Eliciting and Curating Procedural Knowledge in Industry: Challenges and Opportunities

Anisa Rula[2], Gloria Re Calegari[1], Antonia Azzini[1], Davide Bucci[1], Ilaria Baroni[1] and Irene Celino[1]

[2]*University of Brescia, Italy*
[1]*Cefriel - Politecnico di Milano, Italy*

Abstract

The capability of extracting useful information from documents and further transferring it into knowledge is essential for advancing technology innovations in industries. Procedures described within the service manuals provide guidelines as unstructured human-readable documents. Although annotating manuals with metadata makes them searchable, the *real* knowledge is still hidden in the procedural information which provides essential guidance for the operators. Therefore, there is a need to develop data curation techniques in order to build such procedural knowledge. However, creating this knowledge automatically can be hard to explicitly articulate as it refers to abilities and skills that may be hard to explain and describe. Still, manuals and other documentation often can include the description of procedures in terms of steps of a process or predefined plans. In this paper, we provide an overview of the state-of-the-art approaches based on manual and automatic annotations with or without human-in-the-loop involvement. We will discuss the challenges and the opportunities based on representative state-of-the-art work related to the annotation of documents as well as their semantic representation that may support knowledge curation of procedures in the manuals.

Keywords

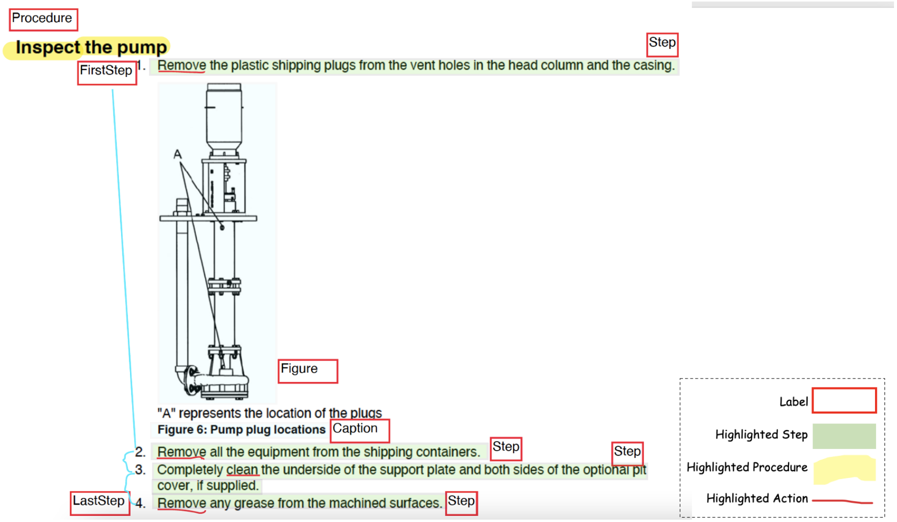
procedure annotation, data acquisition, information extraction, digitization, procedures annotations in service manuals, manual annotation

1. Introduction

The manufacturing industry is advanced by a technological revolution, often referred to as the Industry 4.0 [1], where the future trend lies in the convergence of several technologies including artificial intelligence, smart manufacturing, Internet of Things and web-based knowledge management. With specific reference to the latter, manufacturing companies face the challenge of managing, maintaining and transferring different kinds of knowledge between people and across company functions: product design, process definition, production lines, system maintenance, customer service, etc. Most of this knowledge remains implicit in the head of company employees; when it is made explicit, this knowledge is typically present in documents like user manuals, troubleshooting documents, guidelines, internal processes and so on. Those manuals should ensure optimum comprehensibility by the operator to safely and effectively install, operate, maintain and service the product.

Qurator 2022: 3rd Conference on Digital Curation Technologies, September 19-23, 2022, Berlin, Germany

anisa.rula@unibs.it (A. Rula); gloria.re@cefriel.it (G. Re Calegari); antonia.azzini@cefriel.it (A. Azzini); davide.bucci@cefriel.it (D. Bucci); ilaria.baroni@cefriel.it (I. Baroni); irene.celino@cefriel.it (I. Celino)



Inspect the pump

1. Remove the plastic shipping plugs from the vent holes in the head column and the casing.

"A" represents the location of the plugs

Figure 6: Pump plug locations

2. Remove all the equipment from the shipping containers.
3. Completely clean the underside of the support plate and both sides of the optional pit cover, if supplied.
4. Remove any grease from the machined surfaces.

Figure 1: Example of a step-by-step procedure annotated

A first tentative was to contextualise with metadata these manufacturing manuals and then apply traditional keyword-based search techniques of relevant documents [2, 3]. However, they still contain a lot of hidden value that is included in the description of procedural information which provides essential guidance for the operator. Imagine a virtual assistant which understands the instructions retrieved from those documents and suggests actions. The actions or steps to solve a problem are mentioned in the documents in the form of procedural knowledge, which by definition is the know-how needed to perform some tasks. Guided troubleshooting is an example of such a use case where the knowledge of problems and solutions in the form of step-by-step procedures can enable a systematic guidance to users. Therefore, there is a need to develop data curation techniques such as extraction, linking, annotation and enrichment in order to build this *procedural knowledge* that may be related both to production (e.g. how-to on the production line in the plant) and services (e.g. how-to for troubleshooting during maintenance). This knowledge should be searchable and more sophisticated techniques based on semantic search which combine text with knowledge bases will be required [4, 5, 6]. Figure 1 shows an example of a procedure where an annotation tool is used to highlight not only the steps of the procedures but also the actions performed by each step. This example shows only one of the many styles of how a procedure may be written where the steps are given as an enumerated list and where all the steps are next to each other on one page.

Methods to automatically extract or enhance the structure of various corpora has been a core topic in the context of the Semantic Web [7]. Such processes are often based on Information Extraction methods, which in turn are rooted in techniques from areas such as Natural Language Processing, Machine Learning and Information Retrieval. Semantic Web techniques can be applied to guide the Information Extraction process. The focus is on the extraction and/or linking of the schema elements from an (unstructured or semi-structured) input source that are: concepts and relations. In this work we discuss the problem of procedural knowledge curation, starting from its identification and extraction from documents at different granularity levels and its representation according to standard vocabularies. Most of the existing research has focused

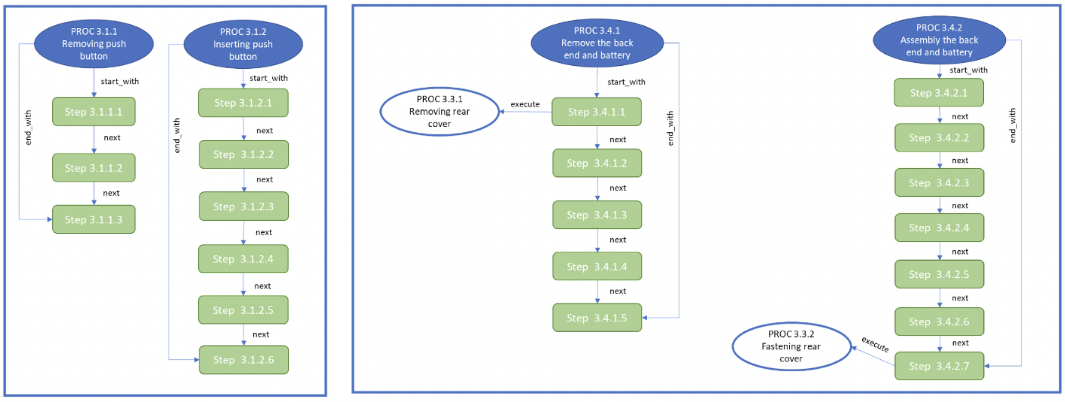


Figure 2: Atomic procedure (left) and composed procedure (right).

on annotating scientific documents but there are also others that focus on different domains such as legal domain, health information or industry 4.0. Different types of documents represent new challenges due to their different style of procedures are described. We are interested to study how the information extraction of procedures is introducing new problems w.r.t the classic problems of information extraction. There are several approaches that extract workflows either fully automatically or fully manually [8, 9, 10] but none discuss the challenges of the combined approach between partially manual annotations ingested in the fully automatically approaches.

The remainder of this paper is organised as follows. Section 2 introduces a definition of the Procedures and the problem of annotating them in the documents. Section 3 discusses different annotation approaches and vocabularies. Section 4 discusses challenges and the opportunities of the annotation of the procedures. Section 5 provides conclusions and discusses directions for future research.

2. Procedure Definition and Annotation

2.1. Procedure Definition

The concept of procedures expressed in natural language that we consider in the manuals can be defined as follows. A procedure is a sequence of steps. An *action* is an operation that triggers the execution of an activity which generates an output used by the next step. In addition to the work in [11] we distinguish two types of procedures:

- *An atomic procedure* is a procedure where each step performs an action at a time. For example, the procedure shown in Figure 2 in the example on the left illustrates a sequence of actionable steps `Step 3.1.1.1`, ...`Step 3.1.1.2` under procedure `PROC 3.1.1 Removing push button`. Each of these steps are performing actions in a coherent manner.
- *A composed procedure* is a procedure where each step performs an action or another procedure at a time to form a higher-level procedure. For example, the procedure shown

in Figure 2 in the example on the right illustrates that `Step 3.4.1.1` is needed to perform the procedure `PROC 3.3.1 Removing rear cover`. Hence, there can be nested procedures in the documents.

2.2. Annotation Procedure

The procedure identification problem consists in spotting the procedures in a document text and subsequently extracting them. This problem can be broken down into three main phases. In the following, we provide a description for each phase.

- *Phase 1.* A document is usually organized as chapters, sections, subsections at various nested levels, and then paragraphs, lists and nested lists. A document like a troubleshooting article may contain only a subset of those elements. To identify both atomic and composed procedures present in the document, a human must identify the hierarchical structure of the document. Typically, the document title is the root and each chapter title, section and subsequent subsections form hierarchical nodes. The content text, lists and paragraphs, form the leaf nodes of this hierarchy.
- *Phase 2.* After a user analyses and understands the structure of the document and the granularity level of the document to be annotated, he can decide to employ a manual or an automatic annotation tool considering the advantages and disadvantages of each tool as resumed in Section 4. These tools should provide a linguistic analysis in order to identify i) the steps of the procedure, ii) the actions of the steps, and iii) the connections between steps.
- *Phase 3.* The last step is related to the representation of the annotations based on standard vocabularies. The annotation represented as structured information can be extracted and linked with other structured representations. In this way, it is possible to share and query our procedural representation.

3. Annotation Approaches and Vocabularies

In this section we divide the state-of-the-art approaches according to i) manual and automatic approaches dealing with the document annotations and ii) vocabulary for the standardisation of the annotations.

3.1. An hybrid annotation approach

The annotation approaches of documents can be divided into manual and automatic approaches. Both techniques represent advantages and disadvantages therefore it is necessary to understand which tool to use and in which case. A list of criteria could help to select the right tool: r1) consider the structure of the document since most of the annotation approaches work on sequential documents (e.g., text documents); r2) allow the creation of relationships between annotations; r3) customizability of labels; r3) allow multi-label annotations; r4) support for ontologies and terminologies; r5) support for extracting and saving annotations r6) the tool should be open-source and r7) simple to set up and use.

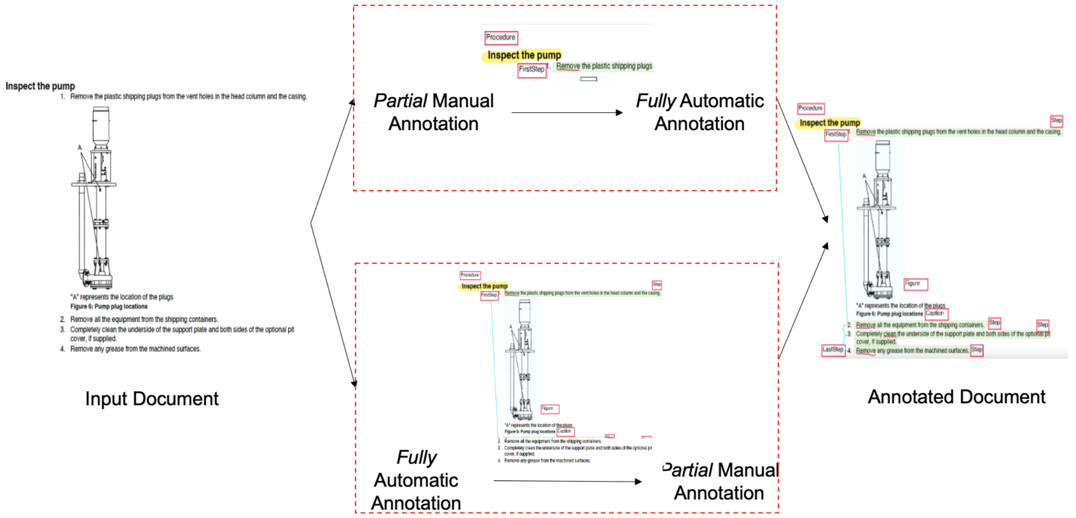

Input Document
Partial Manual Annotation
Fully Automatic Annotation
Fully Automatic Annotation
Partial Manual Annotation
Annotated Document
Inspect the pump
Procedure
FirstStep
Step
LastStep
Figure
Caption

Figure 3: Workflows of the hybrid approach.

To understand the differences two possible workflows can be envisioned as shown in figure 3 i) a fully automatic annotation followed by a manual annotation and validation, and ii) a partial manual annotation followed by a fully automatic annotation and validation. In both cases, a human-in-the-loop approach may be considered either in the manual annotation or the validation phase.

Automatic annotation approaches. The automatic annotation approaches rely on NLP processing techniques. However, most of the tools proposed are trained on scientific papers [2, 3], where the structure of the document considered is different from the structure of other types of documents such as technical manuals. On the other hand, there are automatic approaches very specific to the domain [12, 13] that are difficult to generalise to other use cases. An example of possible tools that satisfy these criteria is GROBID[1], an open source tool trained on scientific papers [2]. However, it is difficult to run this tool on other domains. Other NLP approaches are focused on process extraction from Text [14, 15, 16]. Although there can be similarities between a process [8] and a procedure extraction still they represent two different principles.

Manual annotation approaches. In these approaches, the human has a fundamental role since he has to read a particular pre-selected document and provides additional information in the form of annotations [17]. This is the task that we implicitly always do when we read instructions and manuals. As described in the procedure annotation approach, the first phase includes some activities, such as defining the annotation schema, determining the document structure and the pre-processing of the document according to the task. In the second phase, it is possible to select between the manual annotation tools surveyed in [17] that satisfy most of the criteria. An example of a tool that satisfies these criteria: PDFAnno[2] is an open-source tool for the annotation of PDF documents [18]. The documents can only be uploaded in PDF format, and annotations can be carried out for entities and relationships. It is easy to use and it is possible

[1] https://github.com/kermitt2/grobid
[2] http://github.com/paperai/pdfanno

to customise the tool. PDFAnno visualises the relationships between annotations. On the other hand, it is not possible to define predefined tags to be reused (e.g. step, plan) but they must be entered every time. Creating relationships between a step and a procedure cross-referenced by it if on different pages is difficult to be handled. The annotations generated are difficult to navigate ("flat" drop-down menu without references to pages); PAWLS[3] is particularly suited for mixed-mode annotation and scenarios in which annotators require extended context to annotate accurately [19]. It is possible to define predefined types of annotations (e.g. step, plan) and is easy to navigate between the annotations made. But on the other hand, relationships between annotations are not displayed. It is not possible to modify but only to delete the relations created. In case of multiple annotations on the same piece of text, the annotations overlap. After the manual annotation by a user, the validation process will run on top of accurate annotations. However, it is difficult to understand how to store the relations between annotations in composed procedures.

In the first workflow of the hybrid approach first, the partial manual annotation phase is run which includes the document structure analysis and training step preparation. In the second phase, the annotation algorithm includes the manual annotations in the training and automatically extracts the annotations of the whole document. In the second workflow of the hybrid approach first, the fully automatic annotation approach is executed. In the second phase, a manual annotation and validation phase will include human intervention. Differently to the first workflow where the manual validation step is proposed at the end after the two approaches are performed, in the second workflow, the validation is part of the manual annotation approach. In both cases, the human-in-the-loop paradigm can be exploited to improve the annotations. However, in some cases, it can be time-consuming to have the involvement of the user for both the manual annotation and the validation.

3.2. Vocabularies

There has been a general interest in providing workflow vocabularies to represent scientific experiments [20, 21, 22, 23]. The authors in [23] have explored the publication of workflows as Linked Data to illustrate how the workflow inputs and outputs can be linked to other resources in the Linked Data cloud. Other works focus on the provenance information of workflows in order to help scientists debug their experiments based on a standardised description [20]. A popular vocabulary for describing activities is provided by the Provenance Ontology PROV-O[4]. While PROV-O had a predicate to relate activities to a plan, it did not allow for plans to be described. This was provided by P-PLAN[5], which extended `prov:Plan` to be related to steps, which in turn correspond with activities. Both PROV-O and P-PLAN were then reused by OPMW[6], which allowed to create workflow templates, and instantiations thereof, of scientific processes (publishing an article, generating results, etc.). Not only is OPMW domain-specific, the models that one can create with this ontology focuses on workflows, much like BPMN. OPMW also extends the Open Provenance Model (OPM), a legacy provenance model developed

[3] https://github.com/allenai/pawls

[4] https://www.w3.org/TR/prov-o/

[5] http://purl.org/net/p-plan

[6] http://www.opmw.org/model/OPMW

by the workflow community that was used as a reference to create PROV. DataONE-OPM (D-OPM) is a provenance model that extends OPM. It can represent the workflow structure, traces from workflow executions, data structure, and workflow evolution. ProvONE extends the PROV model with an explicit representation of prospective provenance, thus capturing the most relevant information on scientific workflow processes. It is designed to accommodate extensions for specific scientific workflow systems [21]. ProvONE+[7] is a lightweight and general-purpose specification model for the control-flows in scientific workflows [21]. Another direction of vocabulary standardisation are the works focused on modelling business processes [16, 24, 25]. An example is BPMN 2.0 a state-of-the-art meta-model of business processes [24].

While there are a lot of works on the provenance of the documents and more recently of the workflows with well-known vocabularies there are no works so far on the annotation of procedures with a vocabulary. Furthermore, the current tools and approaches do not focus on annotation based on a standard vocabulary but they all rely heavily on manual curation by users.

4. Challenges and Opportunities

4.1. Challenges

Taking into consideration the annotation approaches and vocabularies mentioned previously, we identify the following challenges that encompass key-related issues:

i) *Heterogeneity of the document structure.* This issue refers to the different ways a procedure is written even within the same document since there is a lack of standardisation. The heterogeneity of the document structure is particularly challenging because of the high variability of the structure and the linguistic style used not only on the manuals of the different companies but also between the manuals of the same company or even within the same manual. Second, the procedures are not just a list of items or steps but the steps may by related through cross-references to other steps or procedures in the entire manual e.g., a non-linear description of procedures. This is more challenging for composed procedures rather than the atomic ones which are usually a set of items closed to each other. The problem with jumping back and forth is to identify the right cross-reference and to keep the relationships between steps within a document or even across documents. Finally, procedures and their steps may be described as a mixed representation with other elements such as images or tables. ii) *Unknown granularity.* This issue refers to the difficulty on deciding which is the appropriate level of granularity for the annotations and the vocabulary to be used. iii) *Limited background knowledge.* This issue refers to the absence of explicit background knowledge included in the document because given for granted or obvious (e.g. terminology or even common sense). iv) *Lack of training data.* This challenge refers to the preparation of the dataset for the training. In the automatic annotation approaches, the preparation of the training set requires a big effort for manually annotating the data. This training set needs to be large enough to train an automatic annotator.

[7] http://purl.org/provone

4.2. Opportunities

In the following we provide some opportunities in transforming procedures into procedural knowledge that are:

i) A better *curation* of this difficult-to-explain type of knowledge (structured description instead of unstructured); ii) *Reuse and understandability* of procedural knowledge as guidance or support tools (e.g. virtual assistants/intelligent agents), both to facilitate the performance of knowledgeable workers and to train new employees; iii) *Integration/linking* of procedural knowledge with other knowledge, by building a knowledge graph (e.g. tools to perform some action, products/systems on which actions are performed, other master data or historical data/logs of previous procedure execution, etc.) iv) Generation of further *supporting documentation* (e.g. FAQ, presentations, videos, tutorials, etc.) starting from the extracted procedures.

5. Conclusion and Future Work

This paper has presented the first discussion of the annotation of procedures in manuals. We discussed the different approaches based on manual and automatic approaches which should be able to annotate both atomic and composed procedures. We also provided a systematic review of the vocabularies used for workflows or business processes which can be adopted for procedures.

In future work, we aim to provide a combination of both manual and automatic annotation approaches in order to determine the best combination. Moreover, we aim to explore techniques to better identify the boundaries of procedures in the presence of other elements or formatting issues in the documents.

Acknowledgments

This work was partly supported by EIT Manufacturing with the project K-Hub.

References

[1] S. S. Kamble, A. Gunasekaran, S. A. Gawankar, Sustainable industry 4.0 framework: A systematic literature review identifying the current trends and future perspectives, Process safety and environmental protection 117 (2018) 408–425.

[2] P. Lopez, L. Romary, GROBID - information extraction from scientific publications, ERCIM News 2015 (2015).

[3] M. Y. Jaradeh, K. Singh, M. Stocker, A. Both, S. Auer, Better call the plumber: Orchestrating dynamic information extraction pipelines, in: ICWE, volume 12706 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 240–254.

[4] H. Bast, B. Buchhold, E. Haussmann, Semantic search on text and knowledge bases, Found. Trends Inf. Retr. 10 (2016) 119–271.

[5] L. Tamine, L. Goeuriot, Semantic information retrieval on medical texts: Research challenges, survey, and open issues, ACM Computing Surveys (CSUR) 54 (2021) 1–38.

[6] F. Lashkari, F. Ensan, E. Bagheri, A. A. Ghorbani, Efficient indexing for semantic search, Expert Syst. Appl. 73 (2017) 92–114.

[7] J. Martínez-Rodríguez, A. Hogan, I. López-Arévalo, Information extraction meets the semantic web: A survey, Semantic Web 11 (2020) 255–335.

[8] P. Bellan, M. Dragoni, C. Ghidini, Process extraction from text: state of the art and challenges for the future, CoRR abs/2110.03754 (2021).

[9] S. Zhou, L. Zhang, Y. Yang, Q. Lyu, P. Yin, C. Callison-Burch, G. Neubig, Show me more details: Discovering hierarchies of procedures from semi-structured web data, ACL, 2022.

[10] Z. Zhang, P. Webster, V. Uren, A. Varga, F. Ciravegna, Automatically extracting procedural knowledge from instructional texts using natural language processing, ELRA, 2012, pp. 520–527.

[11] S. Agarwal, S. Atreja, V. Agarwal, Extracting procedural knowledge from technical documents, 2020. doi:10.48550/ARXIV.2010.10156.

[12] S. Mysore, Z. Jensen, E. Kim, K. Huang, H. Chang, E. Strubell, J. Flanigan, A. McCallum, E. Olivetti, The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures, in: LAW@ACL 2019, ACL, 2019, pp. 56–64.

[13] Z. Dong, S. Paul, K. Tassenberg, G. Melton, H. Dong, Transformation from human-readable documents and archives in arc welding domain to machine-interpretable data, Comput. Ind. 128 (2021) 103439.

[14] F. Friedrich, J. Mendling, F. Puhlmann, Process model generation from natural language text, volume 6741, Springer, 2011, pp. 482–496.

[15] A. Rebmann, H. van der Aa, Extracting semantic process information from the natural language in event logs, in: CAiSE, volume 12751, Springer, 2021, pp. 57–74.

[16] P. Bertoli, F. Corcoglioniti, C. D. Francescomarino, M. Dragoni, C. Ghidini, M. Pistore, Semantic modeling and analysis of complex data-aware processes and their executions, Expert Syst. Appl. 198 (2022) 116702.

[17] M. Neves, J. Seva, An extensive review of tools for manual annotation of documents, Briefings Bioinform. 22 (2021) 146–163.

[18] H. Shindo, Y. Munesada, Y. Matsumoto, PDFAnno: a web-based linguistic annotation tool for PDF documents, ELRA, 2018.

[19] M. Neumann, Z. Shen, S. Skjonsberg, PAWLS: PDF annotation with labels and structure, in: ACL 2021, 2021, pp. 258–264.

[20] W. M. de Oliveira, D. de Oliveira, V. Braganholo, Provenance analytics for workflow-based computational experiments: A survey, ACM Comput. Surv. 51 (2018) 53:1–53:25.

[21] A. S. Butt, P. Fitch, Provone+: A provenance model for scientific workflows, in: WISE, volume 12343 of *LNCS*, Springer, 2020, pp. 431–444.

[22] A. Gangemi, S. Peroni, D. M. Shotton, F. Vitali, The publishing workflow ontology (PWO), Semantic Web 8 (2017) 703–718.

[23] D. Garijo, Y. Gil, Ó. Corcho, Abstract, link, publish, exploit: An end to end framework for workflow sharing, Future Gener. Comput. Syst. 75 (2017) 271–283.

[24] A. Annane, M. Kamel, N. Aussenac-Gilles, Comparing business process ontologies for task monitoring, SCITEPRESS, 2020, pp. 634–643.

[25] R. Singer, An ontological analysis of business process modeling and execution, CoRR abs/1905.00499 (2019). URL: http://arxiv.org/abs/1905.00499. arXiv:1905.00499.