Documentation Gap in Ontology Creation: Insights into the Reality of Knowledge Formalization in a Life Science Company

Marius Michaelis and Olga Streibel

Bayer Business Services GmbH, 51368 Leverkusen, Germany
{marius.michaelis,olga.streibel}@bayer.com

Abstract. To achieve the goal of FAIR — findable, accessible, interoperable, and reusable — data, life science companies employ Semantic Web standards and Linked Data principles. In doing so, they create ontologies that formally represent knowledge. This paper presents the results of a survey among knowledge engineers and domain experts involved in ontology creation for a global life science company. The survey results indicate that the conceptualization phase of the ontology creation process, including knowledge acquisition, remains largely undocumented. The majority of knowledge engineers surveyed begin to document during or after the creation of the formal knowledge model. The authors discuss the risks that may arise from this documentation gap and recommend addressing them by means of joint, timely, and structured documentation.

Keywords: Ontology · Documentation · Knowledge Management

1 Introduction

For over 10 years there have been initiatives to apply semantic technologies in the field of life sciences. In 2008, for instance, the W3C interest group *Semantic Web Health Care and Life Sciences* was founded, which has continued its work as a community group since 2018 [23]. Another example is the international *Semantic Web Applications and Tools for Healthcare and Life Sciences* (SWAT4HCLS) conference which has been taking place annually since 2008 [18]. As the shift towards the use of semantic technologies is becoming more common, the international standardization organization *Health Level Seven International* (HL7) has published a *Linked Data Module* for its standard framework *Fast Healthcare Interoperability Resource* (FHIR) [10]. In addition, pharmaceutical companies and authorities such as the *U.S. Food and Drug Administration* (FDA) are involved in non-profit organizations like *Pharmaceutical Users Software Exchange* (PhUSE). There, the working group *Linked Data & Graph Databases* worked on the use of semantic technologies [11]. At EU level, the intergovernmental organization ELIXIR, which is engaged in the *European Open Science Cloud* (EOSC), encourages semantic integration with its Interoperability Platform to achieve the goal of FAIR life science data [4]. FAIR refers to a set of four principles: data must be *findable*, *accessible*, *interoperable*, and *reusable* [24]. The FAIR strategy is mainly driven by the GO FAIR initiative. Both EOSC and GO FAIR follow the recommendations of the *European Commission expert group on FAIR data* [2,7], which recommends, among others, the use of semantic technologies [3]. In non-for-profit collaborations such as the *Pistoia Alliance*, companies, vendors, publishers, and academic groups are jointly dedicated to the implementation of FAIR data principles in biopharmaceutical R&D [25]. In order to achieve the goal of FAIR data, life science companies employ Semantic Web standards and Linked Data principles. In doing so, they create ontologies that formally represent knowledge. This paper provides insights into the reality of ontology creation in a life science company, focusing on the documentation that takes place during the process.

First, the process of ontology creation and the roles involved are briefly outlined in section 2. Following the description of the applied methodology in section 3, the findings are presented in section 4. They provide information about the company's ontology creators and their approach to documentation. Based on these findings, section 5 evaluates whether the prevailing documentation approach is sufficient. Finally, in section 6, we draw a conclusion on the challenges life science companies face when creating ontologies.

2 Ontology Creation Process

Based on the definitions by Gruber [8] and Borst [1], Studer et al. define ontologies as "a formal, explicit specification of a shared conceptualisation" [20]. There is no single, uniform approach to the structured development of ontologies. Instead, over the last two decades, a variety of so-called *ontology engineering methodologies* have been proposed in literature that describe more or less specific processes for ontology creation. The roles involved may differ from methodology to methodology in terms of their quantity, designations and responsibilities [6,19]. Following the understanding of roles in the life science company, this paper distinguishes between only two roles similar to the ones of the *Unified Process for Ontology Building*: knowledge engineer and domain expert [17]. *Domain experts* (DEs) have expertise in a certain subject area, i.e. DEs are familiar with the main concepts of a domain, their characteristics and relationships. In terms of ontology development, this means DEs are knowledgeable in the domain which is to be represented by the ontology. *Knowledge engineers* (KEs) capture, structure and formalize knowledge so that it can be processed by machines in order to solve certain problems. In terms of ontology development, the KEs are those who build the ontology. In the following, a basic ontology creation process is outlined, as it underlies many methodologies (see figure 1). For the sake of clarity, neither feedback loops nor special cases are discussed.

1. *Ontology specification*: Collection of requirements and definition of framework conditions [5,17,21]. Usually includes collecting so-called *competency questions* (CQs), i.e. questions to be answered by exploring and querying the ontology. CQs are initially expressed informally at the conceptual level, not as formal queries [9,17].

2. *Conceptualization*: Acquisition of knowledge, during which KEs gather the required domain knowledge from non-human as well as human knowledge sources. To do so, they research *explicit* knowledge stored in media outside the human brain (e.g. in the form of databases, documents, vocabularies) on the one hand, and elicit *tacit* knowledge, which is bound to individuals (e.g. practical knowledge in the memory of a long-time employee), on the other hand, by interviewing, observing, and probing DEs [16]. The collected knowledge is conceptually analyzed by the KEs in order to create an informal knowledge model. [5,17,22]
3. *Implementation*: KEs encode the informal knowledge model as an ontology using a formal ontology language. [5,17,22]
4. *Test*: KEs and DEs evaluate the ontology's quality in different dimensions [14]. Basically, the ontology must meet technical standards and the defined requirements so that it can be used to answer the collected CQs. [5,17,22]

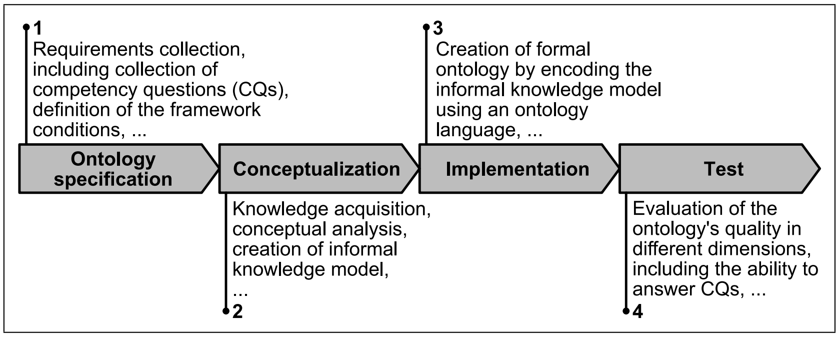

1
Requirements collection, including collection of competency questions (CQs), definition of the framework conditions, ...
3
Creation of formal ontology by encoding the informal knowledge model using an ontology language, ...
Ontology specification
Conceptualization
Implementation
Test
Knowledge acquisition, conceptual analysis, creation of informal knowledge model, ...
2
Evaluation of the ontology's quality in different dimensions, including the ability to answer CQs, ...
4

Fig. 1. Outline of a basic process for ontology creation without loops

3 Methodology

A survey based on two different questionnaires was conducted, one addressed to KEs, the other to DEs. The questionnaires consisted of five questions each. For the purpose of this paper, only 5 of the 10 questions are presented. KEs were asked when and how they document their work on ontologies and what information they consider relevant while creating knowledge models. Besides, they were asked whether fast or resilient results constitute their main goal. DEs were asked what they expect from ontologies. In both cases, only the current situation was enquired, not the desired ideal state. Therefore, the survey results do not necessarily reflect an optimal situation. In other words, just because the KEs work quickly and document barely, this does not mean they consider this to be the best solution. It may be an effect of economic constraints, not a reasonable decision from a professional perspective.

In total, three groups were surveyed: (1) KEs and (2) DEs of a global life science company based in Germany as well as (3) DEs of an international working group, which are referred to as external DEs. The two questionnaires have been designed to be completed quickly and are therefore relatively simple. They have been sent electronically to people known as KE or DE. The response rates were 92.9 % for KEs (13 out of 14), 78.6 % for internal DEs (11 out of 14) and 11.6 % for external DEs (5 out of 43). The participation in the survey was voluntary. As the sample sizes were small for both roles, the survey results do not claim to represent the entirety of the KEs and DEs in the company or the external working group. Nevertheless, they provide valuable insights into corporate reality.

4 Findings

4.1 Relevant Information per Concept

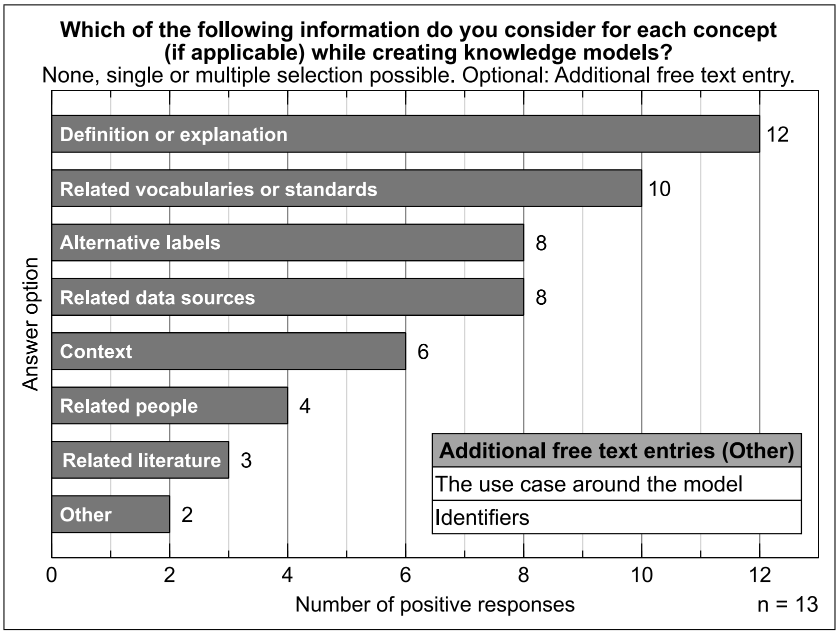
Figure 2 shows what information the KEs surveyed consider relevant for each concept. Each of the answer options offered was represented in the results, complemented by two free text entries added by the respondents. The most frequent choice was *definition or explanation*, which is obvious, since the meaning of concepts must be grasped in order to create ontologies. This is further supported by the results for the answer option *context*. After all, information on context is needed to situate a concept in a semantic network. However, the results also show that not only information directly related to the concept's meaning are considered relevant. KEs take into account related knowledge sources such as *related vocabularies and standards*, *related data sources*, *related people*, and *related literature* as well.

4.2 Expectations towards Ontologies

Figure 3 shows what DEs expect from ontologies for their daily work. Almost all of the DEs surveyed expected *ontologies to work in the background to improve the interaction between IT systems*. More than half of the DEs surveyed, 9 out of 16, expected *to be able to work directly with ontologies to gain knowledge about a domain*, complemented, among other free text entries, by the response "I expect to receive information or explanation of data, which is currently not available".

4.3 Main Goal

Concerning the main goal pursued by KEs, 69.2 % of the respondents aimed for fast results (see figure 4). In return, they accepted less perfect knowledge models.



Additional free text entries (Other)
The use case around the model
Identifiers

Fig. 2. Survey Results KEs: Relevant Information per Concept

4.4 Timepoint of Documentation

Figure 5 shows when KEs start documenting their work on a knowledge model. Only two of the respondents started the documentation *before* creating the formal model. The majority documented *while* or *after* creation of the formal model. In other words, the documentation usually took place after the conceptualization phase and thus after the exchange of knowledge between KEs and DEs (cf. figure 1). One of the KEs surveyed did not create any documentation at all.

4.5 Nature of Documentation

Figure 6 shows how KEs document the exchange with DEs, which takes place primarily in the course of knowledge acquisition. Although most KEs started their documentation in connection with the formal model, only 2 out of 13 KEs documented in a formal way as is possible by using annotation properties. The other 11 KEs documented the insights they acquire by exchanging with DEs informally, i.e. by using natural language. In doing so, the narrow majority of 6 KEs documented unstructured, while the remaining 5 KEs documented in a structured way, for instance by using templates. According to the results for this question, all KEs documented the exchange with DEs. This is not fully coherent with the results regarding the timepoint of documentation where the option “I don’t create a documentation” was selected once.

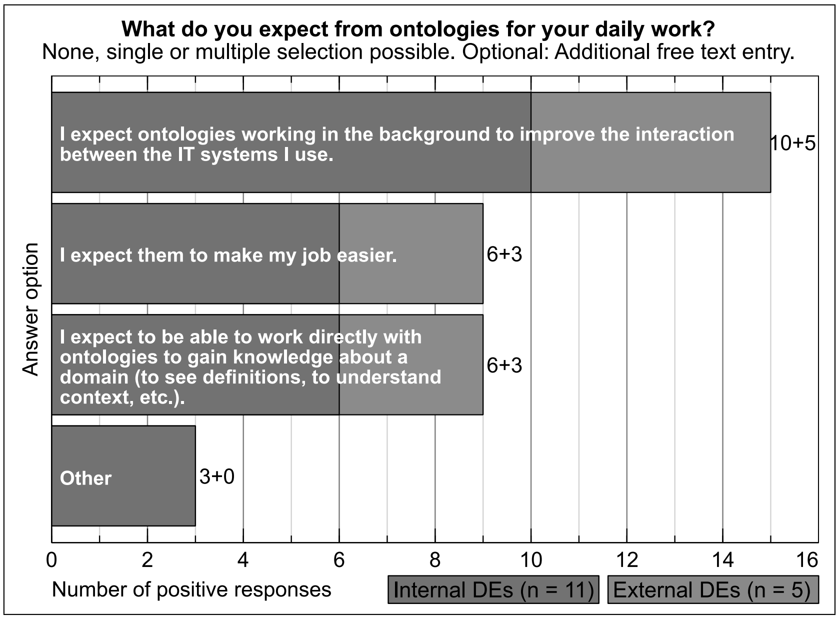
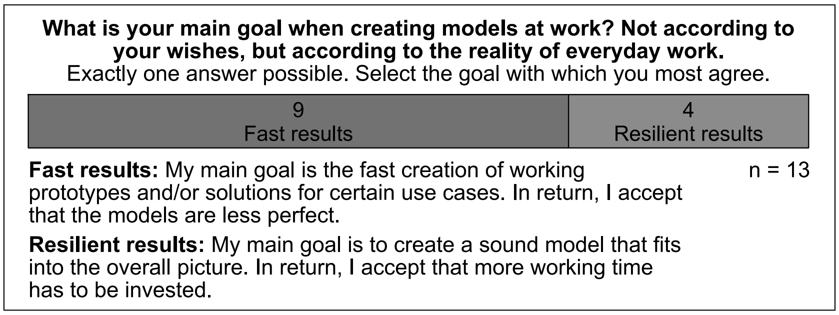


Fig. 3. Survey Results DEs: Expectations towards Ontologies

5 Discussion

According to the survey results, KEs consider information on the meaning of concepts and the associated knowledge resources to be relevant in the course of ontology creation. However, most KEs only begin to document during or after implementation. This means that the conceptualization phase of the ontology creation process remains largely undocumented. This poses a serious problem because in this very phase the knowledge considered relevant is acquired. If the laboriously researched and elicited knowledge is not explicitly recorded, it remains as tacit knowledge in the mind of the respective KE and is therefore difficult to access. As a consequence of this documentation gap, collaboration is impeded and it is more complicated to distribute workload. In addition, there is a risk of knowledge loss through individual and collective oblivion. Hence, **timely documentation** is essential.

If the documentation gap causes knowledge to be lost, this not only complicates the work of the KEs, but also jeopardizes that the DEs' expectations towards ontologies are met. After all, they expect to be able to work directly with ontologies to gain knowledge about a domain. Apart from preserving knowledge, **joint documentation** may also allow to identify synergies and potential misunderstandings at an early stage. Moreover, a shared documentation is a way to put definitions of terms up for discussion early enough. Thus, consensual knowledge as required for the creation of ontologies can already be gathered during knowledge acquisition. Without shared documentation, definitions are initially

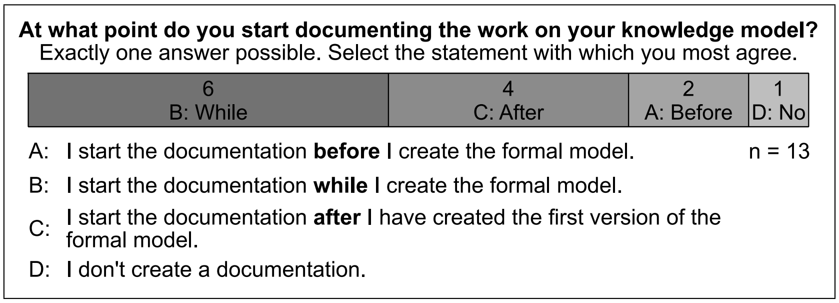


What is your main goal when creating models at work? Not according to your wishes, but according to the reality of everyday work.
Exactly one answer possible. Select the goal with which you most agree.

Fast results: My main goal is the fast creation of working prototypes and/or solutions for certain use cases. In return, I accept that the models are less perfect.

Resilient results: My main goal is to create a sound model that fits into the overall picture. In return, I accept that more working time has to be invested.

Fig. 4. Survey Results KEs: Main Goal



At what point do you start documenting the work on your knowledge model?
Exactly one answer possible. Select the statement with which you most agree.

A: I start the documentation **before** I create the formal model.

B: I start the documentation **while** I create the formal model.

C: I start the documentation **after** I have created the first version of the formal model.

D: I don't create a documentation.

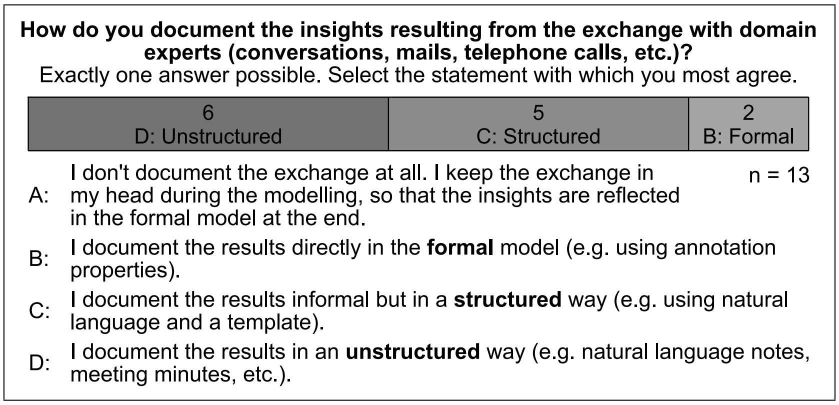
Fig. 5. Survey Results KEs: Timepoint of Documentation

hidden in the personal notes or mind of a KE, which means that consensus building may only begin after the publication of the formal knowledge model.

A possible explanation for the identified documentation gap may be the fact that the majority of KEs in the life science company under investigation strive for fast results, probably at the expense of timely documentation.

With regard to the nature of the documentation, the **structured documentation** approach is recommended, as already adopted by some of the KEs surveyed. *Structured documentation* or *semi-formal documentation* is written in natural language and follows guidelines provided, for instance, by templates. Hence, the documentation is clear and understandable for both KEs and DEs. *Unstructured documentation*, also called *informal documentation*, by contrast, is individual and does not follow guidelines, making it ambiguous and heterogeneous. Creating *formal documentation*, which is machine-readable, requires more effort and specific skills that not all DEs have. Consequently, a joint documentation should neither be formal nor unstructured, but structured and thus easy to handle for all people involved. [12,13]

To illustrate the described consequences of the documentation gap, two fictive scenarios are given below. They are based on personal experiences gained by the authors while working as KEs for the life science company under investigation. In Scenario 1, the KEs document too late and insufficiently which, in



How do you document the insights resulting from the exchange with domain experts (conversations, mails, telephone calls, etc.)?
Exactly one answer possible. Select the statement with which you most agree.

A: I don't document the exchange at all. I keep the exchange in my head during the modelling, so that the insights are reflected in the formal model at the end.

B: I document the results directly in the **formal** model (e.g. using annotation properties).

C: I document the results informal but in a **structured** way (e.g. using natural language and a template).

D: I document the results in an **unstructured** way (e.g. natural language notes, meeting minutes, etc.).

Fig. 6. Survey results KEs: Nature of Documentation

our experience, constitutes the prevailing situation in the company. Scenario 2 represents the desired situation in which the challenges that life science companies face when creating ontologies are addressed by means of joint, timely, and structured documentation.

Scenario 1 Ina[1] does not document acquired knowledge in a timely manner, so that she has forgotten some information by the time of implementation (knowledge loss through individual oblivion). Unfortunately, the knowledgeable colleague is no longer available due to retirement (knowledge loss through organizational oblivion). Until she can ask her KE colleague Cora[1], who hasn't created any documentation either, she has to wait for her to return from vacation (impeded collaboration). If the DE Conan[1] wants to make a definition proposal regarding a concept, he must first write an e-mail to Ina, as there is no structured documentation available in which he can enter information directly (impeded distribution of workload). Ina does not forward Conan's proposal to the other DEs, which is why their disagreement with his definition becomes apparent only after publication of the formal model (delayed consensus building).

Scenario 2 Ina[1], who works as a KE, externalizes knowledge acquired during the conceptualization phase promptly in form of a structured documentation, which can be edited remotely by her colleagues. This allows her KE colleague Cora[1] to see which concepts are already described (collaboration). In addition, the DE Conan[1] is able to add new definitions directly to the documentation without having to contact Ina (distribution of workload). Following this, other DEs can review Conan's definition proposal and initiate a discussion if necessary (consensus building). If Ina forgets something or leaves the company, the documentation can be consulted (knowledge preservation).

[1] The names of the personas are fictitious.

6 Conclusion

In accordance with our personal experience as KEs, the presented survey results suggest that there is a documentation gap between knowledge acquisition and knowledge formalization in the process of ontology creation. Among the surveyed KEs, ontologies are created in various projects for various domains and divisions by international teams consisting of internal and external employees. At the same time, collaborations with external working groups take place. As a result, the challenge is to share knowledge acquired for ontology creation as early as possible in the process. We recommend addressing this challenge by means of structured documentation, which is created jointly and in a timely manner by the KEs and DEs involved. This reduces the risk of knowledge loss while enabling collaboration and distribution of workload. A solution developed for this purpose is the documentation concept proposed by MICHAELIS [15], which enables the company to overcome the documentation gap by providing guidelines in the form of graphical templates on what should be documented by whom, how and when. Further research is needed to determine whether the presented documentation gap constitutes a phenomenon that is specific to the surveyed KEs or represents a general pattern in the life science industry.

7 Acknowledgements

This paper is based on a bachelor thesis cooperation between *Bayer* and the *University of Applied Sciences Potsdam* under supervision of Günther Neher.

References

1. Borst, W.N.: Construction of engineering ontologies for knowledge sharing and reuse. Phd thesis, University of Twente, Enschede, NL (1997)
2. Directorate-General for Research and Innovation: Prompting an eosc in practice: Final report and recommendations of the commission 2nd high level expert group on the european open science cloud (eosc) (2018). https://doi.org/10.2777/112658
3. Directorate-General for Research and Innovation: Turning fair data into reality: Final report and action plan from the european commission expert group on fair data (2018). https://doi.org/10.2777/1524
4. ELIXIR: Interoperability platform (2019), https://elixir-europe.org/platforms/interoperability
5. Fernández, M., Gómez-Pérez, A., Juristo, N.: Methontology: From ontological art towards ontological engineering. In: Farquhar, A. (ed.) Ontological engineering, pp. 33–40. Technical report / American Association for Artificial Intelligence SS, AAAI Press, Menlo Park, Calif. (1997)
6. Fernández-López, M., Gómez-Pérez, A.: Overview and analysis of methodologies for building ontologies. The Knowledge Engineering Review **17**(02) (2002). https://doi.org/10.1017/S0269888902000462
7. GO FAIR: Strategy (2018), https://www.go-fair.org/go-fair-initiative/strategy/
8. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition **5**(2), 199–220 (1993). https://doi.org/10.1006/knac.1993.1008

9. Grüninger, M., Fox, M.S.: Methodology for the design and evaluation of ontologies. Workshop on Basic Ontological Issues in Knowledge Sharing: International Joint Conference on Artificial Intelligence (1995)
10. HL7.org: Fhir release 3 (stu): Fhir linked data module (2019), http://hl7.org/fhir/STU3/linked-data-module.html
11. Kent Innovation Centre: Phuse working groups: Linked data & graph databases (2017), https://www.phuse.eu/linked-data-graph-databases
12. Landes, D., Schneider, K., Houdek, F.: Organizational learning and experience documentation in industrial software projects. International Journal of Human-Computer Studies **51**(3), 643–661 (1999). https://doi.org/10.1006/ijhc.1999.0280
13. Lehmann, A.: A documentation approach for higher education. In: Proceedings of 2018 IEEE Global Engineering Education Conference (EDUCON). pp. 43–50. IEEE, Piscataway, NJ (2018). https://doi.org/10.1109/EDUCON.2018.8363207
14. Lourdusamy, R., John, A.: A review on metrics for ontology evaluation. In: Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018). pp. 1415–1421. IEEE (2018). https://doi.org/10.1109/ICISC.2018.8399041
15. Michaelis, M.: Documentation concept for the exchange of knowledge in the process of creating ontological knowledge models. Bachelor thesis, University of Applied Sciences Potsdam (2019), https://nbn-resolving.org/urn:nbn:de:kobv:525-23611
16. Milton, N.R.: Knowledge acquisition in practice: A step-by-step guide. Decision Engineering, Springer, London and Berlin and Heidelberg (2007). https://doi.org/10.1007/978-1-84628-861-6
17. de Nicola, A., Missikoff, M., Navigli, R.: A software engineering approach to ontology building. Information Systems **34**(2), 258–275 (2009). https://doi.org/10.1016/j.is.2008.07.002
18. Semantic Web Applications and Tools for Healthcare and Life Sciences: About (2019), http://www.swat4ls.org/about/
19. Simperl, E., Luczak-Rösch, M.: Collaborative ontology engineering: a survey. The Knowledge Engineering Review **29**(01), 101–131 (2014). https://doi.org/10.1017/S0269888913000192
20. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: Principles and methods. Data & Knowledge Engineering **25**(1-2), 161–197 (1998). https://doi.org/10.1016/S0169-023X(97)00056-6
21. Suárez-Figueroa, M.C., Gómez-Pérez, A.: Ontology requirements specification. In: Suárez-Figueroa, M.C., Gómez-Pérez, A., Motta, E., Gangemi, A. (eds.) Ontology Engineering in a Networked World, pp. 93–106. Springer Berlin Heidelberg, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-24794-1_5
22. Suárez-Figueroa, M.C., Gómez-Pérez, A., Fernández-López, M.: The neon methodology framework: A scenario-based methodology for ontology development. Applied Ontology **10**(2), 107–145 (2015). https://doi.org/10.3233/AO-150145
23. W3C: Semantic web in health care and life sciences community group (2019), https://www.w3.org/community/hclscg/
24. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3** (2016). https://doi.org/10.1038/sdata.2016.18
25. Wise, J., de Barron, A.G., Splendiani, A., et al.: Implementation and relevance of fair data principles in biopharmaceutical r&d. Drug Discovery Today (2019). https://doi.org/10.1016/j.drudis.2019.01.008