24] and Relation Extraction [1]. Methods to automatically extract or enrich the structure of documents have been a core topic in the context of the Semantic Web [17]; however, those automated methods may not solve the knowledge extraction process entirely. Indeed, extracting complex knowledge from unstructured sources is a challenge [21]: in the industrial domain, for example, troubleshooting documents may contain the description of long and articulated procedures (i.e., sequences of steps to be performed in a precise order and under specific conditions) and those natural language instructions may be represented in very different textual forms, thus making it hard for a knowledge extraction algorithm to correctly identify and structure the relevant information. Oftentimes, automatic extraction is followed by manual revision of domain experts. In any case, all machine-learning based methods require training data which is often not readily available, therefore novel approaches are emerging to exploit interactive dialogues and language models [2].

Even when the extraction is supported by suitable approaches, knowledge still requires to be represented in the structured form of a knowledge graph by means of ontologies. In the case of knowledge extraction from industrial documents, different aspects co-exist: domain concepts and company-specific terms are mixed with procedure/process information. The manufacturing domain is definitely multi-faceted and even recent surveys of the existing semantic vocabularies and ontologies identified a high number of efforts [6]. Therefore, to build knowledge graphs out of industrial documents, multiple ontologies are needed to cover all the relevant elements. In particular, our idea is to propose a set of vocabularies that improve the coverage of document annotations and knowledge extraction thereof, by means of ontology modularization [15], an interesting strategy to

[3] Cf. `https://research-and-innovation.ec.europa.eu/research-area/industry/industry-50_en`

facilitate ontology reuse, since it allows for different ontology modules to cover specific subdomains.

In this paper, we present the K-Hub ontology, a modular conceptual model able to capture the different aspects of manufacturing knowledge management and to support the building of a "knowledge hub" that helps industrial operators like shop floor workers in their daily operations. The K-Hub ontology is made of a set of modules that identify and capture entities and relationships that are relevant for document retrieval and knowledge extraction: an annotation module, that covers the aspects of document analysis and knowledge extraction; a manufacturing module, which contains the most common domain topics that can be found in industrial documents; a procedure module, which addresses the challenges of representing complex process information; and company-specific modules that are necessary every time an enterprise uses dedicated names, terms and acronyms (often even characterized by privacy or confidentiality constraints).

The remainder of this paper is structured as follows: Section 2 illustrates our reference motivational scenario, based on the actual knowledge management needs of two different manufacturing companies; Section 4 describes the methodological approach and its application to the engineering of the K-Hub modular ontology; the details of the K-Hub ontology modules are explained in Section 5; Section 6 demonstrate the use of the K-Hub ontology, both in a document retrieval scenario implemented and tested by shop floor operators via a voice assistant, and in a scenario to support procedure execution; relevant work from state of the art is included in Section 3; finally, we offer our conclusions and delineate future lines of work in Section 7.

2 Motivational scenario

The need for the K-Hub ontology emerged in a cooperative research and innovation project named "Manufacturing Knowledge Hub", with the final purpose to develop a voice assistant solution dedicated to supporting shop floor workers during maintenance processes. In this context, a huge number of documents must be managed and retrieved, in various digital formats: textual documents, pictures, spreadsheets, technical drawings, movies, presentations, etc. In the project, two different manufacturing companies provided their scenarios and specific needs and evaluated the project results.

The first one is Whirlpool, the multi-national home appliances manufacturer; in their maintenance procedures on the production lines, the real challenge is to find the relevant information within this universe of heterogeneous data (sometimes also including documents in paper format or in a scanned digital form), which can create an obstacle for an effective knowledge sharing, but which can represent a key element to take advantage of in the digitalization process. In Whirlpool, different plants, or even different production lines within the same plant, currently, adopt various practices to organize and search for information in the wealth of available documents.

The second involved company is Marposs, a large enterprise specialized in designing and manufacturing products and solutions for measurement, inspection and testing, widely used in very different sectors (e.g., automotive, aerospace, biomedical, energy, consumer electronics); in relation to Marposs' standard products, they already have well-structured documents, but also very long ones, with a lot of information; in this case, during maintenance activities, which employees often perform at the customer plants, the challenge is not only to find the right document but also to identify the relevant information within it, for example, to understand what maintenance or troubleshooting procedure to follow, especially in the case of novice operators.

Within the project, a voice assistant solution was designed and developed to simplify the access to the knowledge for the shop floor operators of both Whirlpool and Marposs. The K-Hub ontology described in this paper is a core part of this solution, with the purpose to facilitate document retrieval and knowledge extraction; as explained in the following, we engineered the K-Hub ontologies with the support of the domain experts from Whirlpool and Marposs, but we generalized their requirements so that our model can be reused in similar scenarios also beyond the two involved companies.

3 Related work

This work involves ontology engineering through the modularization of different related conceptualizations, to combine in a "knowledge hub" relevant concepts and relations. As far as we were able to determine after the initial literature search at the beginning of this ontology development process, as well as during the identification of ontological resources to be reused, this is the first comprehensive and fully documented effort for the generation of a modular ontology "hub" in this area, which is born with the objective of serving further standardisation and community-driven initiatives around this domain [6].

We can mention some previous approaches reported in the literature, where vocabularies of workflows represent scientific experiments [4,7,10]. A popular vocabulary for describing activities is provided by PROV-O which relates activities to a plan but it does not allow for plans to be described. Therefore, P-Plan[4] [9] is proposed as an extension of `prov:Plan`. Other vocabularies such as ProvONE or its extension, ProvONE+[5] are general-purpose specification models for the control-flows in scientific workflows [4]. However, the only vocabulary describing closely the structure of procedures in our scenario is P-Plan. The Web Annotation Data Model[6] provides an extensible, interoperable framework for expressing annotations specifically for Web pages. It is possible to define our TopicAnnotation as a specialization of oa:Annotation for a higher level of expressivity.

In the domain of industrial and manufacturing, there are already a number of available vocabularies, but the critical aspect of this domain is that those

[4] `http://purl.org/net/p-plan`
[5] `http://purl.org/provone`
[6] `https://www.w3.org/TR/annotation-model/`

belong to different areas such as product, systems and supply chain. Therefore, the definitions of the terms are very heterogeneous, as stakeholders view the manufacturing elements differently [6]. However, our extensible design of the K-Hub ontology allows for plugging-in other ontologies as additional modules, like for example SAREF4INMA [23], to cover other elements specific to the industry and manufacturing domain. We plan to register our vocabulary as part of the Industry Ontology Foundry (IOF) Initiative [14] which provides a repository for open reference ontologies to support the manufacturing and engineering industry needs and advance data interoperability.

Other works that aim at creating a modular ontology for the semantic annotation belonging to other domains are reported. The authors in [5] propose a network of ontologies for ICT infrastructures. They solved the problem of interoperability by homogeneously describing the core concepts and properties that are common across configuration and IT Service management databases. Similarly to our approach, the ontology network can be easily extended when new types of items appear. The authors in [22] construct a structure named emerging ontologies, which involves elements of more than one ontology. The idea is to provide a global view of several ontologies in one single structure which is useful for semantic annotation with concepts that come from more than one ontology.

4 Requirements and methodology

We developed the K-Hub Ontology relying on the Linked Open Terms (LOT) Methodology[20], an industrial method for developing ontologies and vocabularies. The LOT methodology enriches the main workflow with Semantic Web-oriented best practices such as the reuse of terms (ontology classes, properties, and attributes) existing in already published vocabularies or ontologies and the publication of the built ontology according to Linked Data principles. The LOT methodology defines iterations over a basic workflow composed of the following activities: (i) ontological requirements specification, (ii) ontology implementation, (iii) ontology publication and (iv) ontology maintenance.

In this section, we focus on the process that we followed for all the steps of the LOT methodology, which are explained in detail in the following while Section 5 describes the contents of the final published ontology.

4.1 Ontology requirements specification

The ontology requirements specification activity was driven by the interviews conducted with domain experts and visits to the operating sites of the two companies. The interviews performed during the visits involved different stakeholders at different levels (management, technical support, end users). We collected information about their processes, their needs and pain points, to identify the main knowledge aspects they manage. This activity can be divided into the following sub-steps:

Use case specification: this activity has the goal to provide a list of use cases. We investigated specific use cases for each case study, with the respective goals to be achieved by the ontology data modeling. In the end, the two companies had similar needs that are captured by the following use cases.

UC1: The user (shop floor worker) wants to retrieve a document for supporting him/her during the maintenance process.
Description: this use case refers to a maintenance situation, focused on retrieving technical documents during specific maintenance activities. The maintenance activities may be based on the management of maintenance data on the shop floor, during daily activities performed by maintenance employees.
Actors: different types of actors are involved in maintenance operations: engineers, expert technicians, maintenance workers or new employees who need access to a specific document.
Flow: in the maintenance scenario, the documents relevant to the project are redacted by documentation workers and maintained within the company (in legacy systems and intranet networks). Engineers and technicians, both experts and novices, access those documents during maintenance activities or interventions related to problems or troubleshooting, for example, to find the most recent version of a document pertaining to a specific topic.

UC2: The user (shop floor worker) wants to be guided step-by-step in the correct execution of a company procedure during the maintenance process, especially if they are not an experienced employee.
Description: this use case refers to a maintenance situation, focused on guiding a shop floor worker in the execution of a specific maintenance process. Such a maintenance process is carried out by the execution of one or more procedures performed step by step, by the maintenance employee.
Actors: different kinds of actors may be involved in maintenance-specific procedures: technicians, maintenance workers and not-experienced employees who need to be guided step by step in the procedure execution.
Flow: in the maintenance scenario, shop floor engineers and technicians, both experts and novices, need to find the operational procedure to be applied to solve the problem at hand; they search for the most suitable procedure; they identify the relevant contextual information (e.g. tools to be used to execute the procedure, spare parts to have at disposal); they follow the procedure, possibly being guided in each step, getting information on what actions to perform in which order and, at the conclusion of each step, what is the next step to be followed. They can find all the relevant information within documents (similarly to UC1) or they can be supported by a digital tool that provides them interactive guidance within the procedure (e.g. an intelligent assistant).

The User Story generated by the identified use case *UC1* is **US1**.
US1: The user wants to retrieve a document and to open it at the most relevant page by specifying one or more topics/characteristics; some examples are:

the type of document (e.g. installation manual), the machine/workstation/component on which the maintenance action will be performed, the action to be executed (e.g. replacement of a component, configuration, repair, etc.), the error to be solved in a troubleshooting.

The User Stories generated by the use case *UC2* are **US2**, **US3** and **US4**. **US2**: The user wants to find a company procedure to be followed, that best suits the specific maintenance activity at hand by specifying one or more topics/characteristics; some examples are: the machine/workstation/component on which the maintenance activity will be performed, the procedure to be executed, the error to be solved.

US3: The user wants to know what the next step is to be executed in the current procedure by specifying the last executed step.

US4: The user wants to know what tools are needed to perform a specific procedure.

Data exchange identification. This activity aims to gather domain documents and resources. In particular, during the interviews conducted with the domain experts from the two companies, we obtained all the relevant information about the domain to be modeled. In particular, we gathered details on the documents a user wanted to retrieve including general aspects of the documents, access constraints, documents' topical content and the main search strategies that people use. During this collection activity, the company stakeholders also provided a list of sample documents (product and service manuals, schematic representations of electrical/mechanical/hydraulic/... components). The analysis of the collected information allowed us to provide a clear definition of the application domain and the appropriate terminology.

Functional ontological requirements. Competency Questions (CQs) are a well-known technique to define ontology functional requirements and in the form of a set of queries that the ontology should answer. On the other hand, preliminary definitions (or facts) are assertions that provide a description of the requirements associated with the considered domain terminology. We cooperated with the domain experts in the definition of both CQs and facts, thanks also to the selection of relevant documents described before. At the end of this stage, we provided the full list of competency questions and facts to the domain experts and user representatives, who performed their validation in terms of accuracy and completeness with respect to the identified use cases and user stories[7]. Some examples of CQs and facts are given, respectively, in Table 1 and Table 2, with their relation to use cases and user stories.

[7] The full list of CQs and facts is available at `https://github.com/cefriel/k-hub-ontology/blob/main/PaperCompetencyQuestions_Facts.xlsx`.

UC	US	Identifier	Competency Question
UC1	US1	Cq-1	Which is the document about topic Z?
UC2	US2	Cq-2	Which is the procedure to do the action X on the component Y?
UC2	US3	Cq-3	Which is the next step to be executed?
UC2	US4	Cq-4	Which are the tools required for the procedure X?

Table 1. Examples of competency questions.

UC	Identifier	Preliminary Definition (Fact)
UC1	Req-1	A document is associated to one or more topic annotation
UC1	Req-2	A topic can be one of: action, component, product, machine, workstation, document type, supplier, trouble, tool, spare part, error, configuration.
UC2	Req-3	A step is associated to the next one.
UC2	Req-4	A procedure requires one or more tools.

Table 2. Examples of preliminary definitions.

4.2 Ontology implementation

The ontology implementation followed the LOT methodology. Our team of ontology engineers analysed the requirements and divided them into modules, since each module contains a subset of concepts and relations identifying an area of specialisation. The creation of modules is useful for facilitating the update and evolution of ontology in the future. After the ontological requirements were identified in the requirements specification process, we created the conceptual models using the Chowlk tool[8], which is a UML-based notation and provides a set of recommendations for ontology diagrams representation. We discussed the conceptual models with the domain experts on the basis of their graphical representation (as it is easier to understand for people with limited or no background on ontologies), then we proceeded to generate the formal representation in the OWL language, again using the capabilities of Chowlk. We carried out the ontology implementation phase iteratively, validating and refining it with some of the same domain experts that were involved in the requirements specification process. The OWL representations of the ontology modules are maintained in the GitHub repository.

We evaluated the ontology throughout the standard LOT process by asking for feedback to stakeholders: two persons from each company (among those interviewed at the beginning) were repeatedly involved to validate use cases, user stories, competency questions and facts. Moreover, when we adopted the ontology in the document retrieval system described in Section 5.1, the users indirectly evaluated the ontology by assessing the search results and comparing them to their expectations. In addition, regarding this methodology, we also used the OOPS tool to evaluate our ontology in terms of common pitfalls.

In the end, we performed a final assessment to verify that the ontology fulfills all the requirements, by checking the compliance between the ontology imple-

[8] Cf. `https://chowlk.linkeddata.es/chowlk_spec`

mentation and the full list of competency questions and facts, and we also verified the absence of syntactic, modeling, or semantic errors.

4.3 Ontology Publication

The aim of the ontology publication activity is to provide an online ontology accessible both as a human-readable documentation and a machine-readable file from its URI, according to the FAIR principles. More specifically, we published the K-Hub Ontology online (cf. Section 5) following the best practice recipes for publishing vocabularies with content negotiation [3]. We documented the ontologies using WIDOCO[9] [8], a wizard that takes as input an ontology to generate a set of linked HTML (draft) pages containing a human-readable description of the ontology from the ontology content. It guides users through the steps to be followed when documenting an ontology, indicating missing metadata that should be included. As recommended by the LOT methodology, we created an extensible and modular ontology with the goal of making it available to support industrial workers covering several aspects; accordingly, we published and documented each of the modules of our ontology. We extended the automatically-generated HTML pages, by adding diagrams and other explanatory information.

As will be explained in more detail in the next section, our modular ontology contains both the knowledge of general use/availability and the specific knowledge of the two companies involved in this ontology engineering effort. With respect to the latter kind of knowledge, it is important to note that a critical requirement is to preserve the privacy of the information represented in the ontology [11]: for example, product names or supplier information may be covered by confidentiality constraints. Therefore, publishing their related details openly on the Web may be impossible, instead, a restricted access will be required. In order to cope with this situation, we managed the open/public and the private modules in partially different ways. For both cases, we adopted a GitHub repository and the content negotiation-based vocabulary publication best practices. While the public modules are maintained in a public GitHub repository with the machine-readable and human-readable representations openly reachable from the respective namespaces, the private modules are maintained in private GitHub repositories and their representations are password-protected and accessible only to those with the proper credentials. This double publication ensures the proper modularization and extensibility of the ontology, while at the same time preserving the business constraints of the two companies. We believe that this (simple) approach can be adopted in many other similar situations, in which privacy or confidentiality is a key requirement.

To complete the ontology publication, we also archived the public modules of the ontology in Zenodo[10], following the usual practices of Open Science.

[9] `https://github.com/dgarijo/Widoco/`
[10] `https://www.doi.org/10.5281/zenodo.7443000`

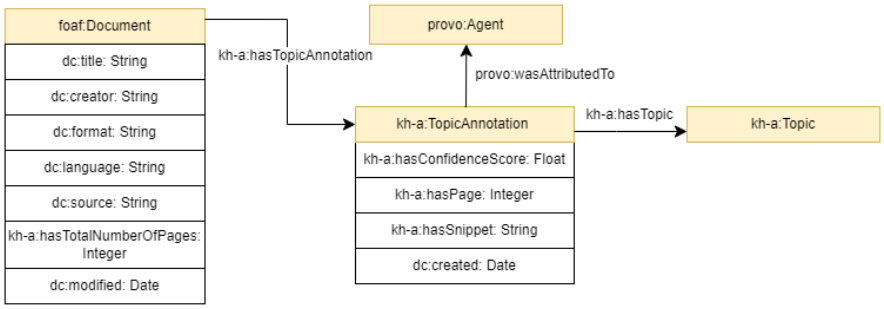

Fig. 1. Graphic representation of the Annotation Module

4.4 Ontology maintenance

Our setup is now prepared for the ontology maintenance phase for the ontology, with the possibility of submitting issues through the GitHub repository (bugs, requests for additions, etc.) for each of the modules in the ontology, so as to facilitate discussions that may arise during future standardisation processes or ontology usage by other organizations that could extend and reuse its modules.

5 The K-Hub ontology modules

In this section, we describe the current version of the implementation of the Knowledge Hub Ontology and its modules, and the main decisions taken during their development.

Annotation Module
`https://knowledge.c-innovationhub.com/k-hub/annotation`
The annotation module of the Knowledge Hub Ontology represents the core of the ontology with concepts and properties used for describing the annotation of documents. This module is composed of 3 main concepts: `Document`, `Topic` and `TopicAnnotation`. The `Document` concept describes the document's information through general data properties such as the author, the edit date, the format, the language, and the url. The `TopicAnnotation` concept describes the semantic annotation of the snippet extracted from the document. The datatype properties describing the annotation include the page number of the document containing the annotation, the snippet containing the information to be annotated, the creator of the annotation, the creation date, and, if present, the annotation's confidence score. The `TopicAnnotation` also connects the document with its content information, expressed with a list of "topics". The `Topic` concept, therefore, refers to any subject, theme, entity or object contained in the document and which the final user may be interested to search. The `Topic` concept is further specialized in the others modules of the ontology and constitutes the main extension point of the ontology. During the analysis of existing vocabularies, we identify and reuse existing concepts and properties in the conceptual

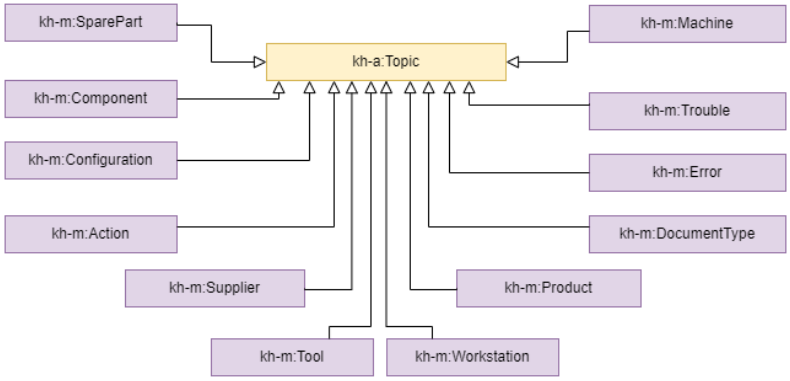

Fig. 2. Graphic representation of the Manufacturing Module

model: *FOAF* ontology for the definition of document concept, the *Dublin Core* ontology for describing the document's properties and the *PROV-O* ontology for modeling the provenance information about the annotations. Figure 1 displays the final version of this module.

Manufacturing Module

`https://knowledge.c-innovationhub.com/k-hub/manufacturing`

The manufacturing module of the K-Hub Ontology defines the specific topics for the domain of interest of the document. During the requirement collection phase, it was possible to define the list of concepts used for the maintenance process in the manufacturing domain. The identified concepts are represented as subclasses of the most general concept `Topic` defined in the Annotation Module, as displayed in Figure 2. These subclasses describe general maintenance elements such as: `Component`, `Configuration`, `Supplier`, `DocumentType`, `Trouble`, `Action`, `Product`, `Machine`, `Workstation`, `SparePart`, `Error` and `Tool`. The implementation of this module was further enriched with a terminology of instances of the aforementioned concepts. This terminology represents a list of entities to be searched in the annotation of documents. The instances defined in this module are specific to the domain of interests given by a single company. The terminology was created thanks to the collaboration of the ontology developer team and the involved industrial partners and it contains a list of terms translated into English and Italian and a list of possible synonyms. The terminology was modeled using the *SKOS* vocabulary [18]. We defined Topic as skos:Concept and the hierarchy of topics as subClassOf Topic. We refer to "topics" as classes with instances and the annotations are expected to refer to those instances (e.g. `annotation1,hasTopic,productX`). The instances provide the indexer with a complete list of terms to be searched in the document for the annotation process.

Procedure Module

`https://knowledge.c-innovationhub.com/k-hub/procedure`

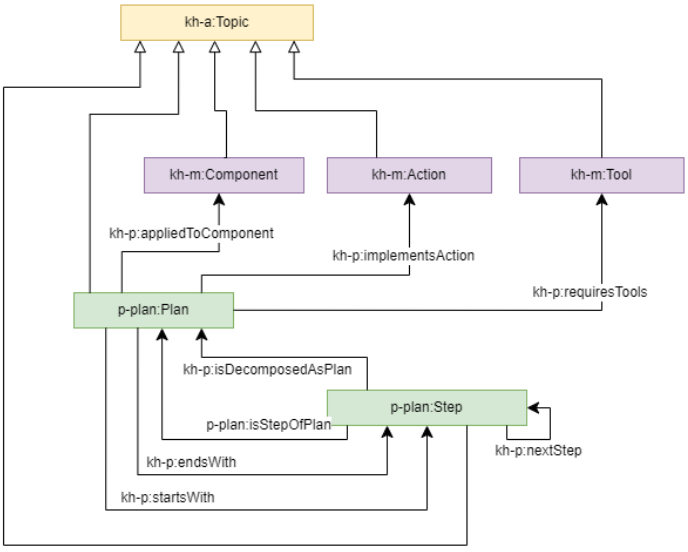

Fig. 3. Graphic representation of the Procedure Module

The procedure module of the K-Hub Ontology defines the concepts and properties for modeling the procedures described in the service manuals, usually composed of multiple atomic steps, for instance, guidelines for maintenance activities. As Figure 3 shows, the ontology consists of 2 main concepts: `Plan` and `Step`, which are also defined as subclasses of `Topic` (from the annotation module). A procedure is an instance of the `Plan` concept and it can be considered as a pattern like "A *Procedure* to do an *Action* on a *Component* with a *Tool*". The object properties `implementsAction`, `appliedToComponent` and `requiresTools` implement these associations between the `Plan` concept and the `Action`, the `Component` and the `Tool` concepts respectively, as defined in the Manufacturing Module. Each atomic activity is an instance of the `Step` concept. A `Plan` is composed of one or multiple `Steps`, which must be executed in a given order. The object property `nextStep` defines the execution order of the steps, whereas the `startsWith` and `endsWith` properties indicate the first and last steps of a `Plan`. A `Plan` may be included as a `Step` of another plan and this association is expressed by the object property `isDecomposedAsPlan`. During the analysis of the existing vocabularies, we identified the P-Plan ontology [9] as a very interesting source for our modeling; therefore the `Plan` and `Step` concepts, as well as some of the properties of this module, reuse the respective P-Plan definitions.

Company Specific Modules

This part of the K-Hub Ontology is intended to model the private concepts and terms of industrial companies that may have privacy/confidentiality issues. For each company, we created a private module that contains context-specific instances related to the company's business. The included terms refer to the

specificity of each company. Some examples are the names of suppliers, the names of specific products or the description of errors that occur on a product. As explained before, those modules are published according to the best practices, but their access is protected.

6 Ontology use

The modular ontology described so far was conceived in the context of the scenario illustrated in Section 2, to improve document management and retrieval in industrial maintenance for the Whirlpool and Marposs companies; this scenario was further detailed in the Use Cases described in Section 4. We fully implemented the entire tooling to support the first use case, which was also evaluated by industry user representatives, while we only started to lay the foundations for the tools and methodologies to support the other use cases.

6.1 Ontology use in document search

The first use corresponds to the use case UC1, in which shop floor operators involved in a maintenance activity want to retrieve the right document for the case at hand. We employed the ontology to build a system that effectively supports this use case (cf. Figure 4).

The ontology was used in the first step of document annotation: the relevant materials provided by the industrial partners were processed by a system that, for each document page, analyse its textual content in order to identify the most relevant `Topic`s (operating, as such, an *entity linking* process): for example, the annotation can discover that a specific document contains information about a certain `Product`, mentions its `Component`s or its `SparePart`s.

In case the automatic document annotation process does not perform correctly or this information is hard to identify automatically or the system (which is usually based on machine learning algorithms) requires a proper training set, the annotation step can also be performed manually by a domain expert[11].

The output of this phase – whether automatic or manual – is a set of `TopicAnnotation`s which indicate, for each document, which topics were identified in which page(s); those annotations are stored and indexed in order to be ready for retrieval. We used a combination of a triple store and a full-text index to manage the annotation storage and provide a web API-based access layer to search applications.

Then, we set up a digital tool to support document retrieval for the shop floor operator: a voice assistant helps the user in finding the right document. This tool exploits the ontology in two ways: first, it elaborates the user requests/utterances in order to understand what is the main retrieval interest (e.g. if the person asks "how can I replace the battery of product X?", the system shall identify the

[11] We are currently working on an ontology-extension of the PAWLS tool for the manual annotation of PDF documents, cf. `https://github.com/cefriel/onto-pawls`

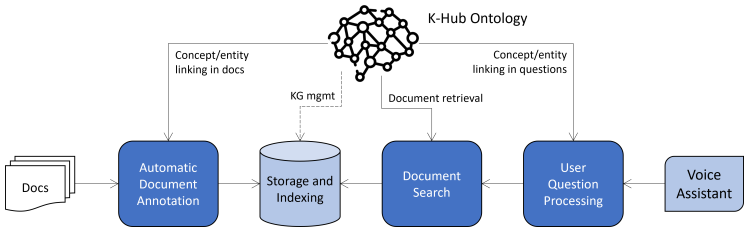

Fig. 4. Use of the K-Hub ontology in the document search scenario

"replace" `Action`, the "battery" `Component` and the "X" `Product`); then, it uses the identified `Topic`s to match the most relevant `TopicAnnotation`s and, consequently, to propose the user with a specific `Document` and to help them opening it and navigating to the right page, to find the answer to the original request.

6.2 Ontology use to support procedure execution

The second usage scenario corresponds to the use cases UC2 and UC3, in which a shop floor operator wants to be guided in the correct execution of a procedure. This scenario could be addressed like the previous one, in case the user is simply given back a relevant document that contains the explanation of a procedure. However, the vision here is to support the operator by giving them *instructions* rather than *documents*. In this sense, the first step of knowledge extraction is key, because the goal is not only to generate `TopicAnnotation`s, but also to reconstruct the specific procedural knowledge (i.e. `Plan` and its `Step`s) and formalise it as structured knowledge, so to reuse it directly in user-supporting applications. As procedural knowledge is hard to formalise in a standard way [16], the knowledge extraction step would probably benefit from the manual annotation approach mentioned before.

The voice assistant application, implemented for the document retrieval scenario, could be exploited instead to provide the user with the exact instructions that they need. Thus, the user can first ask for the right procedure to follow, and then interactively ask the assistant to be supplied with the information required to perform the following step, with commands like "Give me the next step" which navigate the procedural knowledge graph by leveraging the `nextStep` property between `Step`s.

7 Conclusions and future work

In this paper, we presented the K-Hub ontology, a modular conceptual model able to capture the different aspects of manufacturing knowledge management and support industrial operators in their daily activities. In particular, the K-Hub ontology comprises a set of modules that identify and capture entities and relationships that are relevant for document retrieval and knowledge extraction. Our idea is to improve the coverage of document annotations and knowledge

extraction by means of a modular, and hence extensible, ontology "hub" which facilitates the reuse of the conceptual model.

The ontology is available under an open license and can be freely used, reused and further extended with the exception of the company-specific modules, which are intended to have limited access for the reasons explained before. The ontology was created and tested in the real business environments of two large manufacturing companies, Whirlpool and Marposs. A permanent URI and all the resources are completely available online and in GitHub and archived in Zenodo (with a corresponding DOI).

Ontology requirements were collected from the interviews conducted with domain experts and visits to the operating sites of the two companies. The development followed state-of-the-art practices in ontology development – the LOT methodology as well as the best practices to publish vocabularies on the Web – that we are applying in all our ontology development projects.

In terms of impact, therefore, we consider that this work and its results can fill an important gap that has not been addressed sufficiently in the state of the art. This would be as well a resource of interest for the Semantic Web community, in general, demonstrating how ontologies and semantic technologies can be used in an area where knowledge is contained in documents and extracting and representing it by combining different aspects could hence benefit from this type of approach.

We have not demonstrated yet any further reuse of our K-Hub ontology outside our own efforts, given that it has been only created recently. We expect, though, that there may be an interest in the broader context of digitalization in the manufacturing sector, as well as in other sectors with similar requirements. Besides, the way in which the ontology has been structured, together with the rich set of documentation provided for it, should facilitate such reuse and extensibility in the future, even for situations that have not been originally foreseen. For example, the manufacturing related ontologies surveyed in [6] could be reused to provide additional lists of relevant domain concepts to be considered as subclasses of `Topic`, to annotate industry documents; the same approach could be used outside the manufacturing context, by reusing only the annotation module and plugging-in other domain ontologies (biomedical, tourism, commerce, etc.).

Our future work consists of the further maintenance, extension and application of the K-Hub ontology and its employment in document annotation scenarios. In particular, we are interested in exploring its use in further automatic knowledge extraction efforts with machine/deep learning techniques, as well as in manual annotation experiments involving domain experts.

Acknowledgments

This work was partially supported by the K-HUB "Manufacturing Knowledge Hub" project, co-funded by EIT Manufacturing (project id 22330). The authors would like to specifically thank the industrial partners of the project for their invaluable support in both the requirement elicitation and the ontology validation activities.

References

1. Bach, N., Badaskar, S.: A survey on relation extraction. Language Technologies Institute, Carnegie Mellon University **178**, 15 (2007)
2. Bellan, P., Dragoni, M., Ghidini, C.: Process extraction from text: state of the art and challenges for the future. CoRR **abs/2110.03754** (2021)
3. Berrueta, D., Phipps, J.: Best Practice Recipes for Publishing RDF Vocabularies (2008), `https://www.w3.org/TR/swbp-vocab-pub/`
4. Butt, A.S., Fitch, P.: Provone+: A provenance model for scientific workflows. In: WISE. LNCS, vol. 12343, pp. 431–444. Springer (2020)
5. Corcho, Ó., Chaves-Fraga, D., Toledo, J., Arenas-Guerrero, J., Badenes-Olmedo, C., Wang, M., Peng, H., Burrett, N., Mora, J., Zhang, P.: A high-level ontology network for ICT infrastructures. In: The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24-28, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12922, pp. 446–462. Springer (2021)
6. Franc, Y.L.: OntoCommons D3.2 - Report on existing domain ontologies in identified domains (Mar 2022). https://doi.org/10.5281/zenodo.6504553
7. Gangemi, A., Peroni, S., Shotton, D.M., Vitali, F.: The publishing workflow ontology (PWO). Semantic Web **8**(5), 703–718 (2017)
8. Garijo, D.: Widoco: a wizard for documenting ontologies. In: International Semantic Web Conference. pp. 94–102. Springer, Cham (2017), `http://dgarijo.com/papers/widoco-iswc2017.pdf`
9. Garijo, D., Gil, Y.: Augmenting prov with plans in p-plan: scientific processes as linked data. CEUR Workshop Proceedings (2012)
10. Garijo, D., Gil, Y., Corcho, Ó.: Abstract, link, publish, exploit: An end to end framework for workflow sharing. Future Gener. Comput. Syst. **75**, 271–283 (2017)
11. Grau, B.C.: Privacy in ontology-based information systems: A pending matter. Semantic Web **1**(1-2), 137–141 (2010)
12. Jurafsky, D., Martin, J.H.: Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition. Prentice Hall, Pearson Education International (2009)
13. Kamble, S.S., Gunasekaran, A., Gawankar, S.A.: Sustainable industry 4.0 framework: A systematic literature review identifying the current trends and future perspectives. Process safety and environmental protection **117**, 408–425 (2018)
14. Kulvatunyou, B., Wallace, E., Kiritsis, D., Smith, B., Will, C.: The industrial ontologies foundry proof-of-concept project. In: Advances in Production Management Systems. Smart Manufacturing for Industry 4.0: IFIP WG 5.7 International Conference, APMS 2018, Seoul, Korea, August 26-30, 2018, Proceedings, Part II. pp. 402–409. Springer (2018)
15. Le Clair, A., Marinache, A., El Ghalayini, H., Maccaull, W., Khedri, R.: A review on ontology modularization techniques - a multi-dimensional perspective. IEEE Transactions on Knowledge and Data Engineering pp. 1–1 (2022)
16. Li, D., Landström, A., Fast-Berglund, Å., Almström, P.: Human-centred dissemination of data, information and knowledge in industry 4.0. Procedia CIRP **84**, 380–386 (2019)
17. Martínez-Rodríguez, J., Hogan, A., López-Arévalo, I.: Information extraction meets the semantic web: A survey. Semantic Web **11**(2), 255–335 (2020)
18. Miles, A., Bechhofer, S.: Skos simple knowledge organization system reference (2009)

19. Nakashole, N., Tylenda, T., Weikum, G.: Fine-grained semantic typing of emerging entities. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1488–1497 (2013)
20. Poveda-Villalón, M., Fernández-Izquierdo, A., Fernández-López, M., García-Castro, R.: Lot: An industrial oriented ontology engineering framework. Engineering Applications of Artificial Intelligence **111**, 104755 (2022)
21. Rula, A., Re Calegari, G., Azzini, A., Bucci, D., Baroni, I., Celino, I.: Eliciting and curating procedural knowledge in industry: Challenges and opportunities. In: Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022), Berlin, Germany, Sept. 19th-23rd, 2022. CEUR Workshop Proceedings, vol. 3234. CEUR-WS.org (2022), http://ceur-ws.org/Vol-3234/paper4.pdf
22. de Souza, H.C., Moura, A.M.d.C., Cavalcanti, M.C.: Integrating ontologies based on p2p mappings. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans **40**(5), 1071–1082 (2010). https://doi.org/10.1109/TSMCA.2010.2044880
23. Thakker, D., Patel, P., Intizar Ali, M., Shah, T., de Roode, M., Fernández-Izquierdo, A., Daniele, L., Poveda-Villalón, M., García-Castro, R., Thakker, D., Patel, P., Ali, M.I., Shah, T.: Saref4inma: A saref extension for the industry and manufacturing domain. Semant. Web **11**(6), 911–926 (jan 2020)
24. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S.: A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems **32**(1), 4–24 (2021)