Documentation concept for the exchange of knowledge in the process of creating ontological knowledge models

Bachelor Thesis by MARIUS MICHAELIS

Student name:	Marius Michaelis https://orcid.org/0000-0002-6437-7152
Course of studies:	Information and Documentation, Bachelor of Arts
University:	University of Applied Sciences Potsdam, Faculty of Information Sciences
Date of submission:	2019-02-18
Date of defense:	2019-03-18
1st Supervisor:	Prof. Dr. Günther Neher, University of Applied Sciences Potsdam
2nd Supervisor:	Dr. Olga Streibel, Bayer Business Services GmbH
Document URN:	urn:nbn:de:kobv:525-23611

This is a revised version of the bachelor thesis as of 2019-05-15. It includes the following minor changes to the initial version, which was submitted on 2019-02-18:

- Fixes in spelling, grammar and punctuation
- Layout optimization for reading on screen (previously optimized for print)
- Addition of acknowledgements
- Addition of hyperlinks, especially for the DOIs in the bibliography
- Addition of index
- Addition of Creative Commons license
- Addition of author's ORCID iD to the title page
- Addition of document URN to the title page

Acknowledgements

Special thanks are dedicated to all the committed people from industry and academia
who have devoted both their working hours and their free time
to contribute to the realization of this bachelor thesis.

Abstract

This bachelor thesis proposes a documentation concept to address the problem of missing and insufficient documentation during the creation of ontologies in companies. By presenting survey results and assumption personas, the thesis provides insights into the reality of ontology creation in a life science company. These insights indicate that there is a lack of knowledge externalization between the acquisition and formalization of knowledge. To develop an appropriate documentation concept, findings are derived by analyzing the practical insights as well as two ontology engineering methodologies proposed in literature. The outcome is a modular, highly customizable documentation concept designed for the application in companies. Using graphical documentation templates, guidelines are provided on what should be documented by whom, how and when.

Kurzreferat

Diese Bachelorarbeit stellt ein Dokumentationskonzept vor, welches das Problem der fehlenden und unzureichenden Dokumentation bei der Erstellung von Ontologien in Unternehmen adressiert. Anhand von Umfrageergebnissen und Assumption Personas werden Einblicke in die Realität der Ontologieerstellung in einem Life-Science-Unternehmen gegeben. Diese Einblicke weisen auf einen Mangel an Wissensexternalisierung zwischen dem Erwerb und der Formalisierung von Wissen hin. Die praktischen Einblicke sowie zwei in der Fachliteratur vorgestellte Methodologien des Ontology Engineerings werden analysiert, um Erkenntnisse für die Entwicklung eines geeigneten Dokumentationskonzeptes abzuleiten. Das Ergebnis ist ein modulares, hochgradig anpassbares Dokumentationskonzept, das für den Einsatz in Unternehmen konzipiert ist. Anhand von grafischen Dokumentations-vorlagen werden Richtlinien gegeben, was von wem, wie und wann dokumentiert werden sollte.

Table of Contents

Abstract ..... **i**

Table of Contents ..... **ii**

List of Figures ..... **v**

List of Tables ..... **xii**

Abbreviations ..... **xiv**

1 Introduction **1**

1.1 Objectives 4

1.2 Approach 7

2 Theoretical background **8**

2.1 Documentation concept 8

2.2 Knowledge 12

2.2.1 Data, Information, Knowledge 12

2.2.2 Characteristics of knowledge 16

2.3 Ontology 21

2.3.1 Ontologies as controlled vocabulary 22

2.3.2 Structure and components 23

2.3.3 Ontology languages 24

2.3.4 Typology 26

2.3.5 Ontology engineering methodologies 28

2.4 Process 31

3 Selective insights into corporate reality **34**

3.1 Preliminary survey 34

3.1.1 Survey results: Knowledge engineers 35

3.1.2 Survey results: Domain experts 42

3.2 Archetypal knowledge engineers and domain experts 54

3.2.1 Persona method 54

3.2.2 Creation of assumption personas 55

3.2.3 Assumption personas 57

3.3 Findings 69

4 Analysis of ontology engineering methodologies **76**

4.1 Unified Process for Ontology Building UPON 79

4.1.1 Overview 79
4.1.2 Workflows 81
4.1.3 Complete process 83
4.1.4 Findings 85
4.2 NeOn Methodology 91
4.2.1 Overview 91
4.2.2 Findings 95
5 Documentation concept 99
5.1 Documentary reference units 99
5.2 Structure of the documentation concept 102
5.3 Structure of documentation modules 108
5.4 Mandatory module: Primary documentary reference unit 111
5.5 General modules and module extensions 112
5.5.1 Related people 112
5.5.2 Knowledge resources 115
5.5.3 Decisions 117
5.5.4 Tasks 120
5.5.5 Notes 122
5.5.6 Discussion 124
5.5.7 Agreement 127
5.5.8 Source 128
5.5.9 Image 129
5.6 Dimension-specific modules: Domain knowledge 130
5.6.1 Alternative labels 130
5.6.2 Foreign identifiers 132
5.6.3 Domain 133
5.6.4 Definitions 134
5.6.5 Explanations 137
5.6.6 Domain knowledge resources 139
5.6.7 Concept category 143
5.6.8 Relationships to other concepts 144
5.7 Dimension-specific modules: Engineering knowledge 149

5.7.1 Formal representations 149
5.7.2 Engineering knowledge resources 151
5.7.3 Implementation status 154
5.7.4 Test status regarding reasoning 155
5.8 Dimension-specific modules: Management knowledge 156
5.8.1 Related use cases 156
5.8.2 Related competency questions 157
5.8.3 Affiliation 159
5.8.4 Test status regarding relevance 161
5.8.5 Priority 162
6 Discussion 163
7 Conclusion 166
8 Outlook 167
Appendices 169
Appendix A: Ontology engineering methodologies 169
Appendix B: Questionnaires and data of preliminary survey 173
Questionnaire for knowledge engineers 173
Questionnaire for domain experts 175
Survey data of knowledge engineers 177
Survey data of internal domain experts 182
Survey data of external domain experts 187
Appendix C: UPON workflows 192
Requirements workflow 192
Analysis workflow 193
Design workflow 196
Implementation workflow 198
Test workflow 198
Appendix D: Alternative documentary reference units 200
References 201
Index 211
Declaration of Authorship 213

List of Figures

Figure 1: Context of the different documentation terms according to WERSIG (Wersig, 1971, p. 143) 9
Figure 2: Excerpt of the knowledge ladder proposed by NORTH and KUMTA (Cf. North and Kumta, 2018, p. 35) 12
Figure 3: Knowledge spiral at epistemological level in the context of the four modes of knowledge conversion. Own visualization according to Nonaka and Takeuchi (Cf. Nonaka and Takeuchi, 1995, p. 62, p. 72). 20
Figure 4: Diagram of an ontology excerpt showing classes, properties, instances as well as the distinction between TBox and ABox 24
Figure 5: Types of ontologies and their dependencies according to GUARINO (Guarino, 1997, p. 145) 27
Figure 6: Graphical representation of the process of creating ontological knowledge models in context of this thesis 32
Figure 7: Questionnaires and their distribution to surveyed groups in the course of the preliminary survey 34
Figure 8: Survey results knowledge engineers, question 1: Direction 36
Figure 9: Survey results knowledge engineers, question 2: Main goal 37
Figure 10: Survey results knowledge engineers, question 3: Relevant information per concept 38
Figure 11: Survey results knowledge engineers, question 4: Timepoint of documentation. 40
Figure 12: Survey results knowledge engineers, question 5: Exchange with domain experts 41
Figure 13: Survey results internal domain experts, question 1: Motivation 43
Figure 14: Survey results external domain experts, question 1: Motivation 44
Figure 15: Survey results internal domain experts, question 2: Expectations 46
Figure 16: Survey results external domain experts, question 2: Expectations 47
Figure 17: Survey results internal domain experts, question 3: Current usage of ontologies 48
Figure 18: Survey results external domain experts, question 3: Current usage of ontologies 49
Figure 19: Survey results internal domain experts, question 4: Exchange 51

Figure 20: Survey results external domain experts, question 4: Exchange 51
Figure 21: Affiliation of the six assumption personas 56
Figure 22: A comparison of different processes with respect to the IEEE 1074–1995 standard (Nicola et al., 2009, p. 274) 77
Figure 23: UPON process (Nicola et al., 2009, p. 273) on the left side supplemented by the process section to be covered by the documentation concept 84
Figure 24: Scenarios for building ontologies and ontology networks (Suárez-Figueroa, Gómez-Pérez, Fernández-López, 2012, p. 13) 93
Figure 25: Ontology network life-cycle models proposed in the NeOn Methodology framework. (a) Waterfall model family. (b) Iterative-incremental model (Suárez-Figueroa et al., 2015, p. 118). Supplemented by grey markings of the considered phases. 95
Figure 26: Non-ontological resource categorization (Villazón-Terrazas and Gómez-Pérez, 2012, p. 109) 96
Figure 27: Categorization of engineering knowledge resources based on the distinction b) non-ontological resource vs. ontological resource. 97
Figure 28: Categorization of engineering knowledge resources based on the distinctions b) non-ontological resource vs. ontological resource, d) TBox-related reuse vs. ABox-related reuse, and d) reuse without change vs. reuse with change. 98
Figure 29: Visual explanation of the relationship between 'documentary reference unit' and 'documentary unit' using three different example scenarios. 101
Figure 30: Schematic illustration of a documentation consisting of three documentation entries in the minimal configuration without descriptive modules. 102
Figure 31: Visualization of the three different perspectives on a concept, called dimensions, with questions that people may have when they think about the exemplary concept "substance identifier". 104
Figure 32: Dimension-specific modules according to the recommended assignment to dimensions 105
Figure 33: General modules and module extensions 105
Figure 34: Recommended compilation of dimension-specific modules extended by general modules and module extensions 106

Figure 35: Schematic illustration of a single documentation entry with explanations on the left side. On the right, potential documentation aspects are shown that are not covered by the documentation concept. 107
Figure 36: Schematic illustration of a documentation consisting of three documentation entries in the complete configuration using all available module types. 107
Figure 37: Annotated template for the module 'Alternative labels' (full version) 109
Figure 38: Annotated application of the module 'Alternative labels' (full version) for the reference unit "Substance identifier" 110
Figure 39: Loose collection of symbols and hints displayed at the top of the modules. Left: Type of the module. Center: Dimension. Right: Version of the module. 110
Figure 40: Selection from the occurring combinations of recommendations about target group and time of module application 110
Figure 41: Mandatory module 'Primary documentary reference unit' 111
Figure 42: Exemplary application of the mandatory module 'Primary documentary reference unit' 111
Figure 43: General module 'Related people' (full version) 112
Figure 44: General module 'Related people' (light version) 113
Figure 45: Exemplary application of the general module 'Related people' (full version) in the dimension 'Domain knowledge' 113
Figure 46: Exemplary application of the general module 'Related people' (light version) in the dimension 'Engineering knowledge' 114
Figure 47: General module 'Knowledge resources' (full version) 115
Figure 48: General module 'Knowledge resources' (light version) 115
Figure 49: Exemplary application of the general module 'Knowledge resources' (full version) 116
Figure 50: Exemplary application of the general module 'Knowledge resources' (light version) 116
Figure 51: General module 'Decisions' (full version) 117
Figure 52: General module 'Decisions' (light version) 117
Figure 53: Exemplary application of the general module 'Decisions' (light version) in the dimension 'Domain knowledge' 118

Figure 54: Exemplary application of the general module 'Decisions' (full version) in the dimension 'Management knowledge' 119
Figure 55: Exemplary application of the general module 'Decisions' (light version) in the dimension 'Engineering knowledge' 120
Figure 56: General module 'Tasks' (full version) 120
Figure 57: General module 'Tasks' (light version) 121
Figure 58: Exemplary application of the general module 'Tasks' (full version) in the dimension 'Domain knowledge' 121
Figure 59: Exemplary application of the general module 'Tasks' (light version) in the dimension 'Engineering knowledge' 122
Figure 60: General module 'Notes' 122
Figure 61: Exemplary application of the general module 'Notes' in the dimension 'Domain knowledge' 123
Figure 62: Exemplary application of the general module 'Notes' in the dimension 'Engineering knowledge' 123
Figure 63: Exemplary application of the general module 'Notes' in the dimension 'Management knowledge' 124
Figure 64: General module 'Discussion' 125
Figure 65: Exemplary application of the general module 'Discussion' in the dimension 'Domain knowledge' 125
Figure 66: Exemplary application of the general module 'Discussion' in the dimension 'Engineering knowledge' 126
Figure 67: Exemplary application of the general module 'Discussion' in the dimension 'Management knowledge' 126
Figure 68: Modul extension 'Source' (full version) 128
Figure 69: Modul extension 'Source' (light version) 128
Figure 70: Modul extension 'Source' (full version) 129
Figure 71: Modul extension 'Source' (full version) 129
Figure 72: Dimension-specific module 'Alternative labels' (full version) 130
Figure 73: Dimension-specific module 'Alternative labels' (light version) 130
Figure 74: Exemplary application of the dimension-specific module 'Alternative labels' (full version) 131

Figure 75: Exemplary application of the dimension-specific module 'Alternative labels' (light version) 131
Figure 76: Dimension-specific module 'Foreign identifiers' (full version) 132
Figure 77: Dimension-specific module 'Foreign identifiers' (light version) 132
Figure 78: Exemplary application of the dimension-specific module 'Foreign identifiers' (full version) 132
Figure 79: Exemplary application of the dimension-specific module 'Foreign identifiers' (light version) 133
Figure 80: Dimension-specific module 'Domain' 133
Figure 81: Exemplary application of the dimension-specific module 'Domain' 133
Figure 82: Dimension-specific module 'Definitions' (full version) 134
Figure 83: Dimension-specific module 'Definitions' (light version) 134
Figure 84: Exemplary application of the dimension-specific module 'Definitions' (full version) 135
Figure 85: Exemplary application of the dimension-specific module 'Definitions' (light version) 136
Figure 86: Dimension-specific module 'Explanations' (full version) 137
Figure 87: Dimension-specific module 'Explanations' (light version) 137
Figure 88: Exemplary application of the dimension-specific module 'Explanations' (full version) 138
Figure 89: Exemplary application of the dimension-specific module 'Explanations' (light version) 138
Figure 90: Dimension-specific module 'Domain knowledge resources' (full version) 139
Figure 91: Dimension-specific module 'Domain knowledge resources' (light version) 140
Figure 92: Exemplary application of the dimension-specific module 'Domain knowledge resources' (full version) 141
Figure 93: Exemplary application of the dimension-specific module 'Domain knowledge resources' (light version) 142
Figure 94: Dimension-specific module 'Concept category' 143
Figure 95: Exemplary application of the dimension-specific module 'Concept category'. 143
Figure 96: Dimension-specific module 'Relationships to other concepts' (full version) ... 145
Figure 97: Dimension-specific module 'Relationships to other concepts' (light version).. 146

Figure 98: Exemplary application of the dimension-specific module 'Relationships to other concepts' (full version) 147
Figure 99: Exemplary application of the dimension-specific module 'Relationships to other concepts' (light version) 148
Figure 100: Dimension-specific module 'Formal representations (full version) 149
Figure 101: Dimension-specific module 'Formal representations (light version) 149
Figure 102: Exemplary application of the dimension-specific module 'Formal representations' (full version) 150
Figure 103: Dimension-specific module 'Engineering knowledge resources' (full version) 151
Figure 104: Dimension-specific module 'Engineering knowledge resources' (light version) 152
Figure 105: Exemplary application of the dimension-specific module 'Engineering knowledge resources' (full version) 153
Figure 106: Exemplary application of the dimension-specific module 'Engineering knowledge resources' (light version) 154
Figure 107: Dimension-specific module 'Implementation status' 154
Figure 108: Dimension-specific module 'Test status regarding reasoning' 155
Figure 109: Dimension-specific module 'Related use cases' (full version) 156
Figure 110: Dimension-specific module 'Related use cases' (light version) 156
Figure 111: Exemplary application of the dimension-specific module 'Related use cases' (full version) 157
Figure 112: Exemplary application of the dimension-specific module 'Related use cases' (light version) 157
Figure 113: Dimension-specific module 'Related competency questions' (full version) 158
Figure 114: Dimension-specific module 'Related competency questions' (full version) 158
Figure 115: Exemplary application of the dimension-specific module 'Related competency questions' (full version) 158
Figure 116: Exemplary application of the dimension-specific module 'Related competency questions' (light version) 159
Figure 117: Dimension-specific module 'Affiliation' (full version) 159
Figure 118: Dimension-specific module 'Affiliation' (light version) 160

Figure 119: Exemplary application of the dimension-specific module 'Affiliation' (full version) 160
Figure 120: Exemplary application of the dimension-specific module 'Affiliation' (light version) 161
Figure 121: Dimension-specific module 'Test status regarding relevance' 161
Figure 122: Dimension-specific module 'Priority' 162
Figure 123: A competency question and an excerpt of the corresponding use-case model (Nicola et al., 2009, p. 263) 193
Figure 124: Activity of reference lexicon building (Nicola et al., 2009, p. 264) 194
Figure 125: An excerpt of the UML diagrams of the application scenario: (a) use-case diagram; (b) activity diagram; and (c) class diagram (Nicola et al., 2009, p. 265) 195
Figure 126: An overview of the analysis workflow (Nicola et al., 2009, p. 263) 195
Figure 127: Examples of ontological relationships in the eProcurement application (Nicola et al., 2009, p. 269) 197
Figure 128: An overview of the design workflow (Nicola et al., 2009, p. 267) 197
Figure 129: An overview of the implementation workflow (Nicola et al., 2009, p. 270) ... 198
Figure 130: An overview of the test workflow (Nicola et al., 2009, p. 271) 199

List of Tables

Table 1: Survey results knowledge engineers, question 3: Relevant information per concept (free text entries) 38
Table 2: Survey results internal domain experts, question 1: Motivation (free text entries) 43
Table 3: Survey results external domain experts, question 1: Motivation (free text entries) 44
Table 4: Survey results internal domain experts, question 2: Expectations (free text entries) 46
Table 5: Survey results internal domain experts, question 3: Current usage of ontologies (free text entries)..... 48
Table 6: Survey results external domain experts, question 3: Current usage of ontologies (free text entries)..... 49
Table 7: Survey results internal domain experts, question 5: Optional clarification on Question 4 – Other..... 52
Table 8: Survey results external domain experts, question 5: Optional clarification on Question 4 – Other..... 52
Table 9: Attributes of the assumption personas for both target groups – knowledge engineers and domain experts..... 56
Table 10: Findings derived from insights into corporate reality – relevant information per concept 71
Table 11: Findings derived from insights into corporate reality – relevant information per resource 72
Table 12: Demonstration of the properties access control, confidentiality and license for exemplary resources..... 73
Table 13: Findings derived from insights into corporate reality – documentation principles 75
Table 14: List of the analyzed literature regarding the Unified Process for Ontology Building 79
Table 15: Findings derived from the Unified Process for Ontology Building – relevant information per concept 89
Table 16: Findings derived from the Unified Process for Ontology Building – documentation principles 90
Table 17: List of the analyzed literature regarding the NeOn Methodology 91

Table 18: Rating scale for the recording of agreement 127
Table 19: Ontology engineering methodologies and related literature (not exhaustive) 172
Table 20: Survey data knowledge engineers, question 1: Direction 177
Table 21: Survey data knowledge engineers, question 2: Main goal 178
Table 22: Survey data knowledge engineers, question 3: Relevant information per concept 179
Table 23: Survey data knowledge engineers, question 3: Relevant information per concept (free text entries) 179
Table 24: Survey data knowledge engineers, question 4: Time point of documentation 180
Table 25: Survey data knowledge engineers, question 5: Exchange with domain experts 181
Table 26: Survey data internal domain experts, question 1: Motivation 182
Table 27: Survey data internal domain experts, question 1: Motivation (free text entries) 182
Table 28: Survey data internal domain experts, question 2: Expectations 183
Table 29: Survey data internal domain experts, question 2: Expectations (free text entries) 183
Table 30: Survey data internal domain experts, question 3: Current usage of ontologies 184
Table 31: Survey data internal domain experts, question 3: Current usage of ontologies (free text entries) 184
Table 32: Survey data internal domain experts, question 4: Exchange 185
Table 33: Survey data internal domain experts, question 5: Optional clarification on Question 4 – Other 186
Table 34: Survey data external domain experts, question 1: Motivation 187
Table 35: Survey data external domain experts, question 1: Motivation (free text entries) 187
Table 36: Survey data external domain experts, question 2: Expectations 188
Table 37: Survey data external domain experts, question 3: Current usage of ontologies 189
Table 38: Survey data external domain experts, question 3: Current usage of ontologies (free text entries) 189
Table 39: Survey data external domain experts, question 4: Exchange 190
Table 40: Survey data external domain experts, question 5: Optional clarification on Question 4 – Other 191

Abbreviations

ABox	Assertion Box
AL	Application Lexicon
CQ	Competency Question
DE	Domain Expert
DIKW	Data, Information, Knowledge, Wisdom
DILIGENT	DIstributed, Loosely-controlled and evolvInG Engineering of oNTologies
DIN	Deutsches Institut für Normung
DL	Domain Lexicon
ELIXIR	European Life Science Infrastructure for Biological Information
EOSC	European Open Science Cloud
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable
FDA	Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resources
GO FAIR	Global Open FAIR
HCLS	Semantic Web Health Care and Life Sciences
HL7	Health Level 7
ID	Identifier
IEEE	Institute of Electrical and Electronics Engineers
IRI	Internationalized Resource Identifier
ISO	International Organization for Standardization
IT	Information Technology
KE	Knowledge Engineer
NOR	Non-Ontological Resource
NeOn	Networked Ontologies
ODP	Ontology Design Pattern
OPAL	Object, Process, Actor modeling Language
OWL	Web Ontology Language

PhUSE	Pharmaceutical Users Software Exchange
R&D	Research and Development
RA	Regulatory Affairs
RDF	Resource Description Framework
RDFS	RDF Schema
RG	Reference Glossary
RL	Reference Lexicon
SECI	Socialization, Externalization, Combination, Internalization
SKOS	Simple Knowledge Organization System
SME	Subject Matter Expert
SN	Semantic Network
SWAT4LS	Semantic Web Applications and Tools for Healthcare and Life Sciences
SubID	Substance identifier
TBox	Terminology Box
TLO	Top-Level Ontology
TLS	Translational Life Science
UC	Use Case
UML	Unified Modeling Language
UP	Unified Process
UPON	Unified Process for ONtology building
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
W3C	World Wide Web Consortium
XML	Extensible Markup Language

1 Introduction

For at least 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* (HCLS) was founded (Cf. Stephens, 2008) with the objective "to develop, advocate for, and support the use of Semantic Web technologies across health care, life sciences, clinical research and translational medicine" (Semantic Web Health Care and Life Sciences (HCLS) Interest Group, 2018). Since 2018, the interest group HCLS has continued its work as a community group (Cf. W3C, 2018). Another example is the international *Semantic Web Applications and Tools for Healthcare and Life Sciences* (SWAT4HCLS) workshop which has been taking place annually since 2008. It "provide[s] a platform for the presentation and discussion of the benefits and limits of applying Web-based information systems and semantic technologies in the domains of health care and life sciences" (Paschke et al., 2018).

As the shift towards the use of semantic technologies in the life sciences is becoming more common in both the private and public sector, the international standardization organization *Health Level Seven International* (HL7)[1] has published a *Linked Data Module* for its standard framework *Fast Healthcare Interoperability Resource* (FHIR). It represents FHIR resources based on the *Resource Description Framework* (RDF)[2] (Cf. HL7.org, 2017). 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)[3]. There the working group *Linked Data & Graph Databases* is working on the use of semantic technologies (Cf. Kent Innovation Centre, 2017). In accordance with the "important goal" of achieving "an acceptable degree of semantic interoperability", which the FDA declares with its *Study Data Technical Conformance Guide* (U.S. Food and Drug Administration, 2018, p. 36), the FDA also uses RDF internally in its Janus data repository (Cf. Office of the Commissioner, 2018).

---

1 Further information on HL7 can be found in:
Health Level Seven International (2018) *About Health Level Seven International* [Online]. Available at http://www.hl7.org/about/ (Accessed 2019-02-18).

2 Further information on RDF can be found in chapter '2.3.3 Ontology languages' and in:
Schreiber, Guus; Raimond, Yves (2014) *RDF 1.1 Primer: W3C Working Group Note 24 June 2014* [Online], W3C. Available at https://www.w3.org/TR/rdf11-primer/ (Accessed 2019-02-18).

3 Further information on PhUSE can be found in:
Kent Innovation Centre (n.y.) *PhUSE Society FAQs* [Online]. Available at https://www.phuse.eu/faq/ (Accessed 2019-02-18).

At EU level, the intergovernmental organization ELIXIR operates and "brings together life science resources from across Europe" (ELIXIR, 2017a). Accordingly, one project of ELIXIR's interoperability platform is the *Linked Data* project (Cf. ELIXIR, 2017b). Furthermore, being engaged in the *European Open Science Cloud* (EOSC), ELIXIR has FAIR data as its goal (Cf. Niklas Blomberg, 2017, p. 2). FAIR refers to a set of four principles: data must be *findable*, *accessible*, *interoperable*, and *reusable* (Cf. Wilkinson et al., 2016). The FAIR strategy is mainly driven by the *GO FAIR* initiative (Cf. GO FAIR, 2018). Both EOSC and GO FAIR follow the recommendations of the *European Commission expert group on FAIR data* (Cf. Directorate-General for Research and Innovation, 2018a, p. 20; GO FAIR, 2018), which recommends, among others, semantic technologies to achieve FAIR data (Cf. Directorate-General for Research and Innovation, 2018b, p. 42). In non-for-profit collaborations such as the *Pistoia Alliance*[4], life sciences companies, vendors, publishers and academic groups are jointly dedicated to the implementation of FAIR data principles in the field of biopharmaceutical R&D (Cf. Wise et al., 2019).

One of the world's leading life science companies is the *Bayer AG* with its three divisions *Pharmaceuticals*, *Consumer Health* and *Crop Science* as well as an *Animal Health* business unit (Cf. Bayer AG, 2017, p. 5). According to Statista's toplist 2018 for the top 100 pharma and biotech companies, Bayer ranks 8th in terms of overall revenue (Cf. Statista, 2018). Bayer is currently (2018/2019) running several project activities to reorganize its data landscape. The overall framework is determined by Bayer's *Digital Agenda* (Cf. Bayer, 2018), which, among others, aspires to treat *Data as an Asset* [5]. To enable data-driven decisions, the traditional information flow must be transformed. Translational research across all functions and FAIR data are needed (Cf. Bayer, 2018). In course of this, Bayer transfers existing data sets into RDF, formally represents the knowledge of different domains by creating an ontology network and builds the required infrastructure. By using *Semantic Web* standards (Cf. W3C, 2015) and *Linked Data* principles (Cf. Berners-Lee, 2006), data is being integrated.

---

4 Further information on the Pistoia Alliance can be found in:
Pistoia Alliance, Inc. (2019) *About the Pistoia Alliance* [Online]. Available at https://www.pistoiaalliance.org/membership/about/ (Accessed 2019-02-18).

5 Exemplary article describing Bayer's use of data as an asset:
Schenker, Jennifer L. (2018) 'Using Data as an Asset: A plant run by Bayer's Pharmaceuticals division in Italy is considered one of the world's most advanced factories', *The Innovator*, no. 8, p. 31 [Online]. Available at https://kiosque.lesechos.fr/pdf.php?edition=20181105_INNOV.pdf (Accessed 2018-12-29).

The formal representation of knowledge is challenging. Bayer's ontologies are usually 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 such as PhUSE take place. There are no standardized processes for the transformation of informal to formal knowledge within the company. In addition to the lack of uniform processes, there is also a lack of documentation. The fact that the already complicated exchange between the parties involved during ontology creation is barely documented constitutes a serious problem. This is the starting point of this bachelor thesis.

The main objective of this thesis is to develop a documentation concept for the exchange of knowledge in the process of creating ontological knowledge models. The following chapter '1.1 Objectives' describes which objectives are pursued during the development and which problems are addressed by the documentation concept. Chapter '1.2 Approach' explains the procedure for the development of the concept and outlines the thesis chapters. The terms which are fundamental for this thesis – including the terms *documentation concept*, *knowledge*, *ontology* and *process* used in the title – are defined in detail in chapter '2 Theoretical background'.

1.1 Objectives

The documentation concept addresses ***two main issues:*** missing documentation and insufficient documentation.

Missing documentation refers to the lack of *knowledge externalization*. This means that the knowledge in the minds of the people involved is not recorded. The so-called *tacit* knowledge remains bound to individuals and thus difficult to access. A transformation to *explicit* knowledge, which is written down and therefore easily accessible for others, does not take place. This is one of the fundamental problems in the field of knowledge management. According to NONAKA and TAKEUCHI "externalization holds the key to knowledge creation, because it creates new, explicit concepts from tacit knowledge" (Nonaka and Takeuchi, 1995, p. 66). The relevant terms in this context are introduced in detail in chapter '2.2 Knowledge'. The main risk of a lack of knowledge externalization is *knowledge loss*. There are different causes of knowledge loss (Cf. Probst et al., 2010, p. 213), for instance the disbandment of teams, sick leaves or dismissals. In the context of ontology creation, this is particularly relevant for knowledge which does not find its way into the 'final' ontology. If knowledge sources, important modeling decisions or exchanged tacit knowledge are not recorded, collaboration is impeded and knowledge loss can cause projects to fail.

Insufficient documentation relates to the way documentation is done if it takes place. Documentations should be clear, understandable, timely and complete (Cf. Heinrich et al., 2014, p. 249). Considering documentations as user information, they must also be accurate, appropriate and accessible (Cf. Schultz et al., 1993, p. 4). Otherwise, they lose significantly in value. However, as will be shown in chapter '3 Selective insights into corporate reality', ontology creators often document too late and in an unstructured manner. In the worst case, the result is knowledge which is explicit but not available since it is not found or not understood. The effect of unavailability of knowledge is similar to the effect of knowledge loss (Cf. North and Kumta, 2018, p. 292). Consequently, poor documentation is just as futile as missing documentation, even though effort has been invested.

The documentation concept addresses these two issues by encouraging the externalization of knowledge through documentation and thus by contributing to *knowledge preservation* (Cf. Heinrich et al., 2014, p. 322). It consists of guiding documentation templates, which facilitate the documentation and improve its quality. The characteristics of the documentation concept are specified in chapter '2.1 Documentation concept'.

The concept is based on the following assumptions: During the ontology creation within a company there are …

- **No uniform procedures.** The methodology used may differ from individual to individual and from group to group.
- **No uniform objectives**. The objectives pursued may differ from individual to individual and from group to group.
- **No uniform working modes**. The working modes may differ from individual to individual and from group to group.
- **No uniform working conditions**. Working conditions may differ from individual to individual and from group to group.

These assumptions are the result of observations and personal experiences, which are both described in detail in chapter '3 Selective insights into corporate reality'. The objective of this thesis is not to elaborate rules restricting the observed differences. Rather, the objective is a documentation concept that takes these differences into account by supporting different procedures and working modes that pursue different goals under varying conditions.

Overall, the focus is on the creation of ontologies in companies. Community projects or associations that create ontologies with the contribution of volunteers are not the subject of this thesis. The documentation concept may also be of interest to those, but it is primarily oriented towards company contexts.

Furthermore, the documentation concept is designed for the process of creating *new* ontologies. The process considered in this thesis begins once the objectives and framework conditions for the ontology creation have been defined. It ends with the first version of a formal model that meets all previously defined objectives. Chapter '2.4 Process' describes the process in more detail.

The manual or automatic creation of a documentation based on a formally available ontology is not part of this thesis, because, on the one hand, tools for this purpose already exist (e.g. *VoCol*[6], *WIDOCO*[7], *LODE*[8], *OntoxicWiki*[9]). On the other hand, the issues identified refer to the phases of the creation process where no formal model is yet available. Once the knowledge of a domain is represented by a formal model, the advantages of an ontology – e.g. machine processability and inline documentation – can be exploited.

---

[6] VoCol: An Integrated Environment to Support Version-Controlled Vocabulary Development is capable of *documentation generation* and *visualization generation*, among other things.
Further information on VoCol can be found in:
- Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS (n.y.) *VoCol - An Integrated Environment for Collaborative Vocabulary Development* [Online]. Available at https://vocol.iais.fraunhofer.de/ (Accessed 2019-02-18).
- Halilaj, Lavdim; Petersen, Niklas; Grangel-González, Irlán; Lange, Christoph; Auer, Sören; Coskun, Gökhan; Lohmann, Steffen (2016) 'VoCol: An Integrated Environment to Support Version-Controlled Vocabulary Development', in *Knowledge Engineering and Knowledge Management*. 20th International Conference, EKAW 2016. Bologna, Italy, November 19-23, 2016. Cham, s.l., Springer International Publishing, pp. 303–319. DOI: 10.1007/978-3-319-49004-5_20.

[7] WIDOCO: A Wizard for Documenting Ontologies. Further information can be found in:
- Daniel Garijo Verdejo (2018) *WIzard for DOCumenting Ontologies (WIDOCO)* [Online], GitHub. Available at https://github.com/dgarijo/Widoco/ (Accessed 2019-02-18).
- Garijo, Daniel (2017) 'WIDOCO: A Wizard for Documenting Ontologies', in *The Semantic Web – ISWC 2017*. 16th International Semantic Web Conference. Vienna, Austria, October 21-25, 2017. Cham, Springer International Publishing, pp. 94–102. DOI: 10.1007/978-3-319-68204-4_9.

[8] LODE: Live OWL Documentation Environment. Further information can be found in:
- Peroni, Silvio (2013) *Live OWL Documentation Environment (LODE)* [Online]. Available at http://www.essepuntato.it/lode/ (Accessed 2019-02-18).
- Peroni, Silvio; Shotton, David; Vitali, Fabio (2012) 'The Live OWL Documentation Environment: A Tool for the Automatic Generation of Ontology Documentation', in *Knowledge engineering and knowledge management*. 18th international conference, EKAW 2012. Galway City, Ireland, October 8-12, 2012. Berlin, Springer, pp. 398–412. DOI: 10.1007/978-3-642-33876-2_35.

[9] OntoxicWiki: Social Ontology Documentation for Knowledge Externalization. Further information can be found in:
- Aranda-Corral, Gonzalo A.; Borrego-Díaz, Joaquín; Jiménez-Mavillard, Antonio (2010) 'Social Ontology Documentation for Knowledge Externalization', in *Metadata and Semantic Research*. 4th International Conference, MTSR 2010. Alcalá de Henares, Spain, October 20-22, 2010. Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 137–148. DOI: 10.1007/978-3-642-16552-8_14.

1.2 Approach

Before the documentation concept is created, the basic concepts are explained and working definitions are phrased (see chapter '2 Theoretical background').

The development of the documentation concept begins with the collection of well-founded inputs. Two sources are analyzed to identify relevant information that should be documented. On the one hand, observations from practice are considered (see chapter '3 Selective insights into corporate reality'). These are based on survey results as well as on profiles of archetypal ontology creators. The latter are so-called *assumption personas* which are created according to the persona method. On the other hand, two ontology engineering methodologies proposed in literature are considered (see chapter '4 Analysis of ontology engineering methodologies'). Thanks to the diversity of sources, the documentation concept takes into account findings from both: industry and academia.

The findings of the analysis are then used to create the documentation concept in chapter '5 Documentation concept'. The documentation templates are conveyed by means of illustrations similar to forms or graphical user interfaces.

In chapter '6 Discussion' the developed documentation concept is discussed. Finally, the conclusion presented in chapter '7 Conclusion' is followed by an outlook in chapter '8 Outlook'.

2 Theoretical background

In this chapter, the basic concepts of this thesis are explained and put into context. In the course of this working definitions are given.

2.1 Documentation concept

Depending on its use, the term ***documentation*** has different meanings. In order to clarify, which meanings apply in the context of this thesis, a characterization by WERSIG is used. He distinguishes between six concepts (Wersig, 1971, p. 143)[10]:

a) *Methods* or techniques used by a process (the 'documentation technique', German 'die Dokumentationstechnik')

b) The *process* itself (the process of 'documenting', German 'das Dokumentieren')

c) The *result* of the process (the entirety of 'documented' things, German 'das Dokumentierte')

d) The *system* in which these processes take place ('the area of documentation', German 'der Bereich des Dokumentierens')

e) The *whole field of these systems* ('the documentation system', German 'das Dokumentationswesen')

f) The academic discipline dealing with this subject ('documentation science', German 'Dokumentationswissenschaft')

Figure 1 shows an English version of WERSIG's visualization which puts the different terms and meanings into context.

[10] Own translation from German to English

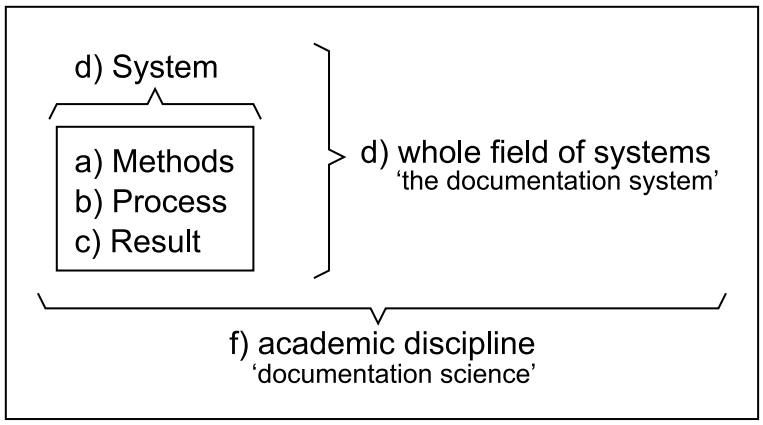

Figure 1: Context of the different documentation terms according to WERSIG *(Wersig, 1971, p. 143)*

As the objective of this thesis is to develop a ***documentation concept*** which can be used to create a documentation in the sense of an artefact, the focus is on the *result* (c). Since the result is based on a documentation *process* (b) using *methods* and techniques (a), these two aspects are considered as well. In other words, a documentation concept according to the understanding in this thesis states:

1) *What* to document (What is part of the *result*?)
2) *How* to document (Which *methods* and techniques are used?)
3) *When* and *by whom* to document (How is the *process*?)

The meanings of the concepts (e) and (f) are not of interest in the context of this thesis. Whether the documentation concept under consideration is a *system* (e) or not depends on the definition of the term *system*. WERSIG distinguishes between two main components of a documentation system, the *material component* and the *conceptual component*. The former includes, among others, material aids, data carriers as well as people and their activities. The latter mainly consists of documentation principles (rules for the execution of documentation activities) and relationships that connect these documentation principles to documentation procedures. (Cf. Wersig, 1971, pp. 216–217)

Applying WERSIGs understanding of *documentation system* to *documentation concept* in the sense of this thesis means that both components are partially covered. Statements are made about people and their activities as well as about documentation principles. However, the thesis neither stipulates a complete documentation process nor a concrete implementation of the concept (e.g. by giving technical specifications for the implementation of a software application).

The priority of the three relevant aspects corresponds to the numbering used above. This means that the documentation states in detail *what* should be documented (1) but refers in a less detailed way to *when* and *by whom* this should be done (3). The reason for this is that as many approaches to ontology creation as possible should be supported by the documentation concept. It therefore does not stipulate but recommends. First, it recommends which aspects should be recorded (e.g. labels for a concept). Then methods are suggested (e.g. 'capture any labels without differentiation' *versus* 'differentiate between preferred labels and alternative labels'). Finally, it recommends who documents this information and when (e.g. domain experts throughout the entire process).

As a result, no strict documentation process is modeled. Rather, a modular system is proposed which can be used completely or in parts, detailed or less detailed, according to the recommended order or arranged differently. When this thesis refers to a *documentation concept*, exactly this modular system is meant.

If one distinguishes between *informal*, *semi-formal* and *formal* documentation, the modular system is a concept for semi-formal documentation. ***Informal*** documentations are individual texts that do not follow any guidelines. Their disadvantage is that they are unstructured, ambiguous and heterogeneous. This type of documentation was identified as insufficient in chapter '1.1 Objectives'. ***Semi-formal*** documentations tackle these disadvantages. They follow guidelines provided by *documentation templates*, also known as *documentation patterns*. The modules of the documentation concept correspond to such templates. They enable the creation of structured documentation using natural language. This improves the quality of the documentation and saves time because the templates are reusable. (Cf. Landes et al., 1999, pp. 650–657; Lehmann, 2018, p. 44)

A completely ***formal*** documentation is not aimed at, due to the high effort of creating such documentations (Cf. Landes et al., 1999, pp. 656–657; Lehmann, 2018, p. 44). Besides, people without the necessary skills should also be able to contribute to the documentation. Documenting the development of an ontology by creating another ontology is therefore inappropriate. This does not exclude the possibility of representing the proposed documentation concept as an ontology and implementing it in a specialized software application with a user interface appropriate for all target groups. On the contrary, this would be one of the possible and desirable implementations of the concept (see chapter '8 Outlook'), but the implementation is not an objective of this thesis.

The documentation concept is designed for a ***constantly changing documentation***. This refers to its scope and stability. It is assumed that the quantity of the documentation units may change during the process of ontology creation. In addition, information once recorded in the documentation can be supplemented, modified and deleted. This should be considered in particular when applying quality criteria such as *accuracy*. Furthermore, it is part of ontology creation to find a common understanding. Meanwhile, the documentation may contain both objective errors and diverging views. Such a 'living documentation' offers an added value already during its creation and evolution – not only after its completion.

Working definition:

The proposed **documentation concept** is a modular system consisting of documentation templates that provide guidelines for creating a semi-formal, structured documentation using natural language. It states what should be recorded by whom, how and when. The modules can be used in parts, at different levels of detail and in a customizable arrangement. The concept is designed for a documentation that may change in scope and content during its creation.

2.2 Knowledge

For the fundamental concepts in the field of information sciences, there is a variety of definitions, definition approaches and models. In the context of this thesis, mainly the three concepts *data*, *information* and *knowledge* are relevant. To explain these concepts, the model of the so-called *DIKW hierarchy* in the form of knowledge ladders is used. The concept of *wisdom* is not discussed here. This also applies to possible extensions of the model, e.g. by *competence, competitiveness* or *truth*, as well as to any stages preceding the data, such as *signals* or *symbols*.

Following the explanation of data, information and knowledge, the characteristics of knowledge are discussed in more detail. This is necessary in order to understand in which area of knowledge management the documentation concept is located.

2.2.1 Data, Information, Knowledge

Models that are based on a ***DIKW hierarchy*** assume a hierarchical relationship between *data*, *information*, *knowledge* and *wisdom*, hence the name DIKW. There are different model variants that follow this approach. They generally assume that the concepts systematically build on one another through the increase in structuring (first physical, then cognitive and finally value-based) and thus in human understanding (Cf. Hobohm, 2016, p. 9). The DIKW hierarchy is controversial (Cf. Frické, 2009; Rowley, 2007) like any of the models, but applicable in the context of this thesis. In concrete terms, an excerpt of the knowledge ladder proposed by NORTH and KUMTA is applied (see figure 2).

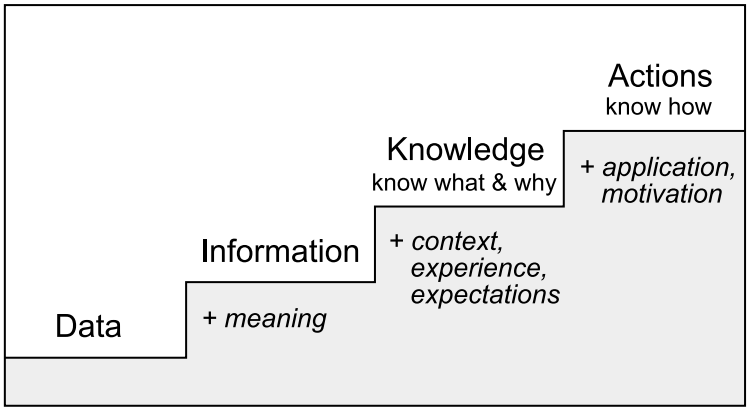

Figure 2: Excerpt of the knowledge ladder proposed by NORTH and KUMTA (Cf. North and Kumta, 2018, p. 35)

According to NORTH and KUMTA, ***data*** are symbols connected by syntax (Cf. North and Kumta, 2018, p. 35). They have no meaning by themselves. In the words of the *European Guide to good Practice in Knowledge Management*, this results in the following working definition:

Working definition

Data are "discrete, objective facts (numbers, symbols, figures) without context and interpretation" (Comité Européen de Normalisation, 2004, p. 9).

If a message is meaningful, it is no longer data, but ***information*** (Cf. Comité Européen de Normalisation, 2004, p. 10; North and Kumta, 2018, p. 36). This approach to defining information is called the *meaning approach* (Cf. Gaus, 2005, p. 30; Wersig, 1971, p. 32). It implies that the recipient of an encoded message can not only decode it but can also understand its meaning. Information is therefore an understood message (Cf. Gaus, 2005, p. 30). The interpretation depends on the interpreting system and may therefore differ from person to person (Cf. North and Kumta, 2018, p. 36). Hence, information is also referred to as "meaning assigned to data by known conventions" (Weik, 1977, p. 184). The conventions refer mainly to a common language and level of knowledge (Cf. Gaus, 2005, p. 30). It is assumed that the recipient is able to assign meaning to data, e.g. to a word consisting of a string (Cf. Wersig, 1971, p. 39). For example, a reader without legal knowledge can read a law text written in his mother tongue. However, he cannot assign any meaning to subject-specific words. Although the text is a message for him, he does not understand it and therefore it is not information.

Working definition

"**Information** is organised data adding meaning to a message. This information is interpreted differently depending on context, experience and the expectations of people." (North and Kumta, 2018, p. 36)

In this thesis, information is thus located at the level of *semantics*, not at the level of *pragmatics*, although pragmatic approaches such as the *effect approach* (Cf. Gaus, 2005, p. 30; Wersig, 1971, pp. 32–33) are dominant in the field of information sciences (Cf. Ingold, 2016, p. 6). According to WERSIG, the definition of the *information* concept must be based on the effect approach (Cf. Wersig, 1971, p. 40). This approach implies that information must have a novelty value, i.e. must reduce uncertainty. If a message does not provide any news, it has no effect on the recipient and is therefore not an information (Cf. Gaus, 2005, p. 30; Wersig, 1971, pp. 32–33). Despite the popularity of the effect approach, the meaning approach is applied in this thesis. Otherwise the decision whether something is an information or not would always depend on the effect it has on the recipient, e.g. the readers of a documentation.

By applying the meaning approach, two already mentioned preconditions are assumed: common language, which in this case is English, and common level of knowledge. This means that this thesis assumes that the readers of the documentation concept are proficient in the English language. It is also assumed that there is a basic common knowledge. Since this is a rather daring assumption, this chapter '2 Theoretical background' is quite detailed in order to establish such a common basic knowledge.

On the knowledge ladder, the concept *information* is followed by the concept ***knowledge***. PROBST, RAUB and ROMHARDT describe knowledge as *the totality of the skills and capabilities that individuals use to solve problems* (Cf. Probst et al., 2010, p. 23). They refer to both the *theoretical* and the *practical* side of knowledge. In this respect, NORTH and KUMTA distinguish between two levels of knowledge: to *know what* as "a result of interiorising information" and to *know how* which refers to the application of the interiorized information (North and Kumta, 2018, p. 36). So, knowledge is the result of information that is put into context, usually with reference to application.

The following excerpt of the definition given in the *European Guide to good Practice in Knowledge Management* illustrates how broad the concept *knowledge* and the influencing factors can be understood:

> *"[Knowledge is] a combination of, for example know-how, experience, emotion, believes, values, ideas, intuition, curiosity, motivation, learning styles, attitude, ability to trust, ability to deal with complexity, ability to synthesize, openness, networking skills, communication skills, attitude to risk and entrepreneurial spirit to result in a valuable asset which can be used to improve the capacity to act and support decision making."* (Comité Européen de Normalisation, 2004, p. 10)

The objective of this thesis is to develop a documentation concept based on which knowledge can be recorded. For this, two forms of knowledge representation are particularly relevant: *tacit* and *explicit* knowledge. *Tacit knowledge* is "the personal knowledge of an individual" which is "very difficult to formulate and to pass on because it is embodied in individuals" (North and Kumta, 2018, p. 46). By contrast, *explicit knowledge* is "knowledge that has been codified […] and can therefore be easily shared and understood" (Comité Européen de Normalisation, 2004, p. 9). Before the characteristics of knowledge are discussed in more detail in the next chapter, the working definition for the concept *knowledge* is given.

Working definition

"**Knowledge** refers to the tacit or explicit understanding of people about relationships among phenomena. It is embodied in routines for the performance of activities, in organisational structures and processes and in embedded beliefs and behaviour. Knowledge implies an ability to relate inputs to outputs, to observe regularities in information, to codify, explain and ultimately to predict" (Carnegie Bosch Institute, 1995, cited in North and Kumta, 2018, p. 36)

2.2.2 Characteristics of knowledge

For the explanation of the characteristics of knowledge, the so-called *knowledge spiral* by NONAKA and TAKEUCHI is used (Cf. Nonaka and Takeuchi, 1995). The focus is on the epistemological dimension of the knowledge spiral, also known as the *SECI model* (*socialization*, *externalization*, *combination*, *internalization*).

NONAKA and TAKEUCHI draw on POLANYI's distinction between ***tacit knowledge*** and ***explicit knowledge*** (Cf. Nonaka and Takeuchi, 1995, p. 59; Polanyi, 1966) defining the two knowledge types as follows:

> "*Tacit knowledge is personal, context-specific, and therefore hard to formalize and communicate. Explicit or 'codified' knowledge, on the other hand, refers to knowledge that is transmittable in formal, systematic language.*" (Nonaka and Takeuchi, 1995, p. 59)

Tacit knowledge is thus *not* just knowledge stored in people's minds, which can be easily articulated. It "consists of mental models, behaviours and perspectives, largely based on experience" (Comité Européen de Normalisation, 2004, p. 12) or as POLANYI put it: "we can know more than we can tell" (Polanyi, 1966, p. 4). This is put together in the description by NORTH and KUMTA, which is used as working definition.

Working definition

"**Tacit knowledge** represents the personal knowledge of an individual. It is based on education, ideals, values and feelings of the individual person. Subjective insights and intuition embody tacit knowledge that is deeply rooted in the actions and experiences of the particular person. […] This form of knowledge is very difficult to formulate and to pass on because it is embodied in individuals." (North and Kumta, 2018, p. 46)

In the definition of NONAKA and TAKEUCHI, *formal* does *not* imply that explicit knowledge can solely be expressed in *formal language* in the sense of a mathematical, machine-processable language (as it is the case in chapters '2.1 Documentation concept' and '2.3 Ontology'). Explicit knowledge can also be expressed using a *natural language*. The decisive statement is that explicit knowledge is not bound to an individual but is "codified […] and can therefore be easily shared and understood" (Comité Européen de Normalisation, 2004, p. 9). This is well illustrated in NORTH's and KUMTA's explanation, which is therefore used as a working definition.

Working definition

"Unlike tacit knowledge, **explicit knowledge** is methodical and systematic and is present in an articulated form. It is stored in the media outside the brain (disembodied knowledge) of an individual and can be transferred and stored by means of information and communication technology. Examples of explicit knowledge are detailed descriptions of processes, patents, organisation trees, quality documents, etc." (North and Kumta, 2018, p. 46)

Tacit and *explicit* are ***forms of knowledge representation***. The knowledge spiral on the epistemological dimension describes the ***conversion of knowledge***, i.e. the *interaction between tacit and explicit knowledge* (Cf. Nonaka and Takeuchi, 1995, p. 61). The conversion of knowledge from one form of representation to another is also called *knowledge transformation* (Cf. Heinrich et al., 2014, p. 485). NONAKA and TAKEUCHI postulate four ***modes of knowledge conversion***:

> "*(1) from tacit knowledge to tacit knowledge, which we call socialization; (2) from tacit knowledge to explicit knowledge, or externalisation; (3) from explicit knowledge to explicit knowledge, or combination (4) from explicit knowledge to tacit knowledge, or internalization*" (Nonaka and Takeuchi, 1995, p. 62)

The following working definitions explain the four modes of knowledge conversion *socialization*, *externalization*, *combination* and *internalization*.

Working definition

Socialization is the "conversion from tacit knowledge of one person to tacit knowledge of another person […]. It is a process of sharing experiences and thereby creating tacit knowledge such as shared mental models and technical skills" (North and Kumta, 2018, p. 47).

According to NONAKA and TAKEUCHI *experience* is key, since tacit knowledge can be acquired "directly from others without using language […] but through observation, imitation, and practice" (Nonaka and Takeuchi, 1995, 62-32).

Socialization takes place when tacit knowledge is exchanged between the people involved in the process of ontology creation. The purpose of the documentation concept is neither to prevent socialization nor to support it directly. Instead, the purpose of the documentation is to codify knowledge. It thus contributes to a ***codification strategy*** instead of a ***personalization strategy***. The latter of these two knowledge management strategies intends that the knowledge remains with the knowledge bearers and is only passed on by means of personal communication if required. Codification, on the other hand, requires knowledge externalization. (Cf. Heinrich et al., 2014, p. 323)

Working definition

Externalization is a "process of articulating tacit knowledge into explicit concepts. It is a quintessential knowledge-creating process in that tacit knowledge becomes explicit, taking the shapes of metaphors, analogies, concepts, hypotheses, or models" (Nonaka and Takeuchi, 1995, p. 64).

Once knowledge is available explicitly, *knowledge combination* and *knowledge internalization* can take place.

Working definition

Combination is a "mode of knowledge conversion [that] involves combining different bodies of explicit knowledge. Individuals exchange and combine knowledge through such media as documents, meetings, telephone conversations, or computerized communication networks. Reconfiguration of existing information through sorting, adding, combining, and categorizing of explicit knowledge (as conducted in computer databases) can lead to new knowledge" (Nonaka and Takeuchi, 1995, p. 67).

Working definition

Internalization is a "process of embodying explicit knowledge into tacit knowledge. It is closely related to 'learning by doing'. When experiences through socialization, externalization, and combination are internalized into individuals' tacit knowledge bases in the form of shared mental models or technical know-how, they become valuable assets" (Nonaka and Takeuchi, 1995, p. 69).

NONAKA and TAKEUCHI state that "[o]rganizational knowledge creation is a continuous and dynamic interaction between tacit and explicit knowledge" (Nonaka and Takeuchi, 1995, p. 70). They locate the ***knowledge spiral*** emerging from this interaction on two dimensions, the *epistemological dimension* (explicit – tacit) and the *ontological dimension* (individual – group – organizational – inter-organizational) (Nonaka and Takeuchi, 1995, p. 57). The *ontological* dimension here refers to "the levels of knowledge creating entities" (Nonaka and Takeuchi, 1995, p. 57) and is not to be confused with the understanding of the concept *ontology* used in this thesis (see chapter '2.3 Ontology' for the applied working definition).

Figure 3 shows the knowledge spiral on the epistemological dimension. It begins in the area of tacit knowledge, because "[t]acit knowledge of individuals is the basis of organizational knowledge creation" (Nonaka and Takeuchi, 1995, p. 72). From then on, the process of *socialization* → *externalization* → *combination* → *internalization* is performed iteratively. The ontological dimension is not illustrated in figure 3. It corresponds to the viewing dimension of the reader of this thesis, i.e. is perpendicular to the figure. This means that the knowledge spiral moves towards the reader. The spiral starts at the individual level and moves up the ontological levels "through expanding communities of interaction, that cross sectional, departmental, divisional, and organizational boundaries" (Nonaka and Takeuchi, 1995, p. 72). In the course of this, "the interaction between tacit knowledge and explicit knowledge will become larger in scale" (Nonaka and Takeuchi, 1995, p. 72).

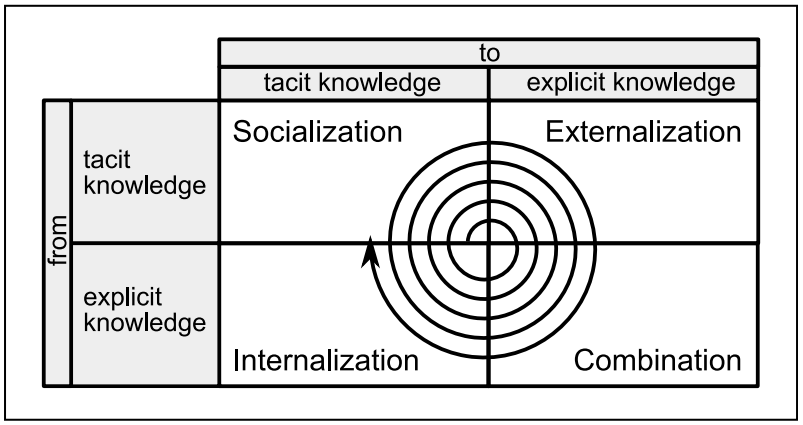

Figure 3: Knowledge spiral at epistemological level in the context of the four modes of knowledge conversion. Own visualization according to Nonaka and Takeuchi (Cf. Nonaka and Takeuchi, 1995, p. 62, p. 72).

In the context of this thesis, the *knowledge spiral* implies the following:

- During ontology creation *tacit knowledge* is shared among the people involved through *socialization*.
- The objective of the documentation concept is to support the *externalization* of knowledge. In the course of creating a documentation using the concept, *tacit knowledge* is codified as *explicit knowledge*.
- By externalizing knowledge, the documentation concept facilitates the easy sharing and *combination* of knowledge. This can lead to new knowledge.
- The explicit knowledge generated by documentation and combination can be acquired by individuals as *tacit knowledge* through *internalization*. This also applies to those individuals who were not involved in the documented ontology creation. Thanks to externalization, they also have access to the previously tacit knowledge.

In these ways, the documentation concept contributes not only to the *preservation of existing knowledge*, but also to the *creation of new knowledge* and its *distribution within the organization*.

2.3 Ontology

In the context of this thesis the term ***ontology*** is defined according to the understanding in computer science. This differs from the original meaning used in the branch of philosophy called *metaphysics*. There, Ontology (Greek *ontologia*) refers to the discipline that deals with 'that which exists' (Greek *ontos*). In computer science the term has been adopted "to mean a formal description of a domain of knowledge, in terms of the entities within it, and their relationships" (Bawden and Robinson, 2012, p. 114). (Cf. Bawden and Robinson, 2012, p. 114; Weller, 2014, p. 209)

The definition of STUDER et al. is used as the working definition.

Working definition:

"An **ontology** is a formal, explicit specification of a shared conceptualisation." (Studer et al., 1998, p. 184)

It is a combination of two definitions given by GRUBER and BORST. The former defines ontology as "an explicit specification of a conceptualization" (Gruber, 1993, p. 199). The latter modifies GRUBER's definition and defines ontology as "a formal specification of a shared conceptualization" (Borst, 1997, p. 12). The combined definition is selected because it is suitable for the application and is cited most frequently alongside GRUBER's initial definition (Cf. Gómez-Pérez et al., 2004, p. 6; Staab and Studer, 2009, p. VIII; Stuckenschmidt, 2011, p. 22)[11]. In addition, it is concisely explained by the authors:

> "*A 'conceptualisation' refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. 'Explicit' means that the type of concepts used, and the constraints on their use are explicitly defined. For example, in medical domains, the concepts are diseases and symptoms, the relations between them are causal and a constraint is that a disease cannot cause itself. 'Formal' refers to the fact that the ontology should be machine readable, which excludes natural language. 'Shared' reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group.*" (Studer et al., 1998, p. 184)

In order to enable a better understanding of the following chapters, ontologies are now explained in more detail.

[11] According to the *Web of Science Core Collection*, the publication by STUDER et al. was quoted 1,211 times, GRUBER's publication 5,000 times (as of 2018-12-14).

2.3.1 Ontologies as controlled vocabulary

Based on the previously presented working definition, ***classifications***, ***taxonomies*** and ***thesauri*** can also be understood as ontologies, albeit as simple ones (Cf. Bawden and Robinson, 2012, p. 114). If one applies a strict information-scientific distinction to the different types of controlled vocabularies, then ontologies are only those vocabularies which go beyond the capabilities of a thesaurus.

Thesauri allow the use of a limited set of relation types. This generally includes relations of *equivalence* (synonyms and quasi-synonyms), *hierarchical* relations (hypernymy/hyponymy and meronymy) and *associative* relations (unspecified connections of concepts) (Cf. Peters and Weller, 2008, pp. 100–101). In any case, the available relations are predefined, for example by standards such as ISO 25964[12] and DIN 1463[13]. Ontologies allow to add *self-defined relations* and *axioms*. With the help of so-called *reasoners*, automatic inferences can be drawn based on the formal definitions of concepts. The expressiveness of an ontology is determined by the chosen *ontology language* (see section '2.3.3 Ontology languages'). (Cf. Weller, 2014)

This thesis is not limited to the creation of controlled vocabularies that exploit the full potential of ontologies. However, it is assumed that the capabilities of taxonomies and thesauri are usually exceeded. Nevertheless, the proposed documentation concept in reduced form is also applicable to the creation of less expressive vocabularies.

---

[12] International Organization for Standardization (2011) *ISO 25964-1:2011: Information and documentation — Thesauri and interoperability with other vocabularies — Part 1: Thesauri for information retrieval.*
International Organization for Standardization (2013) *ISO 25964-2:2013: Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies*.

[13] Deutsches Institut für Normung (1990) *OENORM DIN 1463-1:1990-10-01: Guidelines for the establishment and developement of thesauri; monolingual thesauri.*
Deutsches Institut für Normung (1990) *DIN 1463-2:1993-10: Guidelines for the establishment and developement of thesauri; monolingual thesauri* (withdrawn).

2.3.2 Structure and components

The fundamental elements of an ontology are ***concepts***. The understanding of the term *concept* corresponds to its meaning in documentation science. It is any unit of thought, i.e. a mental image formed by generalization (Cf. Wersig and Neveling, 1976, p. 56). It contains what is typical for a group of similar objects, similar processes, similar ideas etc. (Cf. Gaus, 2005, p. 57). Concepts can refer to *abstract* entities (e.g. intentions, beliefs, feelings) or *specific* entities (e.g. people, computers, tables). They are usually formally represented by ***classes***. Classes are organized hierarchically, i.e. as taxonomies (Cf. Gómez-Pérez et al., 2004, p. 11; Weller, 2014, p. 207). The elements of a class are called ***instances*** or individuals. The characteristics of concepts as well as the relations between concepts are represented by so-called ***properties***. Additional axioms can be modeled using ***restrictions***. (Cf. Weller, 2014, p. 207)

The terminology of an ontology is provided by the classes and properties. This ***ontology model*** contains the knowledge about the vocabulary and the connections within the knowledge domain. Consequently, the so-called ***TBox*** (*Terminology Box*) claims to be universally valid. ***Facts***, by contrast, are represented by instances in the so-called ***ABox*** (*Assertion Box*). The distinction between TBox and ABox is usually helpful, but not always possible and necessary. (Cf. Weller, 2014, p. 213)

Figure 4 illustrates an excerpt of an ontology for the pharmaceutical domain. The upper part of the graphic shows the TBox, the lower part the ABox. TBox and ABox together represent the ontology as a whole.

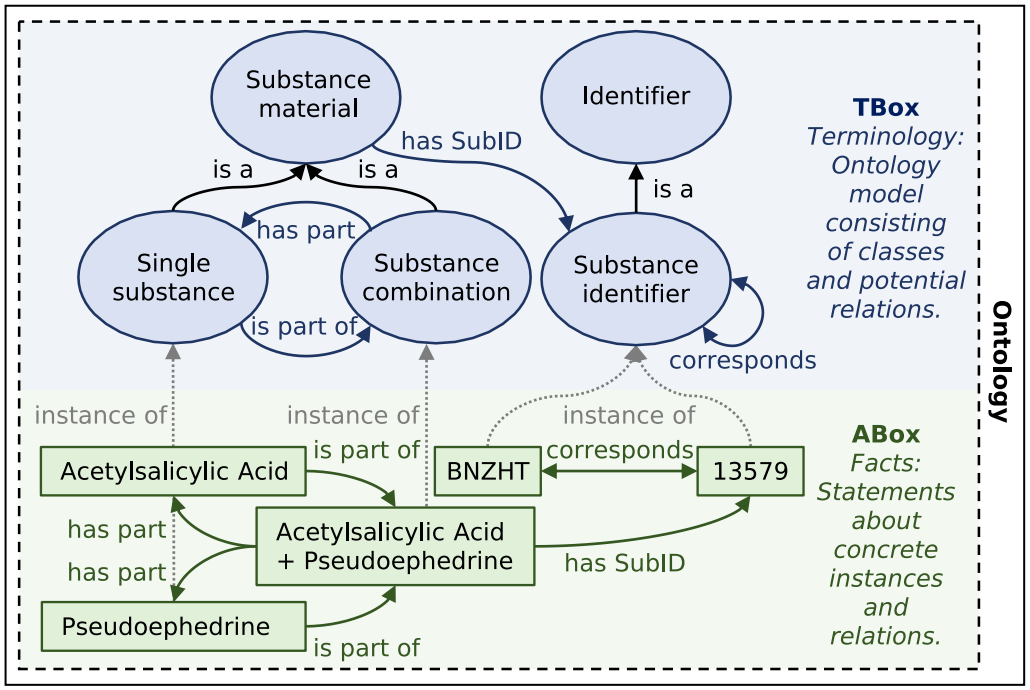

Figure 4: Diagram of an ontology excerpt showing classes, properties, instances as well as the distinction between TBox and ABox

2.3.3 Ontology languages

Ontologies are expressed using machine-readable ***ontology languages***. There are several languages that have evolved over time. A single language or a set of languages can be used to implement an ontology. The choice which language to use depends on several factors. They relate, among others, to expressivity, inference mechanisms, available tools, intended application and interoperability. The two main dimensions for the description of ontology languages are *knowledge representation* and *reasoning mechanism*. (Cf. Gómez-Pérez et al., 2004, pp. 199–204)

A fundamental issue is the trade-off in knowledge representation and reasoning, which LEVESQUE and BRACHMAN describe as follows:

> "*The problem is essentially that it can be more difficult to reason correctly with one representational language than with another and, moreover, that this difficulty increases as the expressive power of the language increases. There is a tradeoff between the expressiveness of a representational language and its computational tractability.*" (Levesque and Brachman, 1985, p. 42)

For the objective of this thesis it is not decisive which formal language is used. Therefore, only two popular examples from the field of the *Semantic Web* are briefly presented here: RDFS and OWL. For a better understanding of these two ontology languages, RDF is explained first as one of the basic standards for RDFS. For further explanations on ontology languages please refer to the corresponding literature[14].

The ***Resource Description Framework*** (RDF) is a standard published by the W3C. It is a "framework for expressing information about resources", which "can be anything, including documents, people, physical objects, and abstract concepts" (Schreiber and Raimond, 2014). The RDF data model is based on simple statements consisting of `<subject> <predicate> <object>`, which therefore are also called ***triples***. Resources are addressed using an *Internationalized Resource Identifier* (IRI)[15], which is the internationalized form of a *Uniform Resource Identifier* (URI). Subsets of URIs are not only *Uniform Resource Locators* (URLs) that identify resources on the web, but also *Uniform Resource Names* (URN) that identify resources outside the web (see example below). A set of connected triples is called a ***graph***. RDF can be expressed by different textual syntaxes, such as the line-based syntax ***N-Triples***[16] or the *Terse RDF Triple Language* called ***Turtle***[17]. (Cf. Schreiber and Raimond, 2014)

---

[14] Literature on ontology languages (not exhaustive):
- Chapter 4 'Languages for Building Ontologies' in Gómez-Pérez, Asunción; Fernández-López, Mariano; Corcho, Oscar (2004) *Ontological Engineering: With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web*, London, Springer-Verlag London Limited.
- Part I 'Ontology Representation Languages' in Staab, Steffen; Studer, Rudi (2009) *Handbook on Ontologies*, 2nd edn, Berlin, Heidelberg, Springer Berlin Heidelberg.
- Chapter 4 'Ontologiesprachen' in Stuckenschmidt, Heiner (2011) *Ontologien: Konzepte, Technologien und Anwendungen*, 2nd edn, Berlin, Heidelberg, Springer Berlin Heidelberg.

[15] Duerst, M.; Suignard, M. (2005) *RFC 3987: Internationalized Resource Identifiers (IRIs)* [Online], The Internet Society. Available at https://www.rfc-editor.org/rfc/rfc3987.txt (Accessed 2018-12-15).

[16] Beckett, David (2014) *RDF 1.1 N-Triples: A line-based syntax for an RDF graph* (W3C Recommendation 25 February 2014) [Online], W3C. Available at https://www.w3.org/TR/2014/REC-n-triples-20140225/ (Accessed 2018-12-15).

[17] Beckett, David; Berners-Lee, Tim; Prud'hommeaux, Eric; Carothers, Gavin (2014) *RDF 1.1 Turtle: Terse RDF Triple Language* (W3C Recommendation 25 February 2014) [Online], W3C. Available at https://www.w3.org/TR/2014/REC-turtle-20140225/ (Accessed 2018-12-15).

For example, a triple could look like the following:

```
<Essays In Science>       <has Creator>     <Albert Einstein> .
```

This corresponds to the natural language statement:

(The book) 'Essays In Science' was created by Albert Einstein.

Identification of the triple components using IRIs:

```
<urn:isbn:6000614454>
                  <http://purl.org/dc/terms/creator>
                                              <http://d-nb.info/gnd/118529579> .
```

A triple from the exemplary ontology excerpt in figure 4 above is the following:

```
<Acetylsalicylic Acid + Pseudoephedrine>          <has Sub ID>       <13579> .
```

RDF on its own is not enough to create a vocabulary. For this purpose the W3C standard ***RDF Schema*** (RDFS) is employed, which "provides a data-modelling vocabulary for RDF data" (Brickley and Guha, 2014). RDFS in turn can be used to design vocabularies which allow describing data using standardized terms. In a nutshell, RDFS facilitates data exchange and machine inferencing. It is a formal language with which simple ontologies can be implemented. (Cf. Yu, 2014, pp. 123–126)

The currently most popular ontology language is the *Web Ontology Language* (OWL) (Cf. Yu, 2014, p. 169). OWL extends RDFS and offers a higher expressiveness, which allows stronger reasoning (Cf. Yu, 2014, p. 169). The latest version is OWL 2, which is a superset of its predecessor OWL. This means that all ontologies created with OWL are still valid (Cf. Yu, 2014, p. 172). Both versions provide different language species which "offer different levels of tradeoff between expressiveness and efficiency, and therefore offer more choices to the users" (Yu, 2014, p. 217). Since OWL is the most common ontology language, it is used in the following chapters for example passages of formal expressions.

2.3.4 Typology

Ontologies can be categorized into different ***ontology types***. The distinction can be made on the basis of different dimensions (Cf. Guarino, 1997, p. 144; van Heijst et al., 1997, p. 192). At this point GUARINO's distinction between top-level ontology, domain ontology, task ontology and application ontology is applied (Cf. Guarino, 1997, p. 145).

According to GUARINO, ***top-level ontologies*** "describe very general concepts like space, time, matter, object, event, action, etc., which are independent of a particular problem or domain" (Guarino, 1997, p. 145). They are also referred to as *generic ontologies* (Cf. van Heijst et al., 1997), *upper ontologies* (Cf. Guarino, 1999), *foundational ontologies* (Cf. Gangemi et al., 2002) or *super theories* (Cf. Borst et al., 1997). Top-level ontologies are "usually equipped with a rich axiomatic layer" (Schmidt et al., 2016, p. 13) and can act as "semantic bridges supporting very broad semantic interoperability between ontologies" (Schmidt et al., 2016, p. 13).

Domain ontologies and ***task ontologies*** specialize concepts defined in the top-level ontology to represent the knowledge "related to a generic domain (like medicine, or automobiles) or a generic task or activity (like diagnosing or selling)" (Guarino, 1997, p. 145).

Finally, ***application ontologies*** contain "all the definitions that are needed to model the knowledge required for a particular application" (van Heijst et al., 1997, p. 193). Depending on the domains and task ontologies, they describe the "roles played by domain entities while performing a certain activity" (Guarino, 1997, p. 145). Usually, application ontologies are not reusable without adjustments (Cf. van Heijst et al., 1997, p. 193).

Figure 5 shows the presented ontology types and their dependencies. In the direction of top-level ontology, the degree of abstraction increases, whereas in the direction of application ontology, the degree of specification increases.

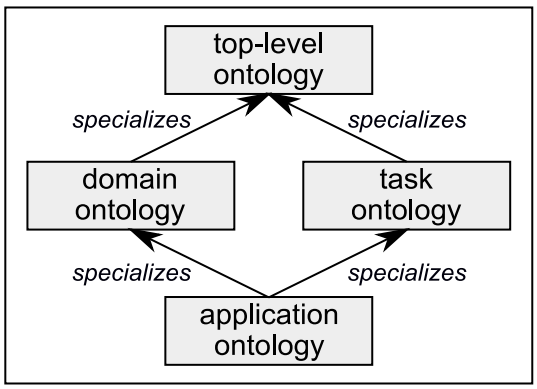

Figure 5: Types of ontologies and their dependencies according to GUARINO (Guarino, 1997, p. 145)

Apart from the classification of ontologies, there are approaches to split ontologies, including top-level ontologies, into ***modules*** (Cf. Teymourian et al., 2010) or ***libraries*** (Cf. van Heijst et al., 1997). Others in turn say "it seems […] reasonable, at least in theory, to have a single, unified top-level ontology" (Guarino, 1997, p. 145).

2.3.5 Ontology engineering methodologies

Various approaches to the structured development of ontologies are proposed in literature. These are called ***ontology engineering methodologies***. The term ***methodology*** is derived from the field of software engineering. The *Software & Systems Engineering Standards Committee of the IEEE Computer Society* explains the term as follows:

> *"A methodology specifies the process to be executed, usually as a set of related activities, tasks and/or techniques, together with the work products that must be manipulated (created, used or changed) at each moment and by whom, possibly including models, documents and other inputs and outputs."* (Software & Systems Engineering Standards Committee of the IEEE Computer Society, 2010, p. 216)

However, this description should not be applied strictly to the field of ontology development, since not all ontology engineering methodologies cover all of the aspects mentioned. Accordingly, FERNÁNDEZ-LÓPEZ and GÓMEZ-PÉREZ attested in 2002: "in the ontological engineering field, the ontology-building process is a craft rather than an engineering activity" (Fernández-López and Gómez-Pérez, 2002, p. 130). There is no common methodology, but many different ones that have been proposed over the last two decades (Cf. Forbes et al., 2018, p. 27). Appendix 'A: Ontology engineering methodologies' contains a selection of known ontology engineering methodologies. In chapter '4 Analysis of ontology engineering methodologies', two of these methodologies are analyzed in order to include the results in the documentation concept.

In addition to proposals of new methodologies, literature also provides comparisons which analyze the different methodologies with regard to different aspects, including for instance Corcho et al. (2003), Fernández-López (1999), Fernández-López and Gómez-Pérez (2002), Iqbal et al. (2013), Simperl and Luczak-Rösch (2014). In the following, two aspects of ontology development are selected for explanation because they are relevant for the objective and understanding of this thesis.

The first aspect refers to the approach that determines the modeling direction or as SIMPERL and LUCZAK-RÖSCH call it, the "Strategy for identifying concepts" (Simperl and Luczak-Rösch, 2014, p. 117). Based on a categorization by USCHOLD and GRUNINGER, a distinction is made between three approaches: *top-down*, *middle-out* and *bottom-up* (Cf. Uschold and Grüninger, 1996).

When applying the ***top-down approach***, the direction of procedure is from general to particular. This means that first a few very general concepts are identified, then they are specialized. The ***bottom-up approach*** follows the opposite direction, i.e. from particular to general. First many very specific concepts are identified, usually by analyzing data, then they are generalized. The ***middle-out approach*** is also called the *combined approach*, since both directions of the two aforementioned approaches are applied. First the most important concepts are identified, then they are generalized and specialized. (Cf. Nicola et al., 2009, p. 268; Uschold and Grüninger, 1996, pp. 107–108; Uschold and King, 1995, p. 8)

USCHOLD, GRUNINGER and KING argue that the middle-out approach is the most promising (Cf. Uschold and Grüninger, 1996, pp. 107–108). However, which approach is most appropriate depends on the planned application, the available explicit knowledge and other conditions influencing ontology development (Cf. Simperl and Luczak-Rösch, 2014, p. 117). Therefore, no approach is treated preferentially in this thesis. All three approaches are considered during the development of the documentation concept.

The second aspect, which is discussed in more detail here, relates to the roles involved during ontology development. A role indicates which task(s) a person involved undertakes during the process of ontology development. One person may undertake several roles. The roles involved may differ from methodology to methodology. In addition, similar roles are designated differently.

Common role designations are *domain expert*, *knowledge engineer*, *knowledge worker*, *ontology engineer* and *ontology user* (Cf. Gómez-Pérez et al., 2004, pp. 107–197; Simperl and Luczak-Rösch, 2014, p. 119). These designations are used differently as the following example illustrates: While the *Unified Process for Ontology Building* (UPON) classifies *ontology engineers* as *knowledge engineers* and *domain experts* (Cf. Nicola et al., 2005, p. 656), the DILIGENT methodology distinguishes between *domain experts*, *ontology engineers*, *knowledge engineers* and *users* (Cf. Pinto et al., 2009, p. 158). In the UPON, the role *knowledge engineer* is a subset of the role *ontology engineer*. In the DILIGENT methodology, by contrast, *knowledge engineer* and *ontology engineer* are two separate roles.

In the context of this thesis, a distinction is made between only two roles, as in the UPON: ***knowledge engineer*** (KE) and ***domain expert*** (DE). On the one hand, this is done to keep complexity as low as possible. On the other hand, these two roles correspond best to personal experience as in practice barely any other roles are employed. In the case of a further distinction, the additional roles would usually be taken over by the same people who hold the initial roles. This makes further distinction less valuable in practice. In return, it is accepted that the roles are less specific. The following working definitions describe the understanding of the roles applied in this thesis.

Working definition

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.

Working definition

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.

Looking at DEs and KEs in collaboration, DEs provide their domain knowledge so that KEs can build ontologies based on it. In other words, DEs are knowledge bearers who hold tacit knowledge and can provide information about explicit knowledge. KEs elicit this knowledge from DEs using various methods, e.g. via interviews. Based on this externalized tacit knowledge as well as knowledge from other sources, e.g. explicit knowledge as a result of a literature research, KEs create conceptual knowledge models. Finally, KEs express those as ontology using a suitable ontology language. (Cf. Simperl and Luczak-Rösch, 2014, pp. 103–104)

This process involving KEs and DEs will be analyzed and described in more detail in chapter ‘4.1 Unified Process for Ontology Building UPON’.

2.4 Process

This thesis refers to *the process of creating ontological knowledge models*. No established definition is applied here. Instead, this thesis outlines a process that constitutes in general only one part of those processes covered by ontology engineering methodologies.

The process considered in this thesis starts with the identification of concepts and ends with the first version of a formal model that meets all previously defined objectives. Usually, this is accomplished when the previously defined *competency questions* (CQs) can be answered by the ontology. If new requirements or CQs arise or the knowledge represented is outdated, this is no longer part of the process.

The process described here focuses on creating a new ontology. It must be distinguished from processes that focus on the evolution, restructuring, mapping or merging of existing ontologies. The simple creation of instances by means of existing vocabularies is excluded as well. Based on these distinctions, the title of the thesis contains the term *creating* instead of *developing* and explicitly refers to *ontological knowledge models*.

The steps listed below may take place before the start of the process:

- **Identification of use cases**. One or more use cases in the broader sense are defined. Usually, use cases result from existing needs.
- **Identification of competency questions**. Collection and documentation of questions to be answered by exploring and querying the ontology. These CQs are initially expressed informally at the conceptual level, not as formal queries (Cf. Grüninger and Fox, 1995, p. 3; Nicola et al., 2009, p. 261).
- **First research on existing vocabularies**. For instance, to answer the question of whether an existing ontology can be reused without need for clarification of concepts.

- **Ontology specification**. A document that, depending on its granularity and the level of knowledge, may contain statements about aspects such as: domain (subject area), discourse (mindset; perception of the domain), purpose and scope, target groups, level of formality, ontology language, division into sub-ontologies, pre-collected terms, existing vocabularies, sources of knowledge, applied methodology or modeling approach, design guidelines, integration (used software, technologies, infrastructure). (Cf. Fernández et al., 1997, p. 36; Suárez-Figueroa and Gómez-Pérez, 2012)

Figure 6 illustrates the understanding of the process in context of this thesis.

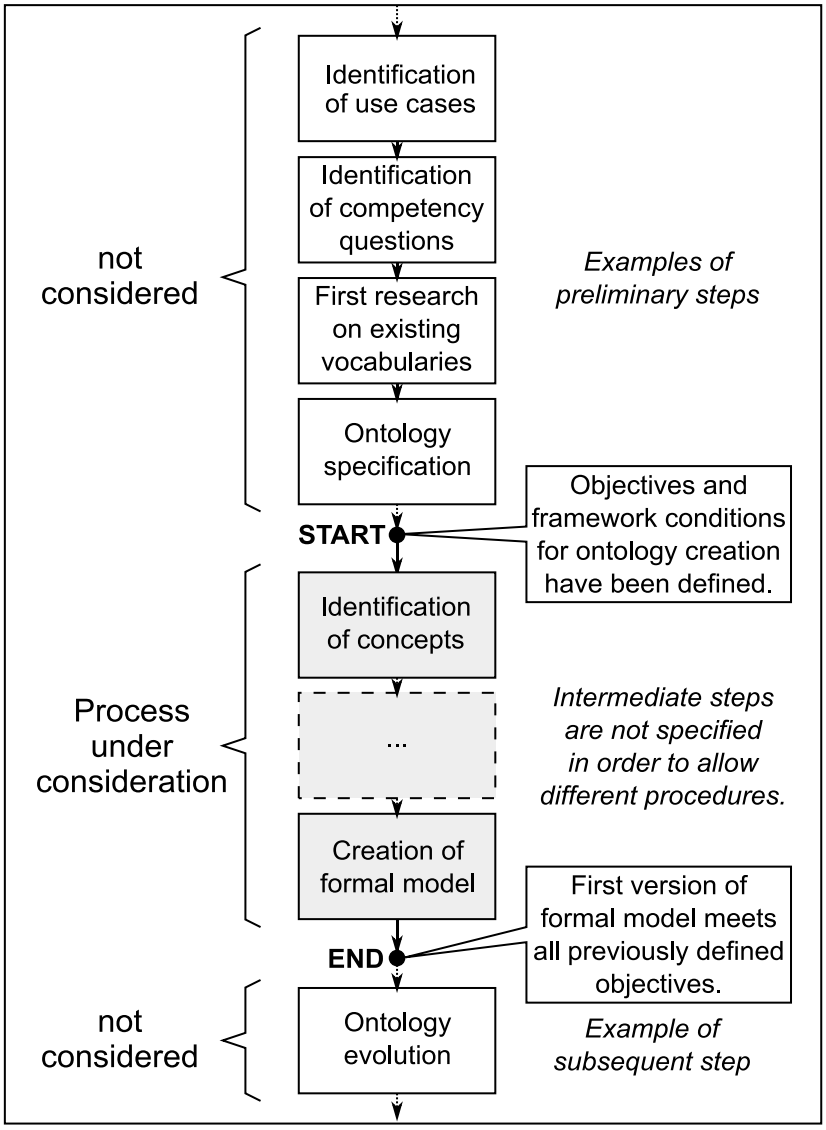

Figure 6: Graphical representation of the process of creating ontological knowledge models in context of this thesis

It is assumed that the ontology is created by one or more KEs in collaboration with one or more DEs. Since this is a matter of exchanging implicit knowledge, the interaction between humans is particularly relevant for the documentation concept. Nevertheless, it is not excluded that a KE creates an ontology without direct exchange with humans, but only based on explicit data, information and knowledge (e.g. raw data, annotated tables or literature). In the latter case, documentation is needed as well since the tacit knowledge of the individual KE must be externalized.

3 Selective insights into corporate reality

This chapter provides insights into how and under what conditions ontologies are created in companies. Firstly, the results of a small survey are presented, which has been carried out in context of this thesis. Subsequently, so-called *assumption personas* are created based on the survey results and personal experiences. They describe archetypal KEs and DEs. In this way, the assumptions are presented explicitly and in an illustrative manner. Finally, findings are derived which must be taken into account when creating the documentation concept.

3.1 Preliminary survey

In the run-up to this thesis a small survey based on two different questionnaires has been conducted. One of the questionnaires is addressed to KEs, the other to DEs. KEs have been asked about their approach to ontology creation and the documentation they create meanwhile. DEs have been asked about their motivation and expectation regarding their contribution to the creation of ontologies. In both cases, the actual state has been enquired concerning the day-to-day work. The desired ideal state has not been asked for. Both questionnaires consist of five questions each and can be found in the appendix 'B: Questionnaires and data of preliminary survey'.

In total, three groups have been surveyed: on the one hand, KEs and DEs of a global life science company based in Germany, on the other hand, DEs of an international working group. The latter are referred to as external DEs from the perspective of the company. Figure 7 illustrates the distribution of the two questionnaires to these three groups.

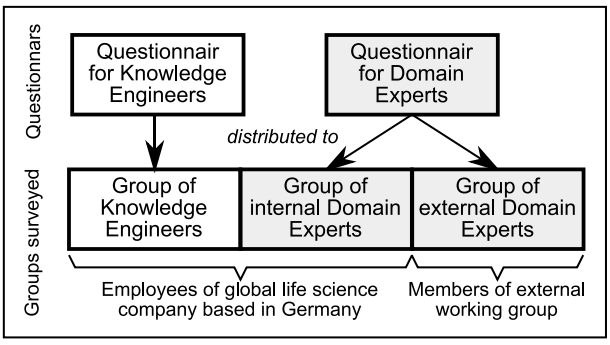

Figure 7: Questionnaires and their distribution to surveyed groups in the course of the preliminary survey

The two questionnaires are designed to be completed quickly and are therefore relatively simple. They have been distributed and completed electronically. The questionnaires have been sent targeted to people known as KE or DE. The response rates are 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 completion of the questionnaires and thus the participation in the survey has been voluntary.

Due to the targeted distribution and voluntary participation, it can be assumed that only those people who have a basic interest in ontologies have taken part in the survey. Besides, only the current situation has been enquired, not the desired target state. So the results of the survey 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.

With n = 13 for KEs, n = 11 for internal DEs and n = 5 for external DEs, the sample sizes are small for both roles. The survey results therefore do not claim to represent the entirety of the KEs and DEs in the company or the external working group. However, they provide valuable insights into corporate reality.

3.1.1 Survey results: Knowledge engineers

In the following, the survey results for the group of KEs are presented using charts and tables. Thereby first findings are derived. The questions and corresponding answer options are part of the charts. Hence, they are not always repeated in their entirety in the text. The survey data can be found in the appendix 'B: Questionnaires and data of preliminary survey'.

Figure 8 shows that all three of the modeling approaches presented in chapter '2.3.5 Ontology engineering methodologies' – *top-down*, *bottom-up* and *middle-out* – are practiced. The documentation concept should therefore take into account the characteristics of all three approaches.

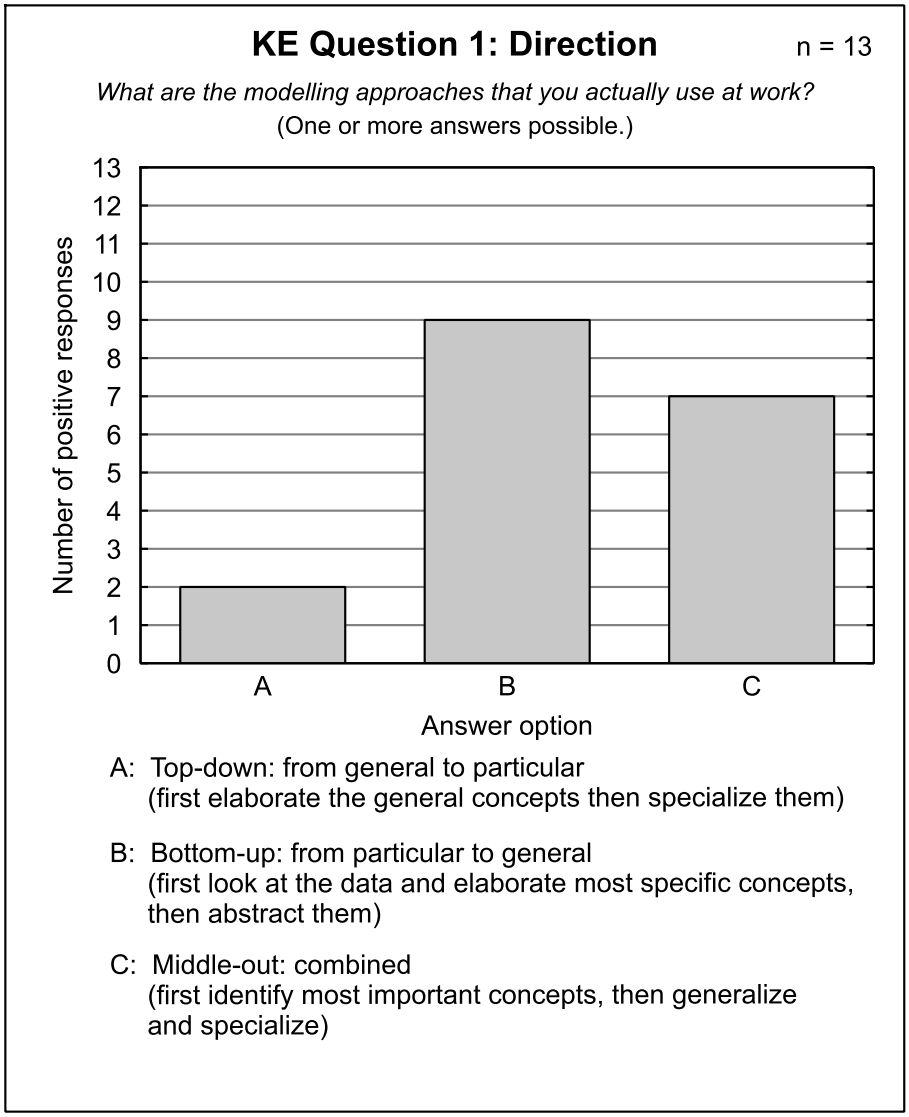

Figure 8: Survey results knowledge engineers, question 1: Direction

With 9 out of 18 responses, the *bottom-up approach* is the most common, followed by the *middle-out approach* with 7 out of 18 responses. This is not surprising since, as described in the introduction, FAIR data and data integration are major objectives triggering the creation of ontologies (see chapter '1 Introduction'). The strategies and knowledge sources typical for the bottom-up approach are therefore of particular relevance for the documentation concept. This means that information on *data sets, data storage locations* and *software applications* as well as information on their *accessibility* and *confidentiality* must be represented.

Regarding the main goal pursued by KEs, 69.2 % of the respondents are aiming for *fast results* (see figure 9). In return, they accept less perfect knowledge models. Under these circumstances, it is to be expected that there is little willingness or opportunity to create extensive documentation. Accordingly, it is necessary to offer a level of granularity that enables fast and yet effective documentation. For the minority of the KEs, who invest more time in *resilient results*, it is assumed that a more detailed documentation concept is appreciated.

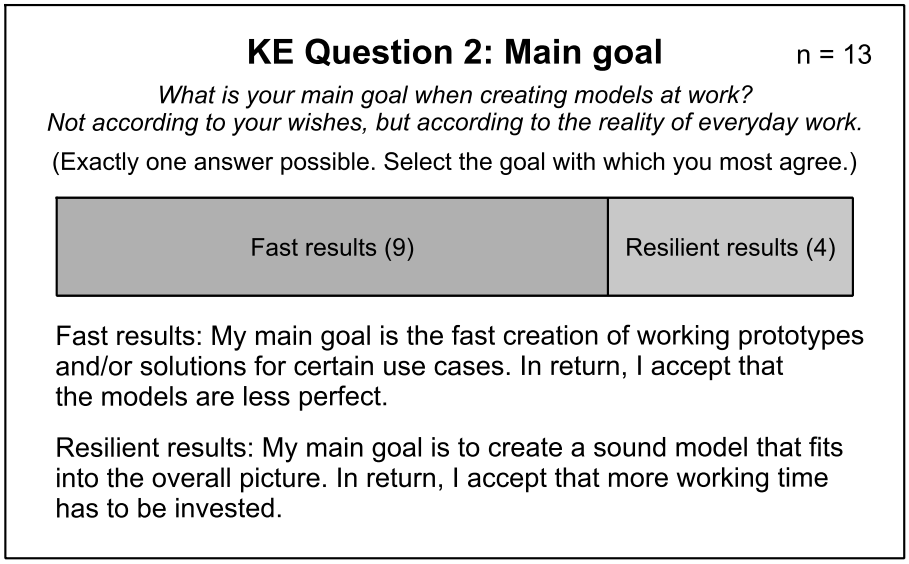

Figure 9: Survey results knowledge engineers, question 2: Main goal

Figure 10 and table 1 show what information the KEs consider relevant for each concept. The first observation is that each of the offered answer options is represented among the results (see figure 10). In addition, further aspects have been added by the respondents (see table 1). The documentation concept must cover all the proposed and added aspects.

The most prominent answers are *definition or explanation*, *alternative labels*, *related data sources* and *related vocabularies and standards*. The most frequent choice has been *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.

The importance of understanding is also shown by the fact that *related vocabularies or standards* have been selected second most frequently. Existing vocabularies and standards may contain helpful explanations and definitions. Concepts that are already formally described in vocabularies can be reused in the sense of *Linked Data*. Technical standards can be consulted on modeling or implementation issues.

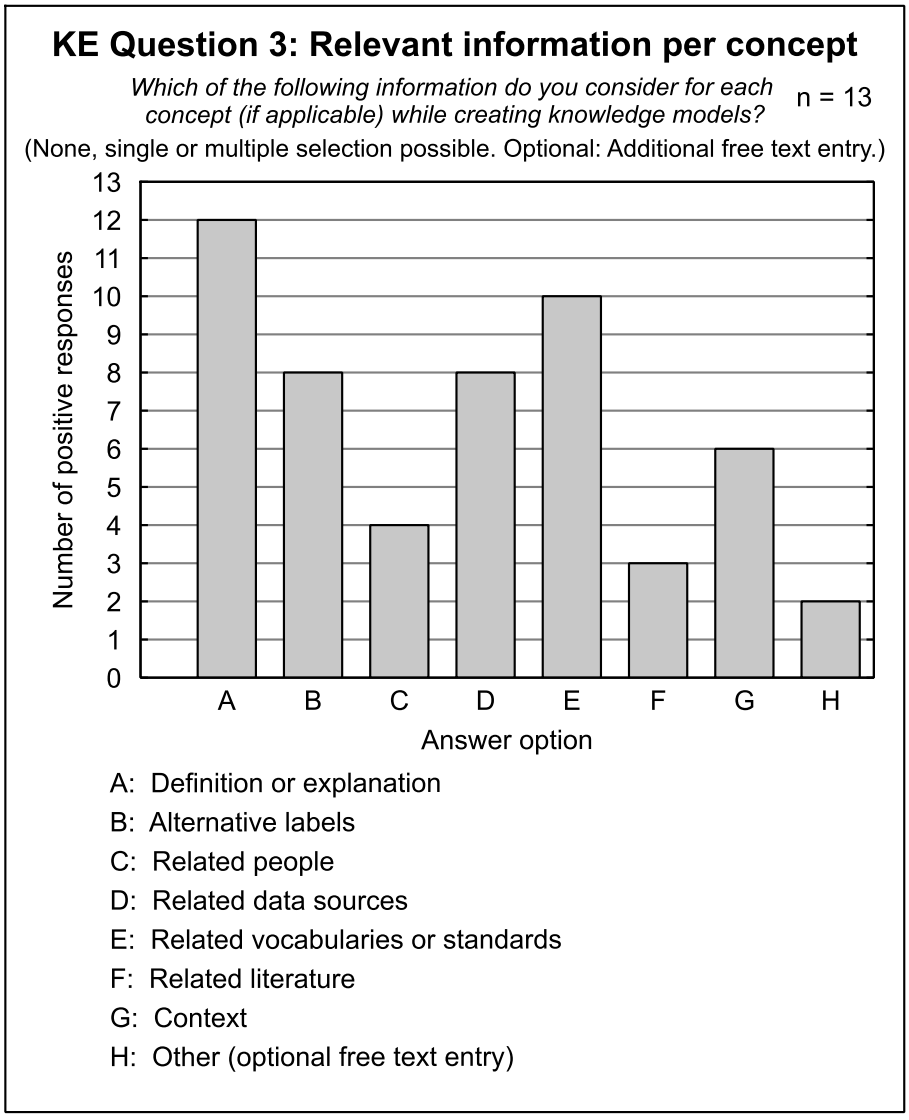

Figure 10: Survey results knowledge engineers, question 3: Relevant information per concept

Optional free text entry for H: Other
The use case around the model
Identifiers

Table 1: Survey results knowledge engineers, question 3: Relevant information per concept (free text entries)

One of the advantages of controlled vocabularies is that they address the problem of equivalence. The results for *alternative labels* indicate the relevance of capturing synonyms and quasi-synonyms (see figure 10). Abbreviations are also expected to be relevant as they are part of everyday work in large companies. For example, the pharmaceutical glossary, one of the abbreviation lists in the company surveyed, contains over 750 entries. The survey results do not provide any information on the relevance of multilingual labels. It is assumed that this depends on the company and the use case. If, for example, the company specifies a uniform business language, it may not be necessary to capture multilingual labels. On the other hand, dealing with multilingualism is another strength of ontologies. Therefore, the documentation concept should consider this aspect.

Related data sources are considered relevant by 8 of the 13 respondents (see figure 10). This fits with the results for *KE Question 1: Direction*, where the *bottom-up approach* has been the most frequently chosen option (see figure 8). Accordingly, *related literature* has been considered less relevant (see figure 10). Literature is a knowledge source that is rather used in the *top-down approach*.

Besides, 4 of the 13 respondents have selected the option *related people* (see figure 10). These related people may be *domain experts* who are knowledgeable about a subject. This also includes *data owners*, *data stewards* or *process owners*. On the other hand, related people may also include *knowledge engineers* who, for instance, are proficient in modeling a specific problem.

Aspects added by the surveyed KEs are *identifiers* and *the use case around the model* (see table 1). *Identifiers* allow to identify and reference individual concepts within a set of concepts, even when the associated label changes. This aspect is a relevant contribution that must be considered when creating the documentation concept. This also applies to information regarding the *use case*. For example, when it comes to specifying for which *use case* and for which *competency question* a certain concept is relevant.

The results for *KE Question 4: Timepoint of documentation* shown in figure 11 confirm the initially stated observation that KEs often documented too late (see chapter '1.1 Objectives'). Only two of the respondents start the documentation *before* creating the formal model. The majority documents *while* or *after* the creation of the formal model. One of the KEs surveyed does not even create a documentation at all. This means that, usually, all knowledge that is acquired and generated is only externalized during and after the creation of the formal model. Until then it is bound to individuals and difficult to access.

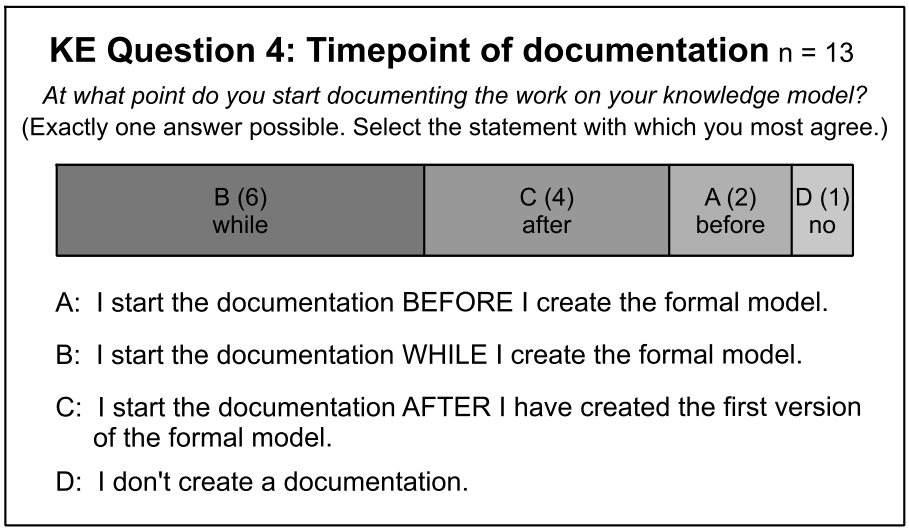

Figure 11: Survey results knowledge engineers, question 4: Timepoint of documentation

Given the fact that documentation usually takes place in connection with the formal model, one might think it at least takes place formally. Annotation properties can be used for this purpose. However, as illustrated in figure 12, only 2 of the 13 KEs use this option to document the knowledge exchange with DEs. The other 11 KEs document the insights they acquire by exchanging with DEs informally. The documentation concept addresses especially the narrow majority of 6 KEs that documents unstructured. The objective is to encourage structured documentation. A positive aspect is that 5 KEs already document the exchange in a structured way. In that case the objective is to improve the documentation quality if necessary. According to the survey results, all KEs document the exchange with DEs. Nobody has chosen the option "I don't document the exchange at all" (see figure 12). This result is positive but does not fully match the results for *KE question 4: Timepoint of documentation* (see figure 11). There the answer "I don't create a documentation" has been chosen once. The questionnaire must therefore be revised before it is reused in order to avoid any misunderstandings that might lead to inconsistencies in the survey results.

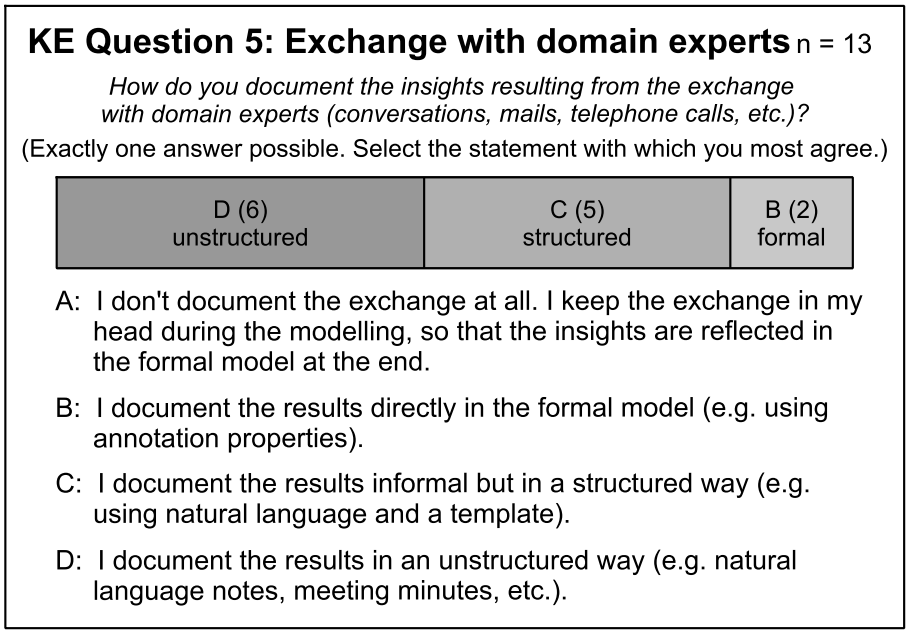

Figure 12: Survey results knowledge engineers, question 5: Exchange with domain experts

3.1.2 Survey results: Domain experts

In the following, the results for both groups of DEs – internal and external DEs – are presented in the same way as previously for the group of KEs. The survey data can be found in the appendix 'B: Questionnaires and data of preliminary survey'.

According to the survey results, the DEs contribute to the creation of ontologies rather because of their intrinsic motivation than because of external influences. Neither the internal nor the external DEs contribute *because they were instructed to do so* or *because apparently everyone does* (see figures 13 and 14). Interpersonal relationships are apparently not a decisive factor either. Only two of the internal DEs (see figure 13) and one of the external DEs (see figure 14) have stated that they are motivated because they *like the people who work on the ontology*. The motivating factor that unites almost all DEs surveyed, 15 out of 16, is that they think *ontologies are a promising technology*. This kind of motivation is appreciated in the context of the documentation concept. It is assumed that convinced DEs are more likely to contribute to documentation than DEs that participate due to external compulsions. In one of the additional free text responses, the intention to pass on domain knowledge to the KEs is even explicitly stated as motivation: "I would like to share my domain knowledge with experts for building ontologies" (see table 2). The other free text entries of the DEs surveyed again reflect the focus on FAIR data and data integration (see tables 2 and 3).

The second major motivation factor is the urge to learn new things, both *about ontologies as a technology* and *about one's own domain in exchange with other DEs* (see figures 13 and 14). Altogether, 10 out of 16 DEs are interested in learning more about ontologies. For the documentation concept, this means that a separation into two documentations by default, one for KEs and one for DEs, is not recommended. Instead, a common documentation should be created that covers both subject-related content and content related to knowledge engineering. In this way, DEs can use the documentation to acquire knowledge about their domain and about ontologies. If the knowledge exchange from KEs towards DEs is not desired by the company, e.g. for efficiency reasons, a separation can be introduced when implementing the documentation concept. For example, it is conceivable that separate views are implemented so that DEs can see components such as alternative labels and definitions, but no components with regard to formalization issues.

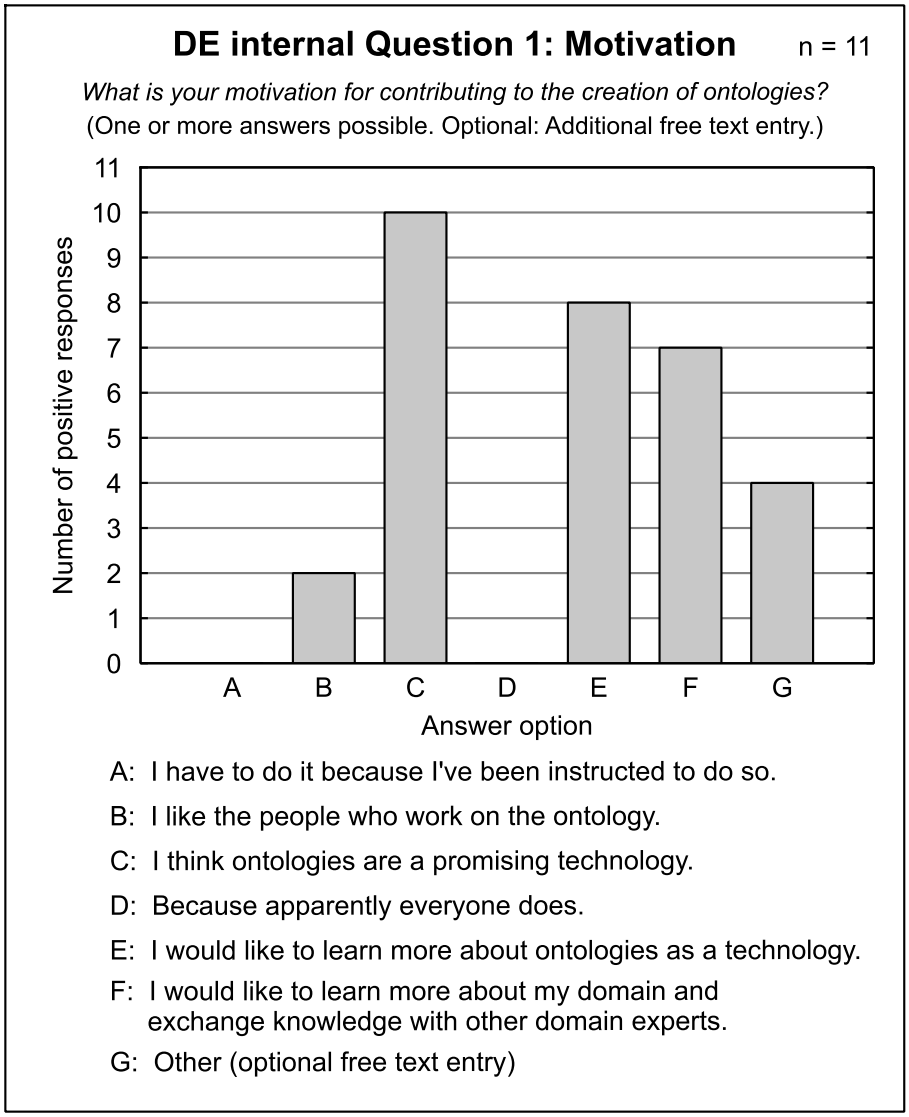

Figure 13: Survey results internal domain experts, question 1: Motivation

Optional free text entry for G: Other
I think they would build an important basis for the very needed data integration.
Best way to achieve digital agenda target of transparency and FAIR data
For integration of clinical studies harmonization of data is essential and ontologies are a promising concept to achieve this harmonization.
I would like to share my domain knowledge with experts for building ontologies.

Table 2: Survey results internal domain experts, question 1: Motivation (free text entries)

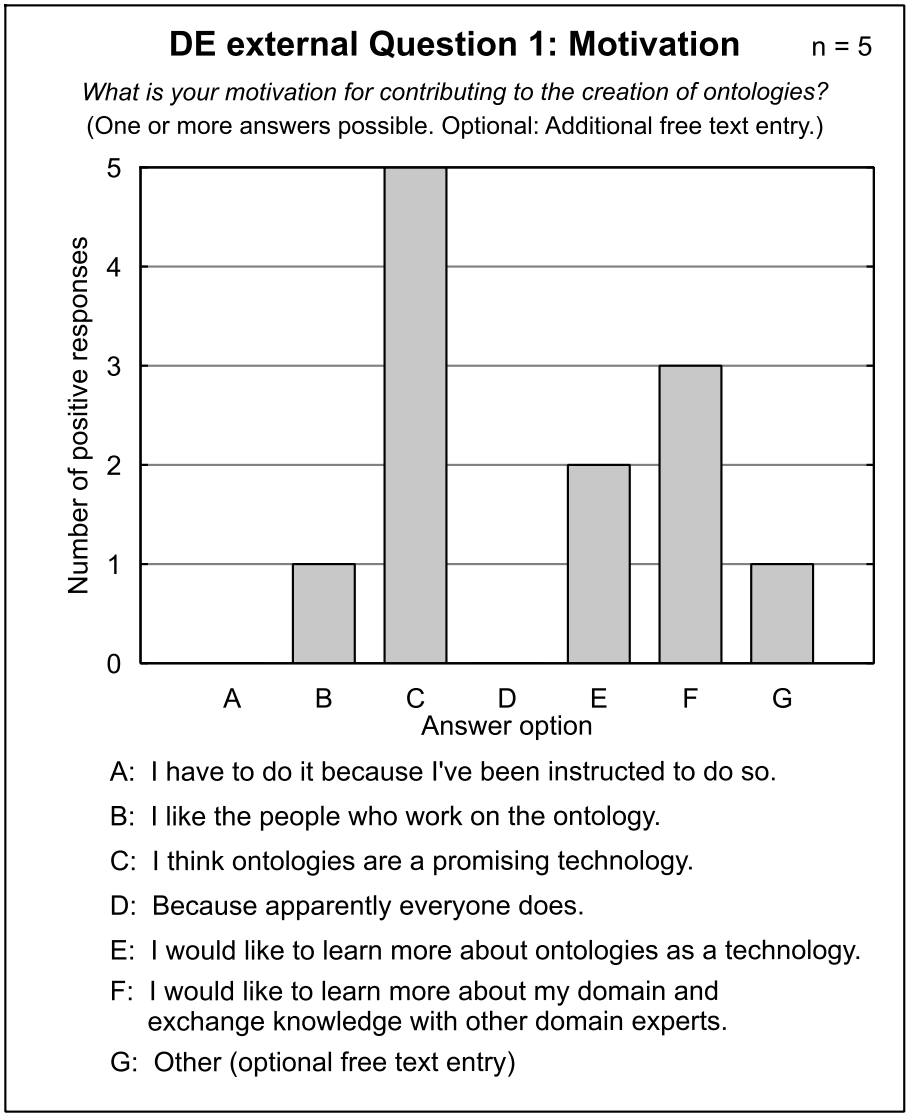

Figure 14: Survey results external domain experts, question 1: Motivation

Optional free text entry for G: Other
Because these models represent the real-world models and facilitate data exchange

Table 3: Survey results external domain experts, question 1: Motivation (free text entries)

Almost all DEs surveyed *expect ontologies to work in the background to improve the interaction between IT systems* (see figures 15 and 16). Once again, the objective of data integration is evident. With regard to the documentation concept, another expectation of the DEs is of interest. More than half of the DEs surveyed, 9 out of 16, *expect to be able to work directly with ontologies to gain knowledge about a domain* (see figures 15 and 16). One of the additional free text entries is: "I expect to receive information or explanation of data, which is currently not available" (see table 4). To meet this expectation, ontologies must provide correct and helpful definitions as well as contextual information. The documentation concept can support this by enabling the recording of such information from the outset. If definitions and context information are gathered, recorded and improved from the start, they can be more easily included in the formal model later.

The other two free text entries in table 4 indicate that DEs expect productive use of ontologies in the distant future rather than in the near future. This fits with the results for *DE Question 3: Current usage of ontologies* shown in figures 17 and 18. None of the DEs surveyed *uses ontologies in a productive way*. 10 of the respondents stated that *ontologies are not part of their daily work as far as they know*. Most of the DEs surveyed *use ontologies within a proof of concept to test whether a productive usage is feasible*. 4 out of 16 DEs *use ontologies to become familiar with the technology*. These results confirm why the focus of this thesis is on the creation of new ontologies (see chapters '1.1 Objectives' and '2.4 Process'). Ontologies are not yet used productively in the groups surveyed. They are currently (2018/2019) being created and tested within proof of concepts. Therefore, the documentation of ontology *evolution* is for the time being less urgent than the documentation of ontology *creation*.

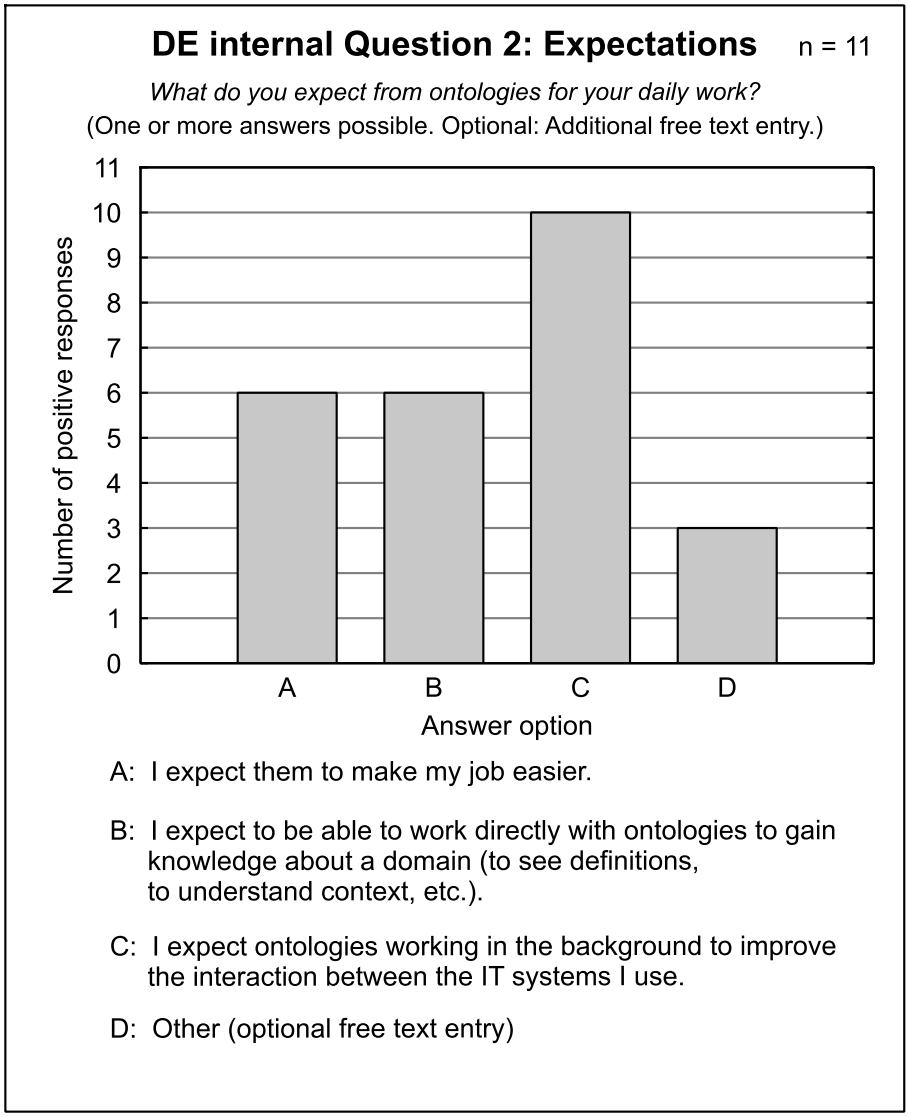

Figure 15: Survey results internal domain experts, question 2: Expectations

Optional free text entry for D: Other
I expect to receive information or explanation of data, which is currently not available.
I expect that it will take much more exploratory work and efforts before ontologies will have a significant impact on our "daily life" as Data Managers, Data Scientists, Data Engineers, etc.
On the long term and in addition with other technologies

Table 4: Survey results internal domain experts, question 2: Expectations (free text entries)

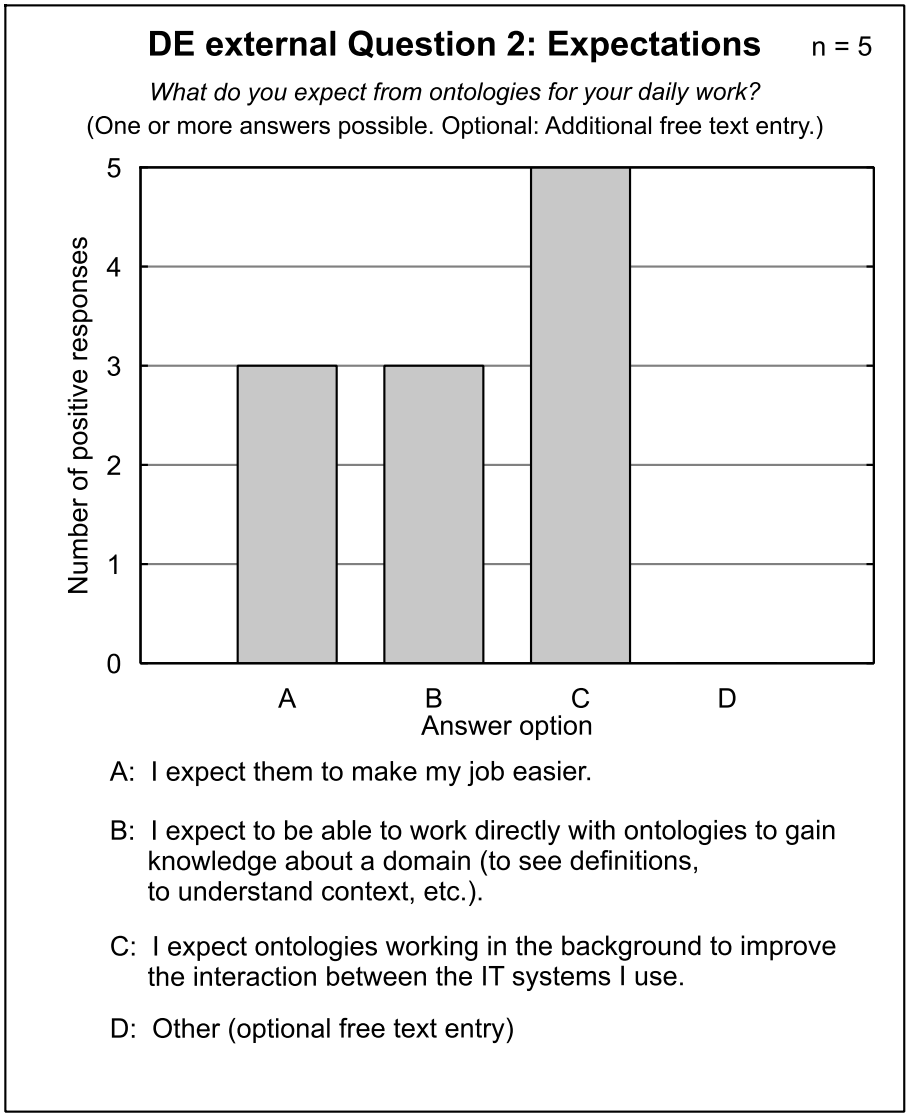

Figure 16: Survey results external domain experts, question 2: Expectations

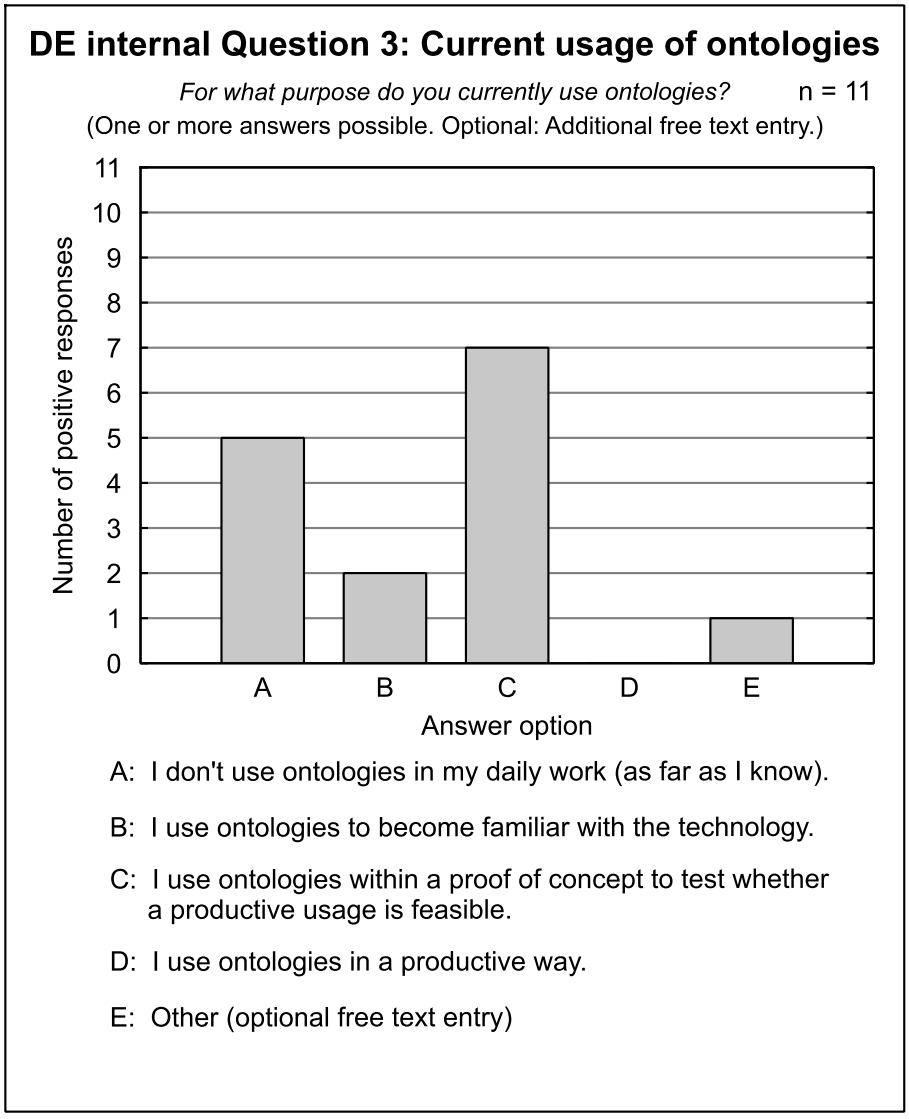

Figure 17: Survey results internal domain experts, question 3: Current usage of ontologies

Optional free text entry for E: Other
I had first contact with ontologies during an AI project ([project name]*) and a [project name]* use case as business contributor.
* Editor's note: Company internal project names have been removed from the original response.

Table 5: Survey results internal domain experts, question 3: Current usage of ontologies (free text entries)

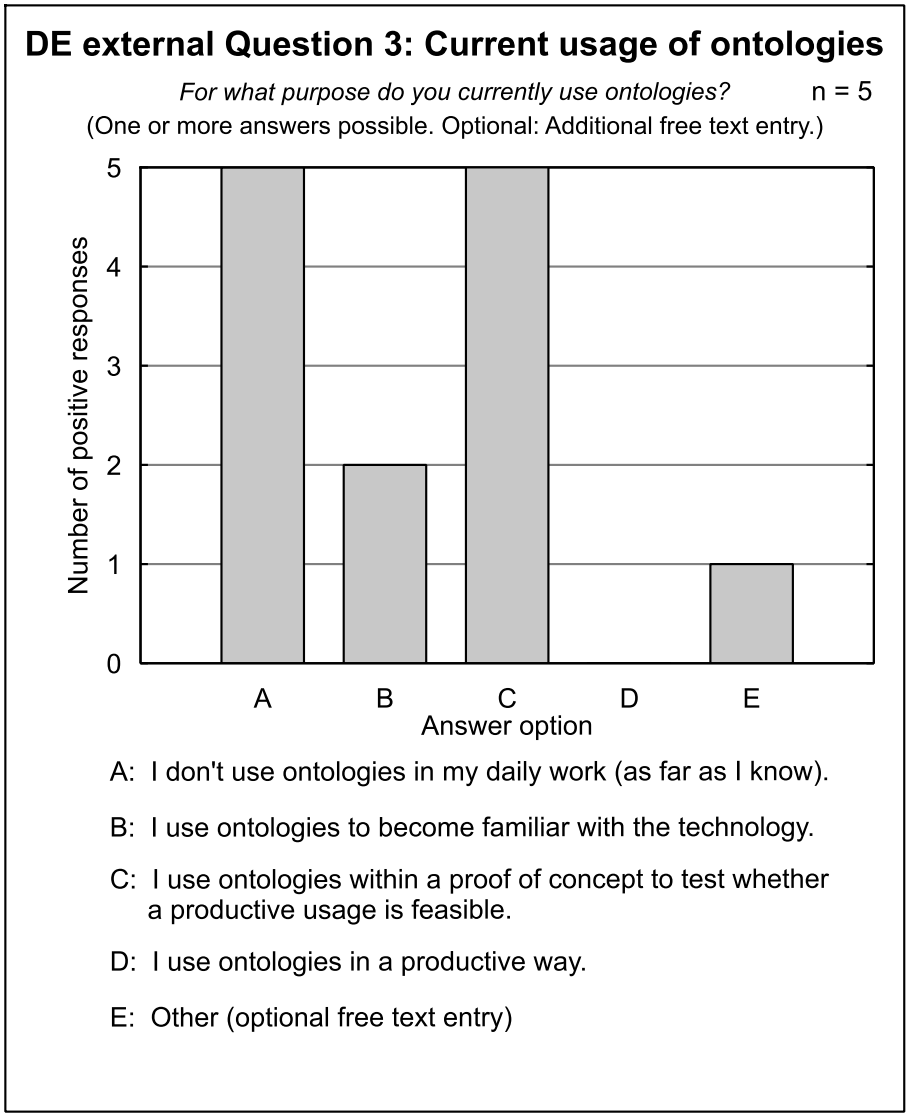

Figure 18: Survey results external domain experts, question 3: Current usage of ontologies

Optional free text entry for E: Other
I develop ontologies as proof of concept, with the move to production.

Table 6: Survey results external domain experts, question 3: Current usage of ontologies (free text entries)

The last two questions that have been presented to the DEs relate to the exchange with KEs. They have been asked to rank the given communication channels according to their personal preference (see figures 19 and 20). In case of missing preferred ways of exchange, the DEs have been asked to add a free text entry (see tables 7 and 8).

Most internal as well as most external DEs prefer *face-to-face meetings* with KEs. This way of exchange leads to some advantages and challenges for the documentation. One advantage is that DEs and KEs can discuss issues directly before adding anything to the documentation. This way, misunderstandings can be clarified before they are recorded explicitly. As described in chapter '2.2 Knowledge', sharing knowledge through socialization is an important step, also for the subsequent externalization of knowledge. In this respect, face-to-face meetings are very useful. At the same time, this means that externalization only takes place when the documentation is created, not during the exchange. This is a challenge as the documentation requires additional effort under these circumstances. Such losses in efficiency may be compensated by reducing face-to-face meetings to an effective minimum. The documentation concept can support this by recommending, for example, that alternative labels should be documented directly by DEs. No meeting with a KE is required for this.

According to personal experience, face-to-face meetings lead to another challenge for documentation. In the course of meetings, handwritten sketches are often made on paper or on whiteboards in order to illustrate relationships between concepts. This way of initially unstructured externalization must somehow be represented in the documentation concept. The challenge here is to weigh up the consistency of documentation and the effort required.

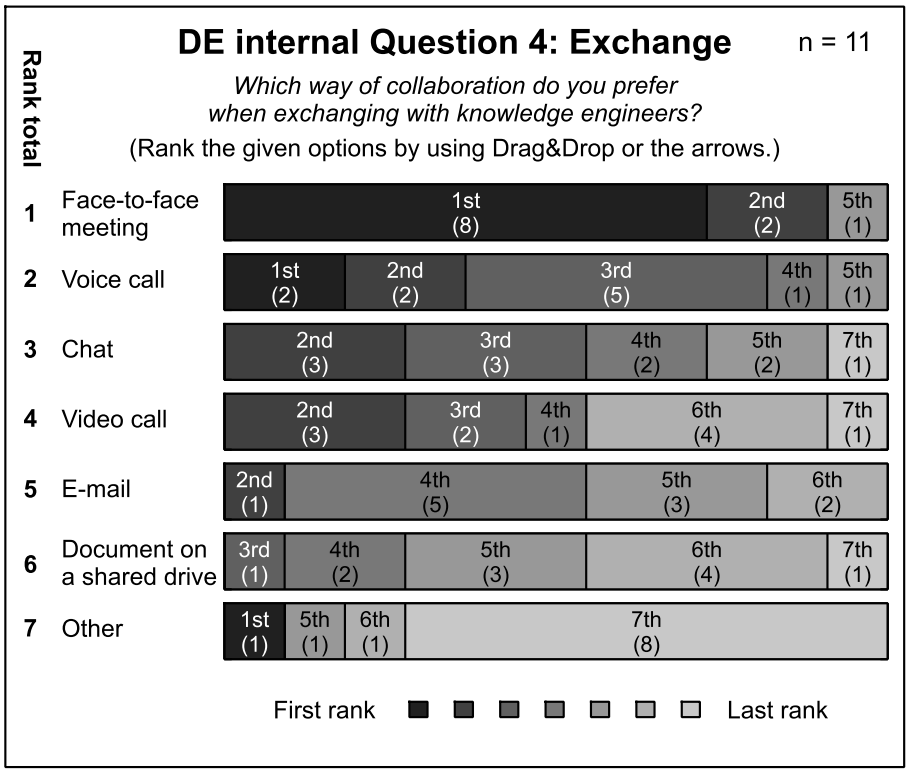

Figure 19: Survey results internal domain experts, question 4: Exchange

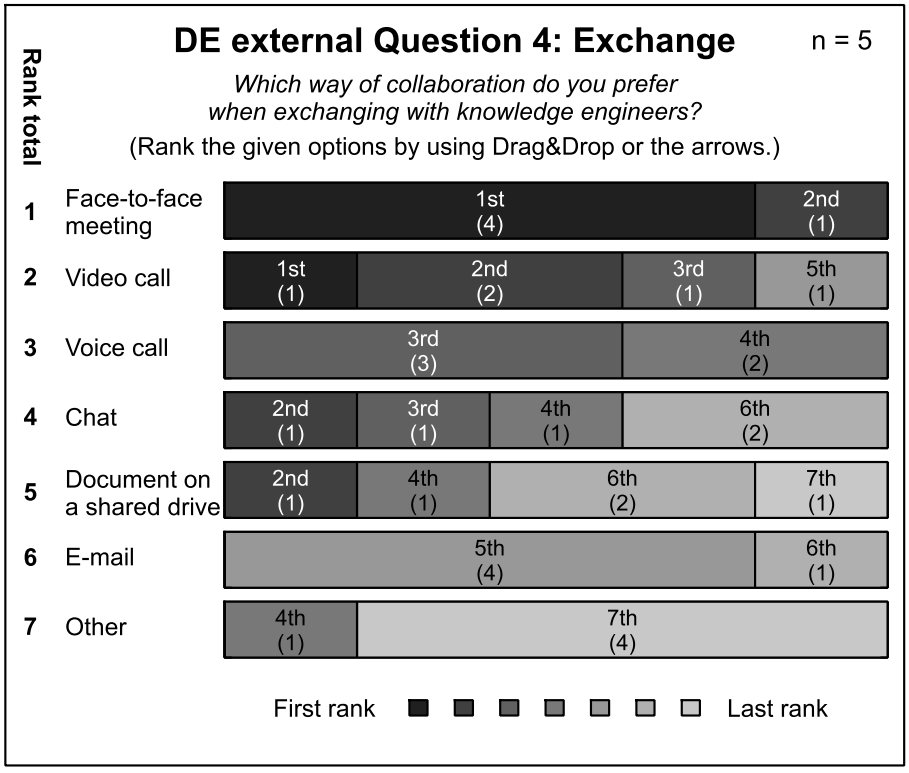

Figure 20: Survey results external domain experts, question 4: Exchange

Optional free text entry
GitHub or similar containing documents and ontologies together would be a preferred exchange platform for me.

Table 7: Survey results internal domain experts, question 5: Optional clarification on Question 4 – Other

Optional free text entry
A collaboratory development environment like github.

Table 8: Survey results external domain experts, question 5: Optional clarification on Question 4 – Other

In terms of popularity, the face-to-face meeting is followed by the three communication channels *voice call*, *chat* and *video call* (see figures 19 and 20). Accordingly, the rather synchronous exchange is more popular than the asynchronous exchange via *e-mail* or *document on a shared drive* (see figures 19 and 20). The latter is usually an undirected communication, while the more popular ways correspond to direct communication between individuals. Electronic chat groups and channels represent an intermediate form, since they allow both the direct addressing of individuals and the communication to a defined group of people.

It is a challenge for documentation that *direct communication* and, in the case of face-to-face meetings and calls, *non-written communication* is preferred. After all, the documentation contradicts these paradigms which are more typical for socialization than for externalization. In other words, documentation is more like a) a document on a shared drive that contains codified knowledge and is accessible to many individuals at any time, than b) a face-to-face meeting that takes place between individuals exchanging tacit knowledge that is not easily accessible to others. The following can be derived from this regarding the implementation of the documentation concept: For an electronic implementation, the documentation should include elements that are similar to chat channels instead of being purely document-based. For example, for each concept a written discussion via chat could be enabled. A chat record is indeed less structured than a document that strictly prescribes how a discussion should be conducted and documented. But a chat record is still better than a phone call or meeting that is not documented at all. In addition, open chats allow people to contribute whose expertise was previously unknown. This is not enabled by peer-to-peer communication.

The additional free text entries in tables 7 and 8 show that some DEs also prefer tools from the area of software development. With *GitHub*[18] they refer to one of the established hosting services for the version control system *Git*[19]. Since not only KEs, but also some DEs operate with such tools, the implementation of the documentation concept may, for instance, enable access via a Git repository or a similar solution.

[18] Further information on GitHub can be found on the company's website:
GitHub (2019) *The world's leading software development platform* [Online]. Available at https://github.com/ (Accessed 2019-02-18).

[19] Further information on Git can be found on the Git community website:
Chacon, Scott; Long, Jason; Git community (n.y.) *git* [Online]. Available at https://git-scm.com/ (Accessed 2019-02-18).

3.2 Archetypal knowledge engineers and domain experts

As the survey results have already shown, there is no such thing as 'the one' knowledge engineer or domain expert. There are different people working on the creation of different ontologies with different objectives and approaches under different conditions. To illustrate this, in the following so-called *assumption personas* are presented.

3.2.1 Persona method

The concept of the persona was introduced by Allan Cooper in the context of his *Goal-Directed Design* method and popularized in his 1999 book '*The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity*' (Cf. Baxter et al., 2015, p. 41; Cooper, 1999; Pruitt and Adlin, 2006, p. 11). In this he describes personas as follows:

> *"Personas are not real people, but they represent them throughout the design process. They are hypothetical archetypes of actual users. Although they are imaginary, they are defined with significant rigor and precision. Actually, we don't so much 'make up' our personas as discover them as a byproduct of the investigation process. We do, however, make up their names and personal details."* (Cooper, 1999, p. 124)

In other words, personas "are detailed descriptions of imaginary people constructed out of well-understood, highly specified data about real people" (Pruitt and Adlin, 2006, p. 3). The persona method is suitable for the development and communication of the documentation concept because of its benefits. PRUITT and ADLIN describe them as follows (Pruitt and Adlin, 2006, p. 14):

- "Personas make assumptions and knowledge about users explicit, creating a common language with which to talk about users meaningfully."
- "Personas allow you to focus on and design for a small set of specific users (who are not necessarily like you), helping you make better decisions."
- "Personas engender interest and empathy toward users, engaging your team in a way that other representations of user data cannot."

Since neither detailed nor extensive data on the target groups KEs and DEs are available, a special form of the persona method is applied: so-called ***assumption personas***, also known as *provisional personas* (Cf. Goodman et al., 2012, p. 484). If the data-driven approach is not feasible, the creation of assumption personas is recommended (Cf. Pruitt and Adlin, 2006, p. 150). Although they are less robust, they still provide advantages as described above. At

first, the assumptions personas presented here are used to make the assumptions explicit. Then they are used to develop, evaluate and communicate the documentation concept.

3.2.2 Creation of assumption personas

The assumptions expressed as personas are based on personal experiences gained while working as a KE in collaboration with DEs and other KEs. In order to increase the validity of the assumption personas, the previously presented survey results are taken into account. In addition, all presented assumption personas are validated as recommended by PRUITT and ADLIN (Cf. Pruitt and Adlin, 2006, p. 152, p. 258). This means that the assumption personas are reviewed by real people who are representatives of the portrayed group. In this process, each assumption persona is presented with one to three representatives. Each representative is inquired individually. In order to get a general feedback, they are first asked: *'Can you imagine that the person described in the persona is a colleague who works in the same function as you?'* Subsequently they are asked to point out all those statements they personally disagree with. If all interviewed representatives reject a statement in the associated assumption persona, the statement is revised according to their feedback. If a statement is considered applicable by at least one representative, it remains unchanged. Essential errors like false facts and spelling errors are corrected immediately.

The attributes described in the assumption personas are based on lists of attributes recommended by literature (Cf. Baxter et al., 2015, p. 42; Goodman et al., 2012, pp. 489–492; Pruitt and Adlin, 2006, pp. 230–232). Those attributes have been selected that are considered relevant and can be described by well-founded assumptions. Application-specific attributes then are added for both target groups. The assumption personas presented here are oriented towards practical application and are not meant to tell a playful story. Therefore, no photos are included, and the focus is on professional instead of private aspects. Table 9 shows the attributes used for the KE and DE assumption personas.

Attributes KE assumption persona	Attributes DE assumption persona
Personal description	Personal description
Working conditions	Working conditions
Relationship to technology	Relationship to technology
Roles and tasks	Roles and tasks
Work-related goals	Work-related goals
Effects of possible errors	Relation to documentation
Relation to documentation	Information reception
Information reception	Exchange with others
Exchange with others	Motivation for contributing to the creation of ontologies
Approach to ontology creation	Expectations from ontologies
Approach to documentation	Current usage of ontologies

Table 9: Attributes of the assumption personas for both target groups – knowledge engineers and domain experts

Six assumption personas are created for the two primary target groups: three for the group of KEs and three for the group of DEs. Secondary and tertiary stakeholders (Cf. Baxter et al., 2015, p. 38) are not considered in this thesis. Since the observations and assumptions as well as the survey results originate from this environment, five of the six assumption personas are working for a global life science company based in Germany. The sixth assumption persona is a member of an external working group that exchanges domain knowledge with the company in question. Figure 21 illustrates the affiliation of the six assumption personas.

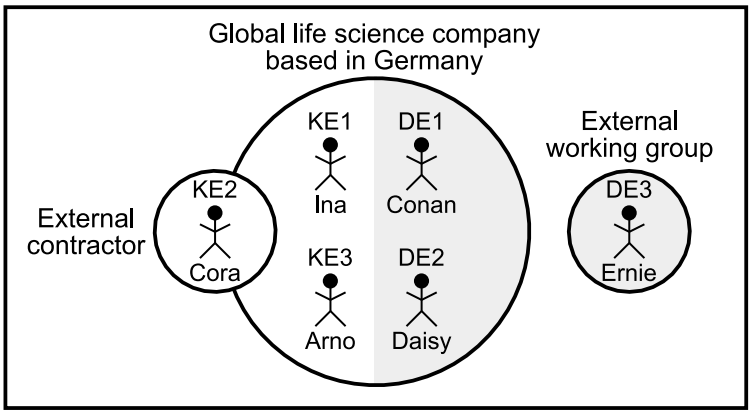

Figure 21: Affiliation of the six assumption personas

3.2.3 Assumption personas

Knowledge engineer 1

Ina the introvert: "I get things done."

Personal description

Ina is a 35-year-old computer scientist. She lives in Berlin, the capital of Germany. After her PhD she founded her own start-up company in the field of semantic technologies. She has been working in the field of IT innovation for a global life science company for three years now.

Working conditions

Ina is an internal employee. She works mainly in the afternoons and evenings at her office in Berlin.

Relationship to technology

Ina is not only experienced in dealing with technology in general. Due to her interests and her education she also has extensive knowledge in the field of computer science.

Roles and tasks

Currently, Ina is the Subject Matter Expert (SME) in a project. Its objective is to carry out a proof of concept on data integration using semantic technologies. She takes over project management tasks and contributes as a Knowledge Engineer (KE).

Work-related goals

Her short-term goal is to deliver fast results and complete the project successfully. In the long run, she strives to implement innovative business solutions as productive applications.

Effects of possible errors

In her role as SME and KE she has responsibility for the course of the project. Mistakes on her part can lead to the failure of the project.

Relation to documentation

The documentation of her work as KE has a low priority for Ina as there are other tasks with higher priority in the project, which is tightly planned and limited in duration.

Information reception

Ina apprehends content quickly by reading it. This also applies to the formal representation of information, such as program code. Reading is therefore also her preferred way to learn new things.

Exchange with others

Ina does not like to interact with others unless there is something important to discuss. She communicates with the project stakeholders if it is necessary for the success of the project. The interaction with Domain Experts (DE) when creating ontologies is limited to the minimum. If real time communication is needed, she prefers voice calls with a shared screen and chat over face-to-face meetings.

Approach to ontology creation

Ina is involved in the creation of three ontologies. The approaches she takes are middle-out and bottom-up. Her focus is on data, the associated data sources and systems. Resources such as literature or people are of less interest to her. Fast results are Ina's main goal, i.e. the fast creation of working prototypes and solutions for certain use cases. In return, she accepts that the resulting knowledge models are less perfect.

Approach to documentation

Ina documents her work on the knowledge model only after she has finished the first version of a formal model. During the previous process steps, she does not write any documentation. Ina's documentation takes place directly in the formal model. For this purpose, she uses especially annotation properties.

Knowledge engineer 2

Cora the communicative: "My models hold water."

Personal description

Cora is 22 years old and has just finished her bachelor's degree in information science at the Cornell University in the state of New York, USA. In her first non-university job she works for a consulting and IT service company.

Working conditions

Cora works 50 % as an external contractor for a life science company based in Germany. She works from her home office in New York between 9am and 5pm. As she is located in the Eastern Standard Time Zone, her working hours differ from those of her colleagues in Germany.

Relationship to technology

Cora is experienced in dealing with technology in general. Her studies have also given her an insight into computer science. However, she is primarily specialized in the interaction between humans and technology, less in computer science.

Roles and tasks

As an external contractor in a project, Cora works exclusively in the role of a Knowledge Engineer (KE). In this role she creates an ontology for the life science company.

Work-related goals

Cora's main goal are resilient results. Through high-quality work in the field of knowledge engineering, she wants to strengthen the reputation of her employer and thus achieve follow-up jobs from her clients. In the long run she wants to contribute to a stable and homogeneous Linked Data landscape in the contracting companies.

Effects of possible errors

Mistakes made by Cora can jeopardize the success of the project she is working for. However, this alone would not be enough to cause the project to fail.

Relation to documentation

Cora primarily creates documentation to record and organize her own work. She wants to record the valuable exchange with domain experts. Also, as an external, she must be able to demonstrate progress in regular meetings as defined by project milestones. Secondarily, she creates documentation for future readers and potential successors.

Information reception

Cora understands complex content best by explaining it to others in conversations.

Exchange with others

Cora seeks direct exchange with Domain Experts (DE) and other KEs. She prefers video or voice calls, ideally with a shared screen, over text-based communication.

Approach to ontology creation

Cora is currently creating a single domain ontology. She applies the middle-out approach. Her focus is on existing vocabularies and standards as well as on input by DEs. Resilient results are Cora's main goal, i.e. the creation of a sound model that fits into the overall picture. In return, she accepts that more time has to be invested.

Approach to documentation

Cora starts the documentation of her work on knowledge models before she creates a formal model. She documents insights informally but in a structured way by using a template and natural language.

Knowledge engineer 3

Arno the architect: "I keep the big picture in mind."

Personal description

Arno is a 55-year-old computer scientist with a PhD. In the first half of his professional life, he worked full-time in research and teaching with focus on semantic technologies. Then he switched to the private sector. For the past five years he has been Enterprise Architect of a global life science company.

Working conditions

Arno works full-time as an internal employee in his office in Berlin, Germany. He likes to start working early in the morning so that he can spend time with his family in the afternoon.

Relationship to technology

Arno is very experienced in dealing with technologies in general and has profound knowledge in the field of computer science. As a former researcher and lecturer, he still keeps track of the current state of research and is able to pass on his knowledge to others.

Roles and tasks

Arno is responsible for the design and implementation of an optimal data and information landscape. This applies both to the technological infrastructure and to changes in the organizational structure and culture. Since he wants to implement semantic technologies and Linked Data principles, he also performs the tasks of a Knowledge Engineer (KE) himself.

Work-related goals

Arnos main goal is the fast enhancement of the data and information landscape by incremental improvements. In the long term, he wants to implement the overall goals, such as enterprise-wide data integration. His mission is to ensure that the right information is available in the right place at the right time.

Effects of possible errors

Since Arno makes fundamental decisions, his actions have a great influence – on the achievement of his own objectives as well as on the achievement of the objectives of others. For example, the success of Ina's project also depends on Arno's infrastructure decisions. Possible major errors by Arno can have a long-term impact on the data and information landscape and thus on the success of the company.

Relation to documentation

Arno creates and encourages documentation to promote the exchange across departments. He considers shared documentation as one way to avoid silo thinking.

Information reception

Arno conceives complex contexts most quickly with the help of visual representations such as diagrams.

Exchange with others

Arno generally seeks exchange with colleagues. Regarding the implementation of semantic technologies, he interacts primarily with decision makers and KEs, less with Domain Experts (DE). He prefers text-based communication, such as chat and e-mail over remote or face-to-face meetings. He considers the former to be more efficient and values asynchronous communication.

Approach to ontology creation

Arno does not focus on individual ontologies or domains. He supervises the network and hierarchy of all ontologies across divisions within the entire company. As KE he works concept-oriented rather than domain-oriented. This means he first examines important concepts individually and then decides in which ontology the concept is to be defined. For fundamental design decisions, he follows the Top-down approach. As an Enterprise Architect, however, he is also responsible for ensuring that the data is properly integrated, for which he uses the bottom-up approach.

Approach to documentation

If Arno himself is working on an ontology, he documents during the creation of the formal model. He records insights from the exchange with others in an unstructured manner, for example in notes or meeting minutes. His documentation can be viewed by others.

Domain Expert 1

Conan the committed: "Ontologies are the solution of choice."

Personal description

Conan is 44 years old. He has a PhD in pharmacy and works in the pharma industry since his graduation. For the last five years he has been working in the Regulatory Affairs department at a life science company in Berlin, Germany.

Working conditions

Conan has a nine-to-five job as an internal employee with an office in Berlin.

Relationship to technology

Conan has basic knowledge in dealing with technologies. When confronted with new applications or paradigms, he needs some time to become familiar with them.

Roles and tasks

In his role as Governance Manager in Regulatory Affairs, Conan must ensure that the information required to register a pharmaceutical product is accurate and complete.

Work-related goals

Conan's main objectives are successful submissions of drug applications as well as the effective communication with health authorities. For him, ensuring compliance has higher priority than rapid innovation. He therefore needs to be ensured that compliance and data quality will not be sacrificed in favor of efficiency in the development of new IT solutions and processes. Only if this prerequisite is fulfilled, Conan supports the replacement of manual processes by semi-automatic ones.

Relation to documentation

Conan actively reads the documentation of the KEs and adds input to the documentation on his own initiative.

Information reception

Conan is accustomed to reading structured information, especially in the form of spreadsheets. In addition, he likes exploring interactive visualizations by himself, without the need of the guidance of a KE. He is interested in the formal representation of the exchanged knowledge but does not intend to fully understand the formal model.

Exchange with others

Conan is happy to answer any questions in face-to-face meetings as well as via chat. He actively introduces new aspects to the discussion and provides advice regarding existent relevant guidelines, such as internal standard operating procedures, external standards or regulatory requirements. Thereby he focuses more on the general concepts and less on the associated data. As a sceptical person, Conan always considers possible exceptions and challenges.

Motivation for contributing to the creation of ontologies

Conan considers ontologies to be a promising technology that he would like to learn more about. Not only does he want to share his knowledge with Knowledge Engineers (KE), he also likes to exchange knowledge with other Domain Experts (DE) to learn more about his own domain.

Expectations from ontologies

Conan expects ontologies to facilitate his daily work, for example by enabling smarter search functions. He is especially interested in data integration, which ideally allows him to easily access data regarding a product over its entire life cycle – from the first research steps via clinical studies to approval and pharmacovigilance. In addition to the use of ontologies in the background of IT systems, he also expects to be able to work directly with them. In other words, he would like to use ontologies to search for definitions and to better understand contexts.

Current usage of ontologies

Currently, Conan does not use ontologies in his daily work. However, he is involved as a DE in a project that creates ontologies as part of a proof of concept. There, he tests together with the other project stakeholders whether a productive usage of ontologies is feasible.

Domain Expert 2

Daisy the Data Scientist: "Ontologies are necessary in the background of our applications."

Personal description

Daisy is 34 years old and has a master's degree in bioinformatics. She has been working for eight years in the field of clinical research for a life science company based in Germany.

Working conditions

Daisy works part-time as an internal employee in her office located in Berlin.

Relationship to technology

Daisy is very experienced in dealing with technology in general. Moreover, she has extensive expertise in statistics and data science.

Roles and tasks

As a data scientist, Daisy applies methods and algorithms to both structured and unstructured data in order to gain insights.

Work-related goals

Daisy's department, Clinical Study Analysis, is analyzing clinical data in order to understand the effects of drugs regarding specific diseases and therapies.

Relation to documentation

Daisy provides brief and concise written input that can be used for documentation. However, she does not add content directly to the documentation on her own initiative.

Information reception

Daisy prefers structured, compact texts to apprehend content. Besides, she likes to apply new content directly in order to understand it in a practical way. Regarding ontologies, she prefers to apply formal models in a sandbox environment by herself rather than just to see a demonstration.

Exchange with others

Daisy shares her knowledge, but only when she is consulted with concrete questions. She expresses her knowledge and requirements from a bottom-up perspective because she works primarily hands-on with data. Since Daisy occasionally uses software development applications, she prefers to access documentation and formal models via a shared Git repository. If written communication is not suitable, she prefers face-to-face meetings over video or voice calls.

Motivation for contributing to the creation of ontologies

Daisy thinks ontologies are a promising technology and would like to learn more about them. The exchange with other Domain Experts (DEs) is less decisive for her.

Expectations from ontologies

Daisy expects ontologies to work in the background to improve the interaction between the IT systems she uses. She is expecting a data integration that will allow her to easily find relevant data and publish data collections she has created for other data scientists within the company. Daisy's ideal is that the data are available according to a common ontology model and in a consistent format like RDF. She does not expect to work directly with ontologies to acquire knowledge about a domain.

Current usage of ontologies

Currently Daisy does not use ontologies in her daily work. However, she is involved as a DE in a project that creates ontologies as part of a proof of concept. There she tests together with the other project stakeholders whether a productive usage of ontologies is feasible.

Domain Expert 3

Ernie the external: "We share our knowledge to jointly turn translational life sciences into reality."

Personal description

Ernie is 62 years old. He has a PhD in chemistry. After his studies, he continued his education in the field of statistical analysis. In this field he has been working for over 10 years for a life science company based in the United Kingdom.

Working conditions

Ernie has a nine-to-five job as an internal employee. He spends 10 % of his working time in the external working group *Translational Life Science* (TLS)[20]. There, he contributes as a Domain Expert (DE) to the integration of clinical data by means of Linked Data principles and semantic technologies. The web conferences of the working group are held during the day. However, most of Ernie's involvement takes place outside regular office hours.

Relationship to technology

Ernie is used to dealing with technology. He is familiar with both computer-based statistical analyses and technical laboratory instruments. In the context of his participation in the TLS working group, he is currently learning how to work with software development tools such as Git.

Roles and tasks

As an analyst in the department Statistical Programming, Ernie analyses clinical data to identify trends and patterns. A part of his work is to transfer raw data into the data models of various established industry standards and vice versa. In the working group TLS, it is Ernie's task to contribute his knowledge.

Work-related goals

Ernie's objective in the TLS working group is to achieve together with the other members results which are beneficial for all participating organizations. His employer expects that the results will enable the company to better utilize their data and exchange data with other organizations in a standardized way.

[20] TLS is a fictional working group. It represents the existing working groups that apply Linked Data principles and semantic technologies in the field of life sciences.

Relation to documentation

Ernie does not actively participate in the documentation of TLS, although the workgroup expects this from its members. Ernie expects the Knowledge Engineers (KEs) to collect and utilize his input. For him, the direct exchange with other DEs comes first. Since documentation is of little interest to Ernie, his personal documentation consists of unstructured notes that he writes on the side.

Information reception

Ernie prefers clear visualizations that are explained to him in conversation.

Exchange with others

Ernie shares his knowledge during weekly web conferences held by the TLS working group. He prefers to explain facts in conversation with the help of presentation slides.

Motivation for contributing to the creation of ontologies

Ernie shares his knowledge because he considers ontologies to be a promising technology to achieve data integration and harmonization. Besides, he would like to learn more about his domain by exchanging with other DEs.

Expectations from ontologies

In the long run, Ernie expects ontologies to contribute to making his work easier. He does not assume that he will work directly with ontologies.

Current usage of ontologies

Currently, Ernie does not use ontologies in his daily work. During meetings with the KEs, he is interested in the overall development of ontologies, to which he contributes as part of a proof of concept. Using the input of the KEs, he also tries to apply parts of these ontologies himself in order to integrate data on a test basis.

3.3 Findings

The assumption personas show that documentation is needed to explicitly preserve and share knowledge. Plainly the different working hours require *asynchronous knowledge exchange*. For instance, Ina and Arno work at different hours of the day, while Cora and Ernie are even located in different time zones.

Besides, there are different levels of experience in dealing with technology. While Ina and Arno are experienced computer scientists, Conan needs time to get used to new software applications. Thus, the implemented documentation concept should also be *accessible to people with little technological experience*.

Different types of information representation are preferred by the assumption personas. Arno and Ernie prefer visualizations, Ina formal statements, Cora structured documentation using natural language. The documentation concept should therefore not only allow natural language text, but in addition also *visual and formal elements*. However, the focus remains on structured documentation using natural language as explained in chapter '2.1 Documentation concept'.

Not all people involved are employees of the company, for which the ontologies are created. In addition to internal employees, external contractors such as Cora and external DEs like Ernie also contribute to the creation of ontologies. For the documentation concept, this is relevant, since different regulations may exist regarding access to related data and information. Potential reasons for this are intellectual property, privacy or IT security. In other words, the documentation concept must provide an option to record any *access controls* or *confidentiality classifications*.

Further restrictions on the use of resources such as data, standards, vocabularies and literature may exist due to applicable *licenses*. These are relevant, for example, when definitions given in literature or excerpts of controlled vocabularies are adopted during ontology creation.

A key observation is that the KEs, on the one hand, do not necessarily work with a focus on one ontology at a time. While Cora concentrates on a single ontology, Ina works on three ontologies at the same time. Arno in turn works concept-oriented in the context of an ontology network. DEs like Daisy and Ernie, on the other hand, are less interested in what particular ontologies they contribute to. They answer questions about concepts and expect their input to be used to create ontologies that will improve data integration. In view of this, the focus of the documentation concept should be on individual concepts. This enables all presented assumption personas to make use of the documentation concept. Two alternative but less suitable approaches are briefly presented in appendix 'D: Alternative documentary reference units'.

The *concept-oriented documentation* has the advantage that it can be applied to all modeling approaches. Concepts are so to say '*the lowest common denominator*' for all DEs and KEs. This granularity allows concepts to be considered individually or in sets. Sets of concepts can be created and split more easily than with the alternative approaches discussed in the appendix. Individual concepts can be part of several sets without having redundant information about a concept. The exact determination of the documentary reference unit is described in chapter '5 Documentation concept'. It also discusses the situation that no concepts have yet been identified, but only a set of terms is available at first.

The following tables list the findings derived from the selective insights into corporate reality. Table 10 contains the *relevant information per concept*, i.e. it states *what* needs to be documented for each concept. Table 11 shows the *relevant information per resource*. A resource in this context is any non-human source of data, information or knowledge. This includes data sets (e.g. csv file, file containing formal vocabulary), data storage locations (e.g. relational database, triple store), software applications (e.g. user interface with predefined views of databases), electronic and printed publications (e.g. websites, journal articles, standards). While Table 11 states *what* needs to be documented for each resource, Table 12 illustrates this for exemplary resources. Table 13 contains general aspects that must be considered when creating the documentation concept. In the broader sense, these are *documentation principles*. They state *how* the documentation should be done.

Finding		**Derived from …**	
Title of finding	Description of finding	Survey results	Assumption personas
Alternative labels	Enable documentation of alternative designators including synonyms, quasi-synonyms, abbreviations and multilingual labels.	Considered relevant by KEs (see figure 10).	
Context	Enable documentation of context information, i.e. relationships to other concepts.	Considered relevant by KEs (see figure 10).	Conan expects to be able to use ontologies to better understand context.
Definition	Enable documentation of definitions.	Considered relevant by KEs (see figure 10) and expected by DEs (see figures 15 and 16).	Conan expects to be able to use ontologies to search for definitions.
Explanation	Enable documentation of explanations which are more detailed than a short definition.	Considered relevant by KEs (see figure 10) and expected by DEs (see figures 15 and 16).	Conan expects definitions and context. Detailed explanations can support this and help to fill any gaps in knowledge.
Identifier	Enable documentation of identifiers.	Considered relevant by KE (see table 1).	
Related data	Enable documentation of related data sets, data storage locations and software applications.	Considered relevant by KEs (see figure 10).	Ina considers data, data sources and systems for the bottom-up approach.
Related people	Enable documentation of related people, e.g. KEs, DEs, data owners, data stewards, process owners.	Considered relevant by KEs (see figure 10).	
Related use case	Enable documentation of related use case and related competency questions.	Considered relevant by KE (see table 1).	
Related literature	Enable the documentation of related literature that is relevant to the subject or the implementation.	Considered relevant by KEs (see figure 10).	
Related vocabularies and standards	Enable the documentation of related vocabularies and standards that are relevant to the subject or the implementation.	Considered relevant by KEs (see figure 10).	

Table 10: Findings derived from insights into corporate reality – relevant information per concept

Finding		**Derived from …**	
Title of finding	Description of finding	Survey results	Assumption personas
Access control	Enable documentation of access conditions for resources which are protected by access control.	The bottom-up approach is most common among the KEs (see figure 8). In the course of this, protected data sets, data storage locations and software applications may be used.	Ina considers data, data sources and systems for the bottom-up approach.
Confidentiality	Enable documentation of confidentiality classification for resources.	The bottom-up approach is most common among the KEs (see figure 8). In the course of this, classified data and information may be used.	Involvement of externals like Cora and Ernie. Different confidentiality rules may apply to internal and external parties.
License	Enable documentation of license conditions for resources.	KEs consider resources such as data, literature, vocabularies and standards as relevant (see figure 10). The permitted use of those resources may be determined by licenses.	

Table 11: Findings derived from insights into corporate reality – relevant information per resource

Resource type	Resource	Access control	Confidentiality	License
Software application	Substance dashboard "*SubDash*"	Available in company network	Internal	-
Data storage location	Substance database "*SubDB*"	Individual permission and completed training required	Restricted	-
Webpage	Intranet article "*Our Top 10 Substances*"	Available in company network	Internal	-
Webpage	Wikipedia article "*Chemical substance*"	No access control	Not classified	CC BY-SA 3.0 Unported License
Electronic publication	Standard "*ISO 11615:2017*"	Available via company-internal standards library	Not classified	Copyright protected. Unauthorized use prohibited. Permissions granted by ISO.
Data set	RDF/XML file "*Friend of a Friend (FOAF) vocabulary*"	No access control	Not classified	Attribution 1.0 Generic (CC BY 1.0)
Print publication	Book "*Pharmaceutical lexicon: a dictionary of pharmaceutical science*"	Available via *New York State Library*, among others	Not classified	Public domain (published 1873)
Data set	Excel file "*substances.xlsx*"	Password-protected	Restricted	-

Table 12: Demonstration of the properties access control, confidentiality and license for exemplary resources

Finding		**Derived from …**	
Title of finding	Description of finding	Survey results	Assumption personas
Additional visual elements	Enable text-based documentation to be supplemented by illustrations.	DEs prefer face-to-face meetings (see figures 19 and 20). In the course of this, sketches are made.	Arno and Ernie prefer visual representations.
Concept-oriented documentation	Use single concept as main documentary reference unit.		The people involved have different approaches and objectives. The reference unit, which is considered by all KEs and DEs, is the single concept.
Direct exchange between people	Enable direct exchange between the people involved, at best following paradigms of electronic chats, in order to reduce non-textual communication to an effective minimum.	DEs prefer direct exchange, either non-text-based such as face-to-face meetings and calls or textual via electronic chats (see figures 19 and 20).	Daisy only contributes to the documentation if she is consulted with concrete questions. Therefore, direct addressing must be possible.
Distribution of workload	Provide recommendations on who can document what independently. The objective is to increase efficiency by reducing the time spent in meetings to an effective level.	DEs prefer face-to-face meetings (see figures 19 and 20). However, they can document some information themselves without having a meeting.	Ina limits exchange with DEs to a minimum. Arno prefers text-based communication because he considers it to be more efficient and values asynchronous communication.
Multilingualism	Consider the option that a multilingual ontology is created. Enable documentation of natural language content in different languages.	Assumption based on the fact that alternative labels are considered relevant by KEs (see figure 10). However, KEs have not been surveyed for multilingualism.	
One documentation for all target groups	No separate documentations for DEs and KEs. Instead, common documentation with optional separation between • subject-related content (DE knowledge) and • content regarding knowledge engineering (KE knowledge).	DEs not only want to gain domain knowledge, but also learn more about ontologies (see figures 13 and 14).	Conan also wants to see formal definitions without claiming to fully understand them. Daisy likes to apply models by herself in a sandbox environment.

Table is continued on the next page.

Continuation of the table on the previous page.

Two levels of granularity	Enable documentation in different levels of detail: • detailed for comprehensive documentation • less detailed for fast documentation	The main goal of KEs can be both fast results and resilient results (see figure 9).	Cora aims for resilient results and documents regularly, while Ina aims for fast results and considers documentation less important.

Table 13: Findings derived from insights into corporate reality – documentation principles

4 Analysis of ontology engineering methodologies

In the following, ontology engineering methodologies are analyzed in order to derive findings for the documentation concept. Out of the variety of methodologies proposed in literature (see appendix 'A: Ontology engineering methodologies'), two are selected: the *Unified Process for Ontology Building* (UPON) and the *NeOn Methodology*.

The UPON is selected because it represents the *software engineering approach to ontology building*. According to personal experience, this approach largely corresponds to the approach followed in companies. Although not always a uniform process is pursued, the ontology creation is nevertheless characterized by the fact that KEs in companies usually have more of a computer science than an information science background. This personal observation is reflected in parts of an empirical analysis by MIHINDUKULASOORIYA, POVEDA-VILLALÓN, GARCÍ-CASTRO and GÓMEZ-PÉREZ with focus on collaborative ontology evolution and data quality:

> "*The analysis shows that in communities […], where industrial participation is dominant and there is a strong commercial interest, the community tends to follow more rigorous editorial processes and governance procedures. Furthermore, it shows that the tools used by such communities, for instance, version control systems, issue trackers, implementation reviews, etc., are largely influenced by the software development tools and processes*" (Mihindukulasooriya et al., 2017, p. 112)

Another benefit of UPON, according to its authors, is that it covers more processes than other methodologies (see figure 22). In this respect, the integral processes *knowledge acquisition*, *evaluation* and *documentation* are of particular relevance for the documentation concept in addition to the *ontology development-oriented processes*.

The NeOn Methodology is selected for three main reasons. Firstly, because it has been developed within a large-scale project involving many contributors. Secondly, because it is designed for the collaborative creation of networked ontologies. Thirdly, because existing methodologies have been taken into account for its development. All three aspects correspond to the approach of the documentation concept. It is designed to support the exchange of knowledge through documentation in the collaborative creation of single or networked ontologies, taking into account as many different approaches to ontology creation as possible.

IEEE 1074-1995 standard processes			Uschold and King	Grüninger and Fox	METHONTOLOGY	On-To-Knowledge	UPON
Project management processes	Project initiation		–	P	–	+	–
	Monitoring and control		–	P	P	+	–
	Quality management		–	–	P	–	–
Ontology development-oriented processes	Pre-development	Environment study	–	–	–	–	P
		Feasibility study	–	–	–	+	–
	Development	Requirements	P	+	+	+	+
		Design	–	+	+	P	+
		Implementation	+	+	+	+	+
	Post-development	Installation	–	–	–	–	–
		Operation	–	–	–	–	–
		Support	–	–	–	–	–
		Maintenance	–	–	P	P	P
		Retirement	–	–	–	–	–
Integral processes	Knowledge acquisition		+	P	+	–	+
	Evaluation		+	–	+	+	+
	Configuration management		–	–	+	–	–
	Documentation		+	–	P	–	+
	Training		–	–	–	–	P

–, unsupported; +, supported; P, partially supported.

Figure 22: A comparison of different processes with respect to the IEEE 1074–1995 standard (Nicola et al., 2009, p. 274)

The NeOn Methodology is one of the results of "the NeOn project (life cycle support for networked ontologies), which was funded by the European Commission's Sixth Framework Programme under grant number FP6-027595" (Suárez-Figueroa, Gómez-Pérez, Motta et al., 2012, p. x). The NeOn project has started in March 2006 and had a duration of 4 years (Cf. Suárez-Figueroa et al., 2010, p. 1). The NeOn consortium has involved "fourteen institutions with extensive experience in research and development from United Kingdom, Germany, Spain, Italy, Slovenia and France" (NeOn Project, n.y.b)[21]. The project has enabled the collaboration of "[s]everal dozens people" from the field of ontology engineering (Suárez-Figueroa, Gómez-Pérez, Motta et al., 2012, p. x) with backgrounds in academia and industry (Cf. NeOn Project, n.y.a). The technologies developed "have been applied to three different case studies in two vast, transnational domains (fisheries and the pharmaceutical industry)" (NeOn Project, n.y.a).

[21] An overview of the 14 institutions that have formed the NeOn consortium can be found at: NeOn Project (n.y.b) *Partners* [Online]. Available at http://neon-project.org/nw/Partners.html (Accessed 2019-02-18).

The NeOn state-of-the-art analysis regarding ontology engineering methodologies[22] has revealed the following: "Nowadays no methodology adequately supports the collaborative and context aspects of networks of ontologies" (Suárez-Figueroa et al., 2007, p. 11). Therefore, the NeOn Methodology has been developed with the goal "to support the collaborative construction and dynamic evolution of networks of ontologies in distributed environments where contextual information is introduced by developers at different stages of the ontology development process" (Suárez-Figueroa et al., 2007, p. 11). For the development of the NeOn Methodology, the three already existing methodologies *METHONTOLOGY*, *On-To-Knowledge* and *DILIGENT* have been used (see appendix 'A: Ontology engineering methodologies') in addition to findings gained in the project (Cf. Suárez-Figueroa et al., 2007, p. 11).

[22] See also chapter '2. State of the Art on Methodologies' in Suárez-Figueroa, Mari C.; Cea, Guadalupe A. de; Buil, Carlos; Dellschaft, Klaas; Fernández-López, Mariano; García, Andrés; Gómez-Pérez, Asunción; Herrero, German; Montiel-Ponsoda, Elena; Sabou, Marta; Villazon-Terrazas, Boris; Yufei, Zheng (2008) *D5.4.1. NeOn Methodology for Building Contextualized Ontology Networks* (Class Deliverable: NEON EU-IST-2005-027595; Document Identifier: NEON/2008/D5.4.1/v1.0) [Online]. Available at http://neon-project.org/web-content/images/Publications/neon_2008_d5.4.1.pdf (Accessed 2019-02-07).

4.1 Unified Process for Ontology Building UPON

The *Unified Process for Ontology Building* (UPON) has been proposed by NICOLA, MISSIKOFF and NAVIGLI. For the analysis, two of their publications are considered, as shown in table 14. However, reference is primarily made to the more recent of the two publications as it is more elaborate.

Analyzed UPON publications
Nicola, Antonio de; Missikoff, Michele; Navigli, Roberto (2005) 'A Proposal for a Unified Process for Ontology Building: UPON', in *Database and Expert Systems Applications.* 16th International Conference, DEXA 2005. Copenhagen, Denmark, August 22-26, 2005. Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 655–664. DOI: 10.1007/11546924_64.
Nicola, Antonio de; Missikoff, Michele; Navigli, Roberto (2009) 'A software engineering approach to ontology building', *Information Systems*, vol. 34, no. 2, pp. 258–275. DOI: 10.1016/j.is.2008.07.002.

Table 14: List of the analyzed literature regarding the Unified Process for Ontology Building

4.1.1 Overview

The UPON is derived from the *Unified Software Development Process* (Cf. Jacobson et al., 1999), called *Unified Process* (UP) for short, and takes advantage of the *Unified Modeling Language* (UML)[23]. UPON is "designed in accordance with the UP method, aimed at guiding ontology engineers in the production of an effective and valuable domain ontology" (Nicola et al., 2009, p. 258). It pursues the following main objectives (Nicola et al., 2009, p. 259):

- "reduction of time and costs in the production of largescale domain ontologies (providing also useful guidelines for small ontologies)"
- "enhancement of the quality of the produced ontology, by progressive validation of the intermediate results"
- "creation of a methodological setting where the two kinds of expertise, KE and DE expertise, are explicitly identified and used at best"
- "clear identification of the activities, with roles and responsibilities of the different experts"

[23] Further information on UML can be found in:
Object Management Group (2019) *What is UML* [Online]. Available at http://www.uml.org/what-is-uml.htm (Accessed 2019-01-28).

- "production of intermediate results that can be readily available to the users of the ontology-based applications (e.g., semantic search)"

As the objectives show, an analysis of the UPON is promising as clear statements are made about the activities and collaboration of KEs and DEs. They can be used to make recommendations regarding the distribution of workload, i.e. *who* should document. The intermediate results described by the process can be analyzed to determine *what* needs to be documented.

The UPON differs from other ontology engineering methodologies by its "*use-case driven*, *iterative*, and *incremental* nature" (Nicola et al., 2009, p. 259) which is described as follows:

> "*UPON is use-case driven since it does not aim at building generic domain ontologies, but ontologies that serve its users, both humans and automated systems […], in a well-defined application area and with accurately defined objectives. […] The nature of the process is iterative because each activity is not only cyclically repeated, typically concentrating on different parts of the ontology being developed, but also incremental, since at each cycle the ontology is further detailed and extended.*" (Nicola et al., 2009, p. 259)

The UPON approach is based on *cycles*, *phases*, *iterations* and *workflows* which relate to each other as follows:

> "*Each cycle consists of four phases (inception, elaboration, construction, and transition) and eventually results in the release of a new version of the ontology. Each phase is further subdivided into iterations. During each iteration, five workflows […] take place: requirements, analysis, design, implementation, and test.*" (Nicola et al., 2009, p. 259)

At this point *cycles*, *iterations* and *phases* are not described any further. The analysis in this thesis focuses on *workflows*. They provide concrete information that is relevant for the documentation concept. Throughout the UPON, the following five workflows are performed consecutively:

1) Requirements workflow
2) Analysis workflow
3) Design workflow
4) Implementation workflow
5) Test workflow

4.1.2 Workflows

In the following, the workflows are briefly described according to their sequence. A more detailed description can be found in appendix 'C: UPON workflows'.

The objective of the ***requirements workflow*** is to capture the requirements (Cf. Nicola et al., 2009, p. 260). Therefore, it takes place prior to the process considered in this thesis (see chapter '2.4 Process'). The latter presumes that the requirements have already been identified and specified. The relevant outputs of the requirements workflow are the *application lexicon*, the *competency questions* and the prioritized *use cases* (Cf. Nicola et al., 2009, p. 262). The ***application lexicon*** (AL) is a set of terms which are "pertain to a given application […] as stated by a community of application experts" (Nicola et al., 2009, p. 262). It is created "by collecting the terminology from DE and application-specific documents" as well as by extracting terminology from previously created storyboards (Nicola et al., 2009, p. 261). ***Competency questions*** (CQs) are, as already described in chapter '2.4 Process', "questions at a conceptual level an ontology must be able to answer" (Nicola et al., 2009, p. 261). According to UPON, CQs are addressed by using ***use-case models*** which contain "a number of use cases that serve as a basis to specify the expected use of the ontology" (Nicola et al., 2009, p. 262). These ***use cases*** "correspond to knowledge paths through the ontology, to be followed for […] answering CQs" (Nicola et al., 2009, p. 262). They are visualized using UML use-case diagrams and "detailed during the analysis and design workflows" (Nicola et al., 2009, p. 262).

The ***analysis workflow*** concerns the conceptual analysis. In the course of this, a ***domain lexicon*** (DL) is created which is a set of terms "validated by a community of domain experts" (Nicola et al., 2009, p. 265). Subsequently, a ***reference lexicon*** (RL) is build "by selectively merging the AL (from application DEs) and the DL (from existing external resources)" (Nicola et al., 2009, p. 263). Next, the application scenario is modeled by "adding to the use-case diagrams, drawn in the Requirements Workflow, the activity and class diagrams" (Nicola et al., 2009, pp. 263–264). The last activity of the analysis workflow is the creation of a ***reference glossary*** (RG). It is "built by using the RL and by adding informal definitions (i.e., natural language sentences) to the terms" (Nicola et al., 2009, p. 264). Each term can be assigned one or more definitions which "should be selected from knowledgeable sources and agreed among DEs and users" (Nicola et al., 2009, p. 264).

The objective of the ***design workflow*** is "to give an ontological structure to the set of glossary entries gathered in the reference glossary" (Nicola et al., 2009, p. 266). In other words, concepts as well as relationships between them are identified. This is carried out according to OPAL (Cf. Nicola et al., 2009, p. 266). The *Object, Process, Actor modeling Language* (OPAL) "is an ontology modelling framework aimed at supporting business experts in building an ontology. [...] OPAL design patterns are formally defined by using OWL, allowing for a compatibility with this popular ontology language" (D'Antonio et al., 2007, p. 346).

In the first activity of the design workflow, *modeling concepts*, each concept is assigned a category (Cf. Nicola et al., 2009, p. 266). The used concept categories are adopted from OPAL, where they are referred to as *kinds* (Cf. D'Antonio et al., 2007, p. 348). The primary concept categories which are inspired by UML modeling constructs are: *business actor*, *business object* and *business process* (Cf. Nicola et al., 2009, p. 266). In addition to the primary categories, there are two complementary categories: *message* and *attribute* (Cf. Nicola et al., 2009, p. 267).

The second activity is *modeling concept hierarchies and domain-specific relationships*. First, the concepts are organized "in a taxonomic hierarchy according to the *generalization* (i.e., IsA) relation" (Nicola et al., 2009, pp. 267–268). Then this taxonomy "can be extended with other relations, i.e., *part-of* and domain-specific relationships" (Nicola et al., 2009, p. 268). The output of the design workflow is a ***semantic network*** (SN) which the authors consider "as the informal counterpart of the ontology that will be produced in the next workflow" (Nicola et al., 2009, p. 268). The SN "is represented according to UML class diagram, in particular using *generalization* (IsA), *aggregation* (part-of) and *association*" (Nicola et al., 2009, p. 268).

The objective of the ***implementation workflow*** is "to encode the ontology in a rigorous, formal language" (Nicola et al., 2009, p. 268). First a formal language is selected. Then the informal SN is represented as a formal ontology as defined in chapter '2.3 Ontology'. (Cf. Nicola et al., 2009, p. 270)

The ***test workflow*** concerns the evaluation of the ontology. To do this, the *semantic quality* "is verified by checking the consistency of the ontology, achieved by using a reasoner" (Nicola et al., 2009, p. 271). With regard to *pragmatic quality*, the test workflow proposes two tests, one concerning coverage and the other CQs. The coverage can be verified by asking a DE "to semantically annotate the UML diagrams, modeling a software application, with the

ontology concepts" (Nicola et al., 2009, p. 272). For the CQs, it is tested whether it is possible to "answer them by using the ontology content" (Nicola et al., 2009, p. 272). In other words, "an ontology is complete if the objectives, defined in the requirements workflow, are reached" (Nicola et al., 2009, p. 272).

4.1.3 Complete process

Figure 23 shows the five workflows presented as a complete process. On the right side of the figure, it is illustrated to what extent the DEs and KEs are involved throughout the process. On the left side, the figure given by the authors is supplemented to show which part of the UPON corresponds to the process considered in this thesis (see chapter '2.4 Process').

The relevant part for the documentations concept begins within the *analysis workflow*. The creation of lexicons (AL, DL and RL) as well as the modeling of the application scenario are considered as preliminary activities at this point. The process to be covered by the documentation concept begins with the recording of definitions of terms, i.e. with *building the reference glossary*. Concerning the previous UPON activities, only their outputs are relevant. For example, CQs can be relevant if the associated CQ is documented for each concept. How the CQ has been identified, however, is not the subject of the documentation concept presented in this thesis. The part of UPON covered by the documentation concept ends with the test workflow. In both cases, the terminating condition is that the previously defined requirements are met.

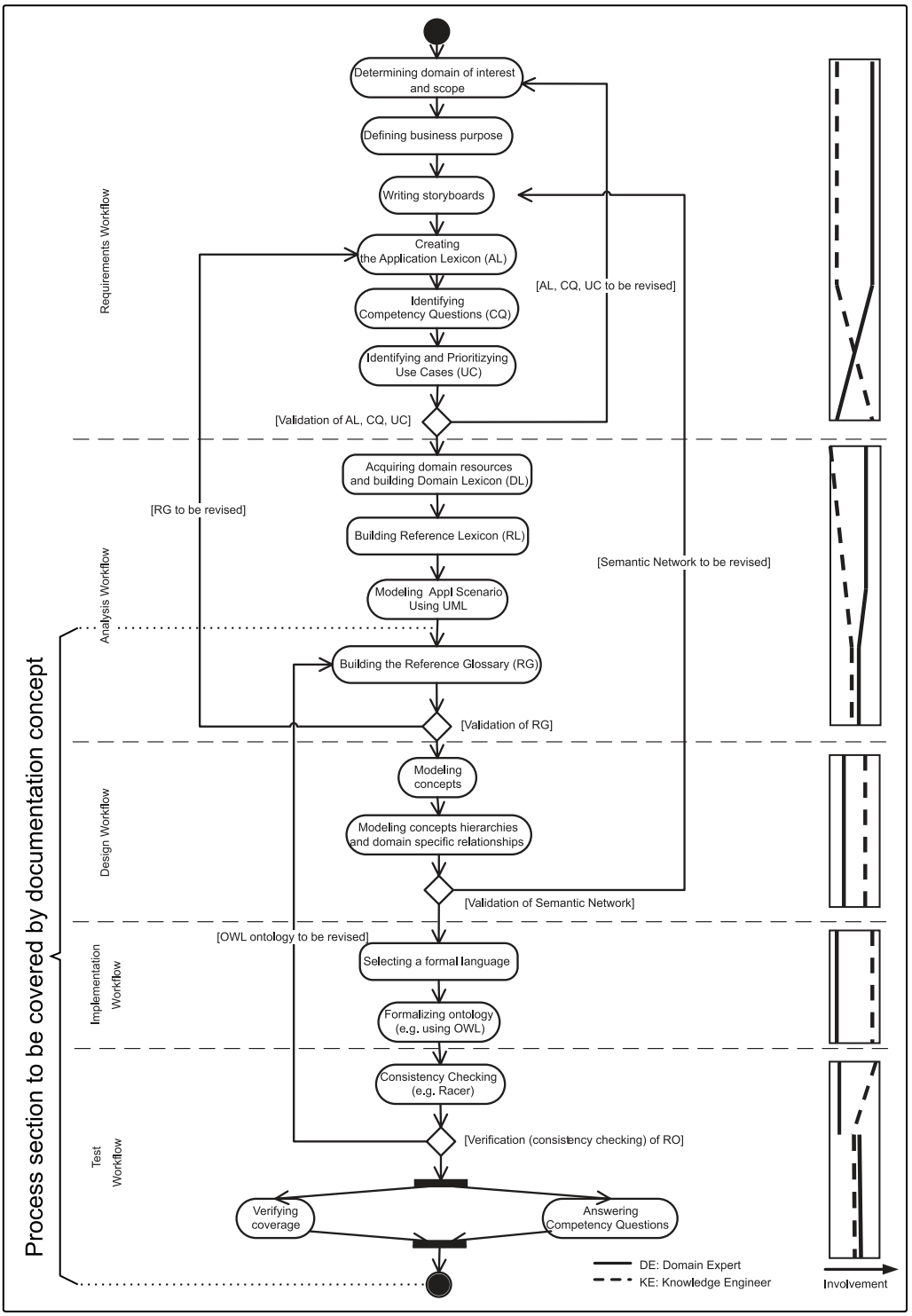

Figure 23: UPON process (Nicola et al., 2009, p. 273) on the left side supplemented by the process section to be covered by the documentation concept

4.1.4 Findings

The UPON is use-case driven and based on the identification of CQs, which are later used to test the pragmatic quality of the ontology. Accordingly, the *use cases and CQs related to a concept* must be considered in the documentation concept.

The UPON describes how a list of terms is transformed into a tested ontology. To do so, the terms in the RL are supplemented by definitions. Consequently, *definitions* must be considered in the documentation concept. Furthermore, it must be considered that the documentary reference unit changes. First, the focus is on *terms*. By assigning definitions, merging synonymous terms and modeling concepts, the *documentary reference unit changes from term to concept*.

During the design workflow, the concepts are assigned to *concept categories*. The UPON adopts these categories from OPAL. It is assumed that the use of OPAL is not very common. Nevertheless, the assignment of generic categories appears useful independently of OPAL and UML. On the one hand, because this facilitates potential connections to a top-level ontology (see chapter '2.3.4 Typology'). For instance, all concepts of the category *actor* can later be subordinated to the corresponding top-level class. On the other hand, the concept is described by the category in its nature. For example, if a concept named *pharmaceutical formulation* is categorized as an *object*, it corresponds to the composition of a drug, the 'recipe' so to speak. If *pharmaceutical formulation* is categorized as a *process*, it is the process of combining different components to produce a drug – not the 'recipe' according to which this is done. This example demonstrates how much meaning can be expressed by simply assigning a category. Therefore, *concept categories* must be considered in the documentation concept.

Next, the concepts are connected by relations to an informal SN. UPON refers to the relation types *generalization*, *aggregation*, *association* and *domain-specific relation*. Therefore, these *relationships between concepts* must be considered in the documentation concept.

In the course of the implementation, the informal SN is formalized. In order to facilitate an efficient distribution of workload between KEs, the *implementation status must be documented*. In this way, KEs can quickly identify which concepts still need to be formalized.

In addition, the *test results must be documented*. On the one hand, whether a formalized concept causes errors during reasoning (semantic quality). On the other hand, whether a concept has proven to be relevant for the intended application (pragmatic quality). In this way, KEs can quickly identify which reasoning errors need to be fixed and of which priority concepts are.

Looking at the involvement of KEs and DEs throughout the entire process, it is evident that both parties are always involved. Recommendations as to *who should document* are therefore not categorical but relate to tendencies. The continuous involvement of both target groups indicates that there should be *common documentation for KEs and DEs*.

According to UPON, definitions should be selected in agreement between the people involved. This corresponds to the understanding of the concept *ontology* in this thesis (see chapter '2.3 Ontology'). According to this, "an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group" (Studer et al., 1998, p. 184). The documentation concept must therefore find a way to *represent agreement*. If DEs can communicate their level of agreement individually and asynchronously, no meeting is required. This can be realized, for example, using a rating scale. For constructive feedback in case of disagreement it is necessary to allow additional free text input. Based on the overall rating and feedback provided by DEs, KEs can assess whether the concept is sufficiently defined or a meeting is actually required.

Moreover, definitions should be selected from knowledgeable sources. To be able to *trace the source of a definition*, it is necessary to document from which source a definition originates. If information originates from a non-human source, the latter corresponds to a resource as defined in chapter '3.3 Findings'. The UPON authors refer to this as *documental resources* and give the following examples: "reports, technical manuals, standards, glossaries, thesauri, legacy computational lexicons, and available ontologies" (Nicola et al., 2009, p. 262). This corresponds to the resources considered in chapter '3 Selective insights into corporate reality'.

An essential part of the UPON is the use of UML. The authors justify this with the fact that "UML has been already shown to be useful in building ontologies" (Cf. Guizzardi et al., 2002; Nicola et al., 2009, p. 259). They argue that "the adoption of the UP and the Unified Modeling Language (UML) makes ontology building an easier task for modellers familiar with these techniques" (Nicola et al., 2005, p. 655). Besides, they state that "diagramming,

documentation, and versioning can be performed with the aid of a variety of UML tools" (Nicola et al., 2009, p. 274). The approach of the documentation concept developed in this thesis contradicts the UPON approach in two aspects. On the one hand, this thesis assumes that a significant part of the people involved in ontology creation is not familiar with UML. On the other hand, the documentation concept is designed to support different approaches to ontology creation including those with less clearly defined use cases. Nevertheless, the documentation concept should *enable the use of additional visual elements* for those users who want to benefit from schematization using UML diagrams.

The following tables list the findings derived from the UPON. Table 15 contains the *relevant information per concept*, i.e. it states *what* needs to be documented for each concept. Table 16 contains general aspects that must be considered when creating the documentation concept. In the broader sense, these are *documentation principles*. They state *how* the documentation should be done.

Finding		**Derived from …**	
Title of finding	Description of finding	UPON workflow	Activity
Concept category	Enable documentation of generic category to which the concept is assigned, e.g. actor, object, process. Recommend documentation by KEs.	Design workflow	Assignment of concept categories during the activity *modeling concepts* with a tendency towards more involvement of KEs.
Definitions	Enable documentation of one or more definitions. Recommend documentation by DEs.	Analysis workflow	Capturing of one or more definitions during *building the reference glossary* with a tendency towards more involvement of DEs.
Implementation status	Enable documentation of the extent to which the concept is formally represented, e.g. using a controlled term list like: • not yet implemented, • partially implemented, • fully implemented. Recommend documentation by KEs.	Implementation workflow	Formalization of informal semantic network to create a formal ontology with a tendency towards more involvement of KEs.
Related competency questions	Enable documentation of related competency questions.	Requirements workflow and Test workflow	CQs are identified in the requirements workflow and used to verify the pragmatic quality of the ontology in the test workflow.
Related use cases	Enable documentation of related use cases.	Complete process	UPON is use-case driven. • *Identifying and prioritizing use cases* in the requirements workflow • *Modeling application scenario* in the analysis workflow • *Verifying coverage* in the test workflow

Table is continued on the next page.

Continuation of the table on the previous page.

Relationships to other concepts	Enable documentation of relationships to other concepts including relationships of the following types: · IsA: generalization (i.e. hyponymy), · part-of: aggregation (i.e. meronymy), · association, · domain-specific relation. Recommend documentation by KEs.	Design workflow	Use of the relation types generalization (IsA), aggregation (part-of), association and domain-specific relation for creation of a semantic network during *modeling concept hierarchies and domain-specific relationships* with a tendency towards more involvement of KEs.
Test status regarding reasoning	Enable documentation of whether the formally implemented concept causes errors when testing with a reasoner, e.g. using a controlled term list like: · not yet tested, · no errors, · causes erros. Recommend documentation by KEs.	Test workflow	Testing of the semantic quality by *checking consistency* using a reasoner with a tendency towards more involvement of KEs.
Test status regarding relevance	Enable documentation of the extent to which the concept has been assesed as relevant in a test, e.g. using a controlled term list like: · not yet tested, · irrelevant, · nice to have, · must have. Recommend documentation by DEs.	Test workflow	Testing of the pragmatic quality by *verifying coverage* and *answering competency questions* with a tendency towards more involvement of DEs.

Table 15: Findings derived from the Unified Process for Ontology Building – relevant information per concept

Finding		**Derived from …**	
Title of finding	Description of finding	UPON workflow	Activity
Additional visual elements	Enable text-based documentation to be supplemented by illustrations.	Complete process	Use of UML diagrams: · creation of use-case diagrams in the requirements workflow, · creation of activity and class diagrams in the analysis workflow, · usage of UML diagrams to verify pragmatic quality in the test workflow.
Documentary reference unit may change from term to concept	Take into account that the reference unit may change from term to concept during the process if the ontology is created based on a list of terms, e.g. a reference lexicon.	Analysis workflow and Design workflow	At the beginning of the analysis workflow, the focus is on terms. When adding definitions and modeling concepts, the focus is on concepts.
One documentation for all target groups	No separate documentations for DEs and KEs. Instead, common documentation with optional separation between · subject-related content (DE knowledge) and · content regarding knowledge engineering (KE knowledge).	Complete process	KEs and DEs are both involved throughout the entire process. In some activities, there are tendencies as to which group is more involved, e.g. KEs in the implementation workflow.
Representation of agreement	Enable documentation of the level of agreement for a documented statement, e.g. for a definition.	Analysis workflow	Definitions captured during *building the reference glossary* should be "agreed among DEs and users" (Nicola et al., 2009, p. 264).
Traceability of sources	Enable documentation of the source from which a statement, e.g. a definition, originates.	Analysis workflow	Definitions captured during *building the reference glossary* "should be selected from knowledgeable sources" (Nicola et al., 2009, p. 264).

Table 16: Findings derived from the Unified Process for Ontology Building – documentation principles

4.2 NeOn Methodology

The NeOn Methodology is one of the results of the NeOn project described above. In preparation for the following analysis, the numerous publications created in context of the NeOn project have been screened, including the *NeOn Book*[24] and the project deliverables[25]. For the analysis, the two publications shown in table 17 are considered. They provide the best overview as they have been published after the completion of the NeOn project and are intended to impart the NeOn methodology.

Analyzed NeOn publications
Suárez-Figueroa, Mari C.; Gómez-Pérez, Asunción; Fernández-López, Mariano (2015) 'The NeOn Methodology framework: A scenario-based methodology for ontology development', *Applied Ontology*, vol. 10, no. 2, pp. 107–145. DOI: 10.3233/AO-150145.
Suárez-Figueroa, Mari C.; Gómez-Pérez, Asunción; Motta, Enrico; Gangemi, Aldo, eds. (2012) *Ontology Engineering in a Networked World*, Berlin, Heidelberg, Springer Berlin Heidelberg.

Table 17: List of the analyzed literature regarding the NeOn Methodology

4.2.1 Overview

SUÁREZ-FIGUEROA, GÓMEZ-PÉREZ and FERNÁNDEZ-LÓPEZ describe the methodology as follows:

> "*The NeOn Methodology framework […] is a scenario-based methodology that provides accurate details about key aspects of the ontology engineering process, paying special attention to the reuse and reengineering of ontological and non-ontological resources*" (Suárez-Figueroa et al., 2015, p. 109)

This scenario-based approach is promising for the analysis in this thesis, because the "framework does not prescribe a rigid procedure, but suggests a variety of scenarios and guidelines for performing different processes and activities" (Suárez-Figueroa et al., 2015, p. 109). In this respect, the approach of the NeOn methodology corresponds to the approach of the documentation concept. However, when considering this methodology, it is more about discovering additional perspectives than about an in-depth analysis. The reason for this is the comprehensiveness of the framework which is evident from a look at its components:

---

[24] NeOn Project (n.y.) *NeOn Book: NeOn Methodology in a Nutshell* [Online]. Available at http://neon-project.org/nw/NeOn_Book.html (Accessed 2019-02-18).

[25] NeOn Project (n.y.) *Deliverables* [Online]. Available at http://neon-project.org/nw/Deliverables.html (Accessed 2019-02-18).

> "*This framework is founded on four pillars: (1) a glossary of processes and activities; (2) a set of nine scenarios for building ontologies and ontology networks; (3) two modes of organizing ontology developments, called ontology life-cycle models; and (4) a set of precise methodological guidelines for performing specific processes and activities.*" (Suárez-Figueroa et al., 2015, p. 109)

New perspectives for the documentation concept open up, especially when looking at the *nine **scenarios** for building ontologies and ontology networks*. They are derived from the following observation:

> "*Based on experience gained from our involvement in different types of projects, we identified three different ontology-building situations: (1) single ontologies; (2) sets of interconnected single ontologies; and (3) ontology networks.*" (Suárez-Figueroa et al., 2015, p. 111)

These experiences are in line with the personal experiences described in chapter '3 Selective insights into corporate reality'. According to NeOn, "Scenarios are sets of ordered processes and activities defined in the NeOn glossary" (Suárez-Figueroa et al., 2015, p. 111). The nine scenarios are as follows (Suárez-Figueroa et al., 2015, pp. 111–115):

- "Scenario 1: From specification to implementation."
- "Scenario 2: Reusing and reengineering non-ontological resources (NORs)."
- "Scenario 3: Reusing ontological resources."
- "Scenario 4: Reusing and reengineering ontological resources."
- "Scenario 5: Reusing and merging ontological resources."
- "Scenario 6: Reusing, merging and reengineering ontological resources."
- "Scenario 7: Reusing ontology design patterns (ODPs)."
- "Scenario 8: Restructuring ontological resources."
- "Scenario 9: Localizing ontological resources."

The authors consider these nine scenarios to be the most common but do not claim to be exhaustive (Cf. Suárez-Figueroa, Gómez-Pérez, Fernández-López, 2012, p. 12). Figure 24 is a visualization given by the authors which shows different possible pathways for the development of ontologies. The encircled numbers indicate the number of the corresponding scenario.

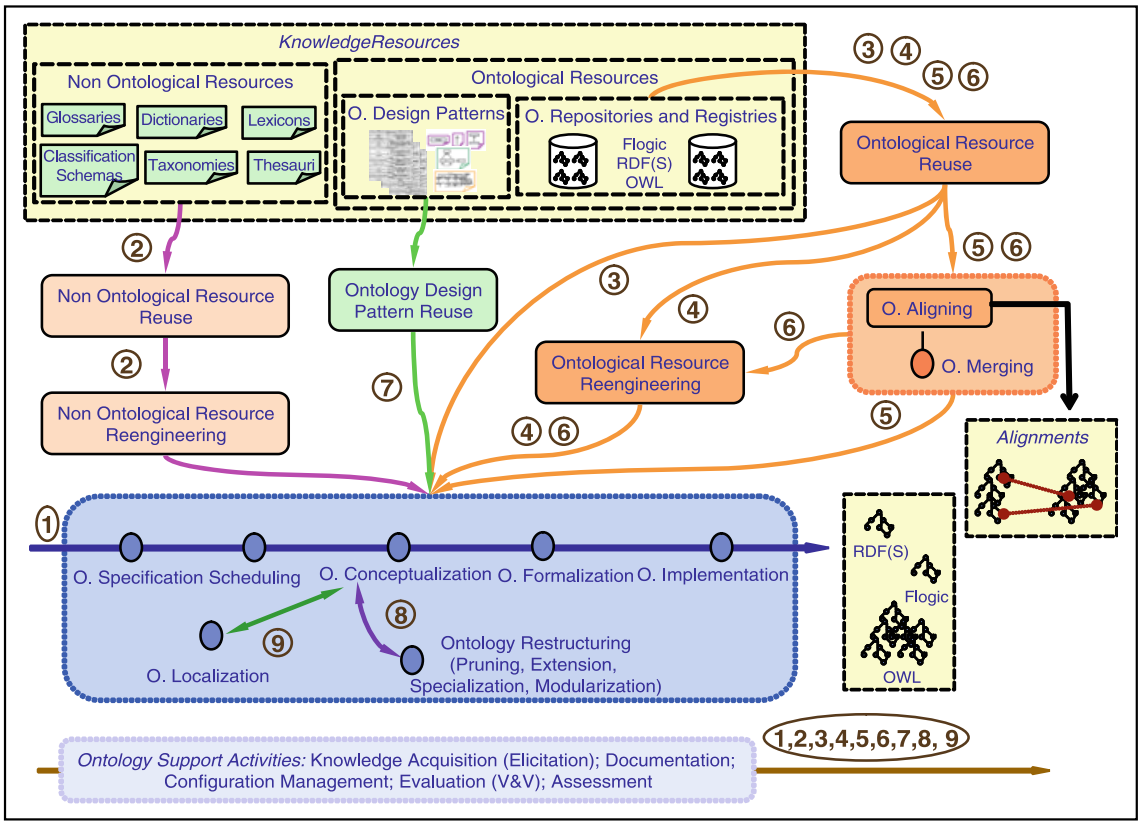

Figure 24: Scenarios for building ontologies and ontology networks (Suárez-Figueroa, Gómez-Pérez, Fernández-López, 2012, p. 13)

The scenarios "can be combined in different and flexible ways, and […] any combination of scenarios should include Scenario 1 because this scenario is made up of the core activities that have to be performed in any ontology development" (Suárez-Figueroa, Gómez-Pérez, Fernández-López, 2012, p. 14). Besides, "Knowledge acquisition, documentation, configuration management, evaluation, and assessment should be carried out during the whole ontology network development" (Suárez-Figueroa, Gómez-Pérez, Fernández-López, 2012, p. 14). The publications considered describe for each scenario the following:

> "*(a) motivation for the scenario; (b) sequence of processes, activities, and tasks to be carried out, where the processes and activities included are taken from the NeOn Glossary of Processes and Activities […]; and (c) outcomes for the scenario.*" (Suárez-Figueroa, Gómez-Pérez, Fernández-López, 2012, p. 14)

The aforementioned ***NeOn glossary***[26] "provides natural language definitions and explanations in English of the 59 key processes and activities potentially involved in ontology construction" (Suárez-Figueroa et al., 2015, p. 110).

[26] The terms of the NeOn Glossary and their definitions can also be found at: http://mayor2.dia.fi.upm.es/oeg-upm/files/pdf/NeOnGlossary.pdf (Accessed 2019-02-07)

With respect to the process of ontology creation as considered in this thesis (see chapter '2.4 Process'), not all nine scenarios are equally relevant. The focus is on the exchange of knowledge in the creation of new ontologies. Thus, the documentation concept primarily addresses *Scenario 1* which "refers to the development of ontologies from scratch" (Suárez-Figueroa et al., 2015, p. 111) and *Scenario 2* which refers to usage of non-ontological knowledge resources. However, it is not excluded that ontological resources may also be used for the representation of some concepts. Consequently, *Scenario 3*, *Scenario 4* and *Scenario 7* are also of relevance. The *aligning* and *merging* as well as *restructuring* of ontologies are not considered, i.e. *Scenario 5*, *Scenario 6* and *Scenario 8* are outside the scope. Since the documentation concept is oriented towards the application in companies (see chapter '1 Introduction'), it is assumed that a common business language can be used. Therefore, *Scenario 9* which refers to *localization* is considered less important.

Besides the scenarios, the methodology proposes *two ontology network* ***life-cycle models***. As the term indicates, they consider the entire life-cycle of ontologies, i.e. they begin before and end after the process to be covered by the documentation concept. Figure 25 shows the two life cycle models with the authors explaining the notation as follows:

> "*the rounded boxes with a solid black border represent the phases of the basic version of the model; the rounded boxes with a dotted black border represent optional phases depending on the model version; the square brackets represent model version-dependent composition options; and, finally, the dashed directed line denotes optional backtracking (in the maintenance phase) in order to fix errors or add missing knowledge. […] The braces […] denote repetition, and the numbers (2 and N) indicate the number of iterations performed.*" (Suárez-Figueroa et al., 2015, p. 118)

In addition, those phases that are of interest for the documentation concept are highlighted in grey. Since the primary objective is to find out *what is to be documented*, the aspect of *iteration* is not discussed further at this point.

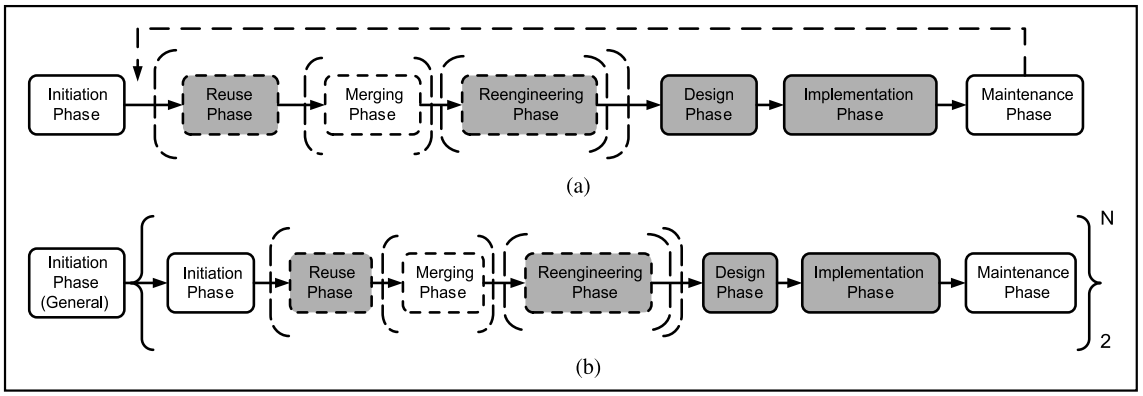

Figure 25: Ontology network life-cycle models proposed in the NeOn Methodology framework. (a) Waterfall model family. (b) Iterative-incremental model (Suárez-Figueroa et al., 2015, p. 118). Supplemented by grey markings of the considered phases.

The following section describes findings gained from an exploratory analysis of the NeOn Methodology. It is assumed that a systematic, in-depth analysis of all 59 processes and activities as well as life cycles and guidelines may provide further insights (see chapter '10 Outlook').

4.2.2 Findings

The NeOn Methodology is characterized by the categorization of knowledge resources into *non-ontological resources* (NORs) and *ontological resources*. The latter are divided into *ontology design patterns* (ODPs) and *ontologies from repositories and registries*. This categorization is suitable to give structure to the documentation concept. The understanding of the concept *resource* is consistent with the understanding applied in chapters '3 Selective insights into corporate reality' and '4.1 Unified Process for Ontology Building UPON'. Another characteristic of NeOn is its focus on reuse. Therefore, the categorization of knowledge resources that is derived for the documentation concept refers to the *resource type* and the *reuse context*.

With the ODPs, a resource type is identified that has not yet been explicitly considered. Moreover, the NeOn authors discuss the characteristics of NORs in detail. As the following figure 26 shows, NORs are characterized by their type, the underlying data model, and their implementation. To document these details for NORs is considered unreasonable in terms of effort and added value. Only those distinctions should be documented that offer added value and are not inherent to the resource or the reuse context.

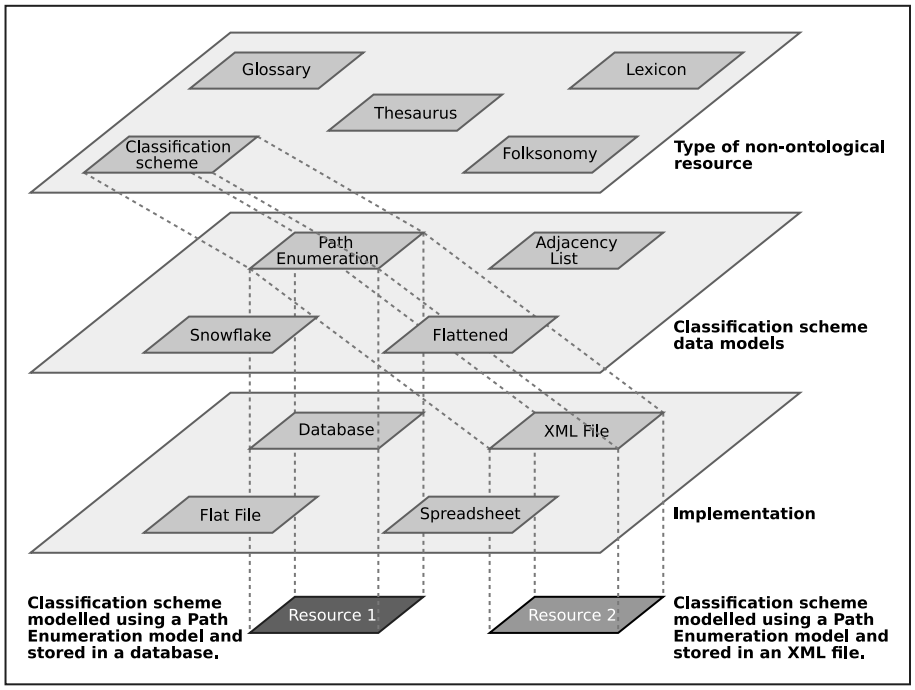

Figure 26: Non-ontological resource categorization (Villazón-Terrazas and Gómez-Pérez, 2012, p. 109)

Primarily, the documentation concept makes the following distinction: a) *domain knowledge resources* versus *engineering knowledge resources*. In this way, DEs who are interested solely in domain knowledge can quickly identify the resources relevant to them. ODPs are usually not domain-specific. Therefore, they must be documented only as engineering knowledge resources. Secondly, the NeOn categorization is adopted: b) *non-ontological resource* versus *ontological resource*.

The third distinction refers to the *reuse context* and is only necessary to document for domain knowledge resources: c) *TBox-related reuse* versus *ABox-related reuse*. Engineering knowledge resources are generally focused on implementation and integration without any direct reference to TBox or ABox. Differentiation (c) refers to whether a resource is used to create the conceptual model or to generate instances. The NeOn Glossary distinguishes between three ways of *knowledge acquisition for ontologies* (Suárez-Figueroa, 2011, pp. 1–2):

- "*Ontology Elicitation. It is a knowledge acquisition activity in which conceptual structures (e.g., T-Box) and their instances (e.g., A-Box) are acquired from domain experts.*"
- "*Ontology Learning. It is a knowledge acquisition activity that relies on (semi-) automatic methods to transform unstructured (e.g., corpora), semi-structured (e.g., folksonomies and html pages) and structured data sources (e.g., data bases) into conceptual structures (e.g., T-Box).*"

- "*Ontology Population. It is a knowledge acquisition activity that relies on (semi-) automatic methods to transform unstructured (e.g., corpora), semi-structured (e.g., folksonomies and html pages) and structured data sources (e.g., data bases) into instance data (e.g., A-Box).*"

Accordingly, reuse of resources for *ontology learning* is *TBox-related* while reuse of resources for *ontology population* is *ABox-related. Ontology Elicitation* refers to knowledgeable people. These are referred to as DEs in the context of this thesis. Resources, on the other hand, are defined as non-human. Thanks to the documentation of the categorization (c), a KE can quickly identify which resource is to be considered for conceptual structures (e.g. a product classification) and which for concrete examples or data to be integrated (e.g. a database table in which article numbers are assigned a notation of the product classification.).

In addition, for ontologies, the following distinction is documented: d) *reuse without change* versus *reuse with change*. It indicates whether an ontology is reused without modification (e.g. importing an ontology module) or modified for reuse (e.g. by re-engineering an ontology). NORs must always be changed, at least regarding formalization and implementation. Therefore, it is not worthwhile to document distinction (d) for NORs. Figure 27 below illustrates the *categorization of engineering knowledge resources* according to the presented distinction (b). Figure 28 illustrates the *categorization of domain knowledge resources* according to the presented distinctions (b), (c) and (d).

In the context of differentiation (d), it is important to note that in this thesis the term *ontology* is defined according to its understanding in computer science (see chapter '2.3 Ontology'). Consequently, a vocabulary that can be characterized as a *formal, explicit specification of a shared conceptualization* is considered an *ontological resource* in the documentation concept. This also applies to vocabularies such as classifications and thesauri. If the vocabulary does not meet these criteria, for example because it is *informal*, it is considered a NOR.

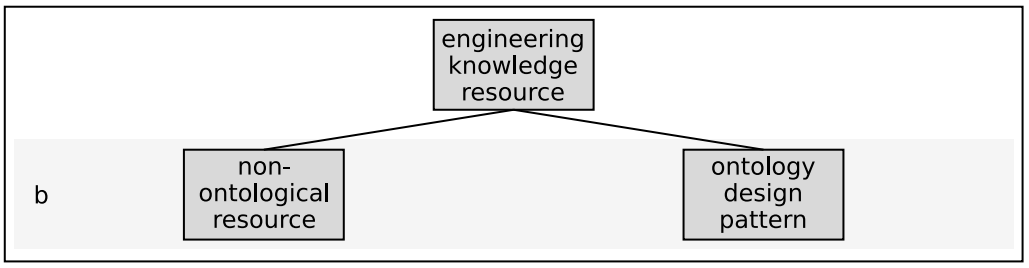

Figure 27: Categorization of engineering knowledge resources based on the distinction b) non-ontological resource vs. ontological resource.

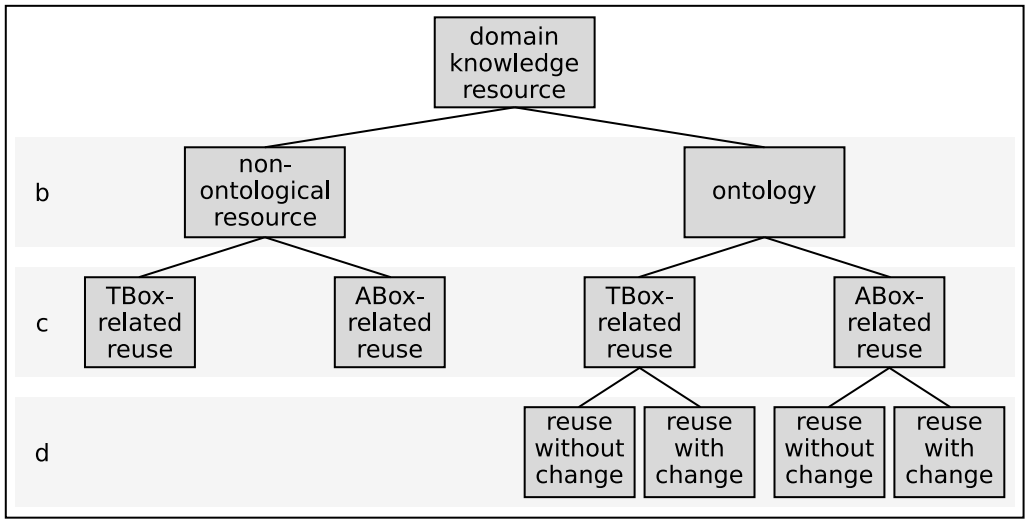

Figure 28: Categorization of engineering knowledge resources based on the distinctions b) non-ontological resource vs. ontological resource, d) TBox-related reuse vs. ABox-related reuse, and d) reuse without change vs. reuse with change.

The NeOn Glossary explicitly distinguishes between the terms *formalization* and *implementation* as follows (Suárez-Figueroa, 2011, p. 2):

- "*Ontology Formalization. It refers to the transformation of a conceptual model into a formal or semi-computable model according to a knowledge representation paradigm (e.g., description logics, frames, and rules).*"
- "*Ontology Implementation. It refers to the activity of generating computable models according to the syntax of a formal representation language (e.g., RDF(S), OWL, and FLogic).*"

This detailed distinction is made neither in this thesis nor in the documentation concept. *Implementation* refers to representation of *informal* statements *by means of an ontology language* as *formal* statements. The intermediate step of creating *formal or semi-computable model according to a knowledge representation paradigm* is not considered to stay with the practice in companies and to keep the documentation effort low.

5 Documentation concept

In the following, the documentation concept is presented which has been developed on the basis of the findings derived in chapters '3 Selective insights into corporate reality' and '4 Analysis of ontology engineering methodologies'. The first step is to explain what the *documentary reference unit* is. Subsequently, the structure of the documentation concept is described. Finally, the components of the documentation concept, called *modules*, are presented. Each module is first presented as a *documentation template* and then *applied for an example*.

5.1 Documentary reference units

In documentation science, a distinction is made between *documentation object*, *documentary reference unit* and *documentary unit*, which is as follows (Cf. Gaus, 2005, p. 12; Wersig, 1971, pp. 185–186; Wersig and Neveling, 1976, p. 98):

- A ***documentation object*** is the object used in the documentation process to produce documentary data. An example of such a *documentation object* could be a concrete physical book.

- Since the *documentation object* does not always enter the documentation process in its entirety, it is distinguished from the *documentary reference unit*. The ***documentary reference unit*** is the amount of data of the *documentation object* that is included as a unit in the documentation process. In other words, it is the part of the *documentation object* whose properties are regarded as the unit to be documented. For example, the *documentation object* could be a concrete physical book, while the *documentary reference unit* is only a chapter of this book. In this case, the properties of the chapter are to be documented, e.g. title and author of the chapter.

- If the values for these properties are now entered in a database, a *documentary unit* is created. A ***documentary unit*** is the set of documentary data that is created as a representative of a *documentary reference unit* during the documentation process. In other words, it's a data set describing the *documentary reference unit*.

To reduce complexity, the distinction between *documentation object* and *documentary reference unit* is not discussed further in the following. The focus is on explaining *documentary reference unit* and *documentary unit* for the documentation concept presented in this thesis. The findings presented in chapter '3 Selective insights into corporate reality' have shown that the documentation must be *concept-oriented*. Accordingly, *concepts* are used as the primary *documentary reference unit*. *Concepts* are understood as *units of thought* as described in chapter '2.3.2 Structure and components'. During the documentation process, a *documentary unit* is created for each concept. In order to reduce the potential for confusion, this *documentation unit* is referred to as ***documentation entry*** in the following.

However, as the analysis in chapter '4.1 Unified Process for Ontology Building UPON' has shown, the knowledge exchange in ontology creation may also commence by analyzing *terms*. Only by assigning definitions, merging synonymous terms and modeling concepts, the *documentary reference unit* changes from *term* to *concept*. ***Consequently, depending on whether concepts have already been identified, the primary documentary reference unit is a term or a concept.***

The term ***primary*** *documentation unit* is used because, strictly speaking, other *documentary reference units* are considered as well. For instance, the documentation concept recommends how information on *knowledge resources* and *people* should be documented. These are, so to speak, *secondary documentary reference units*, depending on how the documentation process is carried out. Since the documentation concept is designed for implementation in companies, this detail is not discussed further.

The following figure 29 explains the relationship between *documentary reference unit* and *documentary unit* in a visual way using three example scenarios:

a) Properties of a book are recorded in a database.

b) A documentation entry is created for a term found by text mining.

c) A documentation entry is created for a concept, as intended by the proposed documentation concept.

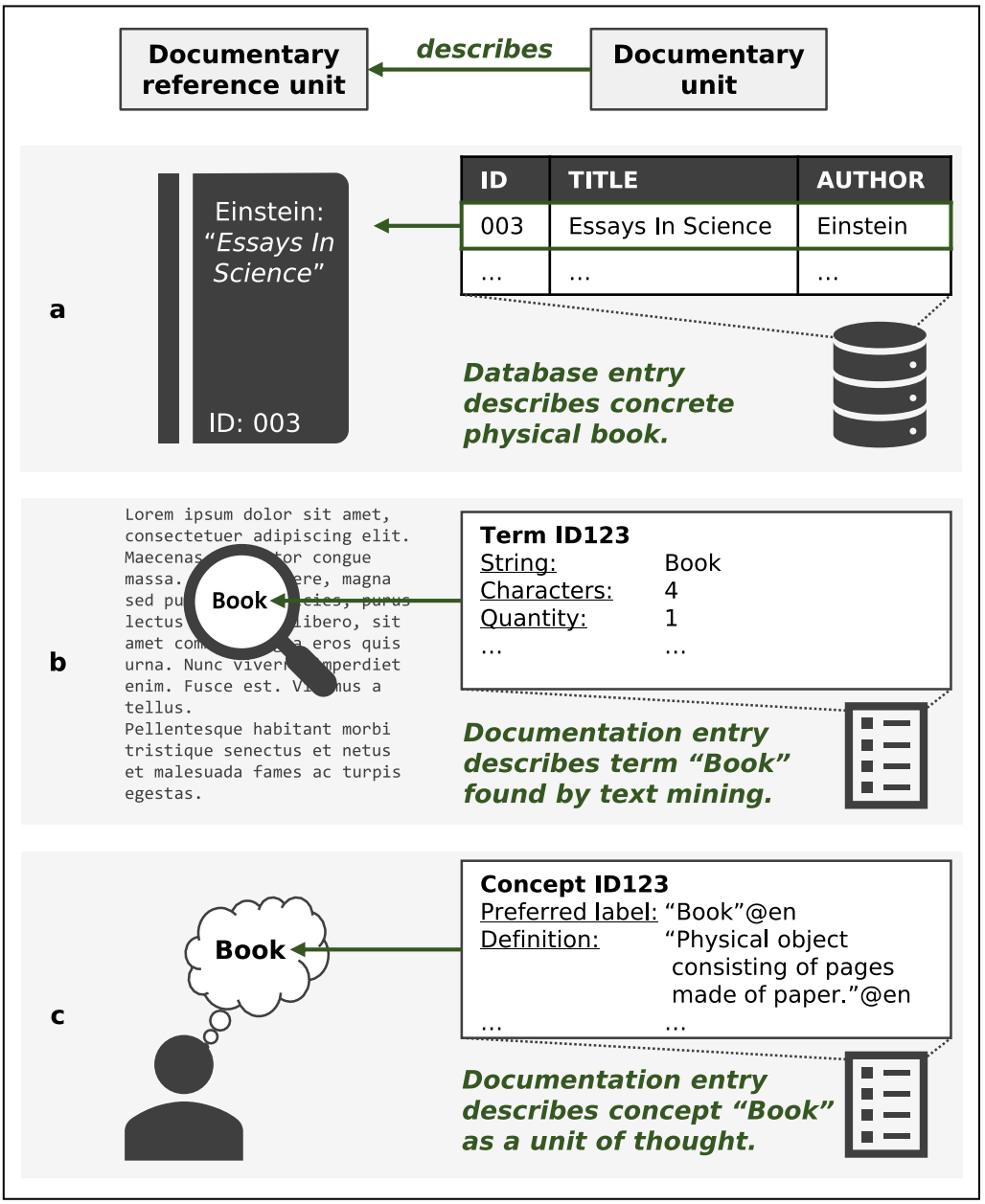

Figure 29: Visual explanation of the relationship between 'documentary reference unit' and 'documentary unit' using three different example scenarios.

5.2 Structure of the documentation concept

As defined in chapter '2.1 Documentation concept', the documentation concept is a modular system consisting of documentation templates. These templates are referred to as ***modules***. There are different types of modules.

The central and therefore mandatory module is the *'Primary documentary reference unit'* module. Since the documentation approach is concept-oriented this template has to be completed first. It is the starting point for a documentation entry. All other modules are used to describe the *primary documentary reference unit*. Usually the entire documentation consists of several documentation entries, each describing an a concrete term or concept. Figure 30 schematically illustrates a documentation consisting of three documentation entries in the minimal configuration without descriptive modules.

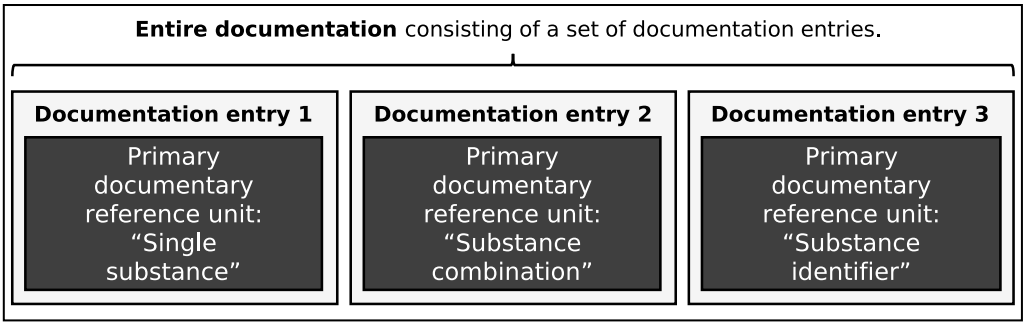

Figure 30: Schematic illustration of a documentation consisting of three documentation entries in the minimal configuration without descriptive modules.

In the course of ontology creation, concepts can be viewed from different perspectives. The documentation concept covers the following three perspectives which are referred to as ***dimensions***:

- The ***domain knowledge dimension*** consideres all the knowledge and knowledge resources that are available for a concept. This includes everything that describes the concept in its nature, i.e. its meaning, its labels, its position in the semantic network, et cetera. This perspective refers to those steps in the process of ontology creation that are referred to in methodologies, for instance, as *analysis*, *design* or *conceptualization* activities. Usually this dimension corresponds to the perspective of DEs whose inputs the KEs use to create the ontology.

- The ***engineering knowledge dimension*** considers all the knowledge and knowledge resources that refer to the implementation of the concept. The focus here is not on the description of the concept's nature, but on how the semi-structured, natural language description of the concept can be formally represented using an ontology language. This may involve formalization matters, but also aspects of integration such as characteristics of the software, technologies or infrastructure used. In methodologies, the associated activities are referred to as *implementation* or *formalization*, for example. Usually, this dimension is only relevant for KEs. However, as the chapter '3 Selective insights into corporate reality' has shown, some DEs are also interested in this perspective.

- The ***management knowledge dimension*** considers the handling of the concept and the associated documentation entry. This includes knowledge about the *priority* and *affiliation* of a concept. In other words, the dimension considers to which concept sets or ontologies a concept belongs, for which use cases and competency questions a concept is relevant, et cetera. Accordingly, *management* does not refer to project management, milestones or the like, but to the management of the concept. The explicit recording of such knowledge is intended to facilitate the handling of the documentation for readers and contributors from both target groups DEs and KEs.

The following figure 31 illustrates the three different perspectives using exemplary questions that may arise when people think about a concept.

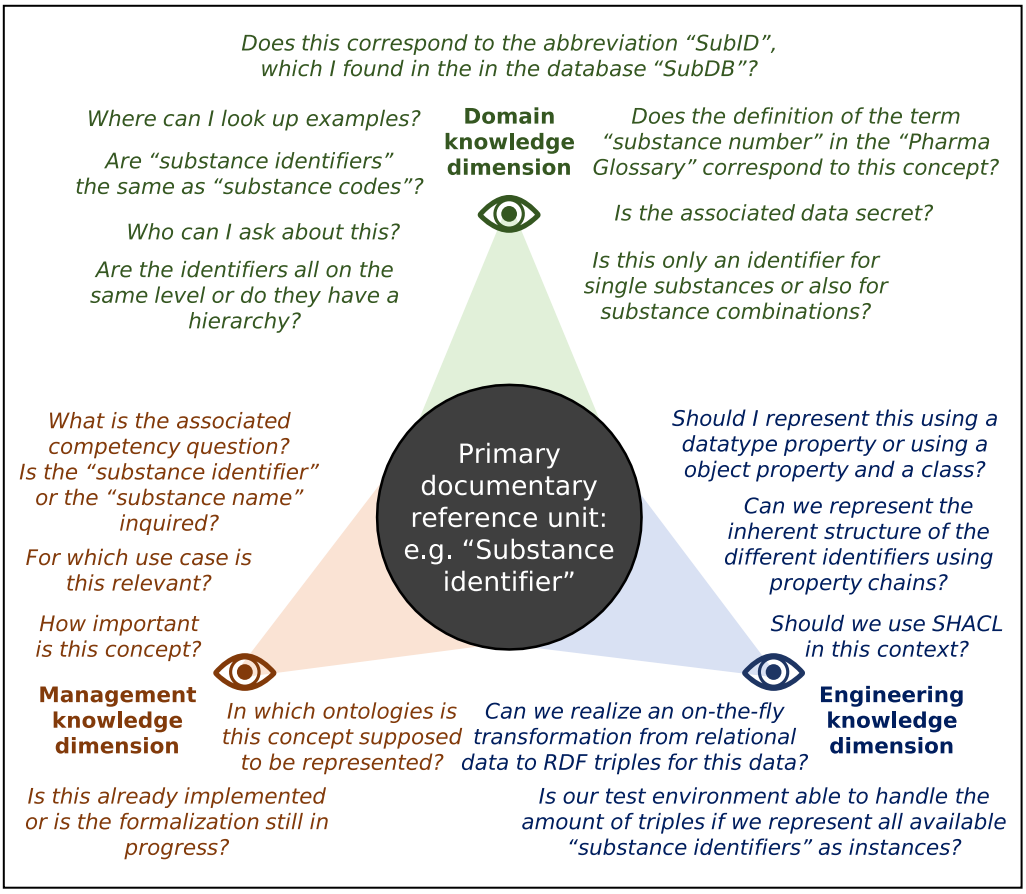

Figure 31: Visualization of the three different perspectives on a concept, called dimensions, with questions that people may have when they think about the exemplary concept "substance identifier".

For each of these three dimensions, the documentation concept provides ***dimension-specific modules*** for describing a concept. Figure 32 shows the recommended assignment of these modules to dimensions. Thanks to the modular documentation approach, the recommended assignment can be adapted by the user of the documentation concept if required. In addition, ***general modules*** are offered. General modules include generic information. Therefore, they are not inherently dimension-specific. They can be used to *extend the set of dimension-specific modules*. ***Module extensions*** can be used to *extend a single module*. Theoretically, module extensions can be applied on virtually any module. However, it is recommended to use module extensions selectively and sparingly in order to keep the documentation effort low and the clarity high. Figure 33 shows the general modules as well as the module extensions. Figure 34 shows the recommended extension of the dimension-specific modules by general modules and module extensions.

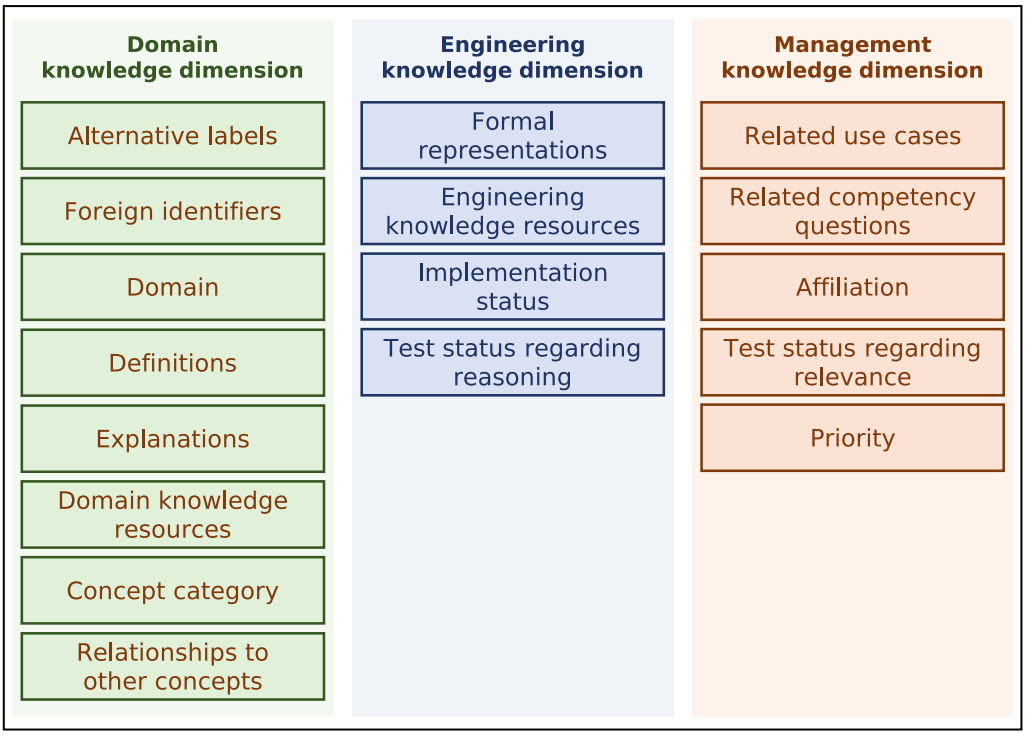

Figure 32: Dimension-specific modules according to the recommended assignment to dimensions

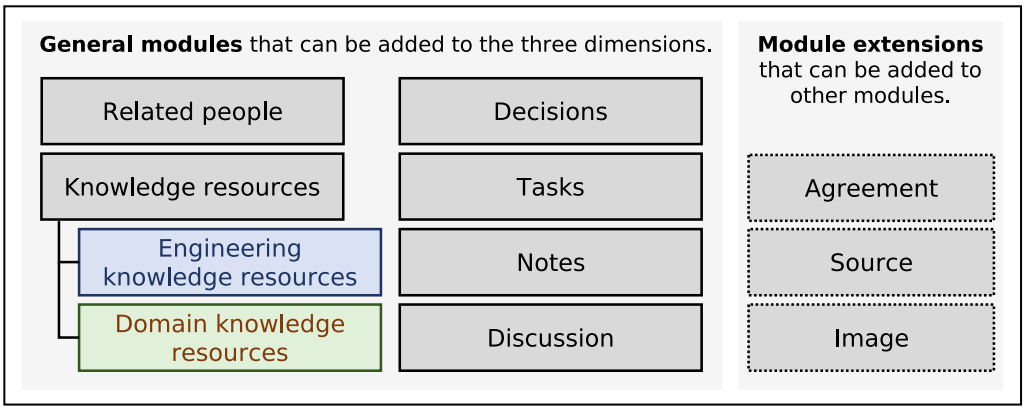

Figure 33: General modules and module extensions

The modules are described individually in the following sections. At this point, only a short note is made for the general module *'Knowledge Resources'* (see figure 33): Its general properties are supplemented by dimension-specific properties in the modules *'Engineering knowledge resources'* and *'Domain knwoledge resources'*. Although it may be used in its general form, it is recommended to use the dimension-specific modules for comprehensive documentation.

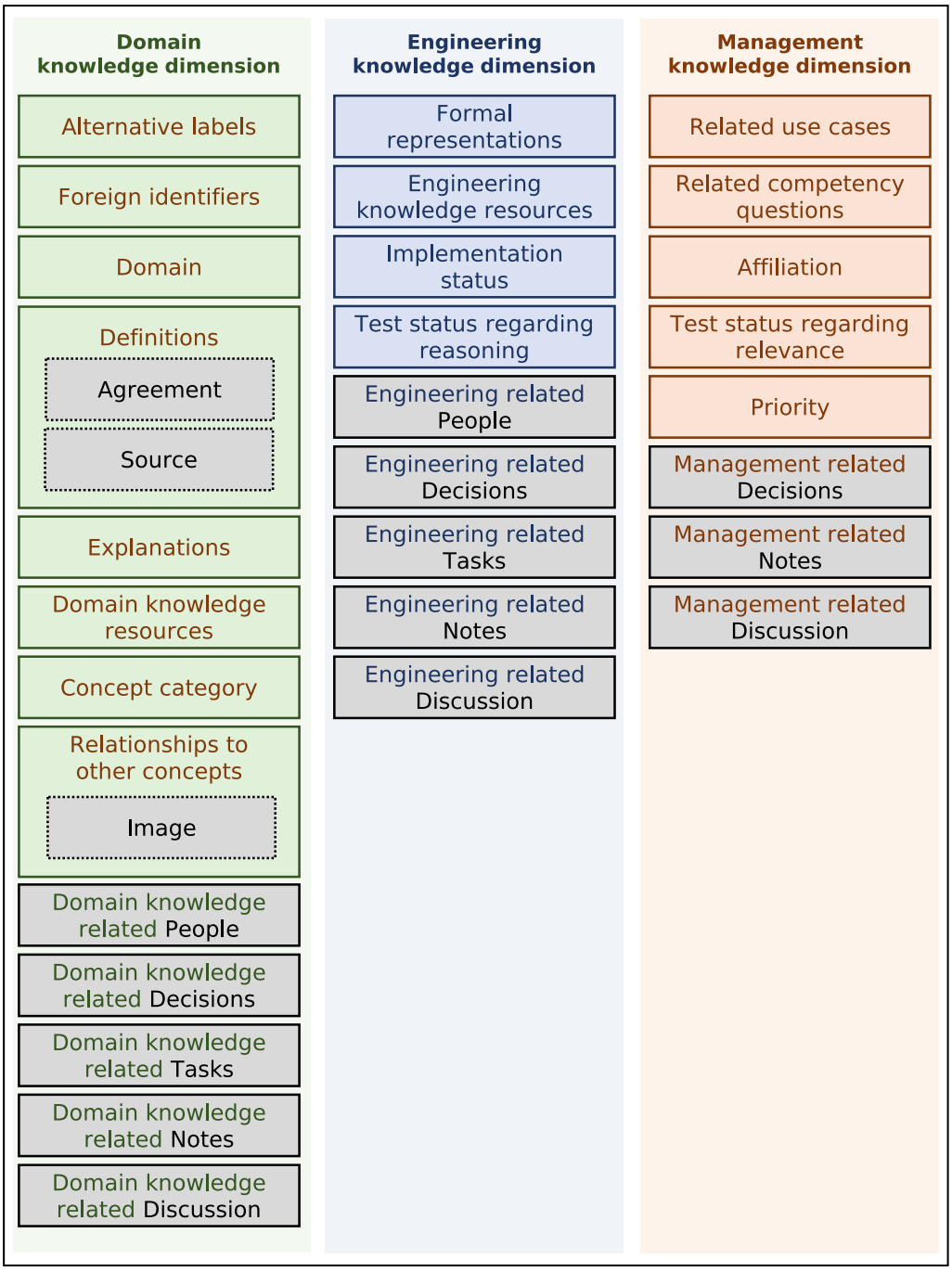

Figure 34: Recommended compilation of dimension-specific modules extended by general modules and module extensions

Figure 35 below shows the composition of a single documentation entry in a compact way. On the right side of the figure, some aspects are listed that are not covered by the documentation concept. The next figure 36 shows what a complete documentation consisting of four documentation entries looks like when the primary documentary reference unit is described by the presented modules.

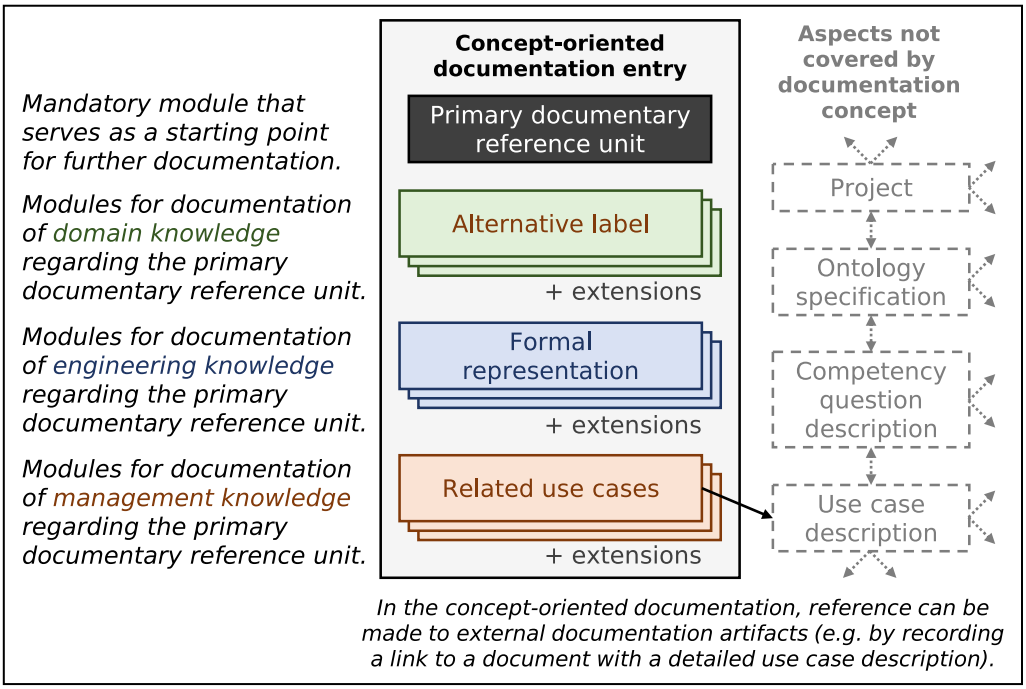

Figure 35: Schematic illustration of a single documentation entry with explanations on the left side. On the right, potential documentation aspects are shown that are not covered by the documentation concept.

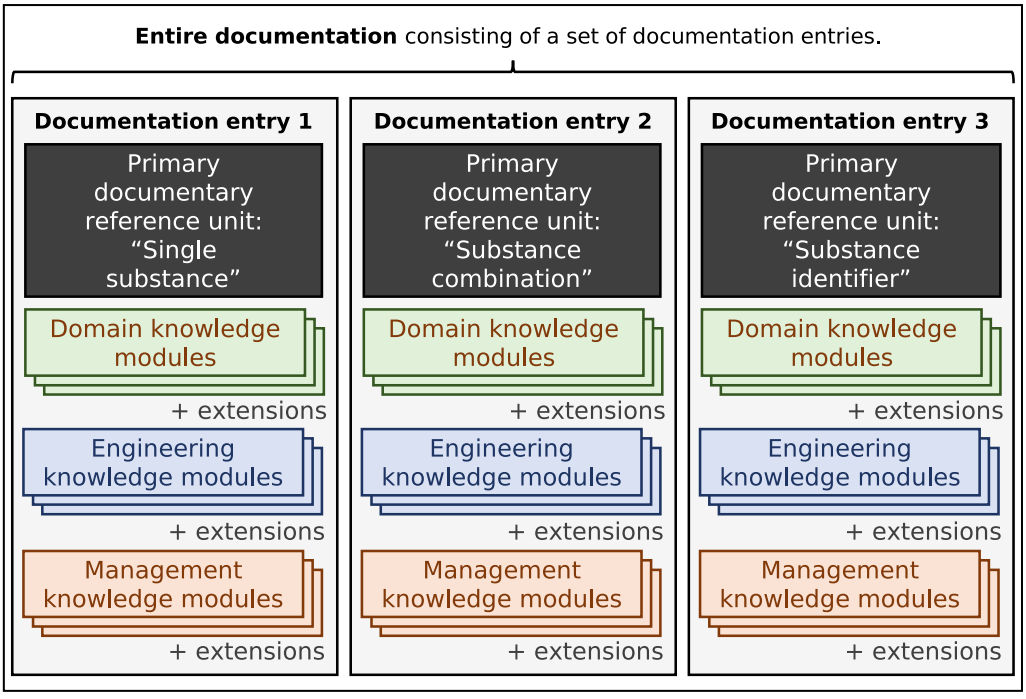

Figure 36: Schematic illustration of a documentation consisting of three documentation entries in the complete configuration using all available module types.

5.3 Structure of documentation modules

The following describes how the documentation modules are structured, i.e. which components they consist of. The insights into corporate reality showed that two levels of granularity are required for documentation (see chapter '3 Selective insights into corporate reality'). For some modules there are therefore two versions: *full* and *light*. Modules in the *full* version are used for *detailed documentation*. Modules in the *light* version are used for *fast documentation*. Thanks to these two versions, the documentation concept can not only be adapted to fast working methods by leaving out modules but also by reducing the level of detail.

Modules consist of *properties* and *statements*. The two terms *property* and *statement* are based on the structure of RDF triples (see chapter '2.3.3 Ontology languages'), however, the informal documentation concept is presented using graphical templates. Each module consists of one or more ***properties***. The full version of the module *'Alternative Labels'*, for example, includes the properties *'Preferred Abbreviation'*, *'Synonym'* and *'Abbreviation'*. A module groups, so to speak, a set of properties. Light versions simply contain fewer properties and are therefore less complex and comprehensive.

In the case of the *'Synonym'* property, the reference unit is the *primary documentary reference unit*. According to the structure of triples, this reference unit corresponds to the *subject*. The property is applied by documenting a concrete value. For example, the property *'Synonym'* is used for the concept "Substance identifier" by entering the synonym "Substance number" in the documentation. By doing so, a ***statement*** is made about the reference unit. If desired and applicable, a ***language tag*** can be assigned to natural language entries. This enables low-level multilingual documentation.

A possible representation of the given example as a triple is the following. The TBox statement represents a part of the module structure while the ABox statement represents the documented fact:

TBox: `<Reference unit>` `<has synonym>` `"Literal"@lang .`

ABox: `<Substance identifier>` `<has synonym>` `"Substance number"@en .`

As the implementation of the documentation concept is not in the focus of this thesis, it is not discussed further at this point but in chapter '8 Outlook'.

Using the full version of the module *'Alternative Label'* as an example, the following figures 37 and 38 explain the structure of modules. Figure 37 shows the module in question as a template. Figure 38 demonstrates its application for the concept "Substance identifier". In both figures, the components *module*, *property* and *statement* are indicated.

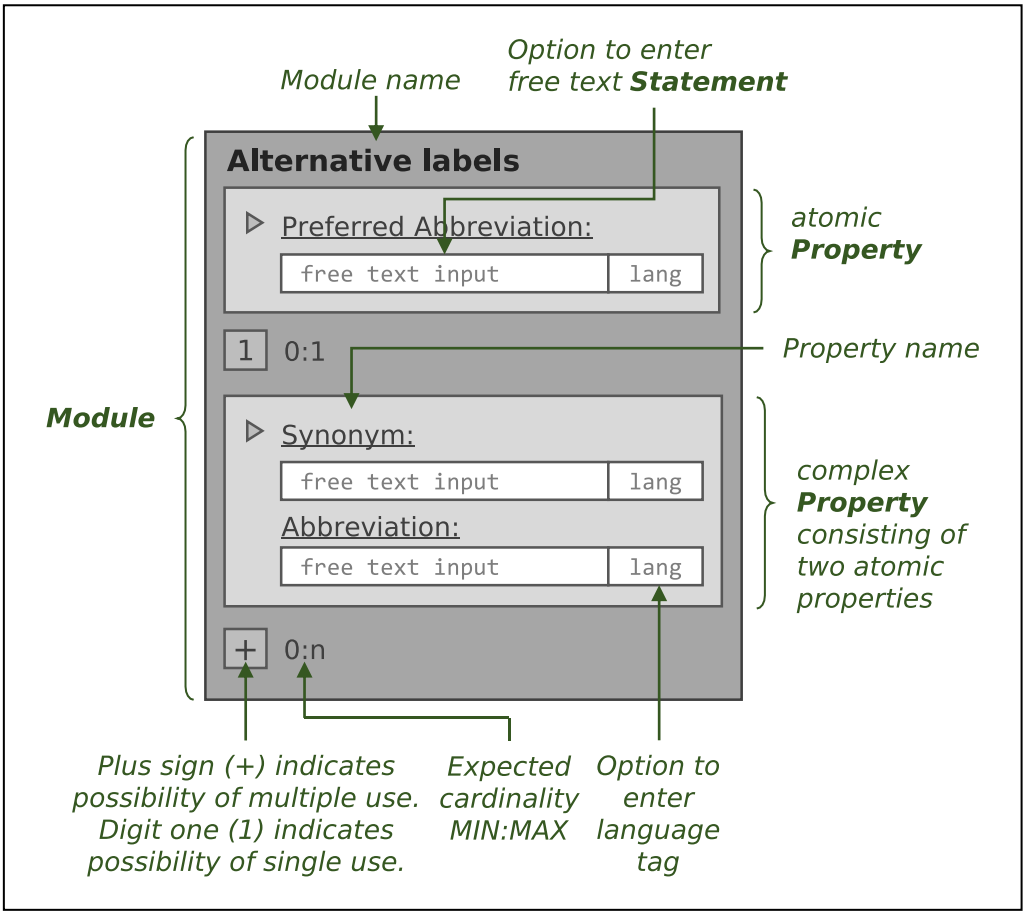

Figure 37: Annotated template for the module 'Alternative labels' (full version)

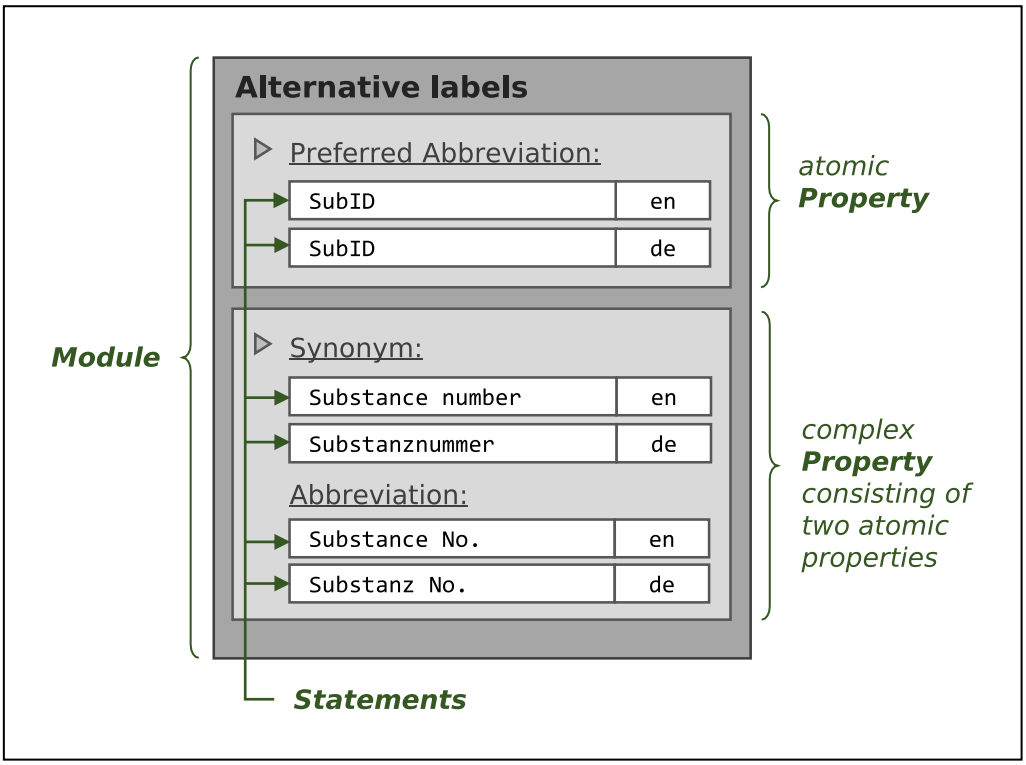

Figure 38: Annotated application of the module 'Alternative labels' (full version) for the reference unit "Substance identifier"

In addition, symbols and notes at the top of the modules indicate the module type, the version and, if applicable, the dimension (see figure 39). At the bottom there are recommendations who should use the module and when (figure 40).

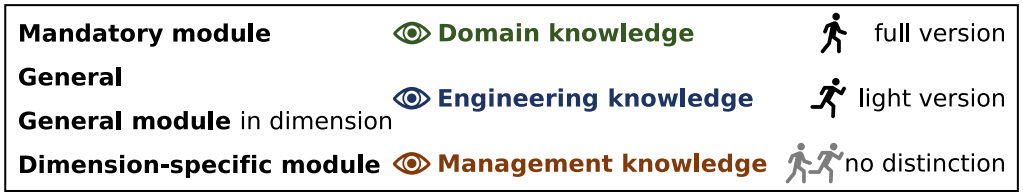

Figure 39: Loose collection of symbols and hints displayed at the top of the modules. Left: Type of the module. Center: Dimension. Right: Version of the module.

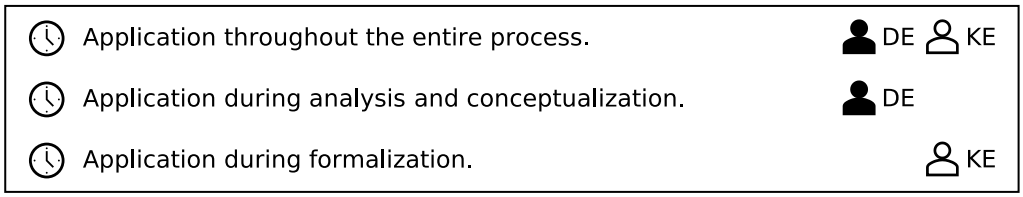

Figure 40: Selection from the occurring combinations of recommendations about target group and time of module application

5.4 Mandatory module: Primary documentary reference unit

As described above, the module *'Primary documentary reference unit'* is the only mandatory module of the concept-oriented documentation. It serves as a reference point for the following modules. Figures 41 and 42 below show the module as a template and in its exemplary application. The *preferred label* is intended for the readers of the documentation as well as for the transfer into the ontology. The ID (*identifier*) allows permanent identification and referencing of the documentation entry, even if the preferred label of the concept changes.

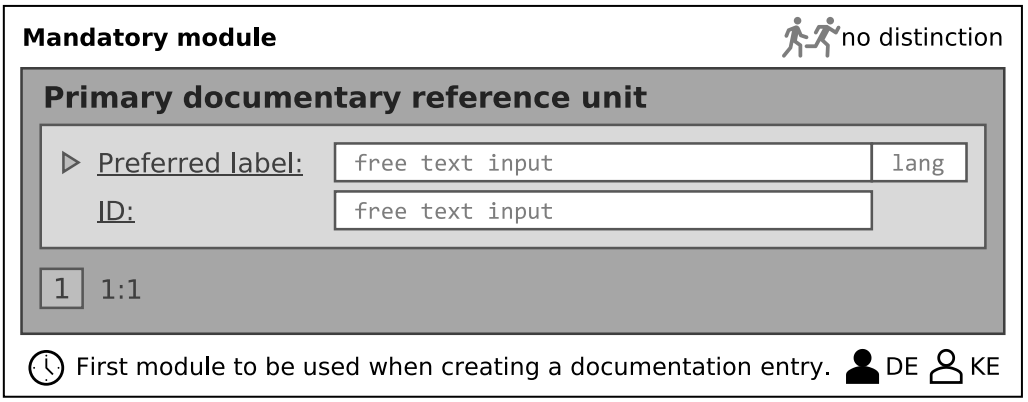

Figure 41: Mandatory module 'Primary documentary reference unit'

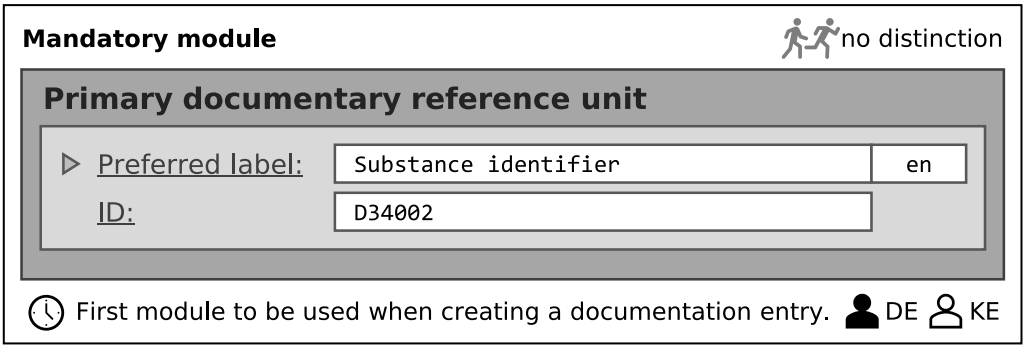

Figure 42: Exemplary application of the mandatory module 'Primary documentary reference unit'

5.5 General modules and module extensions

5.5.1 Related people

Figures 43 and 44 show the general module *'Related people'* in its full and its light version. The full version is applied exemplarily in the dimension *'Domain knowledge'* (see figure 45). The light one is demonstrated in the dimension *'Engineering knowledge'* (see figure 46). The *E-mail* or *Employee ID* is usually unique within a company. They are therefore suitable for referring to a person without much effort. If a corporate register of persons is available, these characteristics can be used to look up additional information such as telephone or room numbers. Besides, the e-mail address can also be used to establish contact. The *Role* can be specified to capture the position or responsibilities of a person in the company; the *Expertise* to indicate due to which knowledge a person is listed in the documentation.

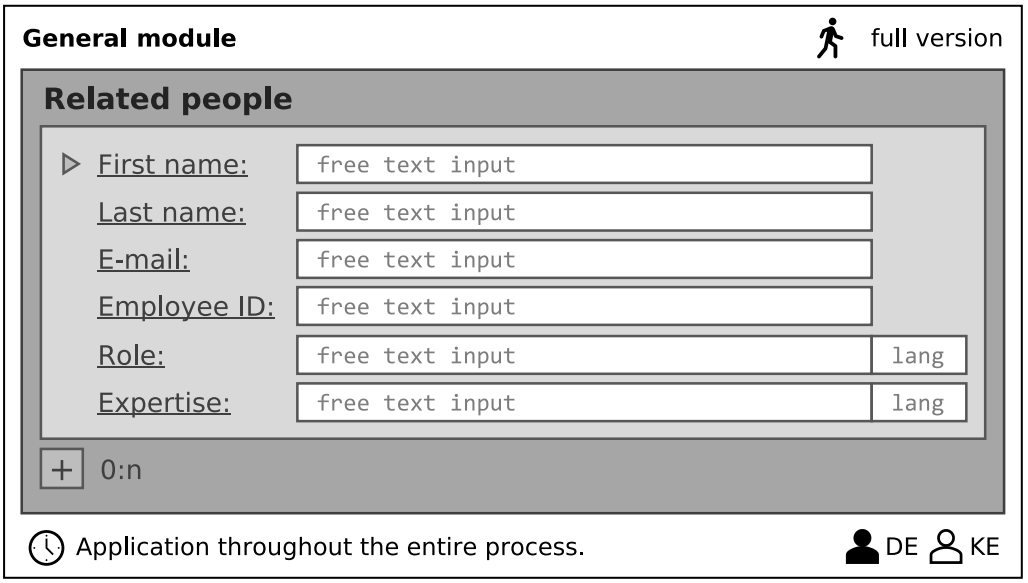

Figure 43: General module 'Related people' (full version)

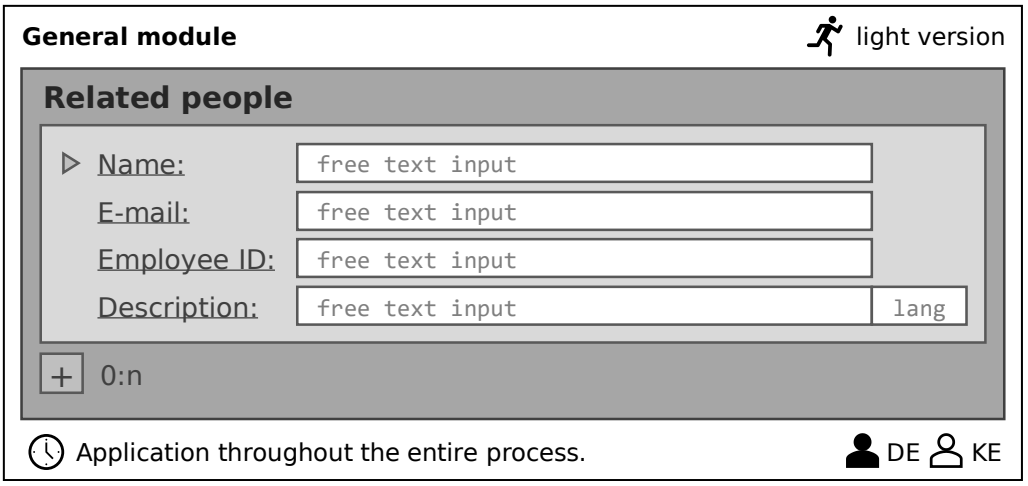

Figure 44: General module 'Related people' (light version)

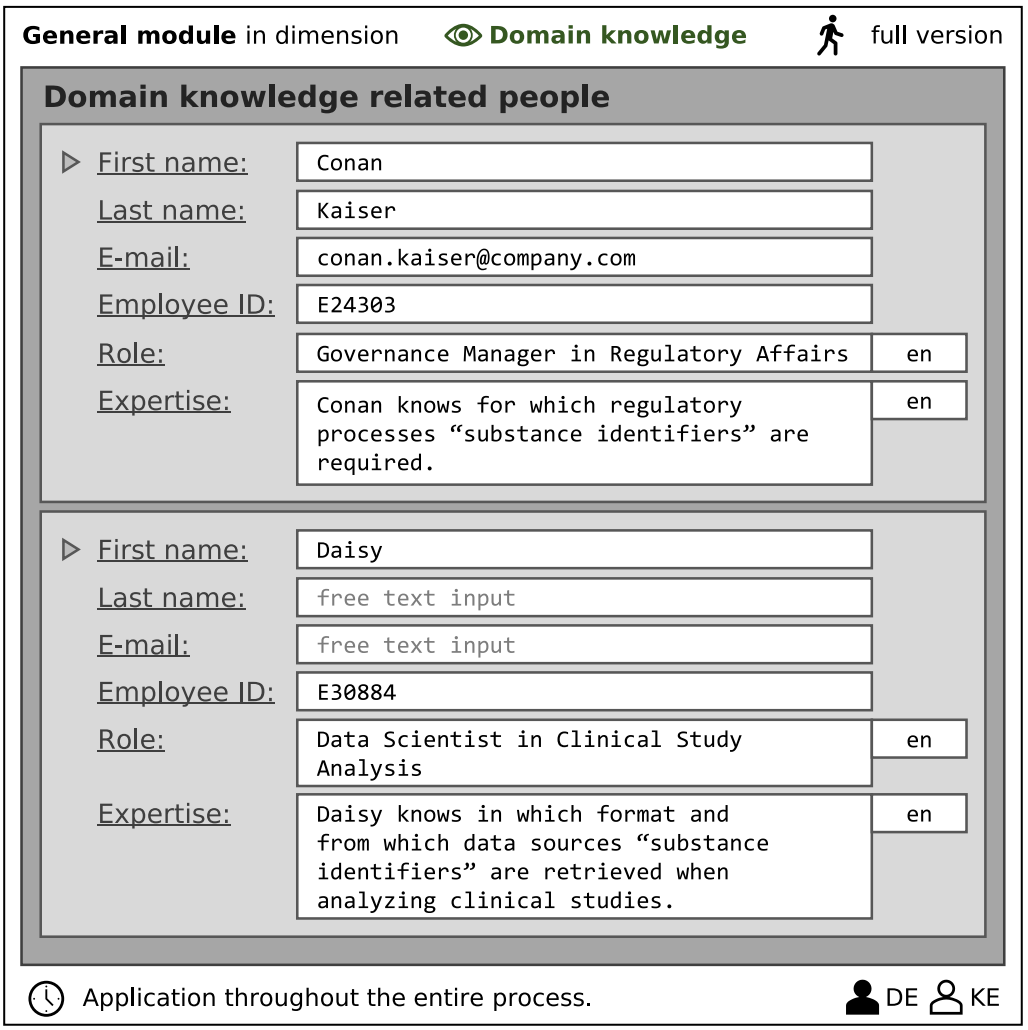

Figure 45: Exemplary application of the general module 'Related people' (full version) in the dimension 'Domain knowledge'

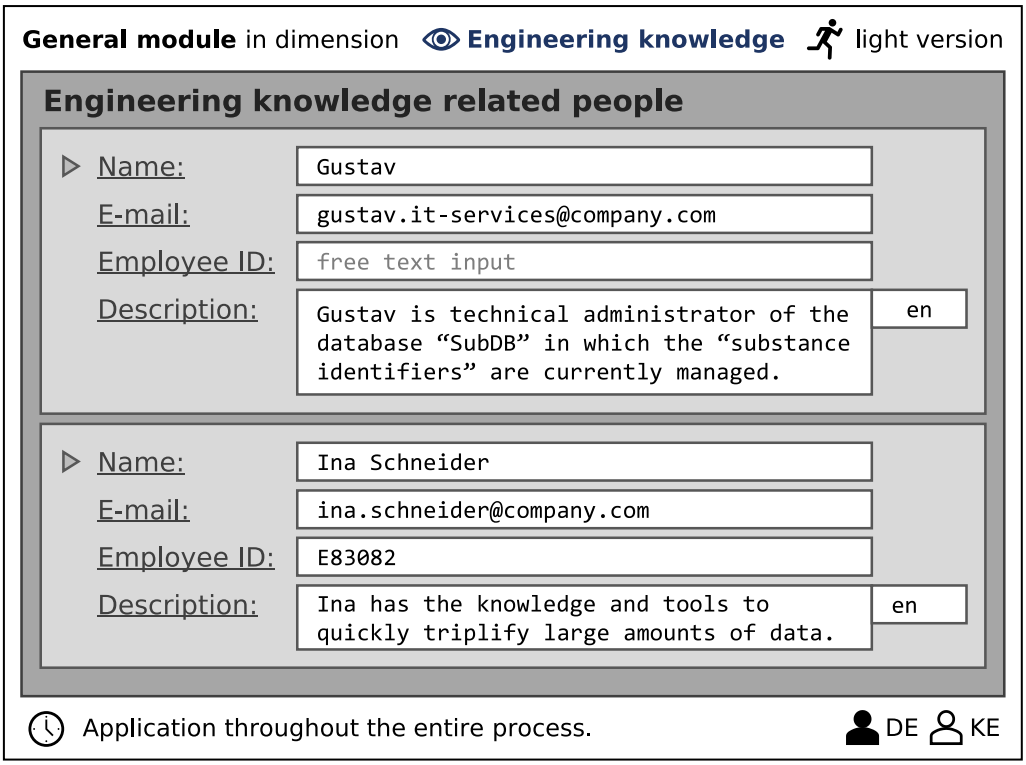

Figure 46: Exemplary application of the general module 'Related people' (light version) in the dimension 'Engineering knowledge'

5.5.2 Knowledge resources

As described in chapter '3 Selective insights into corporate reality', the general module *'Knowledge resources'* contains the properties *Access control*, *Confidentiality* and *License*. The *Location* refers either to a physical location such as a library or to a digital location that is specified by a URL. Figures 47 and 48 present the module in its light and full version. Figure 49 illustrates the full version's application; figure 50 the application of the light version. The module *'Knowledge resources'* is extended by dimension-specific properties in the modules *'Domain knoweldge resources'* and *'Engineering knowledge resources'* .

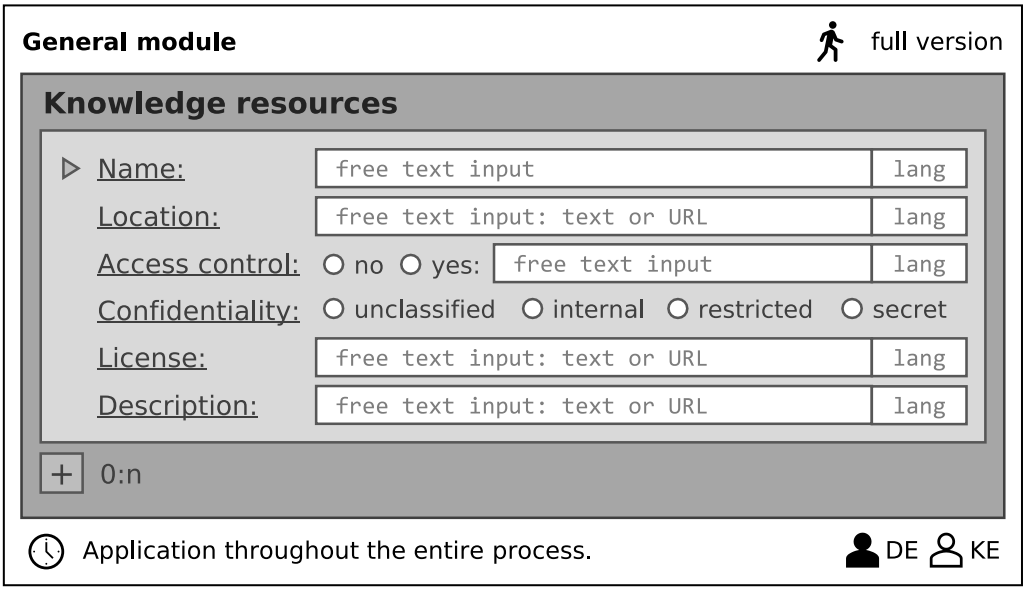

Figure 47: General module 'Knowledge resources' (full version)

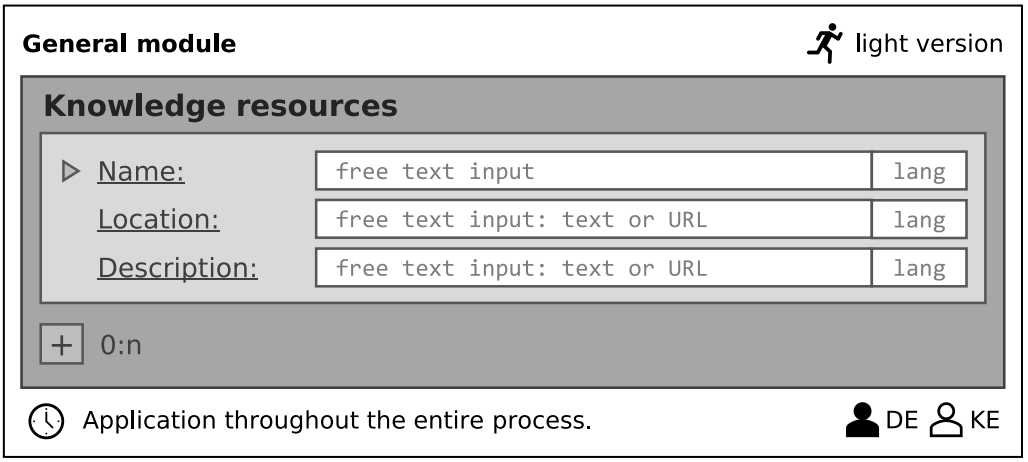

Figure 48: General module 'Knowledge resources' (light version)

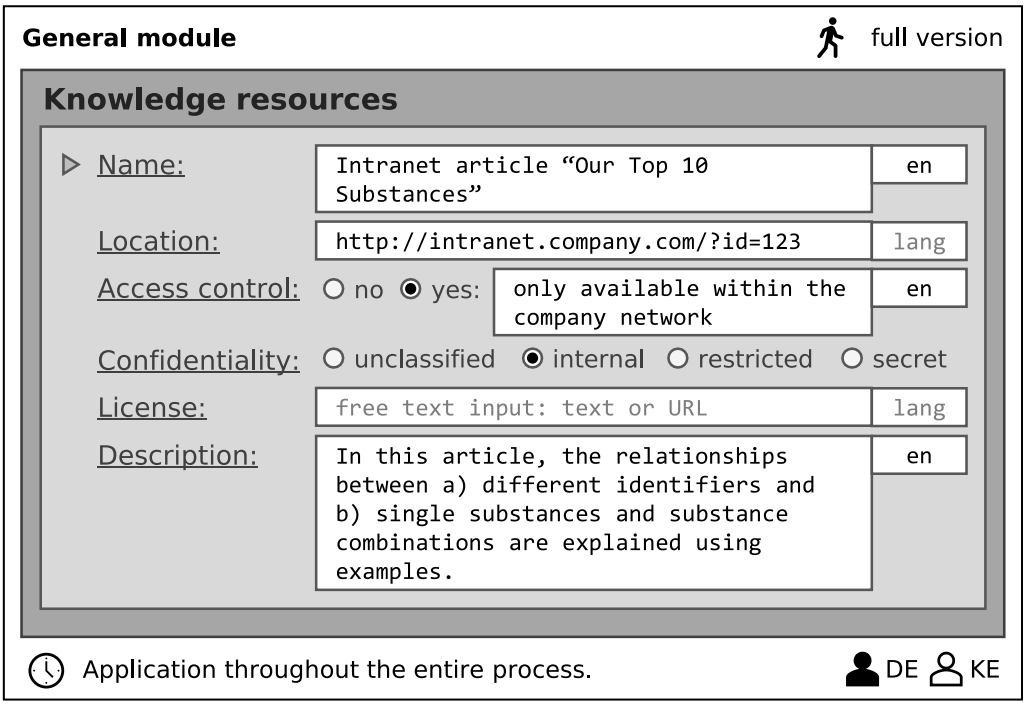

Figure 49: Exemplary application of the general module 'Knowledge resources' (full version)

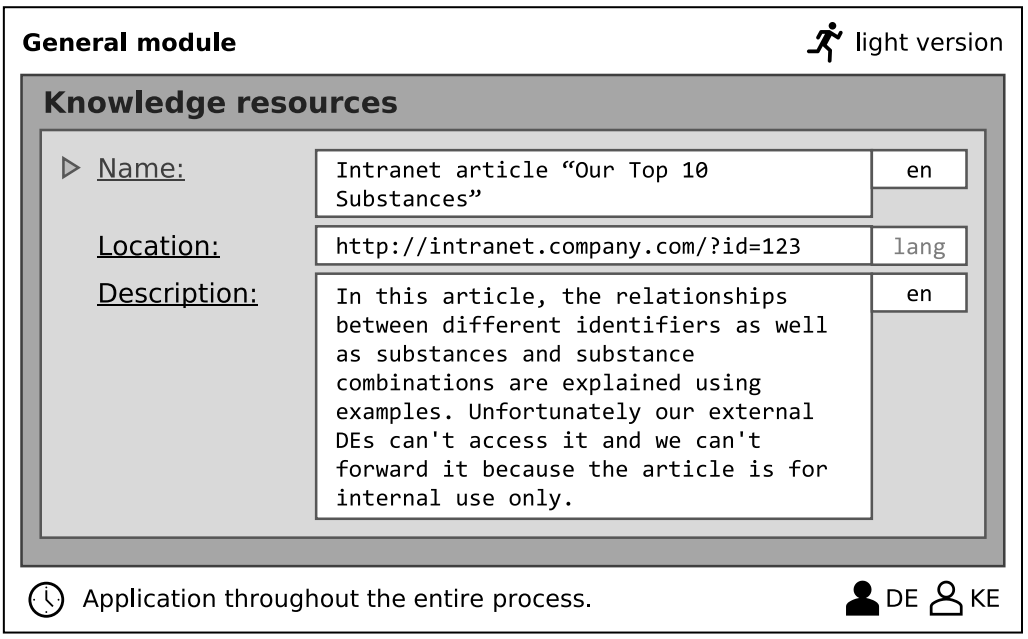

Figure 50: Exemplary application of the general module 'Knowledge resources' (light version)

5.5.3 Decisions

The module *'Decisions'* allows to document decisions in a way that they are comprehensible to others. From the perspective of knowledge exchange, the fields *Options considered* and *Decision reason* are particularly relevant (see full version in figure 51). As shown by the examples in figures 53, 54 and 55, important decisions may occur in all three dimensions.

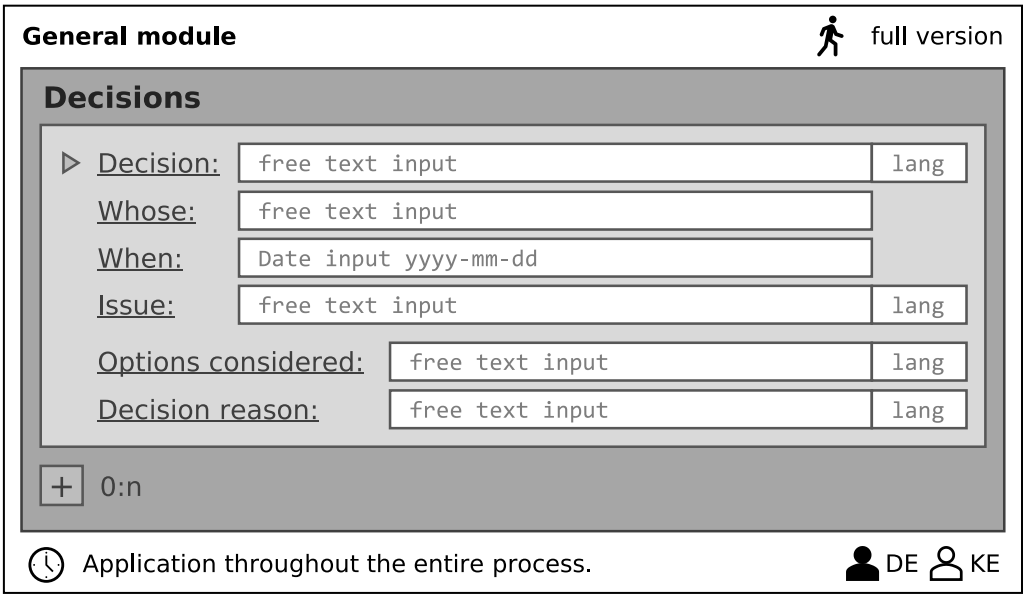

Figure 51: General module 'Decisions' (full version)

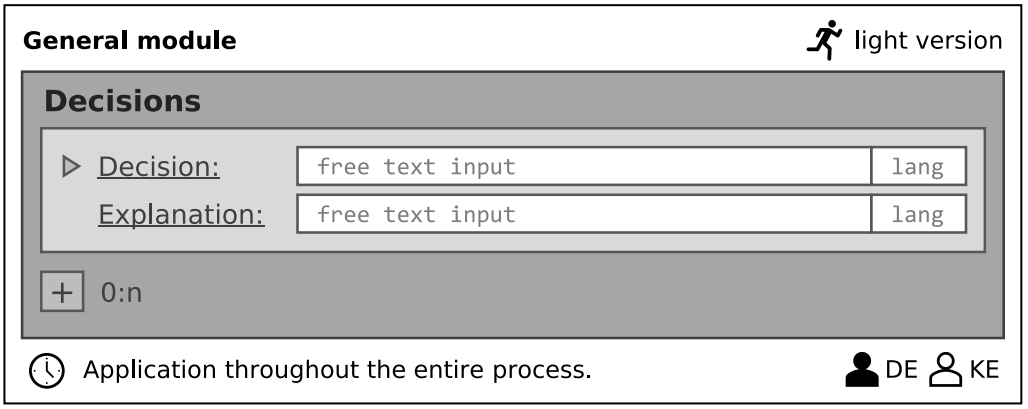

Figure 52: General module 'Decisions' (light version)

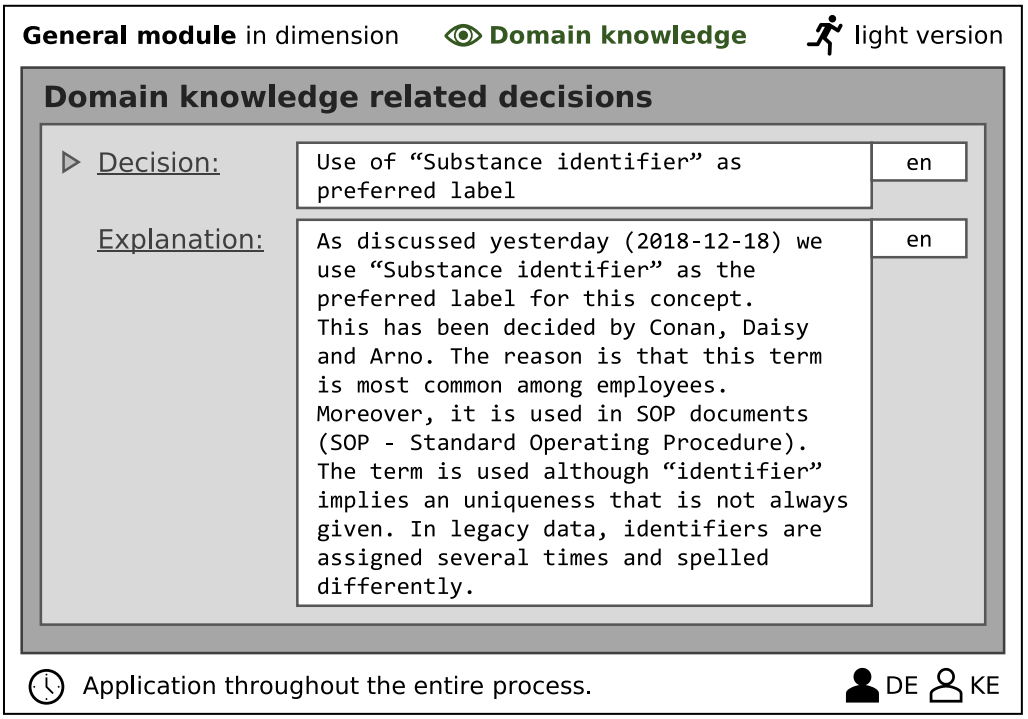

Figure 53: Exemplary application of the general module 'Decisions' (light version) in the dimension 'Domain knowledge'

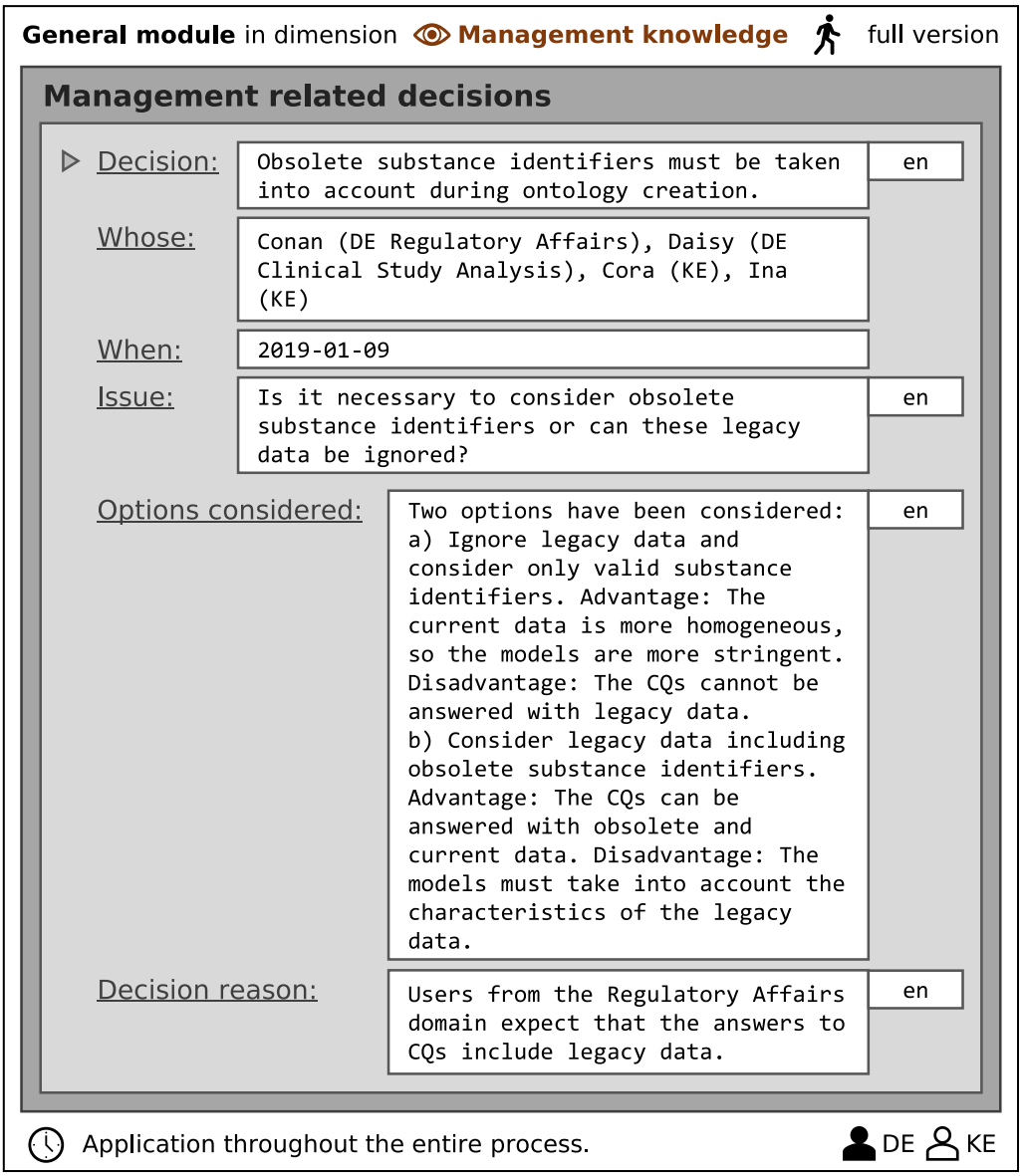

Figure 54: Exemplary application of the general module 'Decisions' (full version) in the dimension 'Management knowledge'

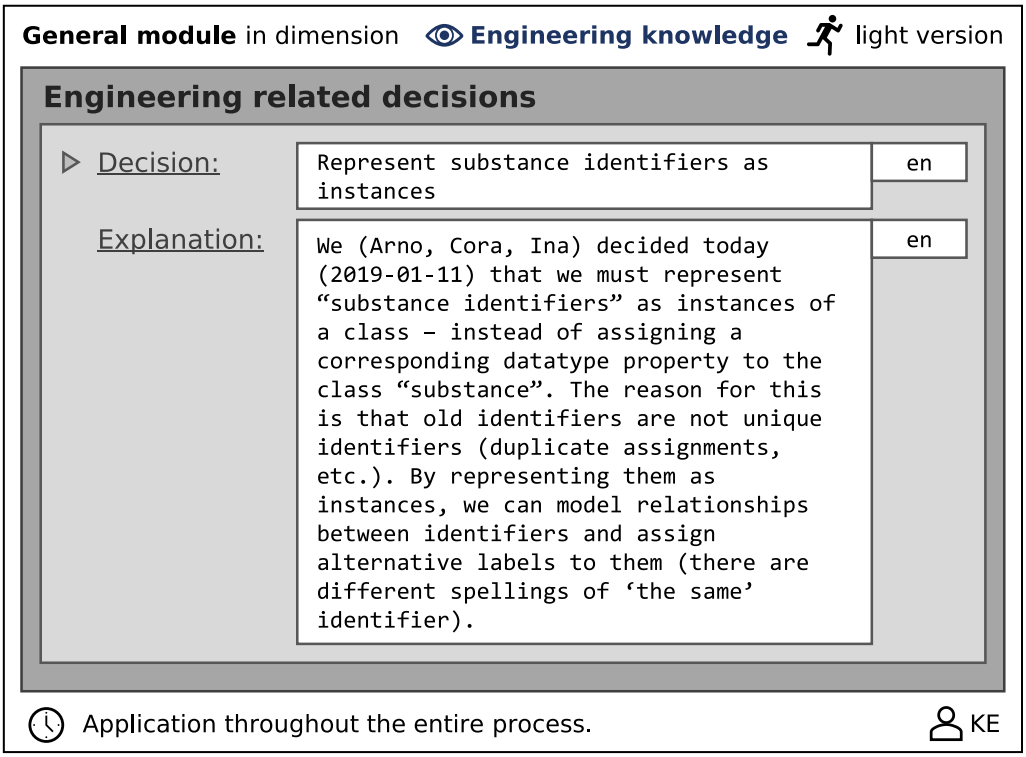

Figure 55: Exemplary application of the general module 'Decisions' (light version) in the dimension 'Engineering knowledge'

5.5.4 Tasks

The module *'Tasks'* may be employed if no other solution, such as a corresponding software application, is used for recording tasks. It is available in a full and light version (see figures 56 and 57) and recommended for the dimensions *'Domain knowledge'* and *'Engineering knowledge'* (see examplary application in figures 58 and 59).

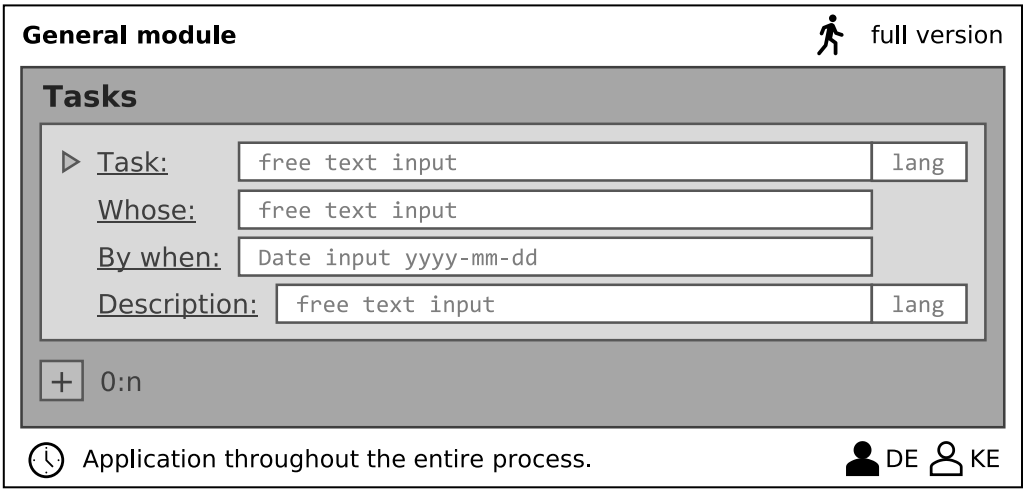

Figure 56: General module 'Tasks' (full version)

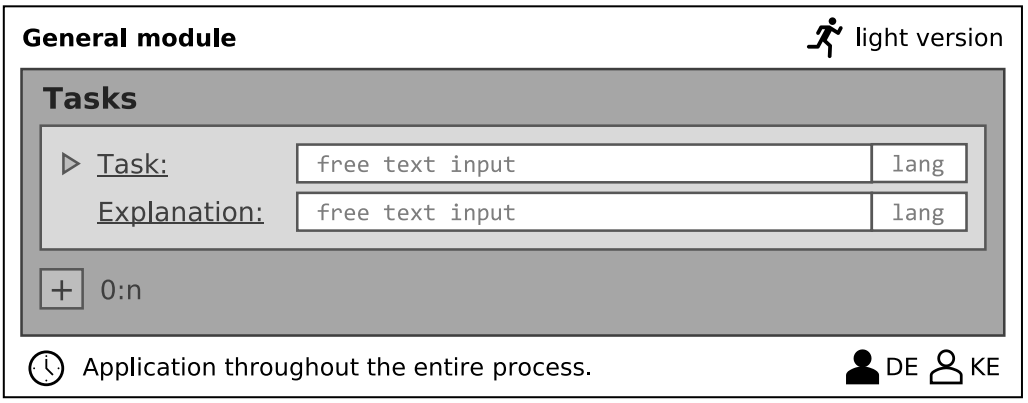

Figure 57: General module 'Tasks' (light version)

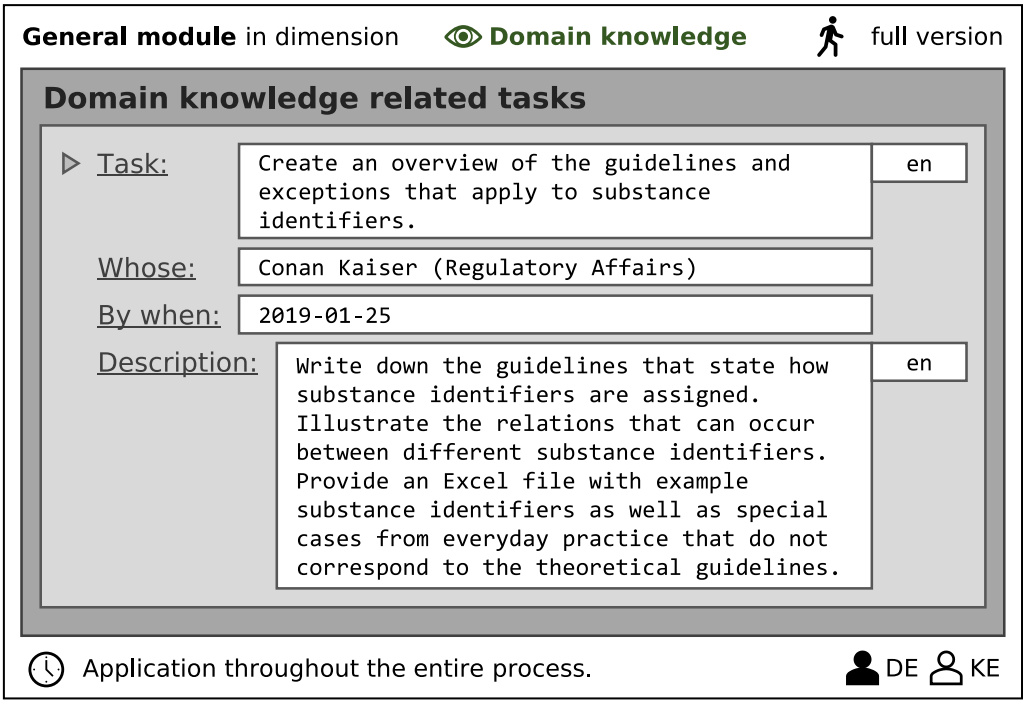

Figure 58: Exemplary application of the general module 'Tasks' (full version) in the dimension 'Domain knowledge'

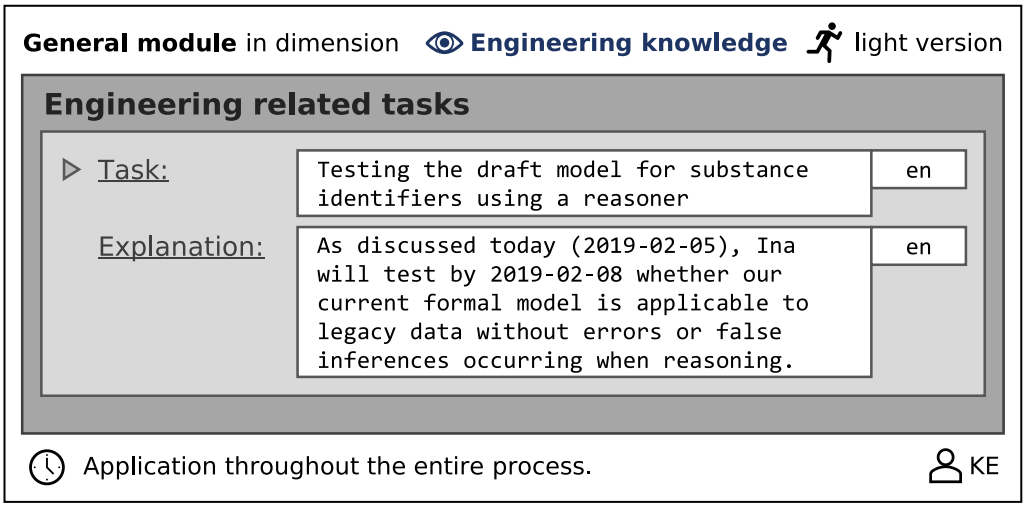

Figure 59: Exemplary application of the general module 'Tasks' (light version) in the dimension 'Engineering knowledge'

5.5.5 Notes

The module *'Notes'* enables all content to be captured that cannot be documented in a structured way using the other modules. Due to its poor structuring, it should be used cautiously. Otherwise, the complete documentation may be entered into this module. As a result, the actual objective of the documentation concept, which is structured documentation, would be missed.

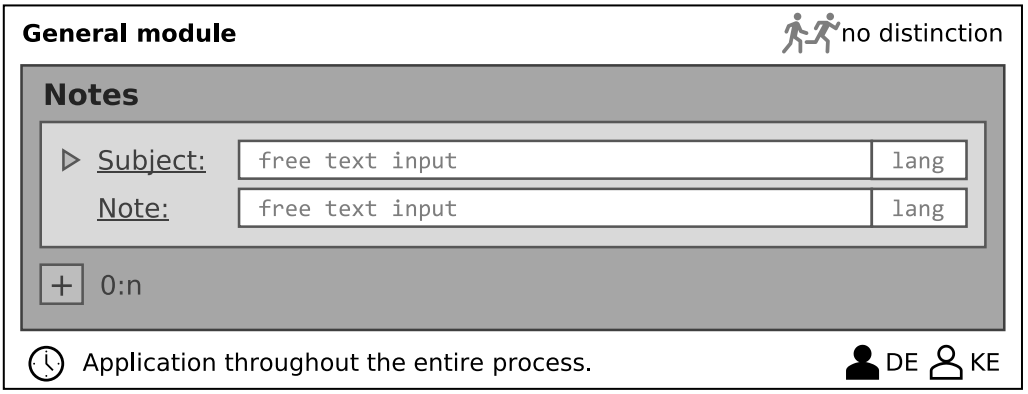

Figure 60: General module 'Notes'

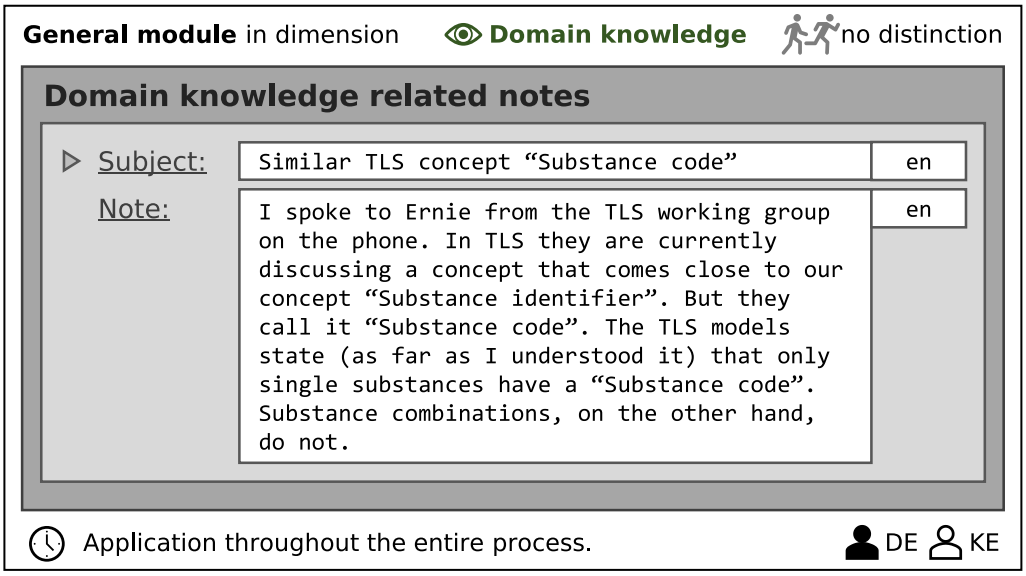

Figure 61: Exemplary application of the general module 'Notes' in the dimension 'Domain knowledge'

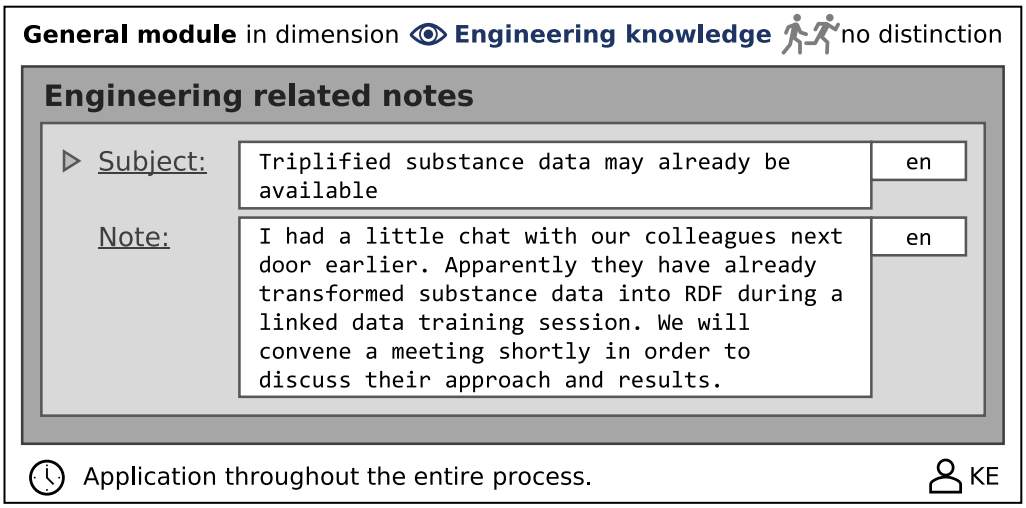

Figure 62: Exemplary application of the general module 'Notes' in the dimension 'Engineering knowledge'

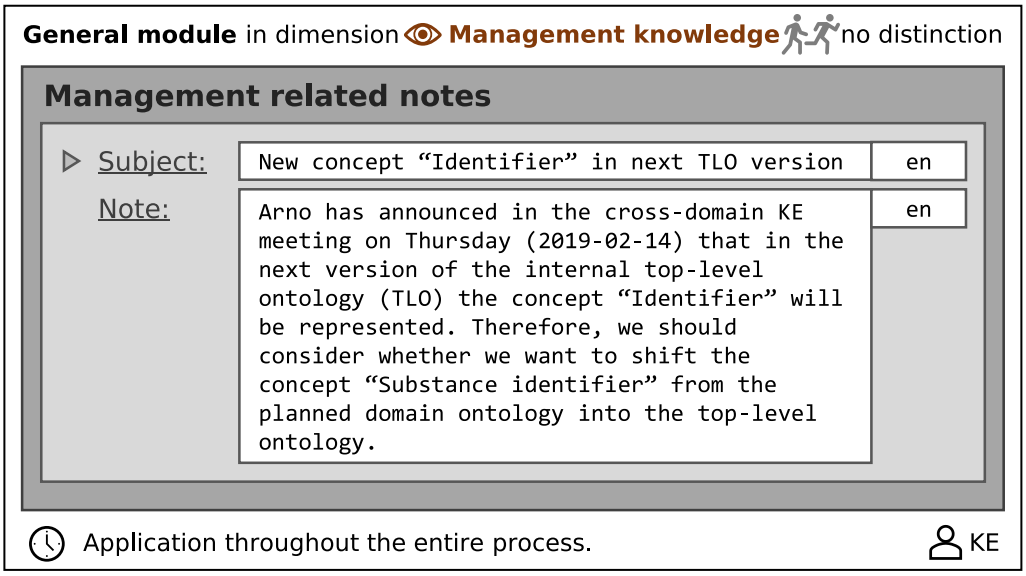

Figure 63: Exemplary application of the general module 'Notes' in the dimension 'Management knowledge'

5.5.6 Discussion

At this point, the module *'Discussion'* is presented by means of a simplified user interface that is characteristic for electronic group chats. It represents the integration of corresponding solutions into the documentation concept. Depending on the implementation, various functions may be integrated such as: Embedding of images, code, links, files; Use of text formatting; Targeted addressing of persons or groups using @ mentions; Ratings of posts; Grouping of posts to a subchannel. Existing solutions such as *Let's Chat*, *Mattermost*, *Microsoft Teams*, *Riot*, *Slack*, *Stackfield* or *Telegram* can may be employed or serve as inspiration.

In the template shown in figure 64 as well as the exemplary applications illustrated in figures 65, 66 and 67, direct addressing of persons is highlighted as an important function. Furthermore, the application example in figure 67 demonstrates a major advantage of open chats as already described in chapter '3 Selective insights into corporate reality': Unknown parties are enabled to participate and thus create synergies.

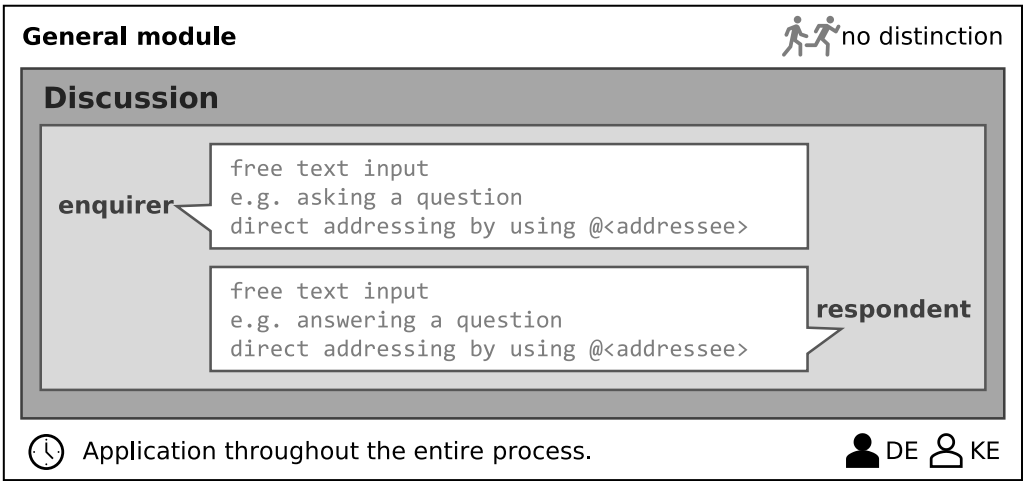

Figure 64: General module 'Discussion'

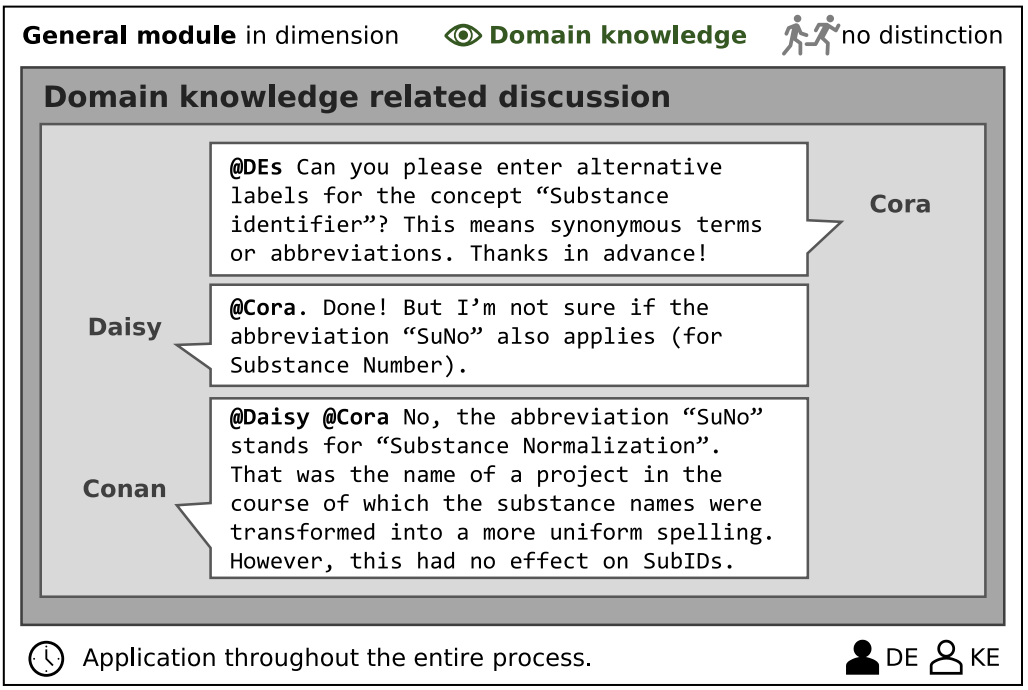

Figure 65: Exemplary application of the general module 'Discussion' in the dimension 'Domain knowledge'

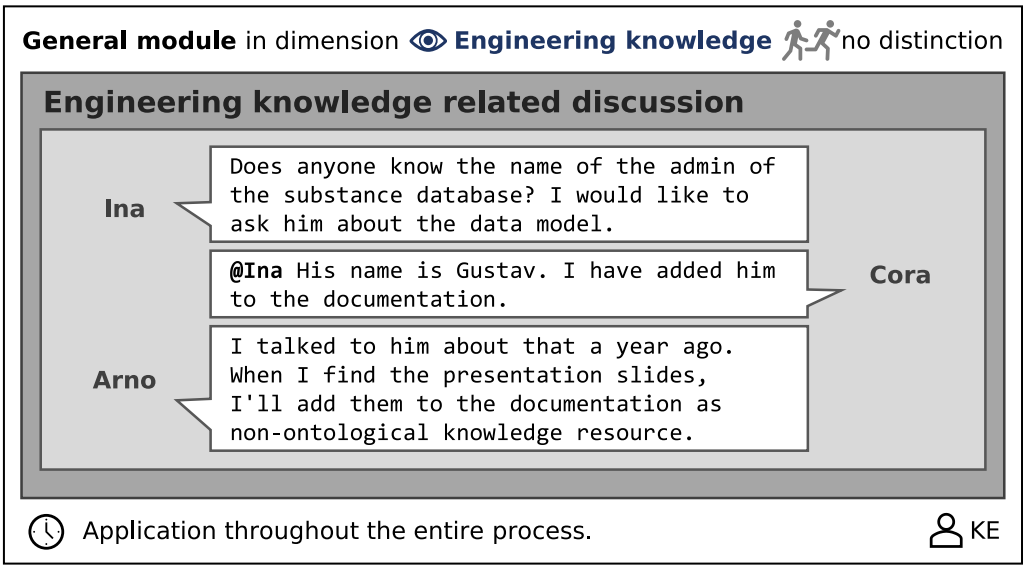

Figure 66: Exemplary application of the general module 'Discussion' in the dimension 'Engineering knowledge'

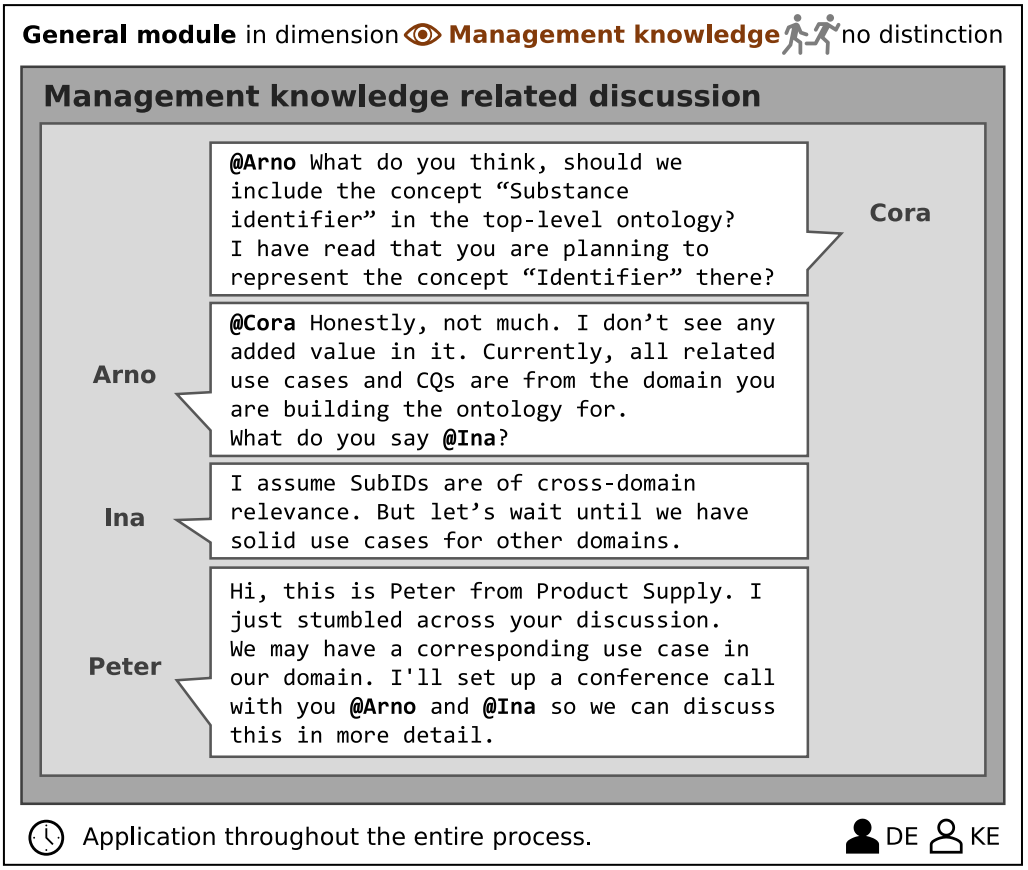

Figure 67: Exemplary application of the general module 'Discussion' in the dimension 'Management knowledge'

5.5.7 Agreement

For the recording of *agreement*, a three-step rating scale similar to a traffic light is proposed. Table 18 below shows the scale along with the meaning of the tree levels. An application of the module extension *'Agreement'* can be found in the module *'Definition'* in the dimension *'Domain knowledge'*. Theoretically, agreement can be measured for virtually any statement. For efficiency reasons, however, it is recommended to use this module extension only on definitions. A distinction between full and light version is not made.

Scale	Label	Intended meaning
🔴	Rejected	I disagree. The statement in question is incorrect.
🟡	Tolerated	I agree with reservations. The statement in question is correct, but insufficient in terms of clarity and/or completeness.
🟢	Approved	I agree. The statement in question is correct, clear and complete.

Table 18: Rating scale for the recording of agreement

In the course of implementation, two questions must be answered:

- What happens to existing ratings if the rated statement is modified?
- Who is enabled to view which ratings and when?

A possible answer to the first question would be to reset the agreement measurement if the statement is changed. This ensures that the ratings always refer to the current version.

With regard to the second question, it is recommended that the KE responsible is allowed to view all ratings. In the case of an electronic implementation, one of the following example options could be selected for the group of DEs:

a) DEs only see their own rating.

b) DEs see the *distribution* of all ratings *after* submitting their own rating.

c) DEs see the *distribution* of all ratings already *before* submitting their own rating.

d) DEs see all *individual ratings* including the respective voting person *after* submitting their own rating.

e) DEs see all *individual ratings* including the respective voting person already *before* submitting their own rating.

Presumably, different social dynamics occur depending on the implementation. For instance, insight into the rating of others may cause more discussion. This would be positive because ideally people are encouraged to share knowledge and reach consensus. On the other hand, insight into other people's ratings may influence individual ratings. How this is handled is in the hands of the applying organization.

5.5.8 Source

The module extension *'Source'* serves to record the source reference. In this way, the origin of a statement can be traced. Theoretically, a source can be captured for virtually any statement. For efficiency reasons, however, it is recommended to use this module extension only on definitions. The corresponding application of the module extension *'Source'* can be found in the module *'Definition'* in the dimension *'Domain knowledge'*. The following figures 68 and 69 show the associated template in its full and its light version.

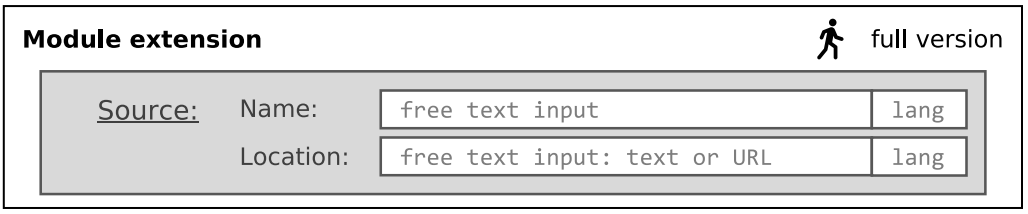

Figure 68: Modul extension 'Source' (full version)

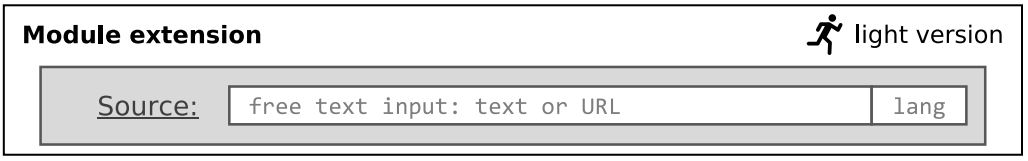

Figure 69: Modul extension 'Source' (light version)

5.5.9 Image

The module extension *'Image'* enables the structured, text-based documentation to be supplemented by visual elements such as schematic computer graphics or photos of whiteboard drawings. Like the previous module extensions, *'Image'* can be used to extend several modules. If it is employed too often, however, the structure of the documentation is compromised as it loses its clarity. In the case of an electronic implementation, for instance, the search function may be less effective. Therefore, the module extension *'Image'* is only recommended to extend the module *'Relationships to other concepts'*. There, it offers added value because complex connections can be conveyed visually. The following figures 70 and 71 show the *'Image'* template in its full and light version. An exemplary application is shown in the light version of the module *'Relationships to other concepts'* in the dimension *'Domain knowledge'*.

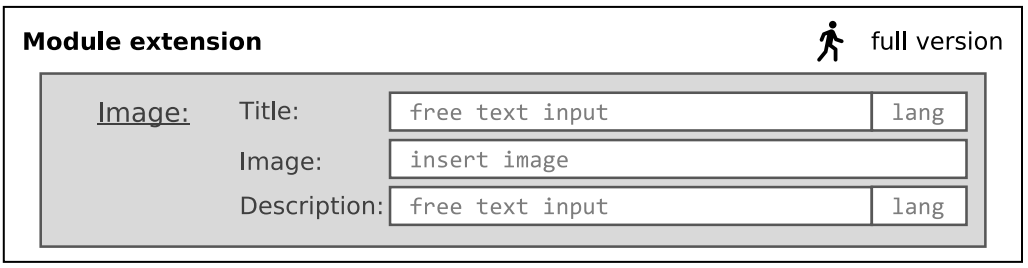

Figure 70: Modul extension 'Source' (full version)

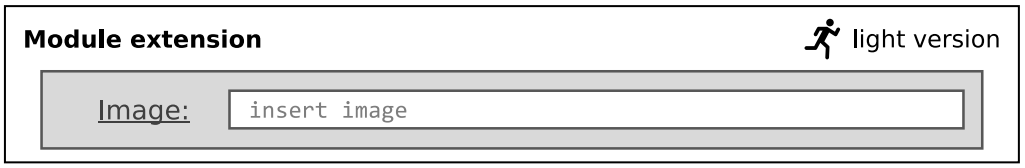

Figure 71: Modul extension 'Source' (full version)

5.6 Dimension-specific modules: Domain knowledge

5.6.1 Alternative labels

The module *'Alternative labels'* in its full version is used to record *synonyms* and *abbreviations* (see figure 72). It is recommended that these are entered by DEs to share the workload. New synonyms or abbreviations may be 'discovered' throughout the entire process. Similar to the preferred label of the concept, the corresponding preferred abbreviation is limited to a maximum of one entry. In the light version of the module, no distinction is made between synonyms and abbreviations (see figure 73). This is also demonstrated by the exemplary applications in figures 74 and 75. Due to lack of space, these are the only figures in this thesis that illustrate the use of multiple languages.

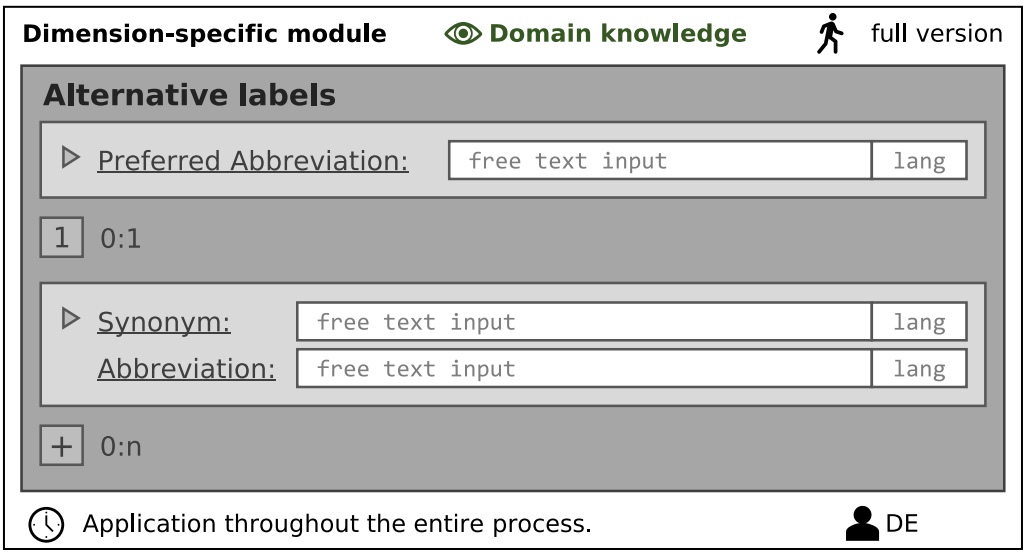

Figure 72: Dimension-specific module 'Alternative labels' (full version)

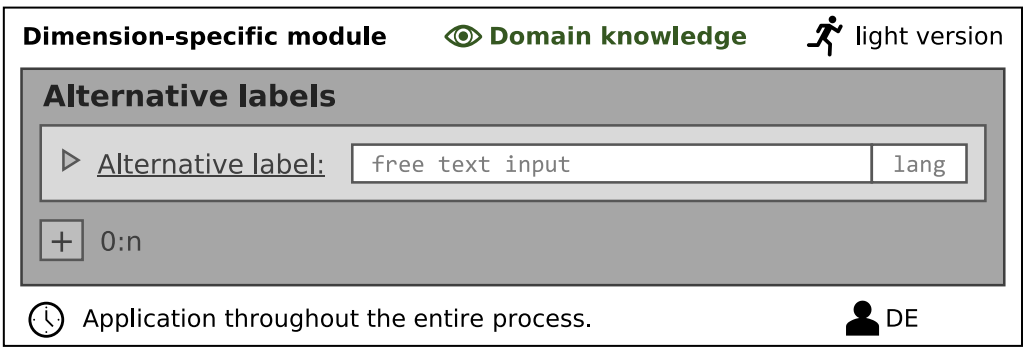

Figure 73: Dimension-specific module 'Alternative labels' (light version)

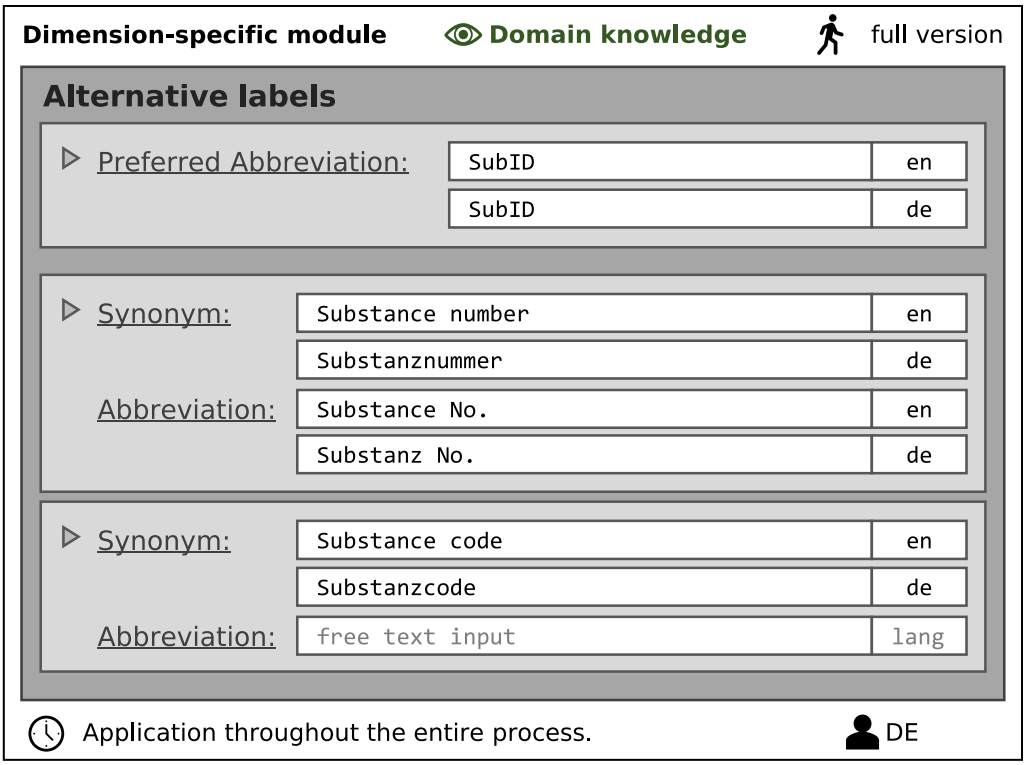

Figure 74: Exemplary application of the dimension-specific module 'Alternative labels' (full version)

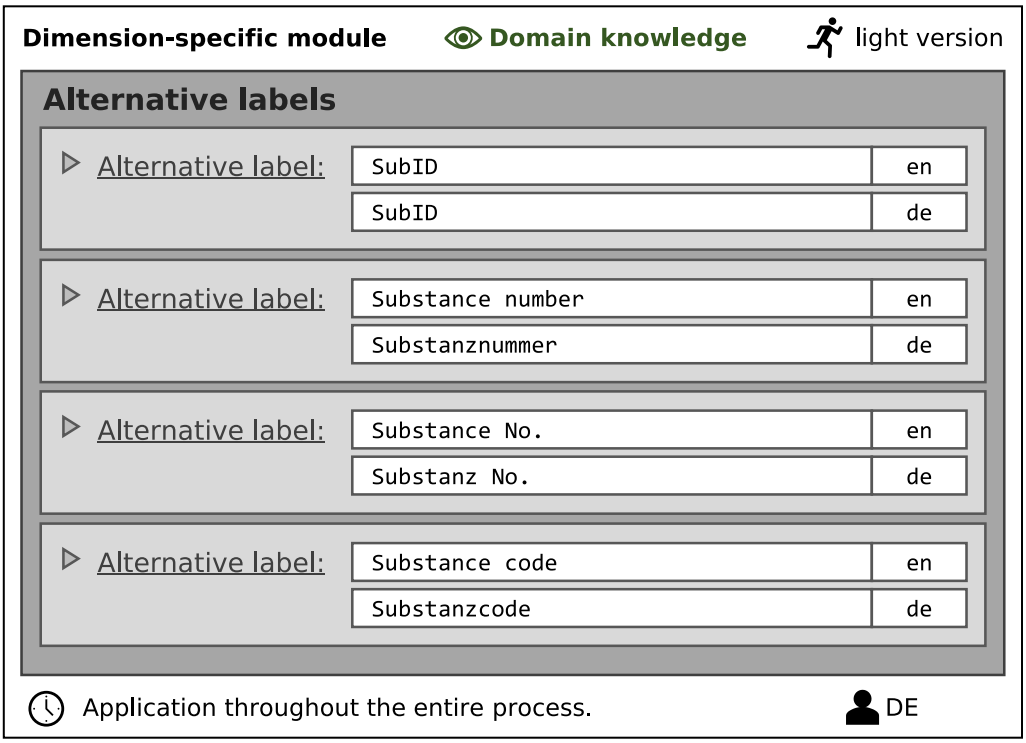

Figure 75: Exemplary application of the dimension-specific module 'Alternative labels' (light version)

5.6.2 Foreign identifiers

The module *'Foreign identifiers'* allows DEs to enter identifiers from other vocabularies corresponding to the concept in question. As with all other recommendations, KEs are not excluded. It is recommended, however, to exploit the knowledge of those DEs who are familiar with controlled vocabularies of the domain. Below is the template in its full and light version (figure 76 and 77) followed by examples illustrated in figures 78 and 79.

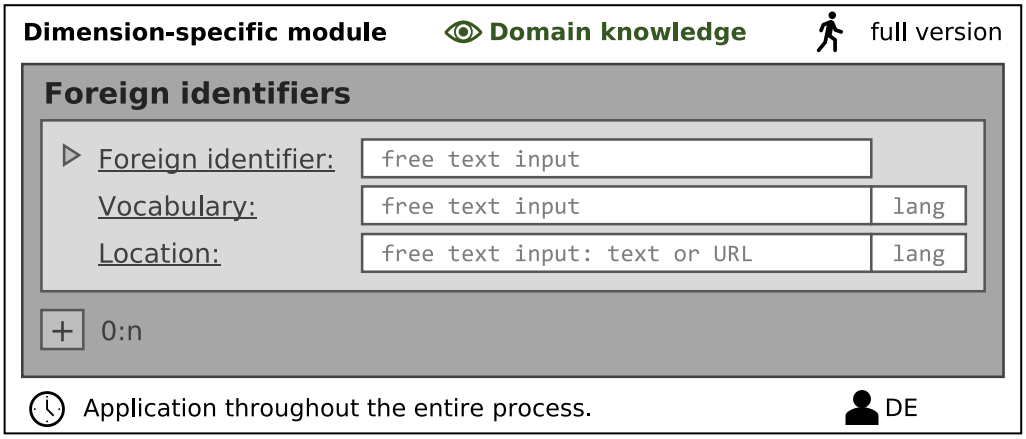

Figure 76: Dimension-specific module 'Foreign identifiers' (full version)

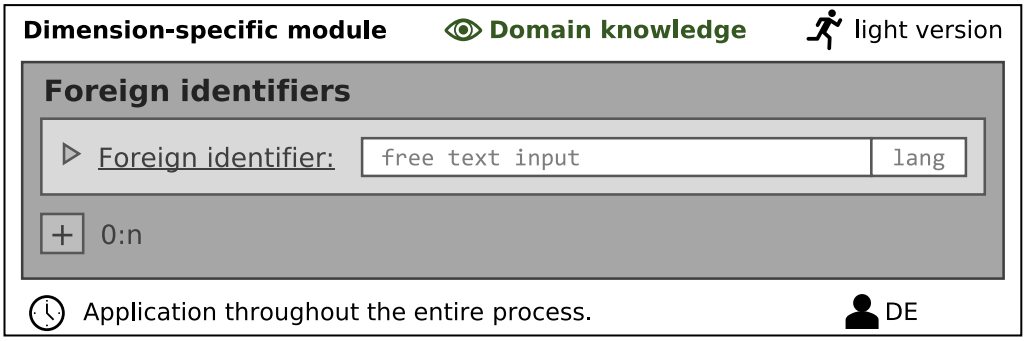

Figure 77: Dimension-specific module 'Foreign identifiers' (light version)

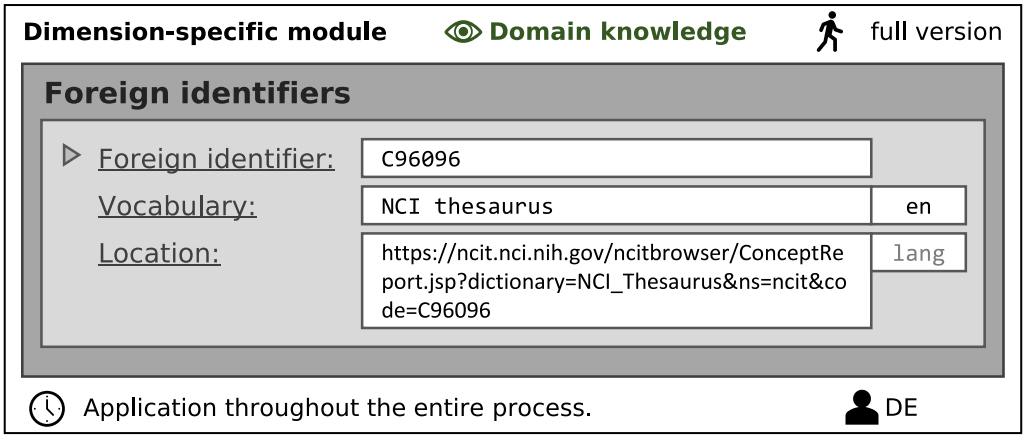

Figure 78: Exemplary application of the dimension-specific module 'Foreign identifiers' (full version)

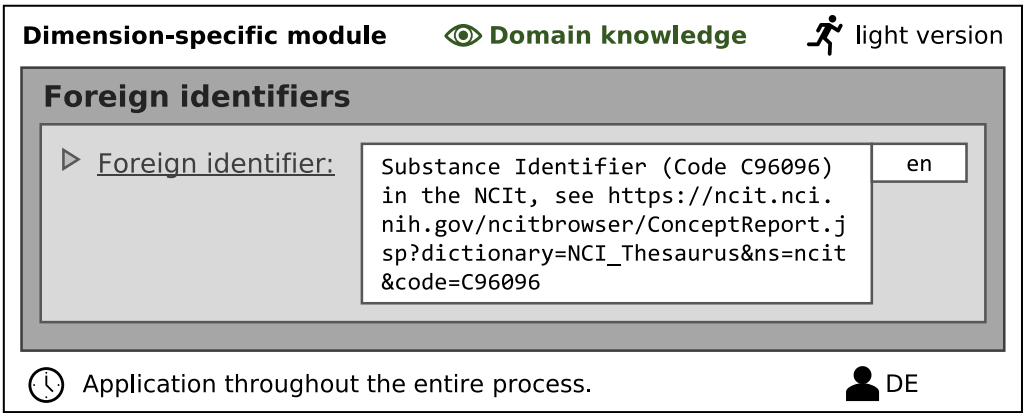

Figure 79: Exemplary application of the dimension-specific module 'Foreign identifiers' (light version)

5.6.3 Domain

The module *'Domain'* is used to assign the concept to a domain. This is particularly appropriate when the same term occurs several times but differs in its meanings (*homonyms*). In this case, the module *'Domain'* corresponds in its function to what is called a *qualifier* in documentation science.

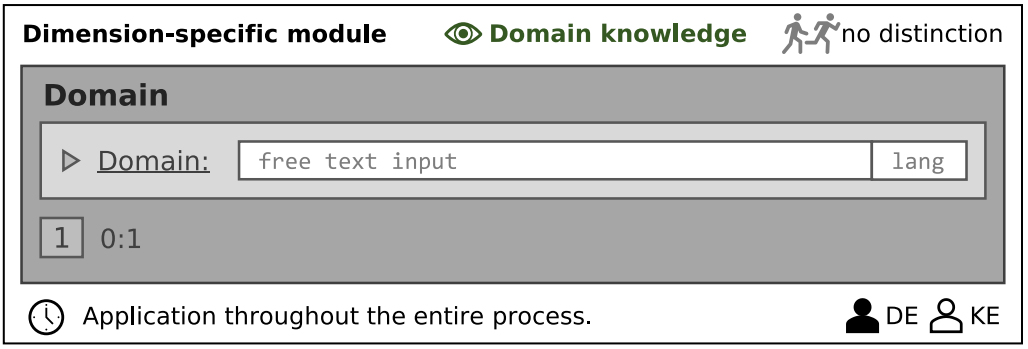

Figure 80: Dimension-specific module 'Domain'

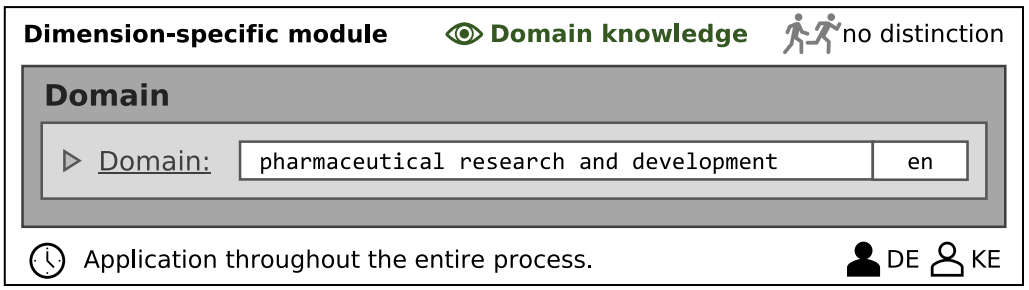

Figure 81: Exemplary application of the dimension-specific module 'Domain'

5.6.4 Definitions

As concepts should be described concisely using natural language, the module *'Definitions'* is one of the central modules. Following the *shared* aspect of ontologies as well as the analysis in chapter '4.1 Unified Process for Ontology Building UPON', definitions have to be agreed upon. Consequently, the module extension *'Agreement'* is used on the full version of the module *'Definitions'* as shown in figures 82 and 84. In the light version, shown in figures 83 and 85, the extension is not applied. The difference between the two versions demonstrates the added value of capturing the level of agreement. The module extension *'Source'*, on the other hand, is used in both versions. In this way, it is possible to trace the origin of a definition.

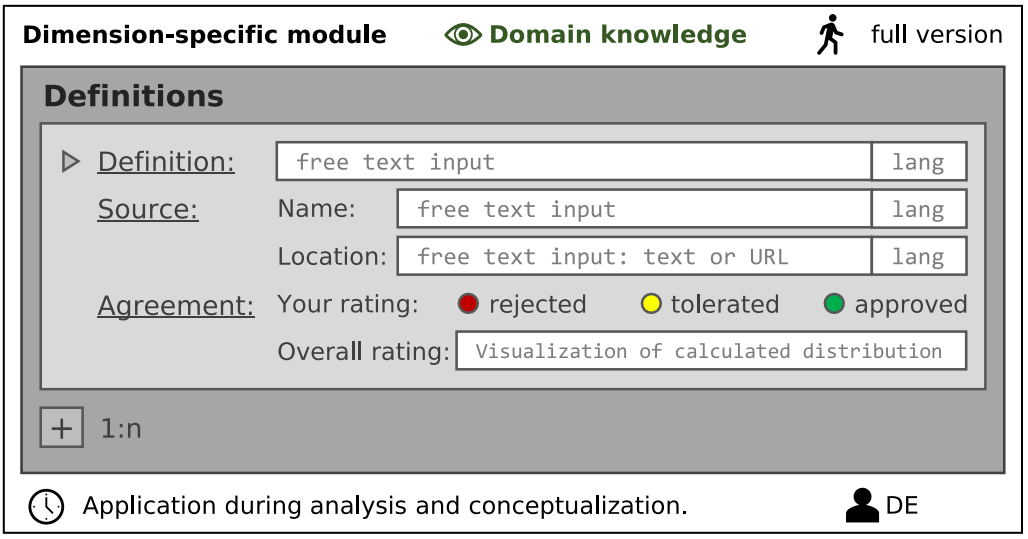

Figure 82: Dimension-specific module 'Definitions' (full version)

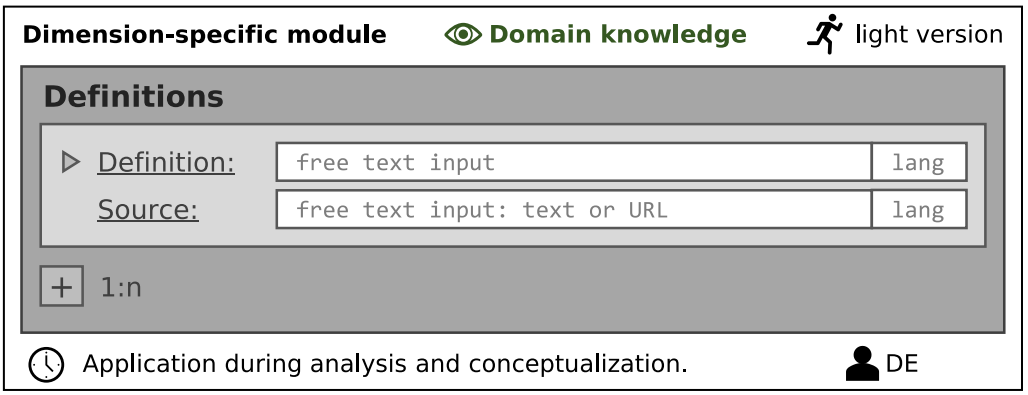

Figure 83: Dimension-specific module 'Definitions' (light version)

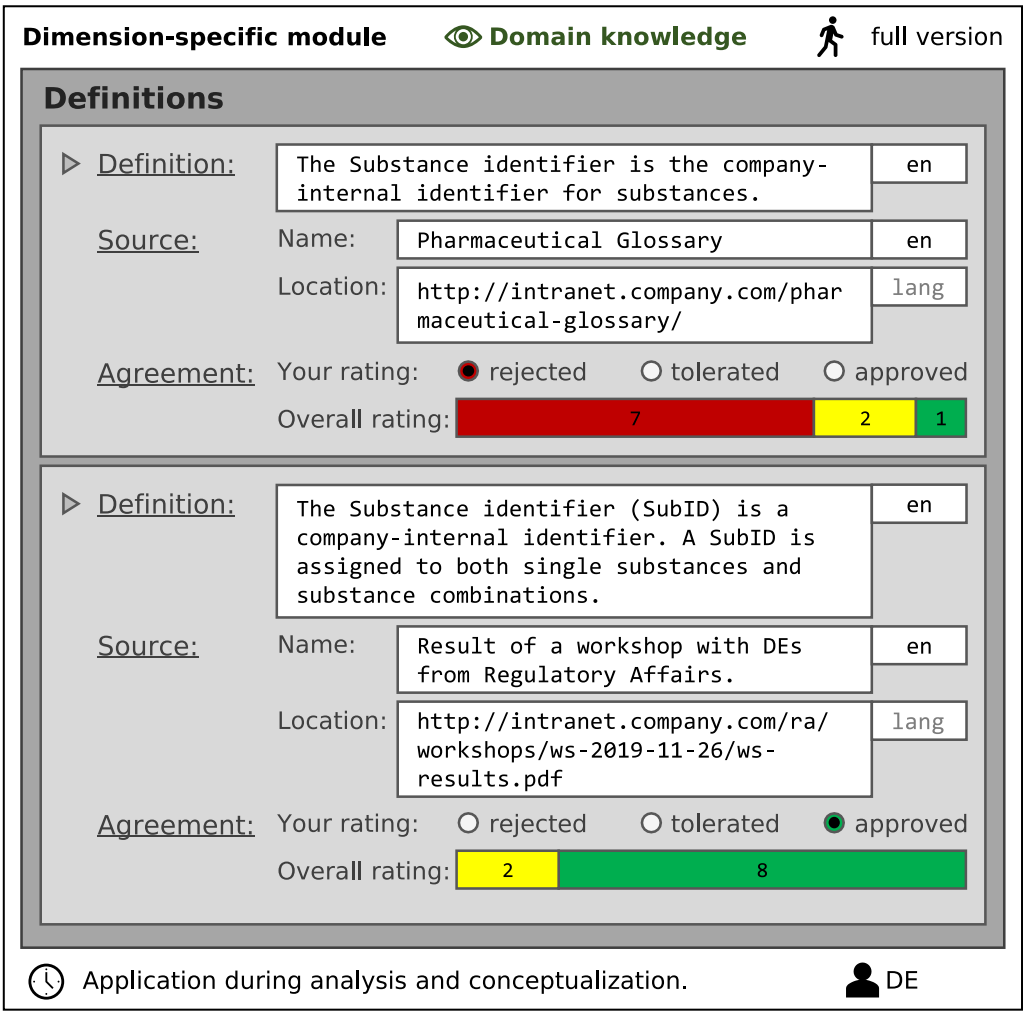

Figure 84: Exemplary application of the dimension-specific module 'Definitions' (full version)

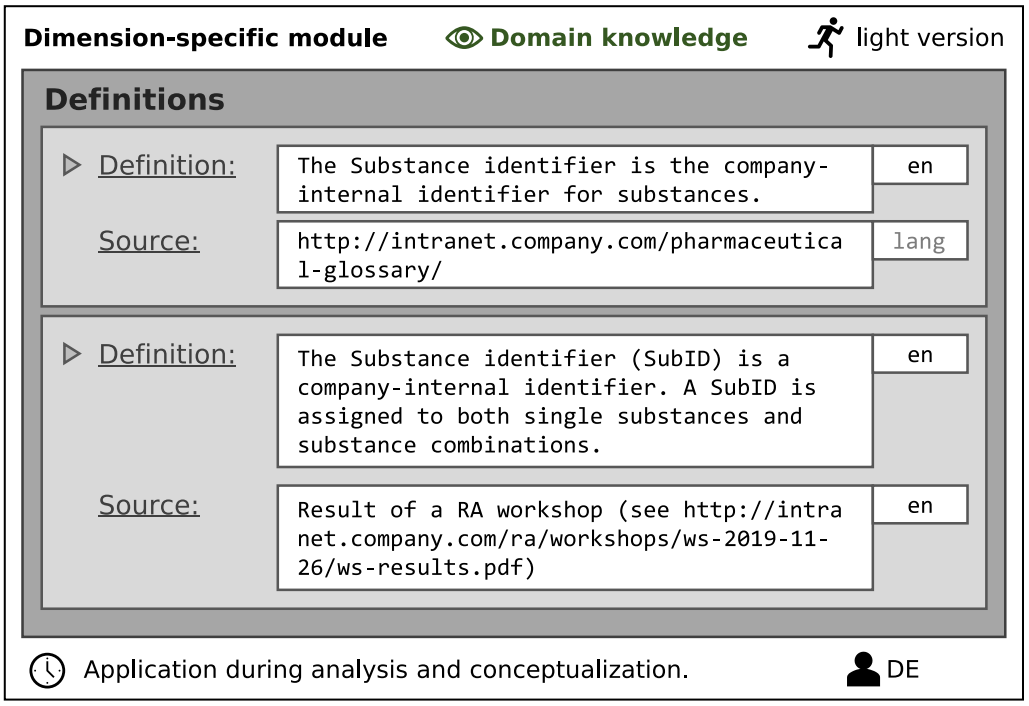

Figure 85: Exemplary application of the dimension-specific module 'Definitions' (light version)

5.6.5 Explanations

The module *'Explanations'* allows to document further information about the concept and its scope. In addition, the full version allows to capture concrete examples (see figure 86). Explanations help KEs and interested DEs to understand a concept better than just using short definitions. Figures 88 and 89 show exemplary explanations. While they are recorded in a structured way in the full version of the module, the light version consists of only one field.

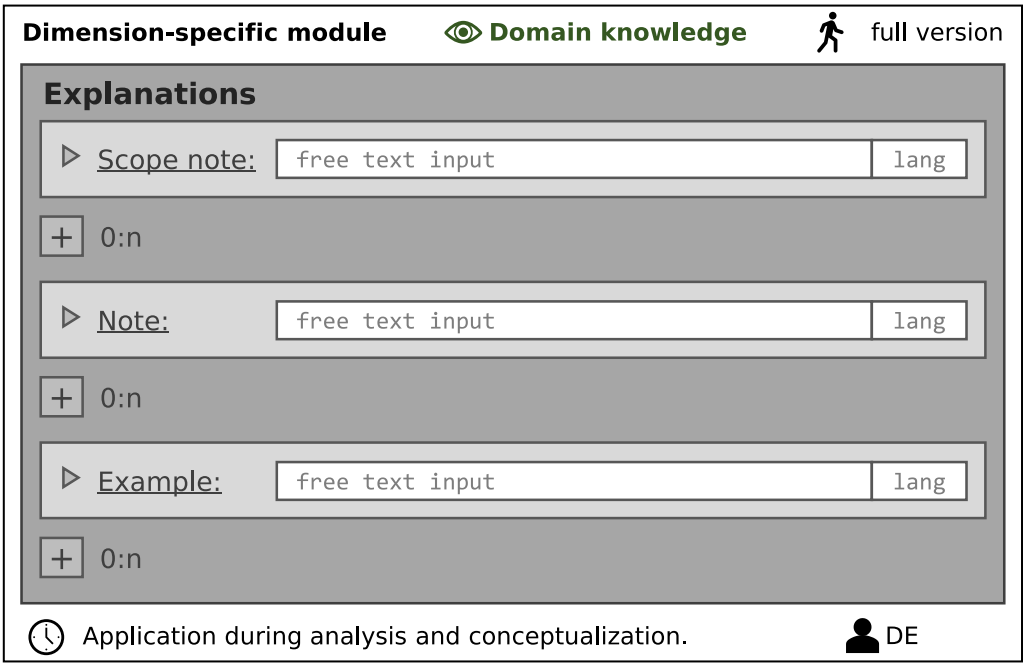

Figure 86: Dimension-specific module 'Explanations' (full version)

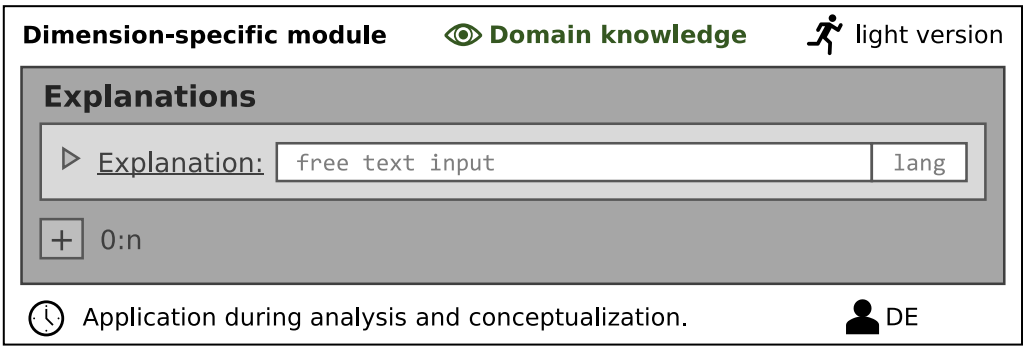

Figure 87: Dimension-specific module 'Explanations' (light version)

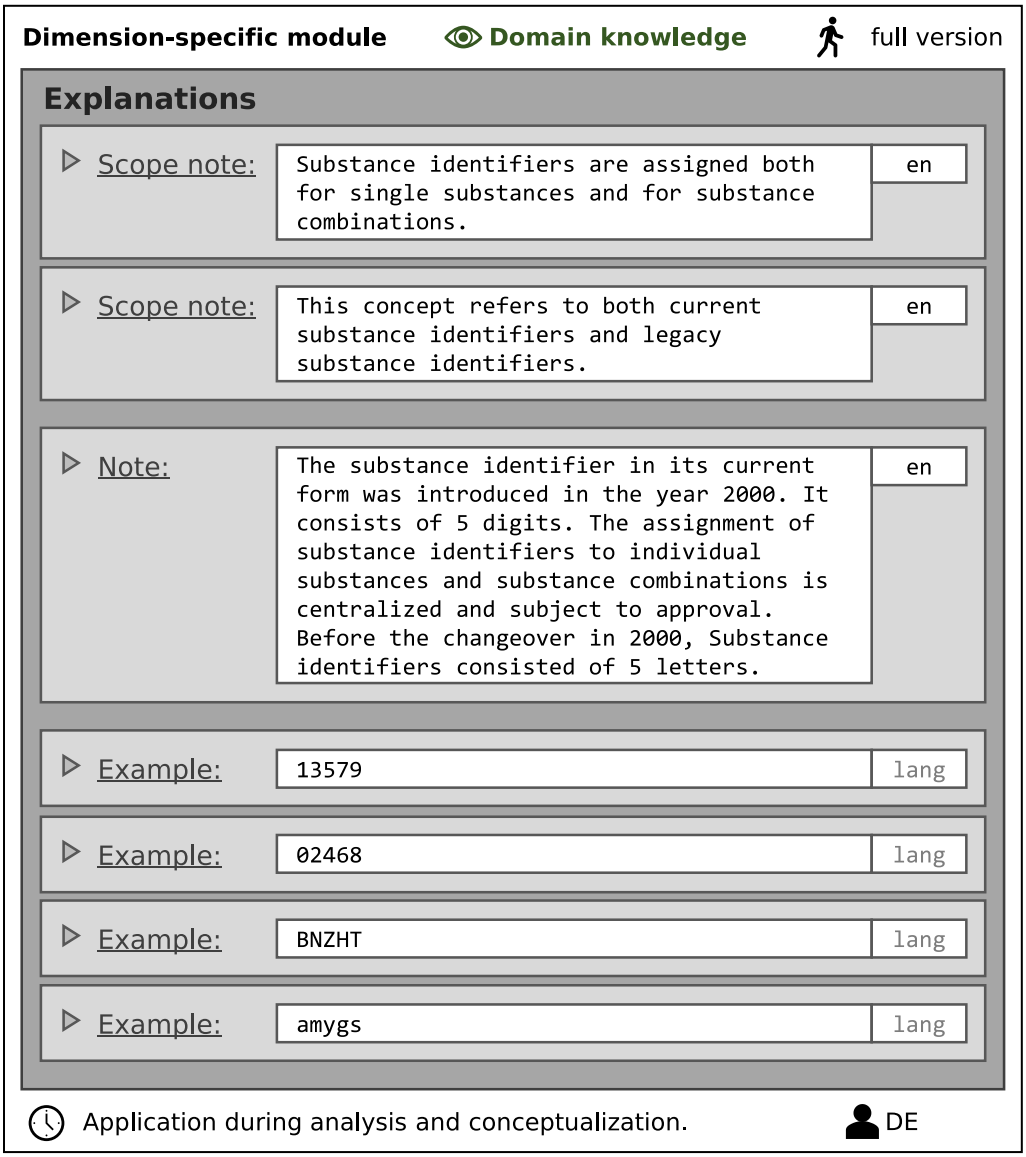

Figure 88: Exemplary application of the dimension-specific module 'Explanations' (full version)

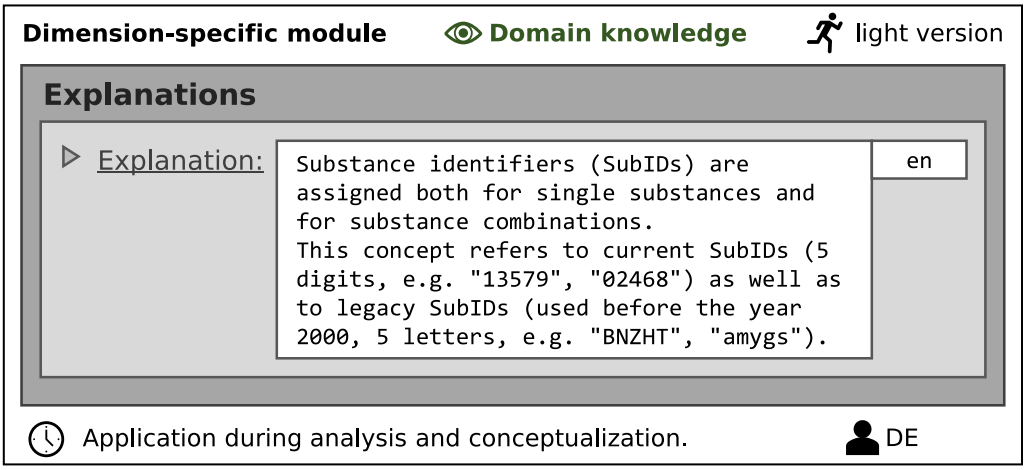

Figure 89: Exemplary application of the dimension-specific module 'Explanations' (light version)

5.6.6 Domain knowledge resources

The module *'Domain knowledge resources'* extends the general module *'Knowledge resources'* by the dimension-specific categorization of knowledge resources. Any kind of resource can be captured, be it a dataset, a printed book or an ontology. For their description, the categorization derived in chapter '4.2 NeOn Methodology' is used. In the light version of the module, the features of the general module are not applied (see figure 91). The application example in figures 92 and 93 shows that checkboxes offer the possibility of multiple selection.

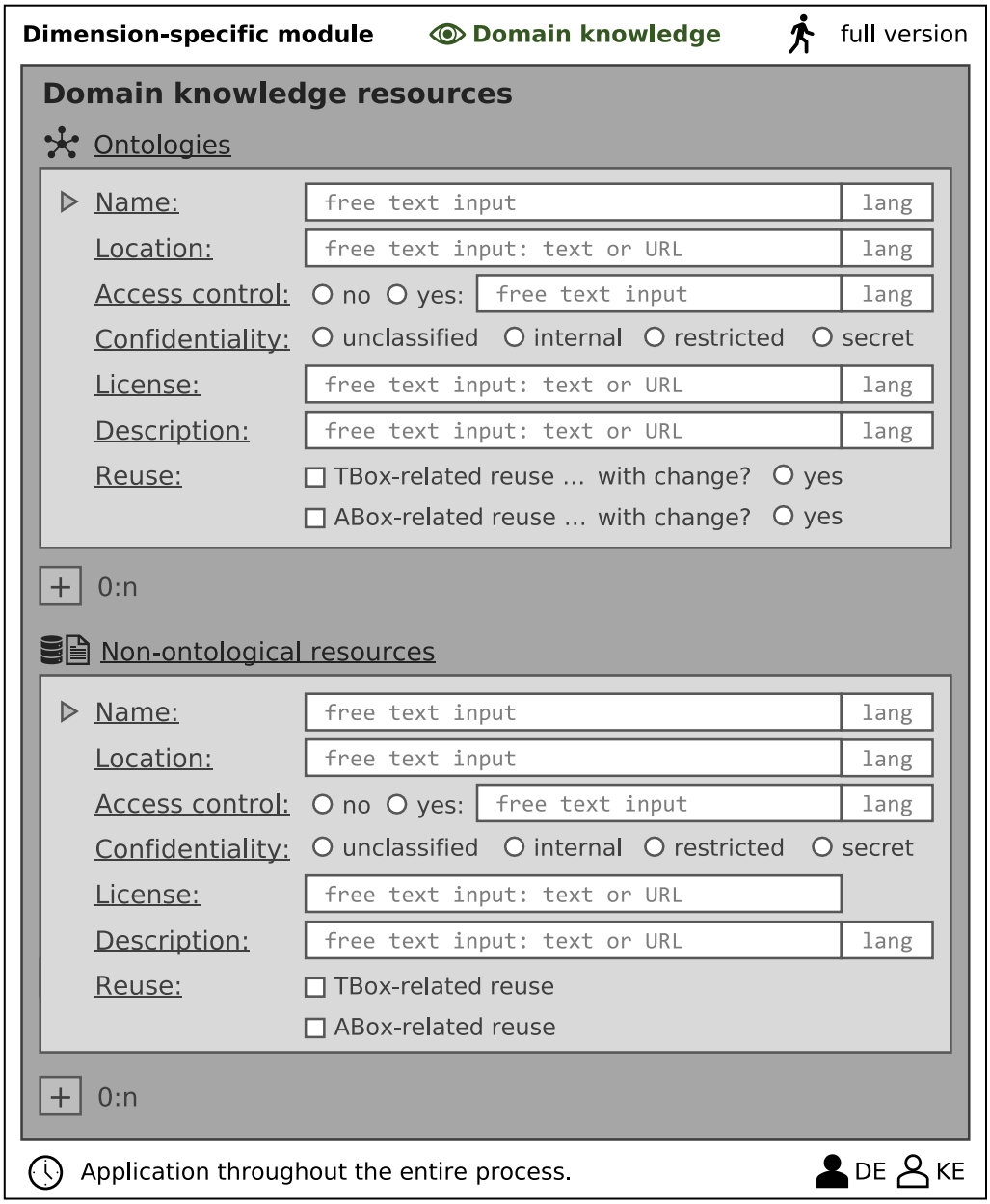

Figure 90: Dimension-specific module 'Domain knowledge resources' (full version)

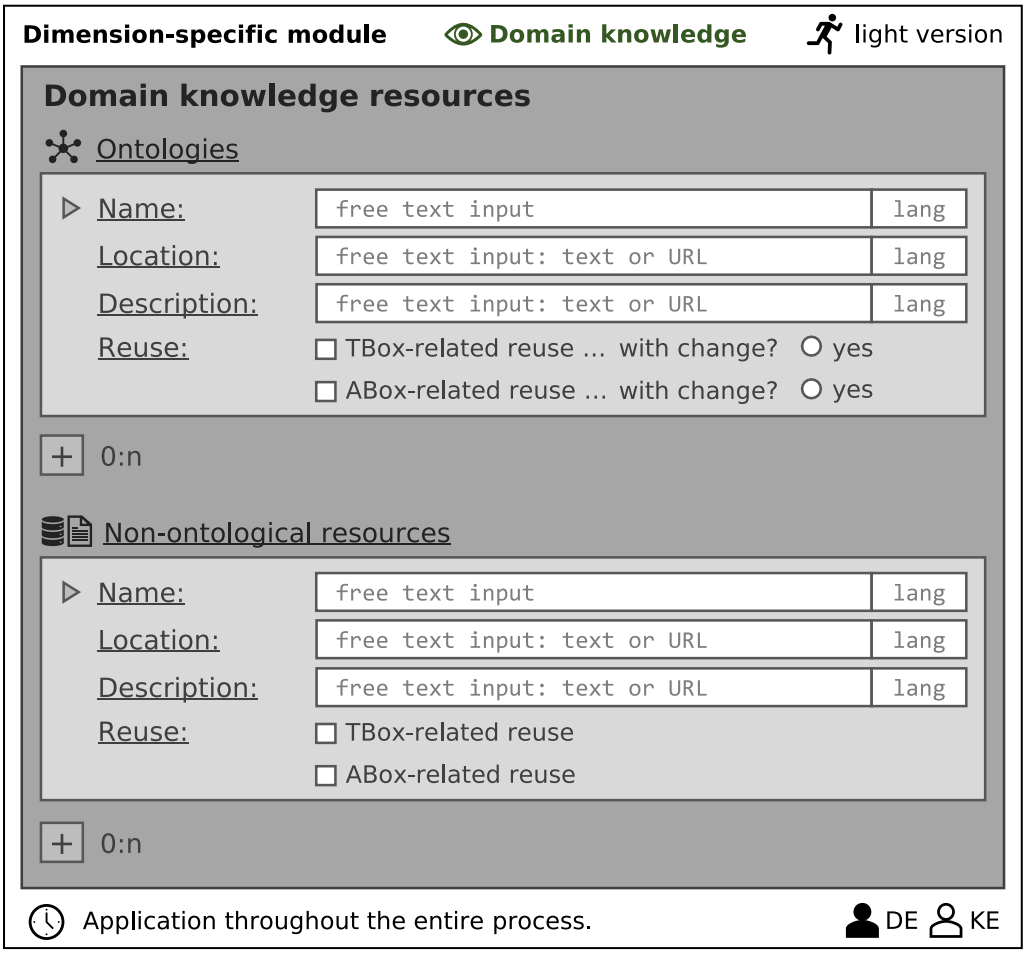

Figure 91: Dimension-specific module 'Domain knowledge resources' (light version)

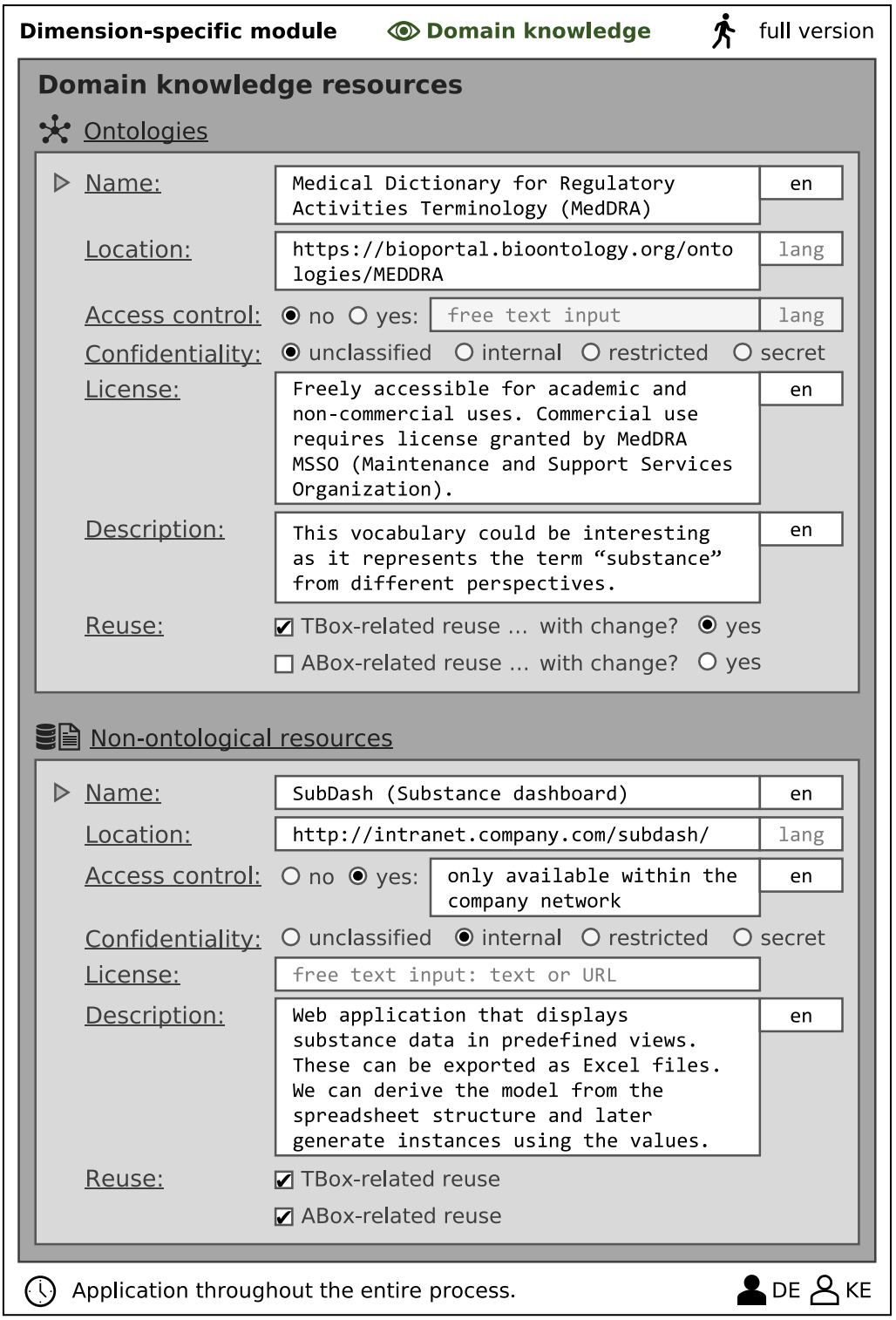

Figure 92: Exemplary application of the dimension-specific module 'Domain knowledge resources' (full version)

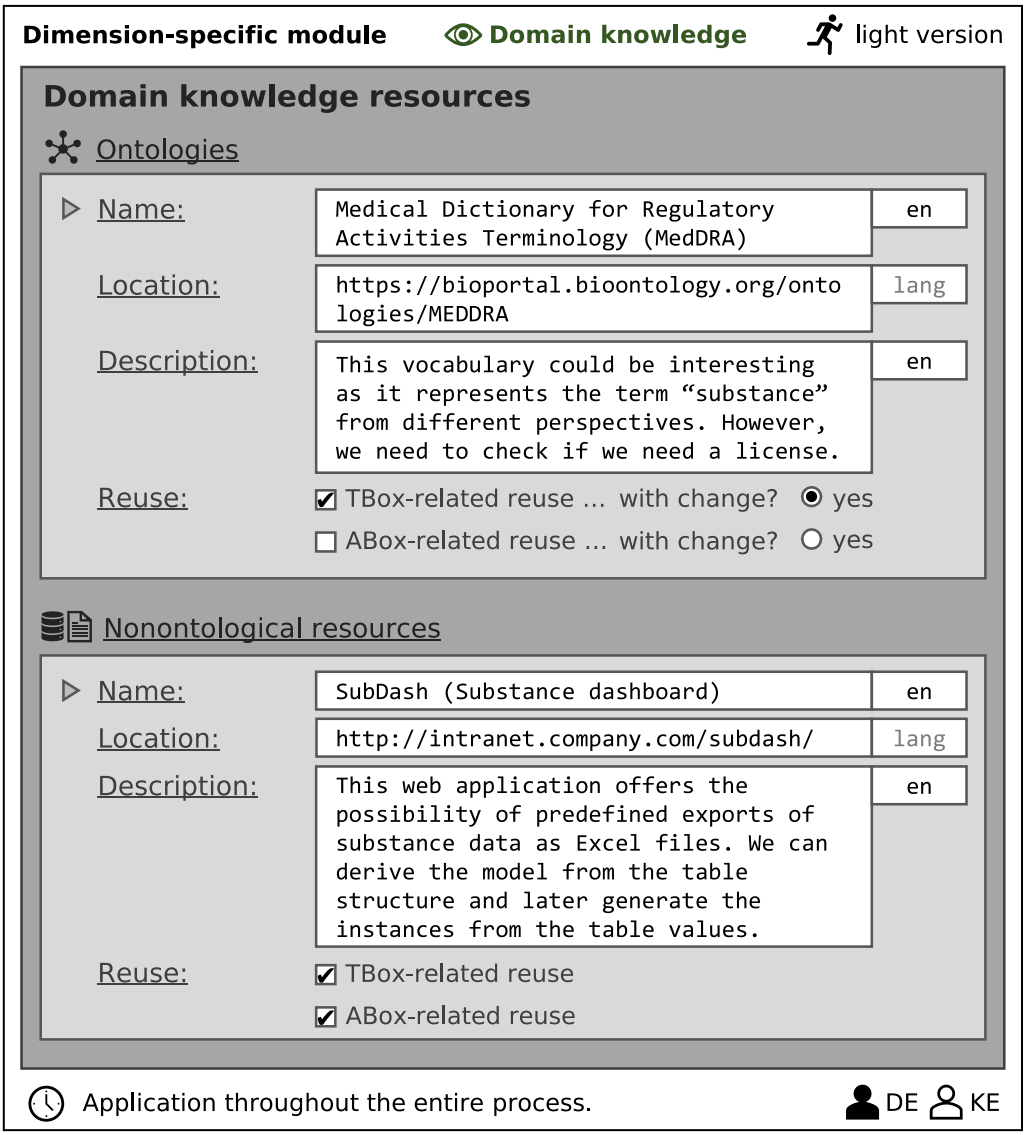

Figure 93: Exemplary application of the dimension-specific module 'Domain knowledge resources' (light version)

5.6.7 Concept category

The module *'Concept category'* allows to assign the concept to a general category. The predefined set of categories is adopted from the methodology described in chapter '4.1 Unified Process for Ontology Building UPON' which in turn is based on OPAL and UML. If desired, the predefined categories can be exchanged, for instance by the general concepts of the top-level ontology used. It is recommended that this categorization is performed by KEs, as those are experienced in abstracting from specific concepts to general levels. The documentation concept assumes and recommends that a concept is assigned to only one general category. Otherwise, the concept in question is probably not a single concept but two different concepts that need to be separated.

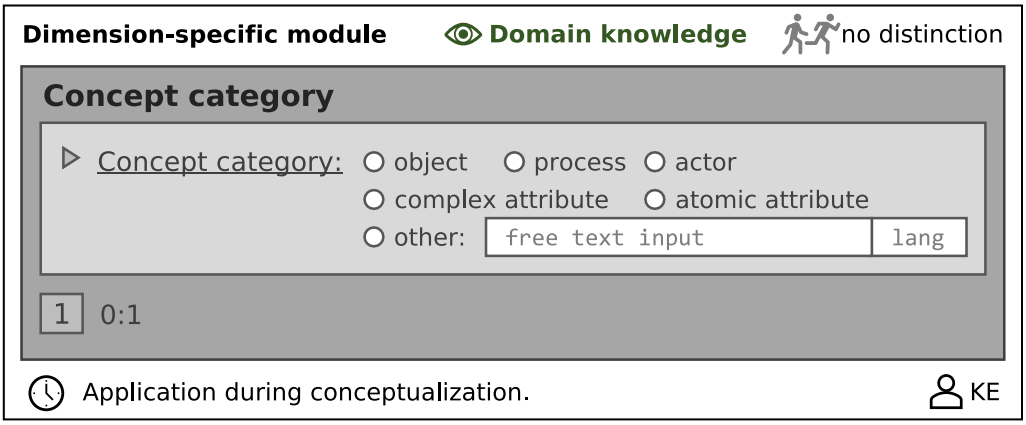

Figure 94: Dimension-specific module 'Concept category'

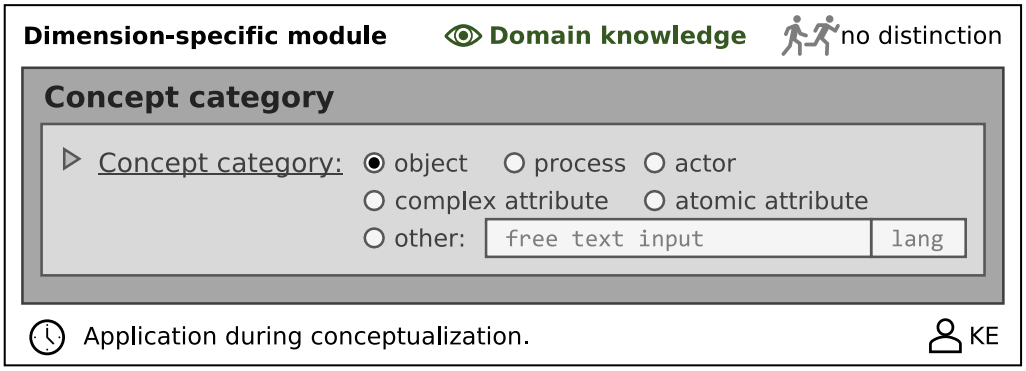

Figure 95: Exemplary application of the dimension-specific module 'Concept category'

5.6.8 Relationships to other concepts

The module *'Relationships to other concepts'* in its full version (see figure 96) is the most comprehensive module. Although it may seem overwhelming at first, its structure aids in categorizing relations. KEs and DEs see the possible options and which inputs are required. This allows to significantly improve the quality of the documentation compared to unstructured notes. A drawback are the numerous options for determining the characteristics of relations which are adopted from the OWL property characteristics. It is therefore recommended that DEs first make general entries, which are then specified by KEs.

The light version of the module allows to capture relations without categorization (see figure 98). To enable the integration of visual elements, it is equipped with the light version of the module extension *'Image'*. The application example in figure 99 shows the integration of a whiteboard drawing. The image allows readers to grasp the relationships between the concepts more quickly than with purely text-based documentation.

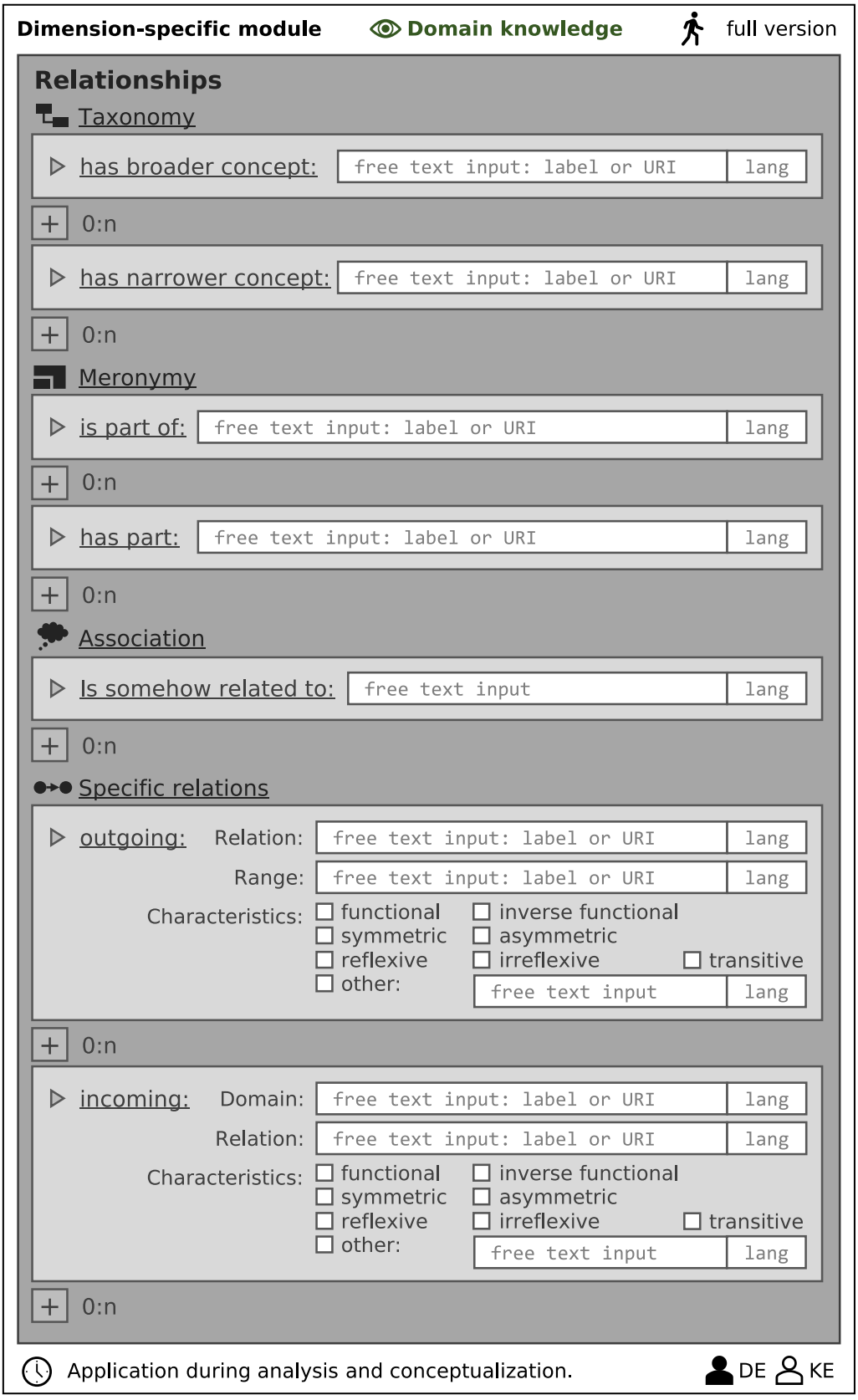

Figure 96: Dimension-specific module 'Relationships to other concepts' (full version)

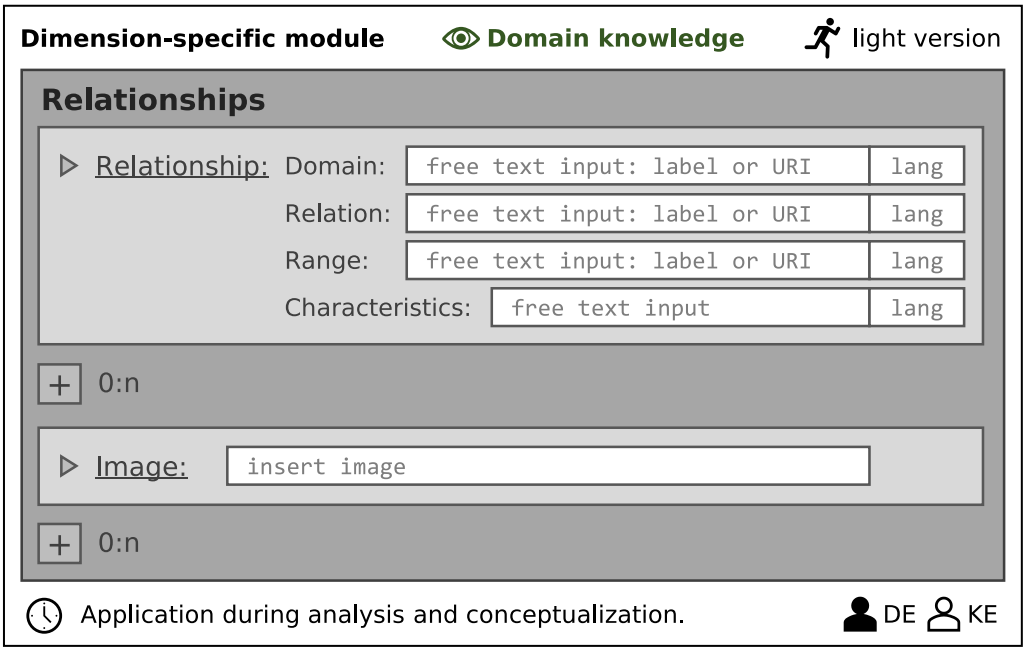

Figure 97: Dimension-specific module 'Relationships to other concepts' (light version)

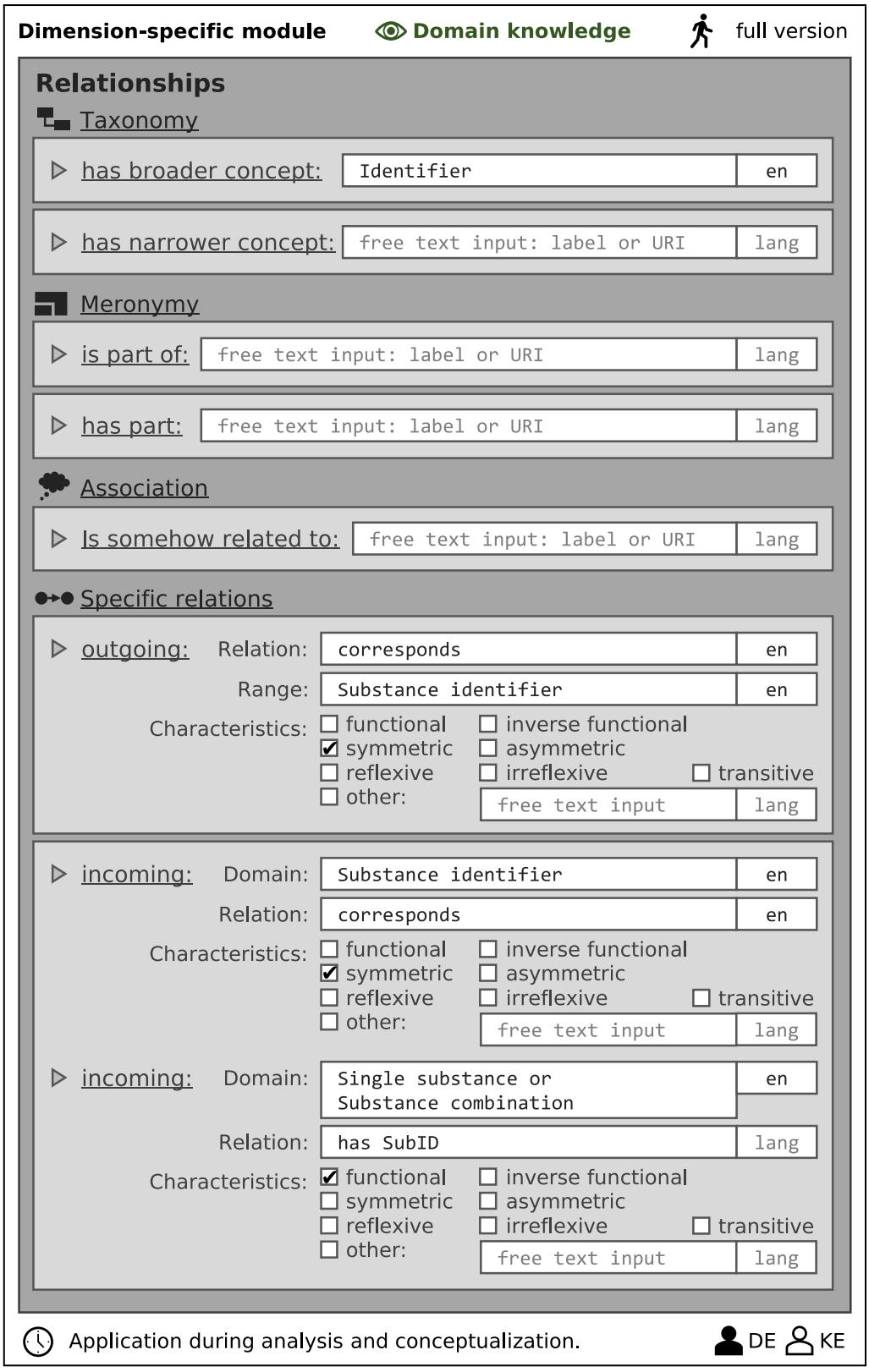

Figure 98: Exemplary application of the dimension-specific module 'Relationships to other concepts' (full version)

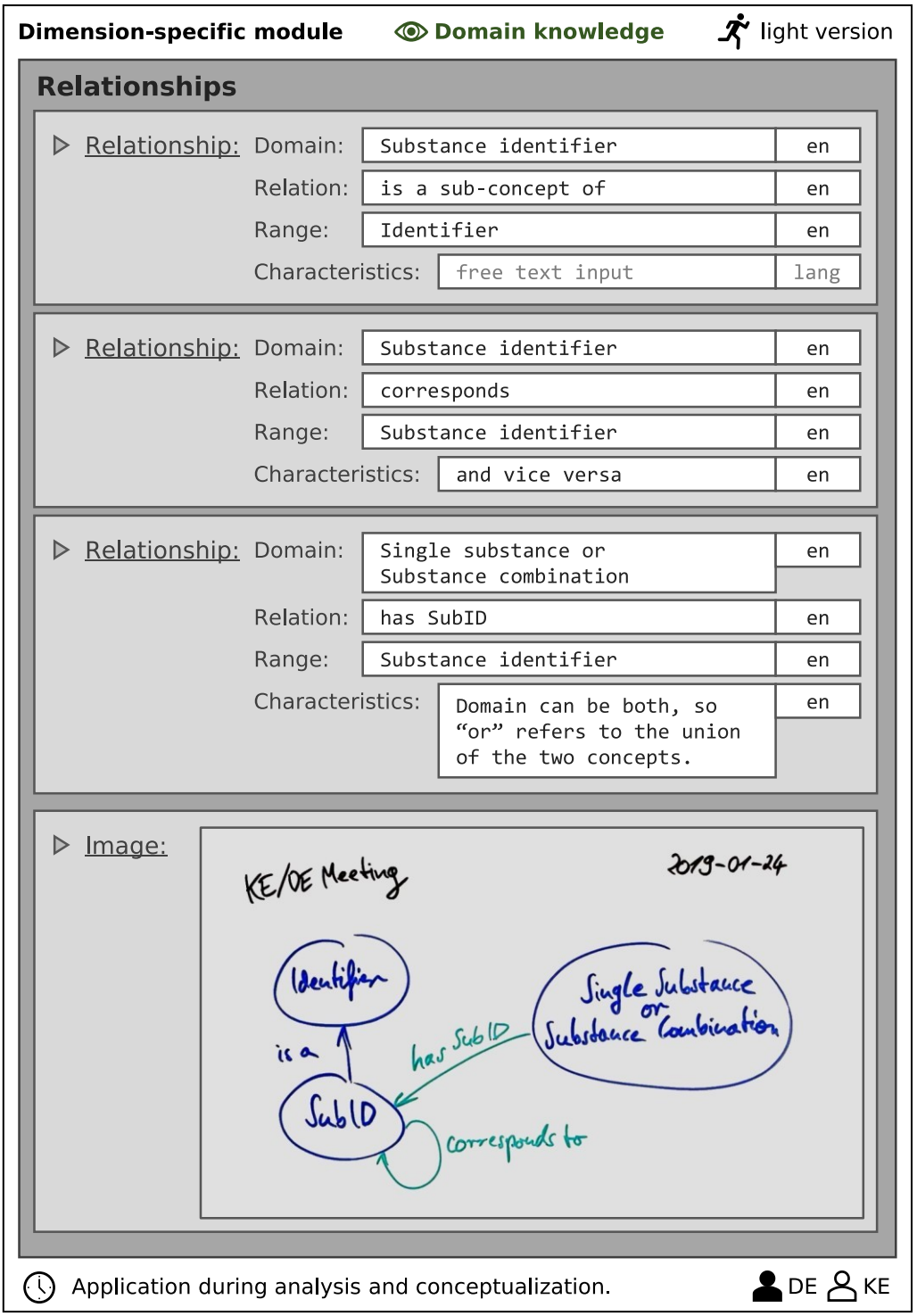

Figure 99: Exemplary application of the dimension-specific module 'Relationships to other concepts' (light version)

5.7 Dimension-specific modules: Engineering knowledge

5.7.1 Formal representations

The module *'Formal representations'* represents, so to speak, the transition from informal documentation to purely formal representation using an ontology language. Full and light version differ only in the capturing of the used ontology language and serialization (see figures 100 and 102). The module is particularly useful for KEs who want to document and share drafts of formal representations. Presumably, this only occurs with complex constructs or fundamental questions. Otherwise, it is assumed that KEs work exclusively in ontology editors. At this point, a link between the purely formal models and the concepts in the documentation is desirable. According to the analysis of the assumption personas in chapter '3 Selective insights into corporate reality', some DEs wish to see formal models.

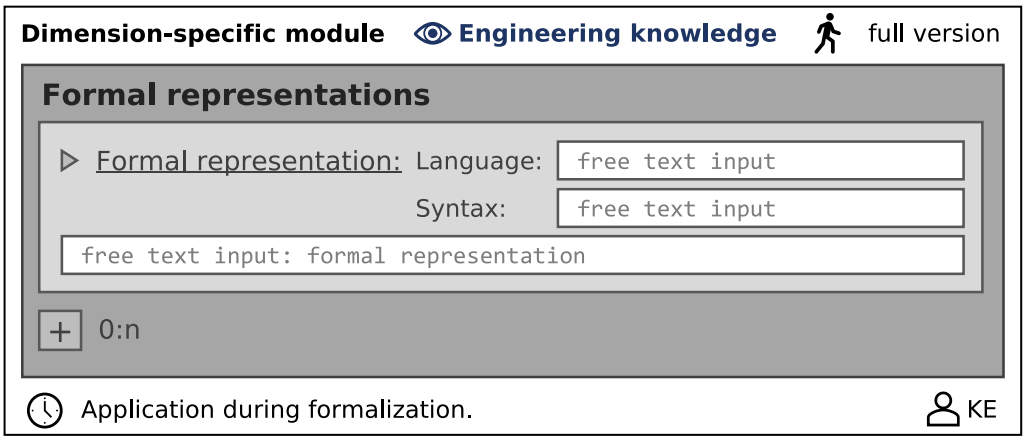

Figure 100: Dimension-specific module 'Formal representations (full version)

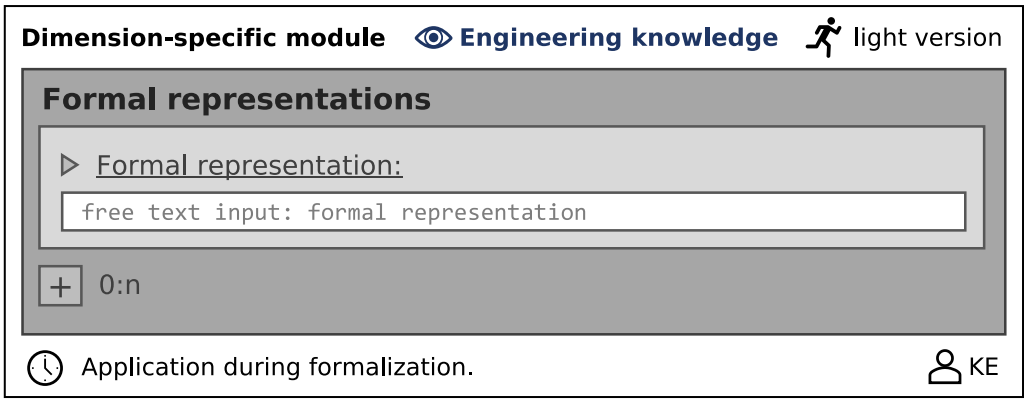

Figure 101: Dimension-specific module 'Formal representations (light version)

Dimension-specific module **Engineering knowledge** full version

Formal representations

▷ Formal representation: Language: Web Ontology Language (OWL)

Syntax: Turtle

```
@prefix : <http://ontology.company.com/1863/> .
@prefix tlo: <http://ontology.company.com/tlo/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@base <http://ontology.company.com/1863/> .

<http://ontology.company.com/1863/>
   rdf:type owl:Ontology ;
   rdfs:label "Formal representation of concept \"Substance
   identifier\""@en ;
   dcterms:creator "Ina Schneider (E83082)" ;
   rdfs:comment "draft, still working on it"@en .

dcterms:creator rdf:type owl:AnnotationProperty .
skos:altLabel rdf:type owl:AnnotationProperty .
skos:prefLabel rdf:type owl:AnnotationProperty .

:corresponds
   rdf:type owl:ObjectProperty , owl:SymmetricProperty ;
   rdfs:domain :SubstanceIdentifier ;
   rdfs:range :SubstanceIdentifier .

:hasSubID
   rdf:type owl:ObjectProperty , owl:FunctionalProperty ;
   rdfs:domain [ rdf:type owl:Class ;
      owl:unionOf ( :SingleSubstance :SubstanceCombination )
   ] ;
   rdfs:range :SubstanceIdentifier ;
   rdfs:label "has Substance identifier"@en .

:SingleSubstance rdf:type owl:Class .
:SubstanceCombination rdf:type owl:Class .

:SubstanceIdentifier rdf:type owl:Class ;
   rdfs:subClassOf tlo:Identifier ;
   skos:altLabel "SubID"@en ,
      "Substance number"@en , "Substanznummer"@de ;
skos:prefLabel "Substance identifier"@en .

tlo:Identifier rdf:type owl:Class .
```

Application during formalization. KE

Figure 102: Exemplary application of the dimension-specific module 'Formal representations' (full version)

5.7.2 Engineering knowledge resources

The module *'Engineering knowledge resources'* extends the general module *'Knowledge resources'* by the dimension-specific categorization of knowledge resources as derived in chapter '4.2 NeOn Methodology'. In the light version of the module the features of the general module are not applied (see figure 104). It is expected that only KEs capture knowledge resources such as those in the examples (see figures 105 and 106).

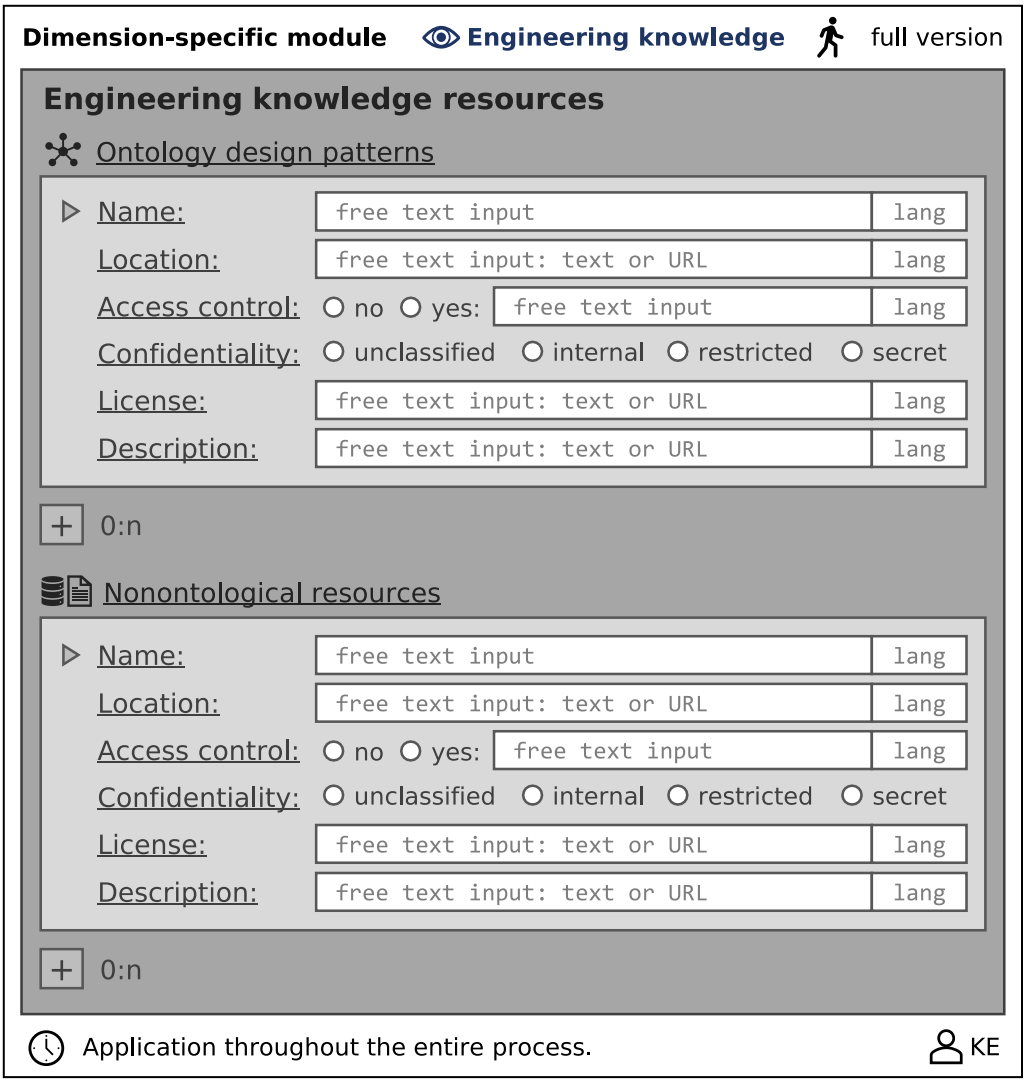

Figure 103: Dimension-specific module 'Engineering knowledge resources' (full version)

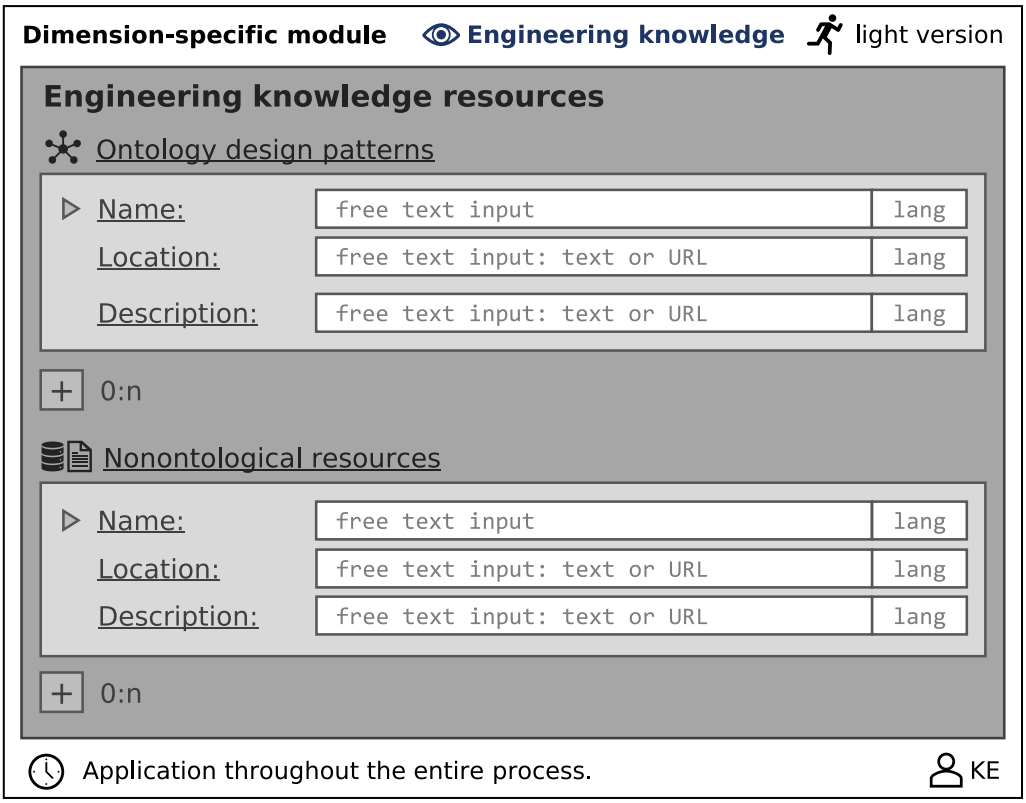

Figure 104: Dimension-specific module 'Engineering knowledge resources' (light version)

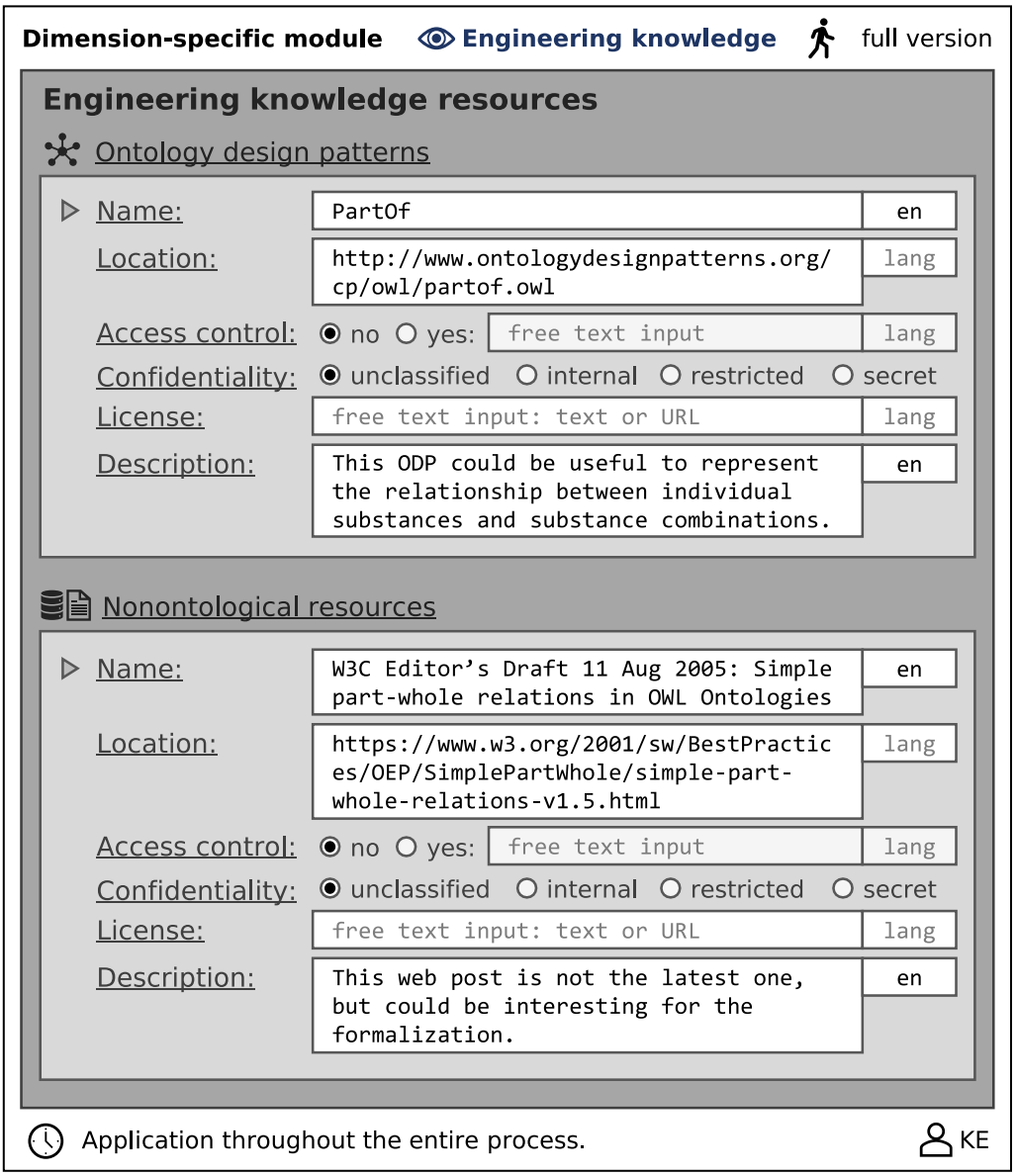

Figure 105: Exemplary application of the dimension-specific module 'Engineering knowledge resources' (full version)

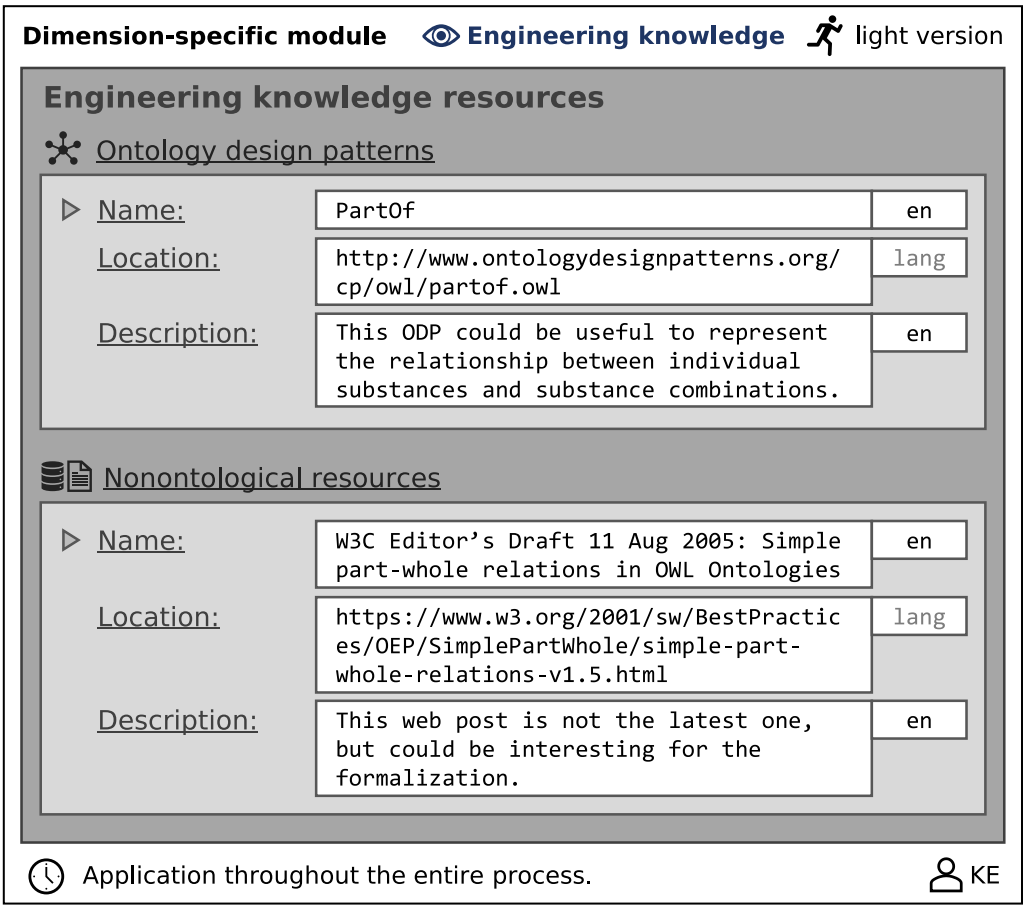

Figure 106: Exemplary application of the dimension-specific module 'Engineering knowledge resources' (light version)

5.7.3 Implementation status

The module *'Implementation status'* suggests three levels to capture the status (see figure 107). If desired, these can be replaced by custom levels. In some cases, it may be necessary to determine exactly when a status has been reached. The application simply involves selecting one of the options.

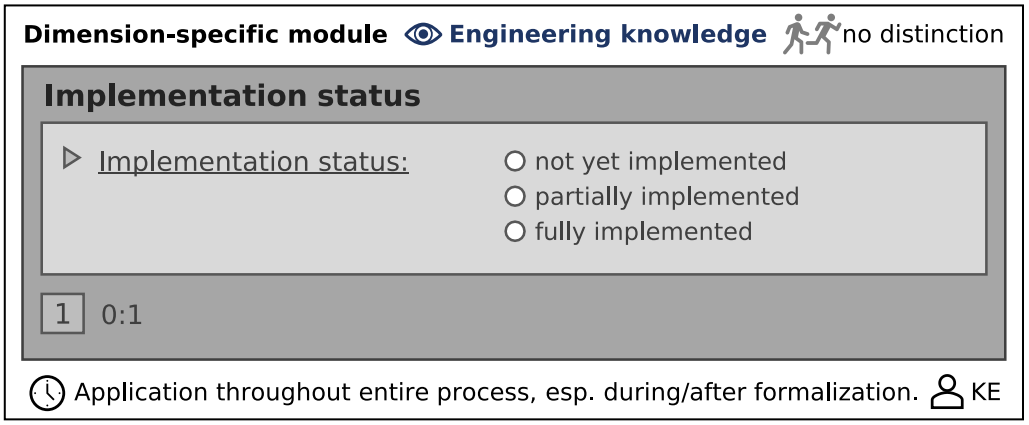

Figure 107: Dimension-specific module 'Implementation status'

5.7.4 Test status regarding reasoning

As shown in figure 108, the module *'Test status regarding reasoning'* offers three levels to choose from. The application simply involves selecting one of the options offered. In case of an error, the error can be described in the free text field, for instance, by pasting an error report.

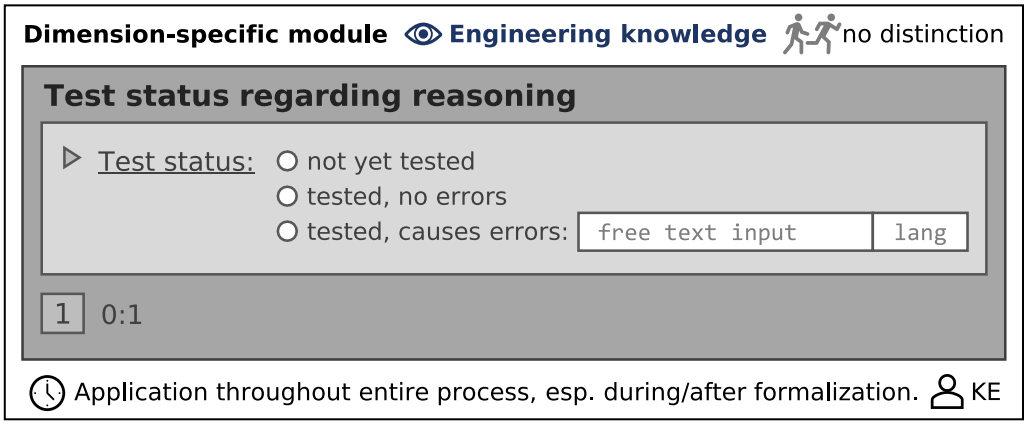

Figure 108: Dimension-specific module 'Test status regarding reasoning'

5.8 Dimension-specific modules: Management knowledge

5.8.1 Related use cases

The module *'Related use cases'* allows to assign concepts to one or more use cases. It is assumed that each concept belongs to at least one use case. The absence of a use case indicates that a concept is of little relevance. Ideally, use cases have an ID that can be used to reference them. In addition, the module offers the option of entering a link to the use case description. The following figures 109 to 112 show the different template versions and their application.

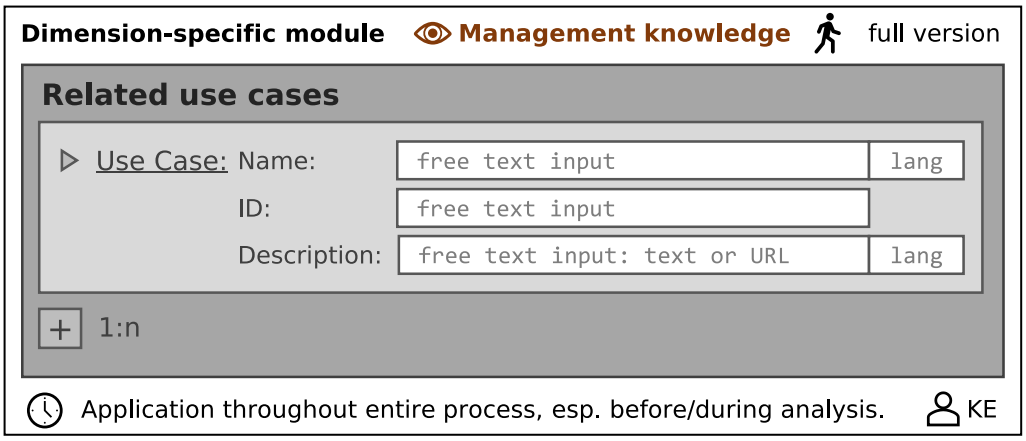

Figure 109: Dimension-specific module 'Related use cases' (full version)

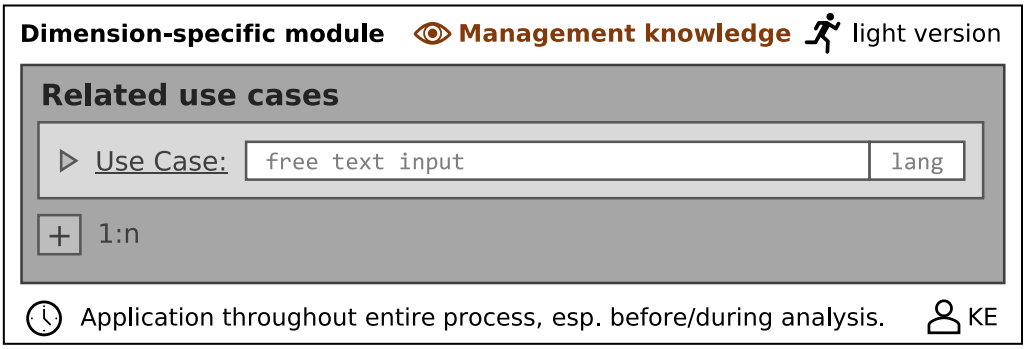

Figure 110: Dimension-specific module 'Related use cases' (light version)

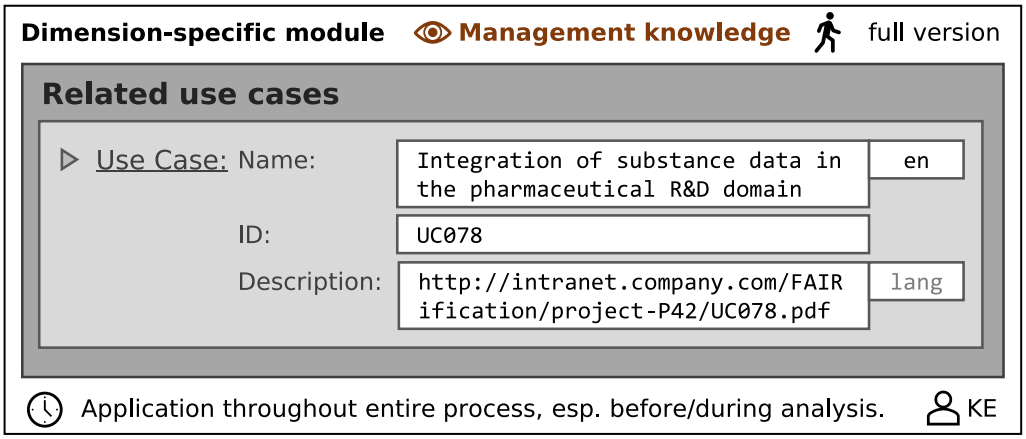

Figure 111: Exemplary application of the dimension-specific module 'Related use cases' (full version)

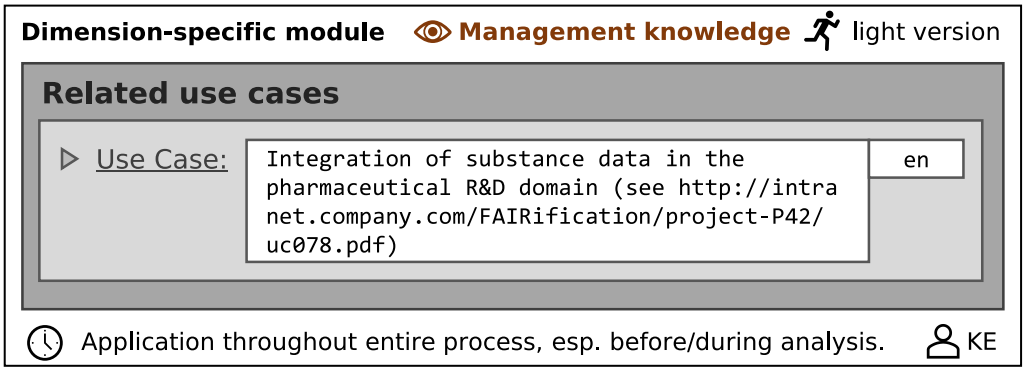

Figure 112: Exemplary application of the dimension-specific module 'Related use cases' (light version)

5.8.2 Related competency questions

The module *'Related competency questions'* allows to assign concepts to one or more CQs. It is assumed that each concept belongs to at least one CQ. The absence of a related CQ indicates that a concept is of little relevance. Ideally, CQs have an ID that can be used to reference them. The module allows to capture the complete question. In addition, a description of the CQ may be linked. The following figures 113 to 116 show the different template versions and their application.

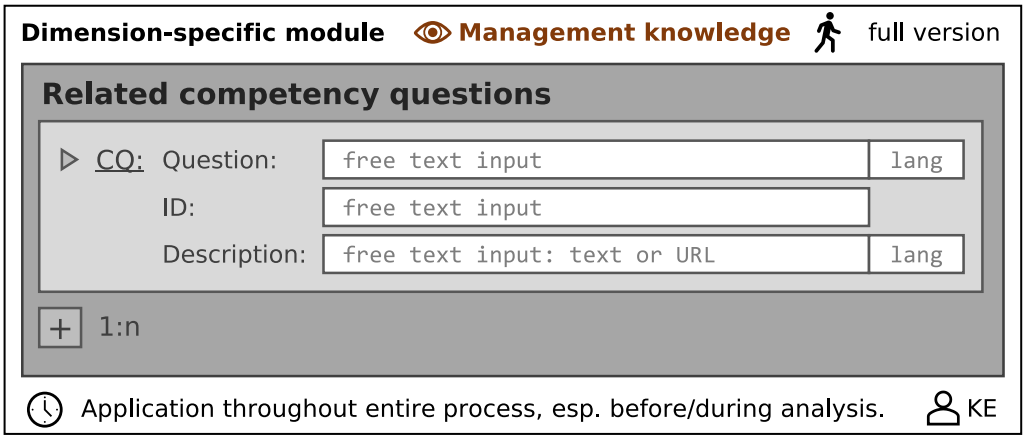

Figure 113: Dimension-specific module 'Related competency questions' (full version)

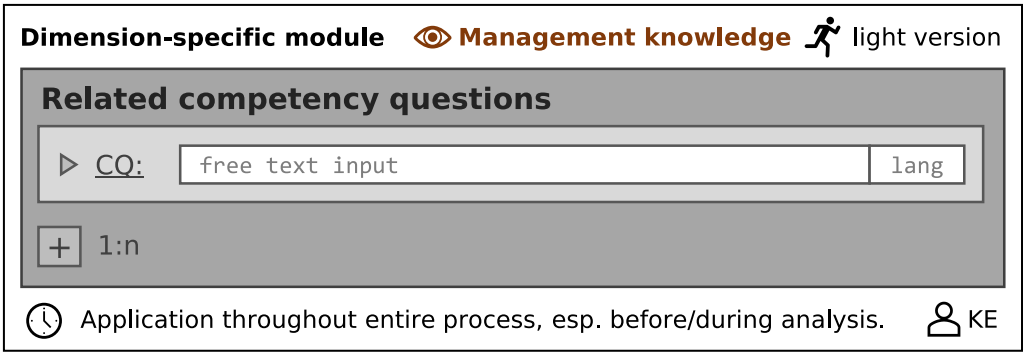

Figure 114: Dimension-specific module 'Related competency questions' (full version)

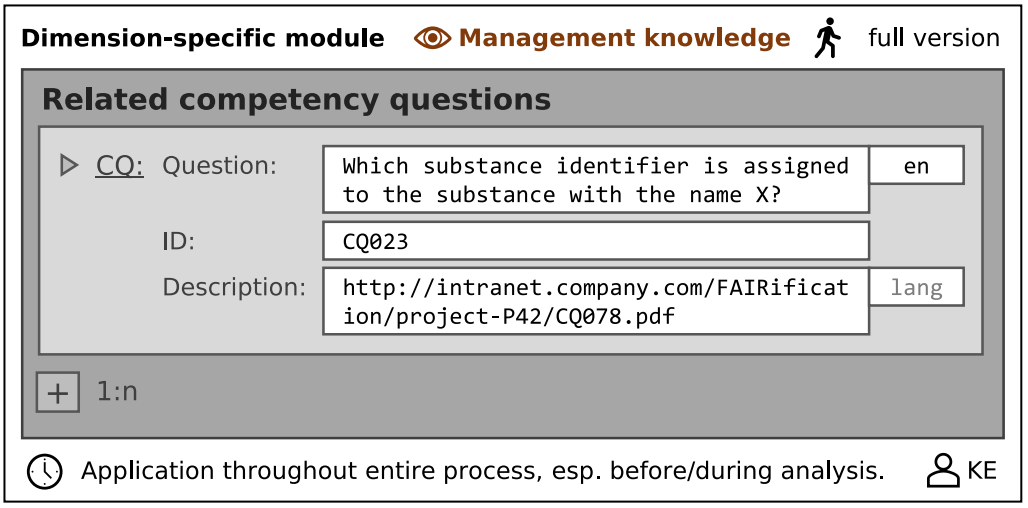

Figure 115: Exemplary application of the dimension-specific module 'Related competency questions' (full version)

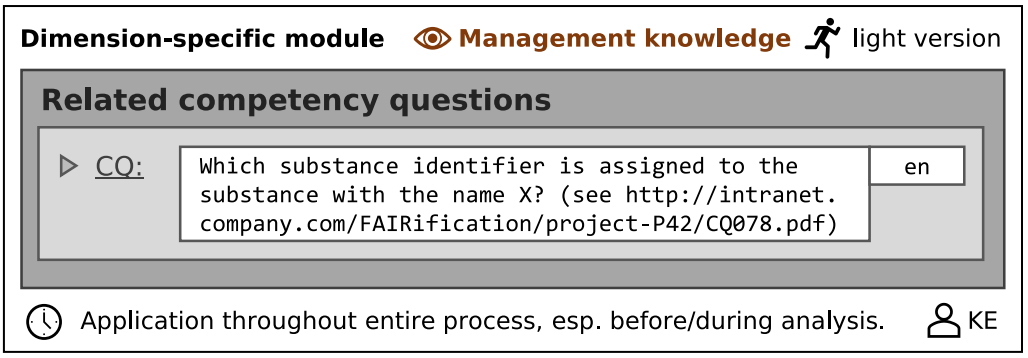

Figure 116: Exemplary application of the dimension-specific module 'Related competency questions' (light version)

5.8.3 Affiliation

The module *'Affiliation'* refers to the assignment of a concept to one or more ontologies or a set of concepts. It is expected that a concept belongs to at least one ontology. Concept sets enable the people involved to create their personal collections or topic-related collections without manipulating the content-related description of the concept. Both cases are demonstrated in figures 119 and 120.

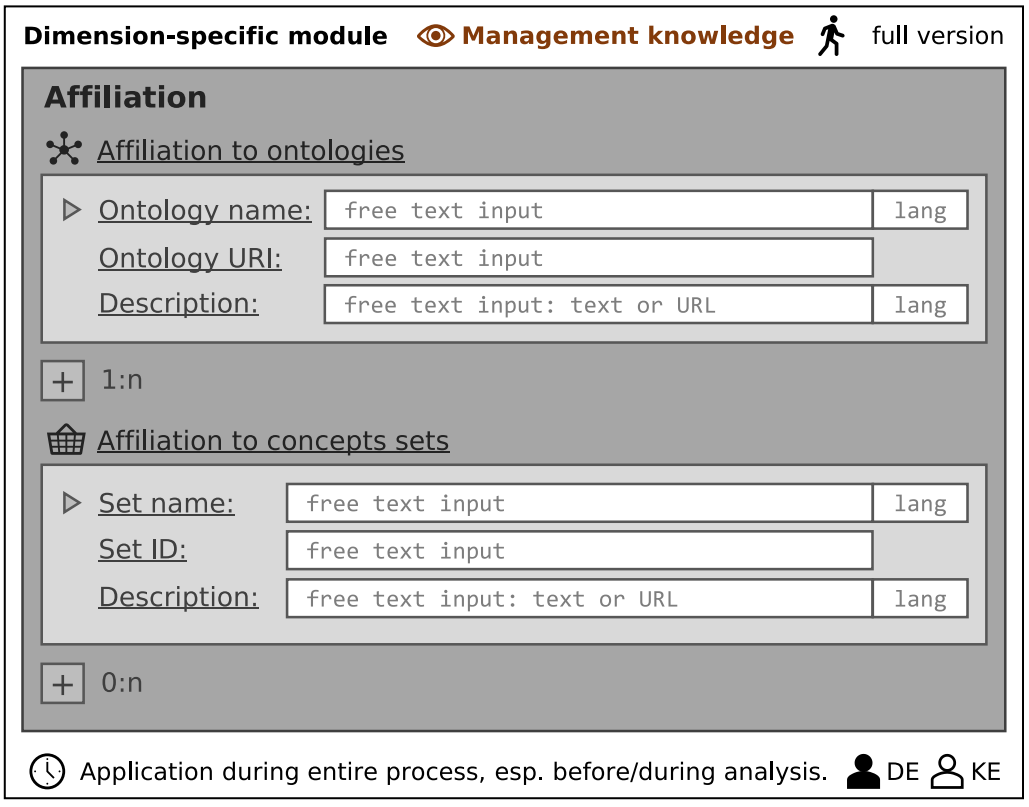

Figure 117: Dimension-specific module 'Affiliation' (full version)

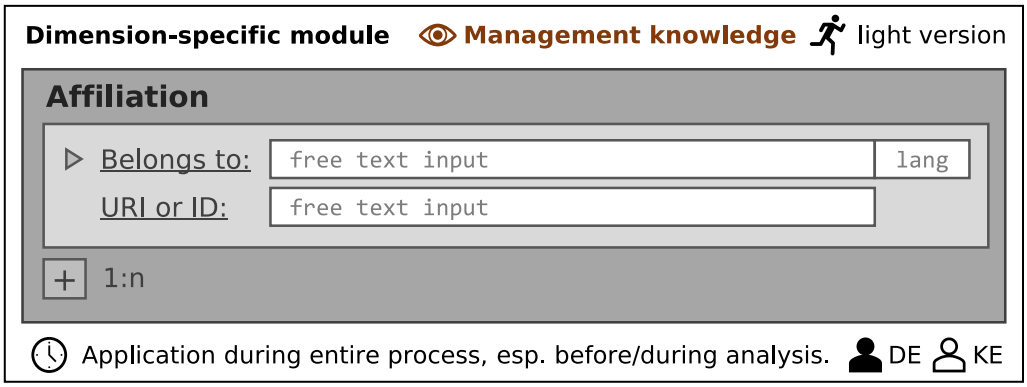

Figure 118: Dimension-specific module 'Affiliation' (light version)

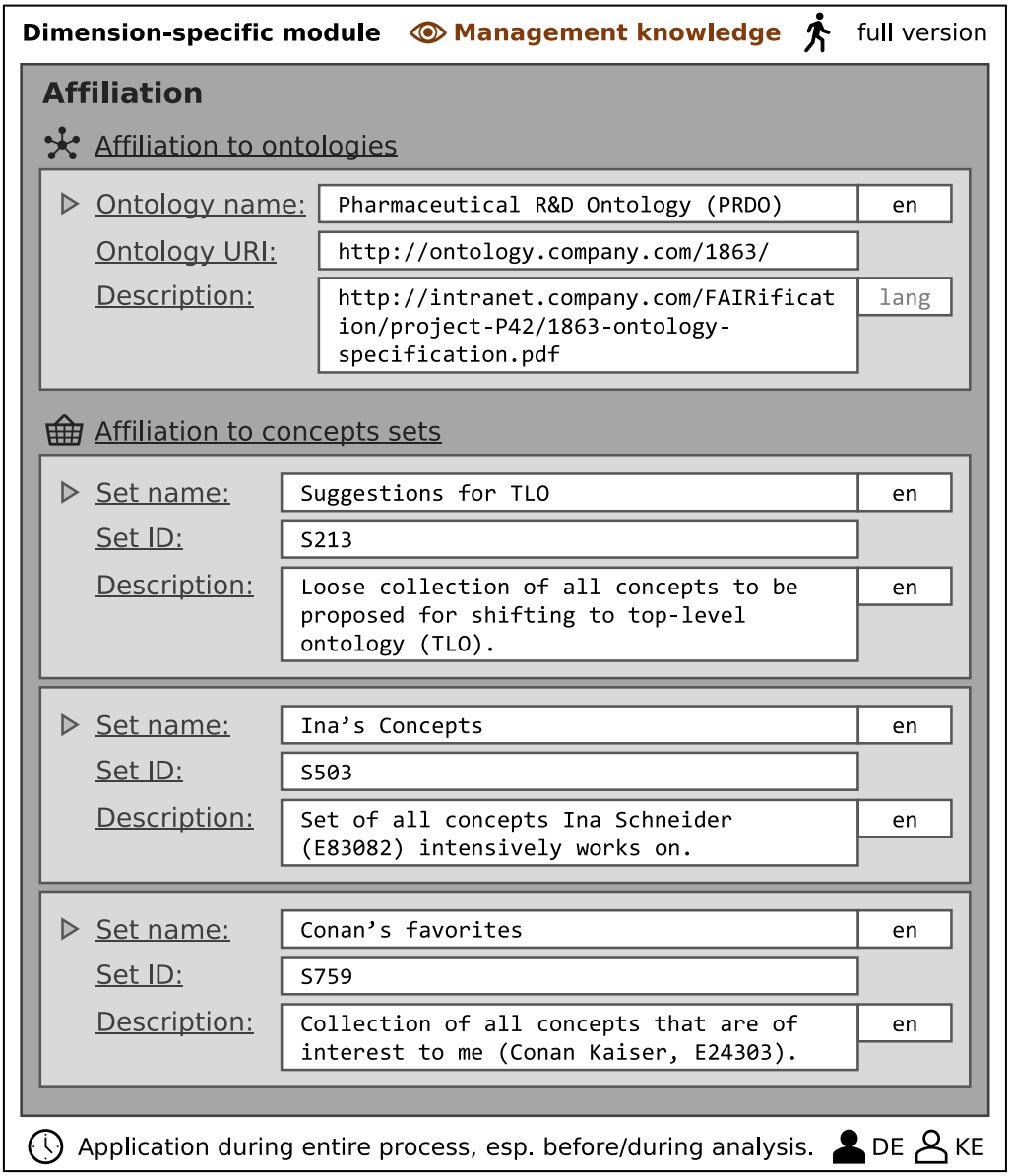

Figure 119: Exemplary application of the dimension-specific module 'Affiliation' (full version)

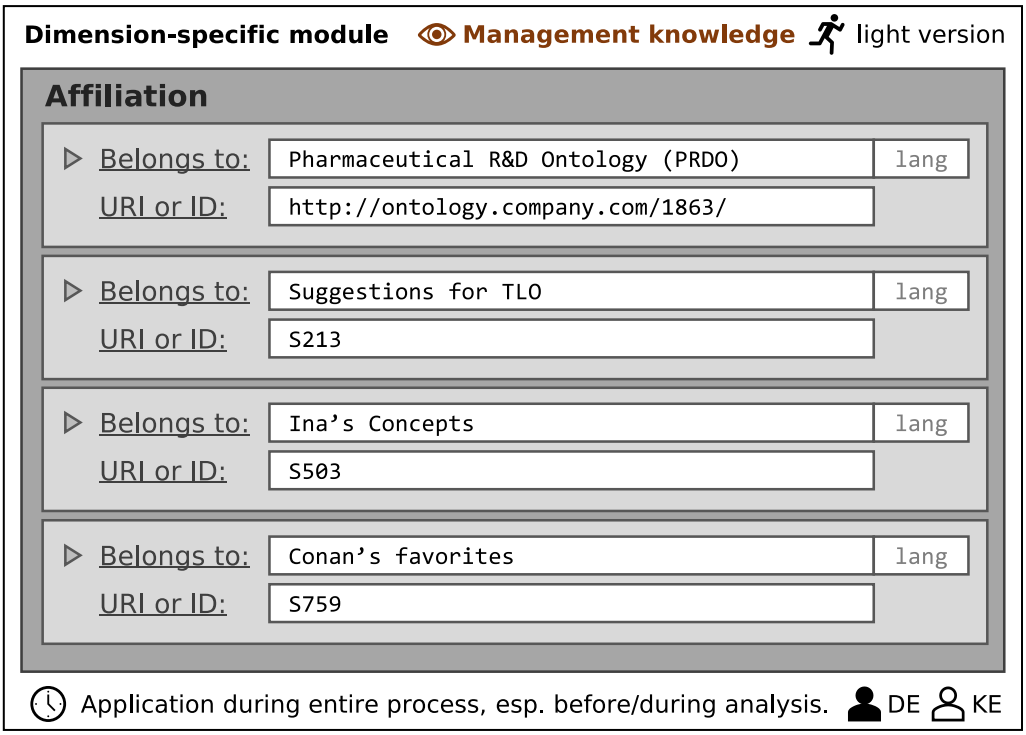

Figure 120: Exemplary application of the dimension-specific module 'Affiliation' (light version)

5.8.4 Test status regarding relevance

As shown in figure 108, the module *'Test status regarding relevance'* offers four levels to choose from. The application simply involves selecting one of the options offered. As it refers to relevance according to test results, this module differs from the next module *'Priority'*.

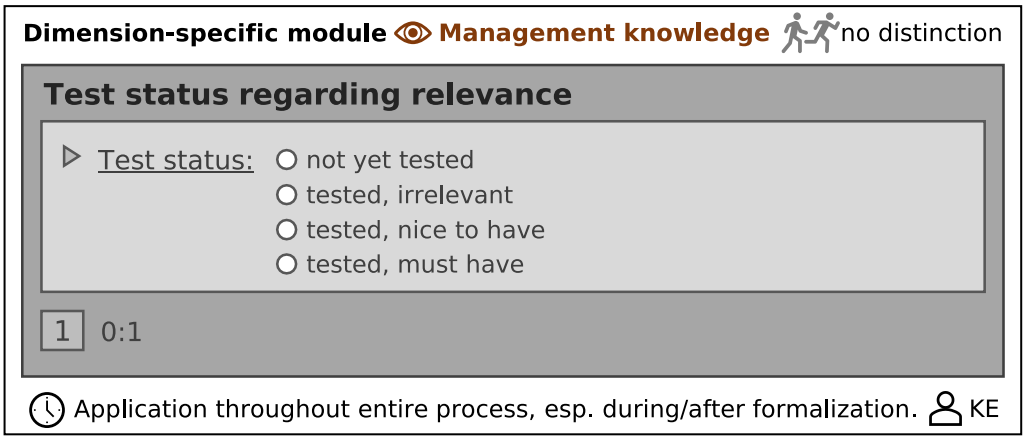

Figure 121: Dimension-specific module 'Test status regarding relevance'

5.8.5 Priority

The module *'Priority'* allows KEs to assign a priority to concepts independently of a test result. For instance, this may occur if KEs cannot consider all concepts due to time constraints, although those may be relevant according to test results. For this case, the module offers a classification based on three categories (see figure 122). If desired, these can be adapted or defined more precisely. The application simply involves selecting one of the categories.

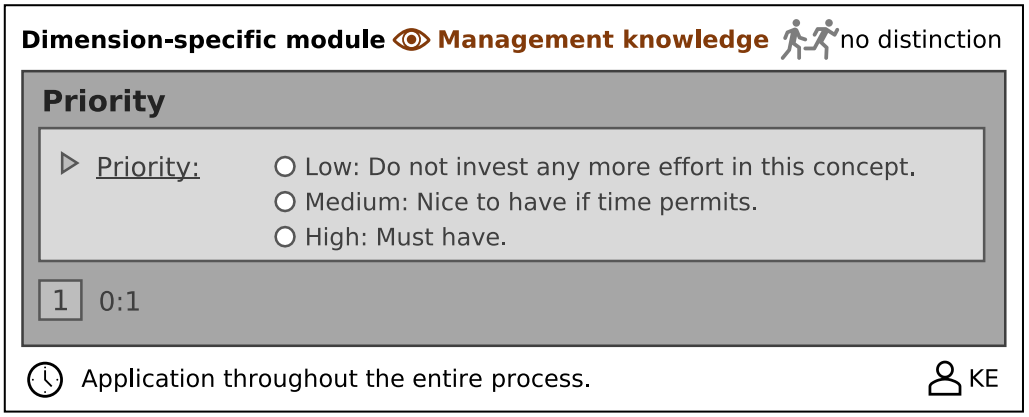

Figure 122: Dimension-specific module 'Priority'

6 Discussion

The outcome of this thesis is a modular documentation concept consisting of documentation templates. They provide guidelines for creating semi-formal, structured documentation using natural language. Each module states what should be recorded by whom, how and when. In the context of this thesis, the modules are solely conveyed by means of illustrations similar to forms or graphical user interfaces. In this way, they are also understandable to users who are not familiar with formal representations such as XML schemas or ontology models.

The documentation concept is designed for companies that create new ontologies. It enables them to overcome the identified problem of insufficient or missing documentation. By applying the documentation concept, knowledge is already externalized before its formalization. As a result, knowledge loss is prevented. Moreover, this contributes to the preservation of existing knowledge as well as to the generation of new knowledge. If the concept is implemented electronically and in a centralized way, the knowledge is easily accessible regardless of place and time.

Thanks to its modular structure, the documentation concept is adaptable to different working procedures, objectives, modes and conditions. It is based on findings from both perspectives: industry and academia. Ontology creators without a documentation approach can select a set of modules based on their needs. In this way, they do not have to start from scratch but can simply configure a suitable documentation template. Ontology creators who are already documenting may compare their documentation approach with the proposed documentation concept. In doing so, they can expand their own documentation, improve its structure or assess it as appropriate.

The structure of the single modules facilitates the creation of formal models. Inputs entered in the documentation templates can easily be identified as statements and transformed into triples. This approach is more efficient than depending on unstructured notes. By recommending which target group should document which information, the modules indicate opportunities to share the workload between DEs and KEs. As a result, knowledge acquisition is more efficient and ideally the number of dispensable meetings is reduced.

The documentation concept encourages timely, shared documentation, taking into account aspects such as discussion and agreement. Timely documentation prevents individual and collective oblivion. Joint documentation allows to identify synergies and potential misunderstandings at an early stage. On the one hand, this is desirable for economic reasons. On the other hand, consensual knowledge is required for the creation of ontologies. Since this consensus can only be formed through the communication between people, it is necessary to promote discussion and the measurement of agreement.

The limitations of the documentation concept are determined by the process under consideration, the underlying analyses, and the characteristics of the concept itself. The documentation concept only covers a certain part of the process that is performed when creating an ontology. For the documentation of the preceding and subsequent activities, the concept must be extended or linked with additional documentation artifacts.

The documentation concept is based on findings derived from survey results and assumption personas. The evaluated survey has been conducted using two questionnaires consisting of 5 questions each. In total, only 29 volunteers took part in the survey – 24 employees of a global life science company and 5 members of an external working group. The survey results are neither representative for the respective organization nor directly applicable to other organizations. Moreover, only the current situation was enquired, not the desired ideal state. Therefore, a more comprehensive and elaborate survey is needed to empirically validate the documentation concept. The same applies to the assumption personas, which have been validated with individuals but are less robust than data-driven personas.

Besides, the documentation concept is based on the analysis of the ontology engineering methodologies *UPON* and *NeOn Methodology*. These two methodologies are only a selection of the many methodologies proposed in the literature. Furthermore, the latter has not been analyzed in depth due to its comprehensiveness. Therefore, the documentation concept does not claim to support all existing methodologies to the same extent.

An essential feature of the documentation concept is its *concept-oriented* approach to documentation. This is a major advantage because concepts are the *'lowest common denominator'* shared by all approaches to ontology creation. A drawback of concept-oriented documentation is that information must be recorded redundantly. For instance, a knowledge resource that is relevant to multiple concepts must be recorded in multiple documentation

entries. This issue can be addressed during the implementation of the documentation concept. For example, by electronically supporting the recording and management of knowledge resources and people. In this case, a resource or person is recorded once. Subsequently, it can be linked to multiple concepts without creating redundancies.

7 Conclusion

Life science companies apply semantic technologies to make data FAIR – *findable*, *accessible*, *interoperable* and *reusable*. They create ontologies to formally represent knowledge and integrate data. The selective insights into corporate reality presented in this thesis pointed out that there is a *lack of knowledge externalization* in the course of ontology creation. Before its formalization, a large part of the acquired knowledge remains in the minds of the people involved. Another large part is externalized, but in the form of unstructured, personal notes. In both cases, the knowledge is difficult to access and can easily be lost.

To address the serious problem of *missing and insufficient documentation*, this thesis proposed a *documentation concept*. By means of graphical documentation templates, guidelines are provided on what should be documented by whom, how and when. Thanks to its modular structure, the documentation concept is highly customizable.

For the development of the documentation concept, sources from industry and academia were analyzed. This thesis presented the *results of a small preliminary survey* conducted among employees of a global life science company and members of an external working group. In addition, this thesis presented so-called *assumption personas*. These are profiles of archetypal ontology creators. Both insights into corporate reality – the survey results and the assumption personas – were analyzed in order to derive findings for the documentation concept. To cover the theoretical perspective as well, this thesis also includes the analysis of two ontology engineering methodologies proposed in literature: The *Unified Process for Ontology Building UPON* and the *NeOn Methodology*.

In this way, this thesis first identified a serious problem that exists in business reality. Then, the issue was approached by analyzing sources from industry and academia in a structured way. Finally, the outcome of this thesis is an appropriate solution proposal in the form of a documentation concept. This documentation concept must now prove itself in practice.

8 Outlook

From the insights, findings and results presented in this thesis arise several opportunities for further research, including the practical question of how best to implement the documentation concept in day-to-day business. One conceivable approach is to first formally represent the documentation modules and subsequently create a software application with a user interface appropriate for DEs and KEs. Following this approach, the knowledge could be formally stored from the outset and the documentation entries later transferred into the first ontology version. A first step in the formal representation of the documentation concept could be mapping the properties of the documentation templates to formal properties. For instance, some of the properties may be represented using the *Simple Knowledge Organization System* (SKOS).

A recommendation for further research is to validate the documentation concept through a comprehensive survey. Given a sufficient number of participants, the results of the comprehensive survey can be utilized for the following:

- **Data-driven personas:** While the personas presented in this thesis are based on the results of a small survey along with observations and assumptions, the results of an extensive survey can be analyzed to create data-driven personas.

- **Profiling** in the sense of creating profiles for certain user types: Using the survey data, users can be clustered, for example into groups that are characterized by similar working modes. Subsequently, those documentation modules that are most relevant for a particular work type can be grouped to a predefined set.

- **Prioritization** by weighting the different documentation modules and individual properties according to the assessment in the survey results. Possible applications: Properties considered important could be a) displayed further up in an input form, b) marked as *recommended* or c) defined as *mandatory*.

- **Streamlining:** Removal of those attributes from the documentation concept that are considered irrelevant by the vast majority without being urgently required by a minority.

- **Extension:** Introduction of new properties or modules into the documentation concept, which were not considered in the context of this thesis but are clearly required according to the survey results.

- **Refinement:** Rephrasing of explanations and adjustment of labels in the documentation concept, for instance because a) other labels are preferred by the majority or b) the free text entries indicate misunderstandings (e.g. request for a property already offered under another label).

- **Management:** The insights into the individual working modes, the collaboration and the expressed requirements can be used beyond the documentation concept as well. For instance, the results may be used to a) evaluate the prevalence of undesired working practices for the purpose of prevention in the future, b) evaluate the prevalence of desired working practices for the purpose of promotion in the future, c) plan a change management strategy for the introduction of a documentation concept. Survey participants should be informed in advance about the intended use of the responses. Therefore, an analysis as suggested in example (a) may have a negative impact on the respondents' honesty and thus the reliability of the answers.

A survey of this kind could be conducted by enquiring the relevance of each property proposed by the documentation concept – with reference to the current situation as well as the desired ideal state.

Further research may focus on the use of speech-to-text solutions in the context of documentation. The insights into the corporate reality revealed that direct, spoken communication is preferred by the people involved. Another research option is to study incentives for DEs to actively contribute to documentation. For instance, the use of gamification approaches that reward DEs for their contribution to a well-evaluated definition could be explored.

Appendices

Appendix A: Ontology engineering methodologies

The table below lists a selection of ontology engineering methodologies proposed in literature. In each case, associated publications are given. Neither the list of methodologies nor the literature given claim to be exhaustive. References between the methodologies are not reflected in the table.

101 methodology
Noy, Natalya F.; McGuinness, Deborah L. (2001) *Ontology Development 101: A Guide to Creating Your First Ontology* [Online], Knowledge Systems Laboratory. Available at http://ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf (Accessed 2019-02-18).
BORO (Business Objects Reference Ontology) methodology
Partridge, Chris (2005) *Business Objects: Re-engineering for Re-use*, 2nd edn, Huntingdon, Boro Centre.
CommonKADS methodology
Schreiber, Guus (2002) *Knowledge engineering and management: The CommonKADS methodology*, Cambridge, Mass., MIT Press.
Cyc methodology
Lenat, Douglas B.; Guha, R. V. (1990) *Building large knowledge-based systems: Representation and inference in the Cyc project*, Reading, Mass., Addison-Wesley.
DILIGENT (DIstributed, Loosely-controlled and evolvInG Engineering of oNTologies) methodology
Pinto, Helena S.; Staab, Steffen; Tempich, Christoph (2004) 'DILIGENT: Towards a fine-grained methodology for distributed, loosely-controlled and evolving engineering of ontologies pages', *Proceedings of the 16th European Conference on Artificial Intelligence, ECAI'2004*, pp. 393–397.
Pinto, Helena S.; Staab, Steffen; Tempich, Christoph; Sure, York (2006) 'Distributed Engineering of Ontologies (DILIGENT)', in Staab, Steffen; Stuckenschmidt, Heiner (eds) *Semantic Web and Peer-to-Peer,* Berlin/Heidelberg, Springer-Verlag, pp. 303–322.
Pinto, Helena S.; Tempich, Christoph; Staab, Steffen (2009) 'Ontology Engineering and Evolution in a Distributed World Using DILIGENT', in Staab, Steffen; Studer, Rudi (eds) *Handbook on Ontologies,* 2nd edn, Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 135–176.
DOGMA (Developing Ontology-Grounded Methods and Applications) methodology
Spyns, Peter; Tang, Yan; Meersman, Robert (2008) 'An ontology engineering methodology for DOGMA', *Applied Ontology*, vol. 3, no. 1-2, pp. 13–39.
EXPLODE (EXtreme Programming for Lightweight Ontology DEvelopment) methodology
Hristova, Maia (2003) *Explode: extreme programming for lightweight ontology development*, PhD Thesis, The University of Melbourne.
HCOME (Human-Centered Ontology Engineering Methodology)
Kotis, Konstantinos; Vouros, George A. (2006) 'Human-centered ontology engineering: The HCOME methodology', *Knowledge and Information Systems*, vol. 10, no. 1, pp. 109–131. DOI: 10.1007/s10115-005-0227-4.

Table is continued on the next page.

Continuation of the table on the previous page.

IDEF5 methodology
Benjamin, Perakath C.; Menzel, Christopher P.; Mayer, Richard J.; Fillion, Florence; Futrell, Michael T.; deWitte, Paula S.; Lingineni, Madhavi (1994) *IDEF5 Method Report,* Knowledge Based Systems, Inc.
JEOE ('Just Enough' Ontology Engineering) methodology
Di Maio, Paola (2011) 'Just enough' ontology engineering', in *Proceedings of the International Conference on Web Intelligence, Mining and Semantics - WIMS '11*. International Conference on Web Intelligence, Mining and Semantics. Sogndal, Norway, May 25-27, 2011. New York, New York, USA, ACM Press. DOI: 10.1145/1988688.1988698.
MENELAS methodology
Bouaud, J.; Bachimont, B.; Charlet, J.; Zweigenbaum, P. (1994) *Acquisition And Structuring Of An Ontology Within Conceptual Graphs*.
Methodology by SURE, STAAB and STUDER
Sure, York; Staab, Steffen; Studer, Rudi (2002) 'Methodology for development and employment of ontology based knowledge management applications', *ACM SIGMOD Record*, vol. 31, no. 4, p. 18. DOI: 10.1145/637411.637414.
Sure, York; Staab, Steffen; Studer, Rudi (2009) 'Ontology Engineering Methodology', in Staab, Steffen; Studer, Rudi (eds) *Handbook on Ontologies,* 2nd edn, Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 135–152.
Methodology by USCHOLD
Uschold, Mike (1996) 'Building Ontologies: Towards a Unified Methodology', in *Proceedings of Expert Systems '96*. 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems. Cambridge, UK, December 16-18, 1996, pp. 16–18.
Methodology by USCHOLD and GRÜNINGER
Uschold, Mike; Grüninger, Michael (1996) 'Ontologies: principles, methods and applications', *The Knowledge Engineering Review*, vol. 11, no. 2, pp. 93–136. DOI: 10.1017/S0269888900007797.
Methodology by USCHOLD and KING
Uschold, M.; King, M. (1995) 'Towards a methodology for building ontologies', *Workshop on Basic Ontological Issues in Knowledge Sharing: International Joint Conference on Artificial Intelligence*.
METHONTOLOGY
Fernández, Mariano; Gómez-Pérez, Asunción; Juristo, Natalia (1997) 'METHONTOLOGY: From Ontological Art Towards Ontological Engineering', in Farquhar, Adam (ed) *Ontological engineering: Papers from the 1997 AAAI Symposium, March 24-26, Stanford, California,* Menlo Park, Calif., AAAI Press, pp. 33–40.
Mikrokosmos methodology
Mahesh, Kavi (1996) *Ontology Development for Machine Translation: Ideology and Methodology,* Computing Research Laboratory, New Mexico State University.
NeOn (Networked Ontologies) methodology
NeOn Project (n.y.) *NeOn Book: NeOn Methodology in a Nutshell* [Online]. Available at http://neon-project.org/nw/NeOn_Book.html (Accessed 2019-02-18).
Suárez-Figueroa, Mari C.; Gómez-Pérez, Asunción; Motta, Enrico; Gangemi, Aldo, eds. (2012) *Ontology Engineering in a Networked World*, Berlin, Heidelberg, Springer Berlin Heidelberg.
Suárez-Figueroa, Mari C.; Gómez-Pérez, Asunción; Fernández-López, Mariano (2015) 'The NeOn Methodology framework: A scenario-based methodology for ontology development', *Applied Ontology*, vol. 10, no. 2, pp. 107–145. DOI: 10.3233/AO-150145.

Table is continued on the next page.

Continuation of the table on the previous page.

ONIONS (ONtologic Integration Of Naïve Sources) methodology
Gangemi, A.; Steve, G.; Giacomelli, F. (1996) 'ONIONS: An Ontological Methodology for Taxonomic Knowledge Integration', *Workshop on Ontological Engineering, ECAI'96. Budapest, 1996*.
OntoEdit methodology
Sure, York; Erdmann, Michael; Angele, Juergen; Staab, Steffen; Studer, Rudi; Wenke, Dirk (2002) 'OntoEdit: Collaborative Ontology Development for the Semantic Web', in *The Semantic Web — ISWC 2002. Proceedings.* First International Semantic Web Conference. Sardinia, Italy, June 9-12, 2002. Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 221–235. DOI: 10.1007/3-540-48005-6_18.
Ontolingua methodology
Farquhar, Adam; Fikes, Richard; Rice, James (1997) 'The Ontolingua Server: a tool for collaborative ontology construction', *International Journal of Human-Computer Studies*, vol. 46, no. 6, pp. 707–727. DOI: 10.1006/ijhc.1996.0121.
OTK (On-To-Knowledge) methodology
Staab, Steffen; Studer, R.; Schnurr, H.-P.; Sure, Y. (2001) 'Knowledge processes and ontologies', *IEEE Intelligent Systems*, vol. 16, no. 1, pp. 26–34. DOI: 10.1109/5254.912382.
Sure, York; Staab, Steffen; Studer, Rudi (2004) 'On-To-Knowledge Methodology (OTKM)', in Staab, Steffen; Studer, Rudi (eds) *Handbook on Ontologies,* Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 117–132.
PhysSys methodology
Borst, Pim; Akkermans, Hans; Pos, Anita; Top, Jan (1995) 'The PhysSys Ontology for Physical Systems', *Working Papers Ninth International Workshop on Qualitative Reasoning QR'95*, pp. 11–21.
PLINIUS methodology
Mars, Nicolaas J.I.; Jong, H. de; Speel, P.-H.; ter Stal, W. G.; van der Vet, P. E. (1994) 'Semi-automatic knowledge acquisition in Plinius: an engineering approach', in *Proceedings of the 8th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop.* Alberta, Canada, January 30 - February 4, 1994, pp. 4.1-4.15.
RapidOWL methodology
Auer, Sören; Herre, Heinrich (2007) 'RapidOWL — An Agile Knowledge Engineering Methodology', in *Perspectives of Systems Informatics.* 6th International Andrei Ershov Memorial Conference, PSI 2006. Novosibirsk, Russia, June 27-30, 2006. Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 424–430. DOI: 10.1007/978-3-540-70881-0_36.
SENSUS methodology
Swartout, Bill; Patil, Ramesh; Knight, Kevin; Russ, Tom (1997) 'Toward Distributed Use of Large-Scale Ontologies', in Farquhar, Adam (ed) *Ontological engineering: Papers from the 1997 AAAI Symposium, March 24-26, Stanford, California,* Menlo Park, Calif., AAAI Press, pp. 138–148.
TOVE (TOronto Virtual Enterprise) methodology
Grüninger, Michael; Fox, Mark S. (1995) 'Methodology for the Design and Evaluation of Ontologies', *Workshop on Basic Ontological Issues in Knowledge Sharing: International Joint Conference on Artificial Intelligence.*
UPON (Unified Process for ONtology) methodology
Nicola, Antonio de; Missikoff, Michele; Navigli, Roberto (2005) 'A Proposal for a Unified Process for Ontology Building: UPON', in *Database and Expert Systems Applications.* 16th International Conference, DEXA 2005. Copenhagen, Denmark, August 22-26, 2005. Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 655–664. DOI: 10.1007/11546924_64.
Nicola, Antonio de; Missikoff, Michele; Navigli, Roberto (2009) 'A software engineering approach to ontology building', *Information Systems*, vol. 34, no. 2, pp. 258–275. DOI: 10.1016/j.is.2008.07.002.

Table is continued on the next page.

Continuation of the table on the previous page.

XP.K (eXtreme Programming of Knowledge-based systems) methodology
Knublauch, Holger (2002) *An agile development methodology for knowledge-based systems including a Java framework for knowledge modeling and appropriate tool support*, Dissertation, Universität Ulm.

Table 19: Ontology engineering methodologies and related literature (not exhaustive)

Appendix B: Questionnaires and data of preliminary survey

This appendix contains the questionnaires of the preliminary survey described in chapter '3.1 Preliminary survey'. There are two different questionnaires, each consisting of five questions. One is addressed to KEs, the other to DEs. The distribution and completion were done electronically. For both questionnaires, the following is given below: The headline, the introduction text, the questions together with the answer type and the possible answer options. The respondents were invited by e-mail, which contained additional information on the deadline, person in charge, usage of the results, et cetera. This information is not included here.

In addition, this appendix contains tables with the survey data. The respective columns are annotated beneath each of these tables.

Questionnaire for knowledge engineers

Headline

"Creation of knowledge models by Knowledge Engineers"

Intro text

"This survey is aimed at people who create (among other things) ontological knowledge models at work.
Please answer the questions according to the ACTUAL STATE of your current work situation.
Please do NOT describe an ideal-typical approach, but the one that is used in your everyday work.
Nobody will be judged if one answers honestly instead of academically perfect – on the contrary.
The more realistic your answer, the more helpful it is! :-)"

Questions and answers options

1. Direction
What are the modelling approaches that you actually use at work?
[One or more answers possible.]

- □ Top-down: from general to particular (first elaborate the general concepts then specialize them)
- □ Bottom-up: from particular to general (first look at the data and elaborate most specific concepts, then abstract them)
- □ Middle-out: combined (first identify most important concepts, then generalize and specialize)

2. Main goal

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.]

- o 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.
- o 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.

3. Relevant information per concept

Which of the following information do you consider for each concept (if applicable) while creating knowledge models? If you are considering information that is not listed here, please write these in the field "other"!
[None, single or multiple selection possible. Optional: Additional free text entry.]

- □ Definition or explanation
- □ Alternative labels
- □ Related people
- □ Related data sources
- □ Related vocabularies or standards
- □ Related literature
- □ Context
- □ *Other* (optional free text entry)

4. Time point of documentation

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.]

- o I start the documentation BEFORE I create the formal model.
- o I start the documentation WHILE I create the formal model.
- o I start the documentation AFTER I have created the first version of the formal model.
- o I don't create a documentation.

5. Exchange with domain experts

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.]

- o 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.
- o I document the results directly in the formal model (e.g. using annotation properties).
- o I document the results informal but in a structured way (e.g. using natural language and a template).
- o I document the results in an unstructured way (e.g. natural language notes, meeting minutes, etc.).

Questionnaire for domain experts

Headline

"Domain experts – Providing knowledge for the creation of ontologies"

Intro text

"This survey is aimed at so-called domain experts. Domain experts share their knowledge about a domain so that knowledge engineers can create ontologies based on it. Therefore, domain experts are essential for the creation of knowledge models."

Questions and answer options

1. Motivation

What is your motivation for contributing to the creation of ontologies?

[One or more answers possible. Optional: Additional free text entry.]

- □ I have to do it because I've been instructed to do so.
- □ I like the people who work on the ontology.
- □ I think ontologies are a promising technology.
- □ Because apparently everyone does.
- □ I would like to learn more about ontologies as a technology.
- □ I would like to learn more about my domain and exchange knowledge with other domain experts.
- □ *Other* (optional free text entry)

2. Expectations
What do you expect from ontologies for your daily work?
[One or more answers possible. Optional: Additional free text entry.]

- □ I expect them to make my job easier.
- □ I expect to be able to work directly with ontologies to gain knowledge about a domain (to see definitions, to understand context, etc.).
- □ I expect ontologies working in the background to improve the interaction between the IT systems I use.
- □ *Other* (optional free text entry)

3. Current usage of ontologies
For what purpose do you currently use ontologies?
[One or more answers possible. Optional: Additional free text entry.]

- □ I don't use ontologies in my daily work (as far as I know).
- □ I use ontologies to become familiar with the technology.
- □ I use ontologies within a proof of concept to test whether a productive usage is feasible.
- □ I use ontologies in a productive way.
- □ *Other* (optional free text entry)

4. Exchange
Which way of collaboration do you prefer when exchanging with knowledge engineers?
[Rank the given options by using Drag&Drop or the arrows.]

- ↕ Face-to-face meeting
- ↕ Video call
- ↕ Voice call
- ↕ Chat (e.g. Skype, Microsoft Teams)
- ↕ E-mail
- ↕ Document on a shared drive (e.g. on a SharePoint)
- ↕ Other

5. Optional clarification on Question 4 – Other
If you have ranked "Other" among the Top 3 for the previous question, please use the free text field below for explanation.
What other way of exchanging with knowledge engineers do you prefer?
[Free text field, optional.]

optional free text entry

Survey data of knowledge engineers

1. Direction

ID	A	B	C
01	0	0	1
02	0	1	0
03	0	1	1
04	0	1	0
05	0	1	0
06	0	1	1
07	0	1	0
08	0	0	1
09	1	1	0
10	1	1	1
11	0	1	0
12	0	0	1
13	0	0	1

Table 20: Survey data knowledge engineers, question 1: Direction

ID: Identification number of responding person

A: Answer = "*Top-down: from general to particular (first elaborate the general concepts then specialize them)*"

B: Answer = "*Bottom-up: from particular to general (first look at the data and elaborate most specific concepts, then abstract them)*"

C: Answer = "*Middle-out: combined (first identify most important concepts, then generalize and specialize)*"

Values: The value "`0`" corresponds to *False*, i.e. the answer option was *not* selected. The value "`1`" corresponds to *True*, i.e. the answer option was selected.

2. Main goal

ID	A	B
01	1	0
02	1	0
03	1	0
04	1	0
05	1	0
06	1	0
07	1	0
08	0	1
09	1	0
10	0	1
11	1	0
12	0	1
13	0	1

Table 21: Survey data knowledge engineers, question 2: Main goal

ID: Identification number of responding person

A: Answer = "*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.*"

B: Answer = "*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.*"

Values: The value "0" corresponds to *False*, i.e. the answer option was *not* selected. The value "1" corresponds to *True*, i.e. the answer option was selected.

3. Relevant information per concept

ID	A	B	C	D	E	F	G	H
01	1	0	0	1	1	0	1	0
02	1	1	1	1	1	1	1	1
03	1	1	0	0	1	0	0	0
04	1	0	0	1	1	0	0	0
05	1	0	0	0	0	0	0	0
06	1	1	1	0	1	0	1	0
07	1	1	0	0	1	0	0	0
08	1	1	1	1	0	0	1	1
09	1	0	0	0	0	0	0	0
10	1	1	0	1	1	1	0	0
11	0	0	0	1	1	0	1	0
12	1	1	1	1	1	1	1	0
13	1	1	0	1	1	0	0	0

Table 22: Survey data knowledge engineers, question 3: Relevant information per concept

ID	Optional free text entry for H: Other
02	The use case around the model
08	Identifiers

Table 23: Survey data knowledge engineers, question 3: Relevant information per concept (free text entries)

ID: Identification number of responding person

A: Answer = "*Definition or explanation*"

B: Answer = "*Alternative labels*"

C: Answer = "*Related people*"

D: Answer = "*Related data sources*"

E: Answer = "*Related vocabularies or standards*"

F: Answer = "*Related literature*"

G: Answer = "*Context*"

H: Answer = "*Other (optional free text entry)*"

Values: The value "0" corresponds to *False*, i.e. the answer option was *not* selected. The value "1" corresponds to *True*, i.e. the answer option was selected.

4. Time point of documentation

ID	A	B	C	D
01	0	1	0	0
02	0	1	0	0
03	0	0	1	0
04	0	1	0	0
05	1	0	0	0
06	0	1	0	0
07	0	1	0	0
08	0	0	1	0
09	0	1	0	0
10	0	0	1	0
11	0	0	0	1
12	1	0	0	0
13	0	0	1	0

Table 24: Survey data knowledge engineers, question 4: Time point of documentation

ID: Identification number of responding person

A: Answer = "*I start the documentation BEFORE I create the formal model.*"

B: Answer = "*I start the documentation WHILE I create the formal model.*"

C: Answer = "*I start the documentation AFTER I have created the first version of the formal model.*"

D: Answer = "*I don't create a documentation.*"

Values: The value "`0`" corresponds to *False*, i.e. the answer option was *not* selected. The value "`1`" corresponds to *True*, i.e. the answer option was selected.

5. Exchange with domain experts

ID	A	B	C	D
01	0	0	0	1
02	0	1	0	0
03	0	0	1	0
04	0	1	0	0
05	0	0	0	1
06	0	0	1	0
07	0	0	1	0
08	0	0	0	1
09	0	0	0	1
10	0	0	1	0
11	0	0	0	1
12	0	0	1	0
13	0	0	0	1

Table 25: Survey data knowledge engineers, question 5: Exchange with domain experts

ID: Identification number of responding person

A: Answer = "*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: Answer = "*I document the results directly in the formal model (e.g. using annotation properties).*"

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

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

Values: The value "0" corresponds to *False*, i.e. the answer option was *not* selected. The value "1" corresponds to *True*, i.e. the answer option was selected.

Survey data of internal domain experts

1. Motivation

ID	A	B	C	D	E	F	G
01	0	1	1	0	1	0	0
02	0	1	1	0	1	1	1
03	0	0	1	0	0	0	1
04	0	0	0	0	0	0	1
05	0	0	1	0	1	1	1
06	0	0	1	0	1	1	0
07	0	0	1	0	1	1	0
08	0	0	1	0	1	0	0
09	0	0	1	0	1	1	0
10	0	0	1	0	1	1	0
11	0	0	1	0	0	1	0

Table 26: Survey data internal domain experts, question 1: Motivation

ID	Optional free text entry for G: Other
02	I think they would build an important basis for the very needed data integration.
03	Best way to achieve digital agenda target of transparency and FAIR data
04	For integration of clinical studies harmonization of data is essential and ontologies are a promising concept to achieve this harmonization.
05	I would like to share my domain knowledge with experts for building ontologies.

Table 27: Survey data internal domain experts, question 1: Motivation (free text entries)

ID: Identification number of responding person

A: Answer = "*I have to do it because I've been instructed to do so.*"

B: Answer = "*I like the people who work on the ontology.*"

C: Answer = "*I think ontologies are a promising technology.*"

D: Answer = "*Because apparently everyone does.*"

E: Answer = "*I would like to learn more about ontologies as a technology.*"

F: Answer = "*I would like to learn more about my domain and exchange knowledge with other domain experts.*"

G: Answer = "*Other (optional free text entry)*"

Values: The value "0" corresponds to *False*, i.e. the answer option was *not* selected. The value "1" corresponds to *True*, i.e. the answer option was selected.

2. Expectations

ID	A	B	C	D
01	0	0	1	1
02	1	1	1	0
03	1	1	1	0
04	1	0	0	0
05	0	1	1	1
06	1	0	1	0
07	0	0	1	0
08	0	0	1	0
09	0	1	1	0
10	1	1	1	0
11	1	1	1	1

Table 28: Survey data internal domain experts, question 2: Expectations

ID	Optional free text entry for D: Other
01	I expect to receive information or explanation of data, which is currently not available.
05	I expect that it will take much more exploratory work and efforts before ontologies will have a significant impact on our "daily life" as Data Managers, Data Scientists, Data Engineers, etc.
11	On the long term and in addition with other technologies

Table 29: Survey data internal domain experts, question 2: Expectations (free text entries)

ID: Identification number of responding person

A: Answer = "*I expect them to make my job easier.*"

B: Answer = "*I expect to be able to work directly with ontologies to gain knowledge about a domain (to see definitions, to understand context, etc.).*"

C: Answer = "*I expect ontologies working in the background to improve the interaction between the IT systems I use.*"

D: Answer = "*Other (optional free text entry)*"

Values: The value "0" corresponds to *False*, i.e. the answer option was *not* selected. The value "1" corresponds to *True*, i.e. the answer option was selected.

3. Current usage of ontologies

ID	A	B	C	D	E
01	1	0	1	0	0
02	0	0	1	0	0
03	0	0	1	0	0
04	0	0	1	0	0
05	1	1	1	0	0
06	0	0	0	0	1
07	0	1	0	0	0
08	1	0	1	0	0
09	0	0	1	0	0
10	1	0	0	0	0
11	1	0	0	0	0

Table 30: Survey data internal domain experts, question 3: Current usage of ontologies

ID	Optional free text entry for E: Other
06	I had first contact with ontologies during an AI project ([project name]*) and a [project name]* use case as business contributor.
* Editor's note: Company internal project names have been removed from the original response.	

Table 31: Survey data internal domain experts, question 3: Current usage of ontologies (free text entries)

ID: Identification number of responding person

A: Answer = "*I don't use ontologies in my daily work (as far as I know).*"

B: Answer = "*I use ontologies to become familiar with the technology.*"

C: Answer = "*I use ontologies within a proof of concept to test whether a productive usage is feasible.*"

D: Answer = "*I use ontologies in a productive way.*"

E: Answer = "*Other (optional free text entry)*"

Values: The value "0" corresponds to *False*, i.e. the answer option was *not* selected. The value "1" corresponds to *True*, i.e. the answer option was selected.

4. Exchange

ID	A	B	C	D	E	F	G
01	1	7	3	2	4	5	6
02	2	6	1	3	4	5	7
03	1	2	5	3	4	6	7
04	5	6	1	2	4	3	7
05	1	4	3	2	6	5	7
06	1	3	2	5	4	6	7
07	1	2	3	4	5	6	7
08	2	6	4	3	5	7	1
09	1	2	3	5	6	4	7
10	1	6	3	7	2	4	5
11	1	3	2	4	5	6	7

Table 32: Survey data internal domain experts, question 4: Exchange

ID: Identification number of responding person

A: Answer = "*Face-to-face meeting*"

B: Answer = "*Video call*"

C: Answer = "*Voice call*"

D: Answer = "*Chat (e.g. Skype, Microsoft Teams)*"

E: Answer = "*E-mail*"

F: Answer = "*Document on a shared drive (e.g. on a SharePoint)*"

G: Answer = "*Other*"

Values: The values express a ranking, whereby "1" corresponds to the first rank (most preferred) and "7" to the last rank (least preferred). The intermediate values ("2", "3", "4", "5", "6") represent intermediate ranks.

5. Optional clarification on Question 4 – Other

ID	Optional free text entry
08	GitHub or similar containing documents and ontologies together would be a preferred exchange platform for me.

Table 33: Survey data internal domain experts, question 5: Optional clarification on Question 4 – Other

ID: Identification number of responding person

Values: This value is an optional free text entry from those respondents who have ranked the answer option "Other" among the Top 3 in question 4.

Survey data of external domain experts

1. Motivation

ID	A	B	C	D	E	F	G
01	0	0	1	0	1	1	1
02	0	0	1	0	0	0	0
03	0	0	1	0	0	1	0
04	0	0	1	0	1	1	0
05	0	1	1	0	0	0	0

Table 34: Survey data external domain experts, question 1: Motivation

ID	Optional free text entry for G: Other
01	Because these models represent the real-world models and facilitate data exchange

Table 35: Survey data external domain experts, question 1: Motivation (free text entries)

ID: Identification number of responding person

A: Answer = "*I have to do it because I've been instructed to do so.*"

B: Answer = "*I like the people who work on the ontology.*"

C: Answer = "*I think ontologies are a promising technology.*"

D: Answer = "*Because apparently everyone does.*"

E: Answer = "*I would like to learn more about ontologies as a technology.*"

F: Answer = "*I would like to learn more about my domain and exchange knowledge with other domain experts.*"

G: Answer = "*Other (optional free text entry)*"

Values: The value "0" corresponds to *False*, i.e. the answer option was *not* selected. The value "1" corresponds to *True*, i.e. the answer option was selected.

2. Expectations

ID	A	B	C	D
01	1	1	1	0
02	0	0	1	0
03	0	1	1	0
04	1	0	1	0
05	1	1	1	0

Table 36: Survey data external domain experts, question 2: Expectations

ID: Identification number of responding person

A: Answer = "*I expect them to make my job easier.*"

B: Answer = "*I expect to be able to work directly with ontologies to gain knowledge about a domain (to see definitions, to understand context, etc.).*"

C: Answer = "*I expect ontologies working in the background to improve the interaction between the IT systems I use.*"

D: Answer = "*Other (optional free text entry)*"

Values: The value "`0`" corresponds to *False*, i.e. the answer option was *not* selected. The value "`1`" corresponds to *True*, i.e. the answer option was selected.

3. Current usage of ontologies

ID	A	B	C	D	E
01	0	1	1	0	1
02	0	0	1	0	0
03	0	0	1	0	0
04	1	0	0	0	0
05	0	1	1	0	0

Table 37: Survey data external domain experts, question 3: Current usage of ontologies

ID	Optional free text entry for E: Other
01	I develop ontologies as proof of concept, with the move to production.

Table 38: Survey data external domain experts, question 3: Current usage of ontologies (free text entries)

ID: Identification number of responding person

A: Answer = "*I don't use ontologies in my daily work (as far as I know).*"

B: Answer = "*I use ontologies to become familiar with the technology.*"

C: Answer = "*I use ontologies within a proof of concept to test whether a productive usage is feasible.*"

D: Answer = "*I use ontologies in a productive way.*"

E: Answer = "*Other (optional free text entry)*"

Values: The value "0" corresponds to *False*, i.e. the answer option was *not* selected. The value "1" corresponds to *True*, i.e. the answer option was selected.

4. Exchange

ID	A	B	C	D	E	F	G
01	2	1	3	4	5	6	7
02	1	3	4	6	5	2	7
03	1	2	4	3	5	6	7
04	1	2	3	6	5	7	4
05	1	5	3	2	6	4	7

Table 39: Survey data external domain experts, question 4: Exchange

ID: Identification number of responding person

A: Answer = "*Face-to-face meeting*"

B: Answer = "*Video call*"

C: Answer = "*Voice call*"

D: Answer = "*Chat (e.g. Skype, Microsoft Teams)*"

E: Answer = "*E-mail*"

F: Answer = "*Document on a shared drive (e.g. on a SharePoint)*"

G: Answer = "*Other*"

Values: The values express a ranking, whereby "1" corresponds to the first rank (most preferred) and "7" to the last rank (least preferred). The intermediate values ("2", "3", "4", "5", "6") represent intermediate ranks.

5. Optional clarification on Question 4 – Other

ID	Optional free text entry
04	A collaboratory development environment like github.

Table 40: Survey data external domain experts, question 5: Optional clarification on Question 4 – Other

ID: Identification number of responding person

Values: This value is an optional free text entry from those respondents who have ranked the answer option "Other" among the Top 3 in question 4.

The answer belonging to ID "04" was entered although the answer option "Other" in question 4 was ranked 4th, i.e. not among the Top 3.

Appendix C: UPON workflows

In the following, the five workflows of the *Unified Process for Ontology Building* (UPON), which are briefly summarized in chapter '4.1.2 Workflows', are described in more detail.

Requirements workflow

The objective of the first workflow is to capture the requirements (Cf. Nicola et al., 2009, p. 260). Therefore, it takes place prior to the process considered in this thesis (see chapter '2.4 Process'). The latter presumes that the requirements have already been identified and specified. Nevertheless, the activities of the requirements workflow are listed in order to understand which outputs are used by the subsequent workflows (Nicola et al., 2009, p. 260):

- "determine the domain of interest and the scope"
- "define business purpose"
- "writing one or more storyboards"
- "creating an application lexicon"
- "identifying the competency questions"
- "modeling the related use cases"

The relevant outputs are the *application lexicon* (AL), the *competency questions* (CQs) and the prioritized *use cases* (UCs) (Cf. Nicola et al., 2009, p. 262).

The AL is a set of terms which are "pertain to a given application […] as stated by a community of application experts" (Nicola et al., 2009, p. 262). It is created "by collecting the terminology from DE and application-specific documents" and by extracting terminology from the previously created storyboards (Nicola et al., 2009, p. 261). The extraction of terms from documents can be supported by automatic tools (Cf. Nicola et al., 2009, p. 261).

CQs are, as already described in chapter '2.4 Process', "questions at a conceptual level an ontology must be able to answer" (Nicola et al., 2009, p. 261). They are addressed "by using use-case models" each of which contains "a number of use cases that serve as a basis to specify the expected use of the ontology" (Nicola et al., 2009, p. 262).

According to UPON, "use cases correspond to knowledge paths through the ontology, to be followed for achieving business operations and answering CQs. Use cases will be detailed during the analysis and design workflows" (Nicola et al., 2009, p. 262). Figure 23 shows an excerpt of such a use case model.

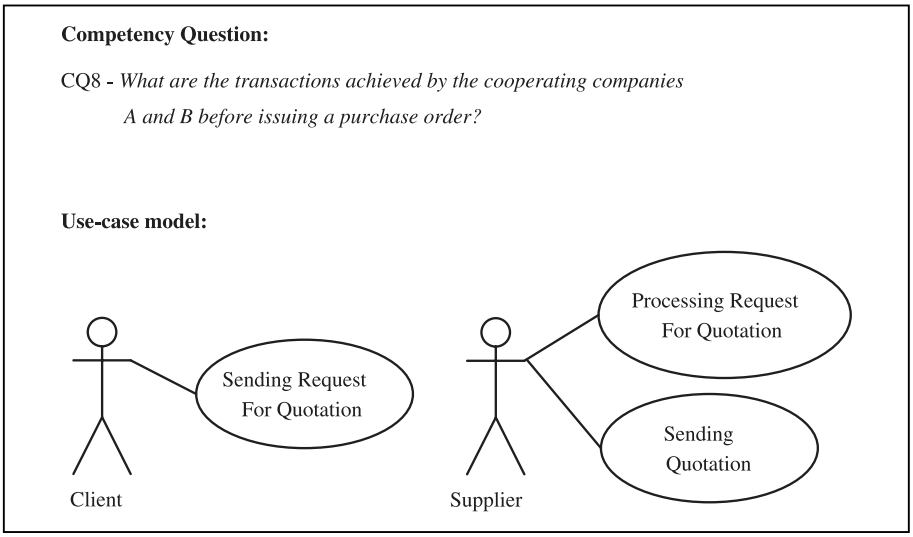

Figure 123: A competency question and an excerpt of the corresponding use-case model (Nicola et al., 2009, p. 263)

Analysis workflow

The second workflow concerns the conceptual analysis. In the course of this, the "AL is enriched through a more general *domain lexicon* (DL)" (Nicola et al., 2009, p. 262) which is a set of terms "validated by a community of domain experts" (Nicola et al., 2009, p. 265). The creation of the DL is described as follows:

> "*The DL is built by gathering the terminology used in the domain of interest, mainly extracted by analysing existing documental resources, such as reports, technical manuals, standards, glossaries, thesauri, legacy computational lexicons, and available ontologies. This step, like in the case of the AL, can be supported by automatic tools for text mining.*" (Nicola et al., 2009, p. 262)

In the next step, a *reference lexicon* (RL) is created "by selectively merging the AL (from application DEs) and the DL (from existing external resources)" (Nicola et al., 2009, p. 263). The activity of building the RL is illustrated in figure 24 and is described by the authors as follows:

> "*During the merge of the two lexicons, the terms are grouped into three major areas: one intersection area and two disjoint areas, application specific and domain specific […]. To build the RL the following 'inclusion policy' is used: the RL should include all the terms coming from the intersection area and, after the users and DEs approval, some terms belonging to the disjoint areas. Therefore the intersection area will be extended, on the one hand, with domain terms, considered useful for a better specification of the application at hand, and, on the other hand, with a part of the remaining application*

> *terms that are considered relevant, even if not extensively used by other applications. The output is a RL.*" (Nicola et al., 2009, p. 263)

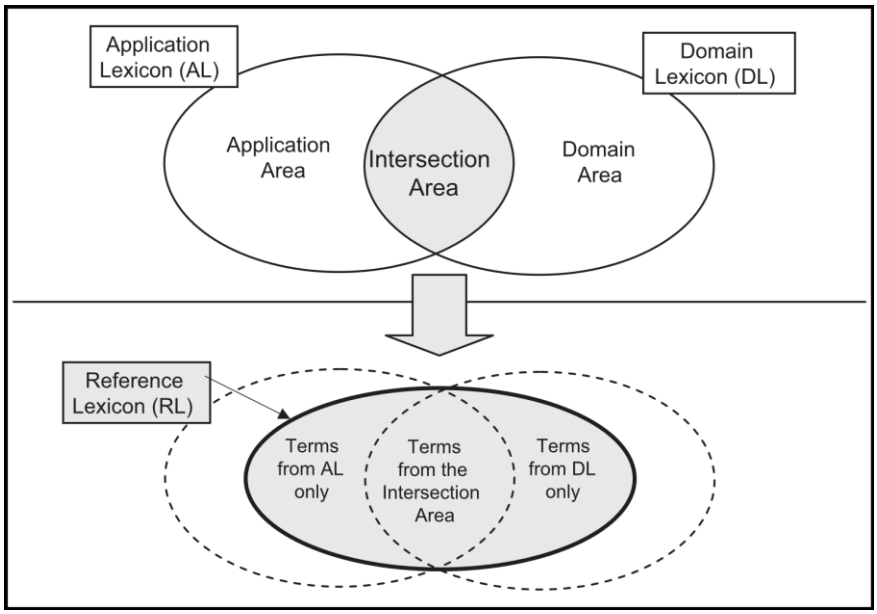

Figure 124: Activity of reference lexicon building (Nicola et al., 2009, p. 264)

The next activity is to model the application scenario using UML diagrams as shown in figure 25 below.

> "*The goal of this activity is to model the application scenario, adding to the use-case diagrams, drawn in the Requirements Workflow, the activity and class diagrams. UML diagrams represent a model of the application and will be used for the validation of the ontology. All classes, actors, and activities modeled in UML must have a corresponding concept in the ontology.*" (Nicola et al., 2009, pp. 263–264)

The last activity of the analysis workflow is the creation of a *reference glossary* (RG). It is "built by using the RL and by adding informal definitions (i.e., natural language sentences) to the terms" (Nicola et al., 2009, p. 264). Each term can be assigned one or more definitions which "should be selected from knowledgeable sources and agreed among DEs and users" (Nicola et al., 2009, p. 264).

Figure 26 below shows the analysis workflow as well as the associated inputs and outputs. Furthermore, it illustrates the extent to which DEs and KEs are involved throughout the workflow.

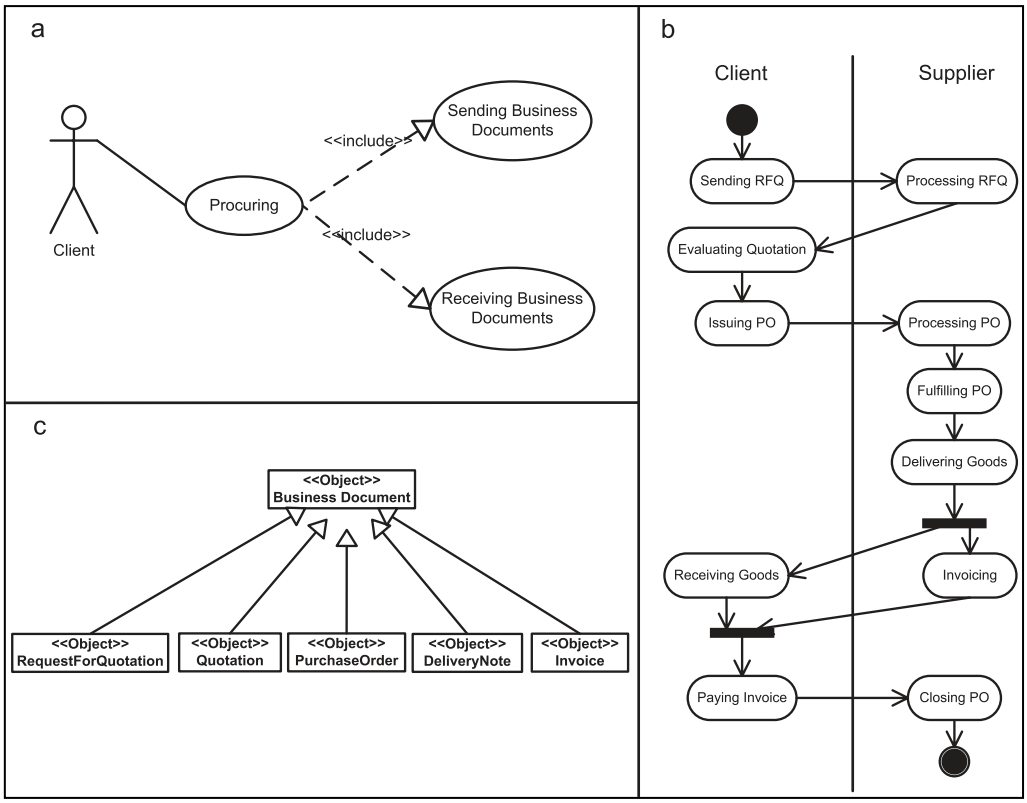

Figure 125: An excerpt of the UML diagrams of the application scenario: (a) use-case diagram; (b) activity diagram; and (c) class diagram (Nicola et al., 2009, p. 265)

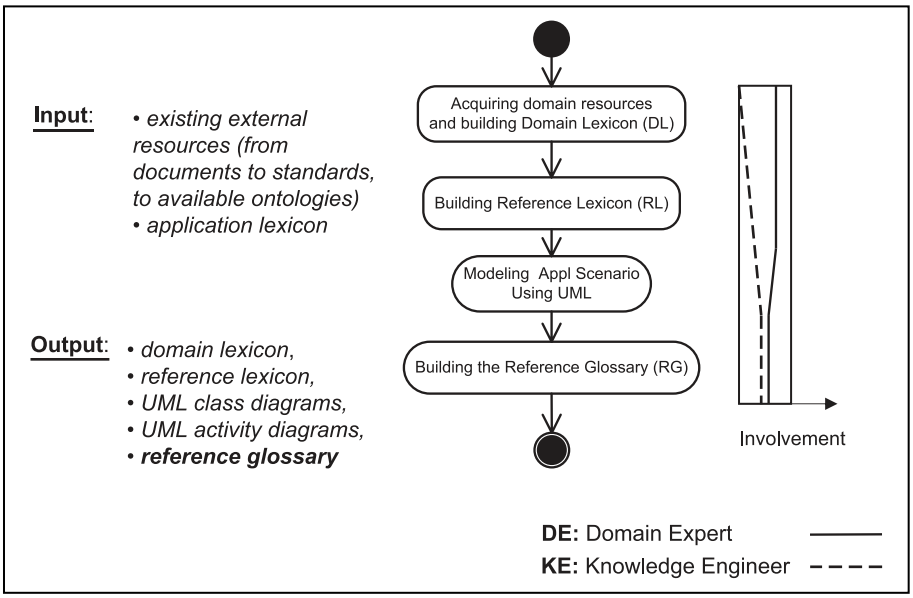

Figure 126: An overview of the analysis workflow (Nicola et al., 2009, p. 263)

Design workflow

The objective of the third workflow is "to give an ontological structure to the set of glossary entries gathered in the reference glossary" (Nicola et al., 2009, p. 266). The terms are organized "according to conceptual hierarchies" and structured "with attributes and axioms" (Nicola et al., 2009, p. 266). In other words, concepts as well as relationships between them are identified.

The design workflow is carried out according to OPAL (Cf. Nicola et al., 2009, p. 266). The *Object, Process, Actor modeling Language* (OPAL) "is an ontology modelling framework aimed at supporting business experts in building an ontology. [...] OPAL design patterns are formally defined by using OWL, allowing for a compatibility with this popular ontology language" (D'Antonio et al., 2007, p. 346).

In the first activity of the design workflow, *modeling concepts*, each concept is assigned a category (Cf. Nicola et al., 2009, p. 266). The used concept categories are adopted from OPAL, where they are referred to as *kinds* (Cf. D'Antonio et al., 2007, p. 348). The primary concept categories which are inspired by UML modeling constructs are: *business actor*, *business object* and *business process* (Cf. Nicola et al., 2009, p. 266). In addition to the primary categories, there are two complementary categories: *message* and *attribute* (Cf. Nicola et al., 2009, p. 267). They are described as follows:

> "*A message represents the information exchanged during an interaction (e.g., request, response) between processes […] A message is characterized by a content that is typically a BOD [business object document] (e.g., a RFQ-message, carrying a request for quotation). […] Attributes characterize the information structure of a concept. In OPAL there are atomic attributes, modeling elementary information (e.g., street name), and complex attributes, modeling structured information (e.g., address). Essentially, a complex attribute is defined as an aggregation of lower level complex and/or atomic attributes.*" (Nicola et al., 2009, p. 267)

The second activity of the design workflows is *modeling concept hierarchies and domain-specific relationships*. First, the concepts are organized "in a taxonomic hierarchy according to the *generalization* (i.e., IsA) relation" (Nicola et al., 2009, pp. 267–268). Then this taxonomy "can be extended with other relations, i.e., *part-of* and domain-specific relationships" (Nicola et al., 2009, p. 268). The output of the design workflow is a *semantic network* (SN) which the authors consider "as the informal counterpart of the ontology that will be produced in the next workflow" (Nicola et al., 2009, p. 268). The SN "is represented according to UML class diagram, in particular using *generalization* (IsA), *aggregation* (part-of) and *association*" (Nicola et

al., 2009, p. 268). Figure 27 shows an example given by the authors for the representation of ontological relations using UML. The complete design workflow including the involvement of KEs and DEs is illustrated in figure 28.

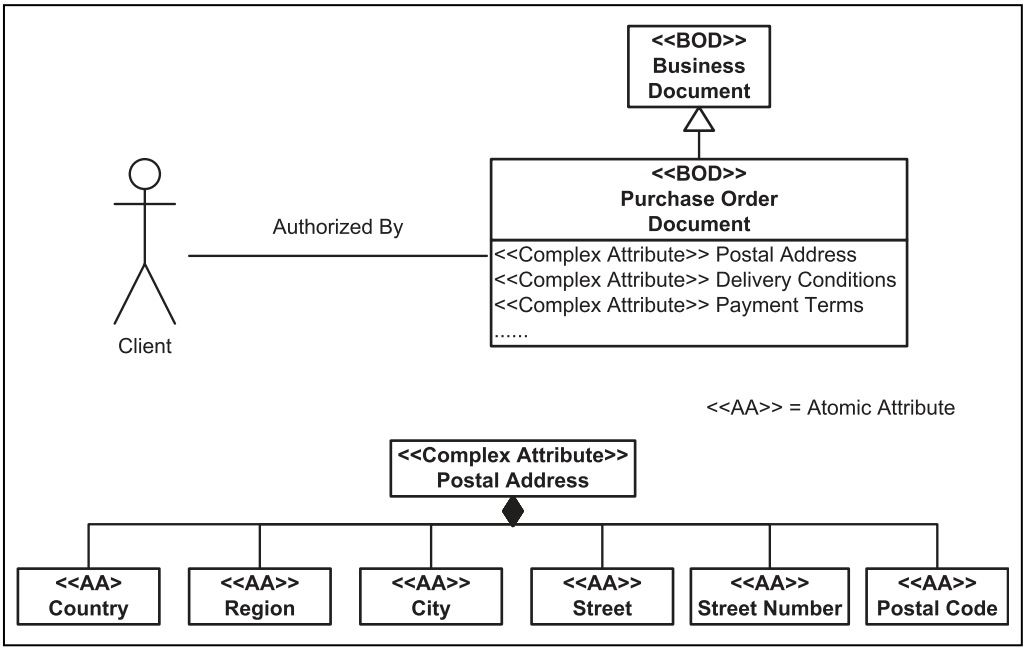

Figure 127: Examples of ontological relationships in the eProcurement application (Nicola et al., 2009, p. 269)

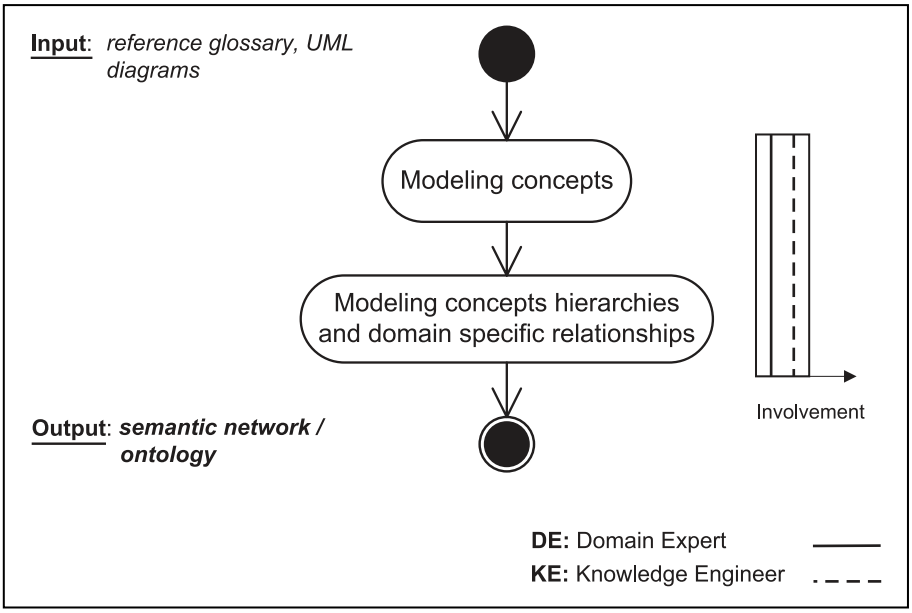

Figure 128: An overview of the design workflow (Nicola et al., 2009, p. 267)

Implementation workflow

The objective of the fourth workflow is "to encode the ontology in a rigorous, formal language" (Nicola et al., 2009, p. 268). Figure 29 shows the implementation workflow in the course of which the informal SN is represented as formal ontology as defined in chapter '2.3 Ontology'.

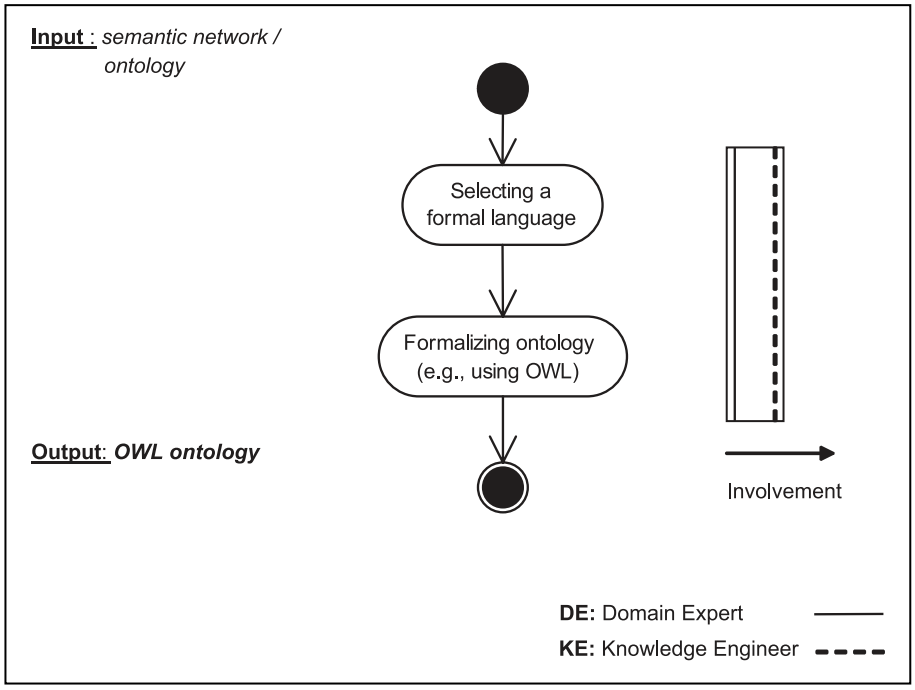

Figure 129: An overview of the implementation workflow (Nicola et al., 2009, p. 270)

Test workflow

The fifth and last workflow concerns the evaluation of the ontology. To do this, two of the four characteristics proposed by BURTON-JONES, STOREY, SUGUMARAN and AHLUWALIA for assessing the quality of ontologies (Cf. Burton-Jones et al., 2005) are considered:

> "*The test workflow […] is conceived to verify the semantic and pragmatic quality of the ontology, since syntactic quality is checked in the previous workflow and social quality can be checked only after its publication. In particular, syntactic quality is guaranteed during the OWL coding and social quality is assured by the interaction with different teams of KEs and cross-DEs.*" (Nicola et al., 2009, pp. 269–271)

The *semantic quality* "is verified by checking the consistency of the ontology, achieved by using a reasoner" (Nicola et al., 2009, p. 271). *Pragmatic quality* refers to three aspects: *fidelity*, *relevance* and *completeness* (Cf. Nicola et al., 2009, p. 272). They are addressed as follows:

> "*Fidelity can be measured by checking if the claims an ontology makes are true in the target domain. This task can be achieved, for instance, by verifying the references to the sources used in the descriptions of terms. […] Relevance is checked in conjunction with completeness, verifying the correct implementation of ontology's requirements, gathered in the first workflow.*" (Nicola et al., 2009, p. 272)

With regard to pragmatic quality, the test workflow proposes two tests, one concerning coverage and the other competency questions (see figure 30). The coverage can be verified by asking a DE "to semantically annotate the UML diagrams, modeling a software application, with the ontology concepts" (Nicola et al., 2009, p. 272). For the CQs, it is tested whether it is possible to "answer them by using the ontology content" (Nicola et al., 2009, p. 272). In other words, "an ontology is complete if the objectives, defined in the requirements workflow, are reached" (Nicola et al., 2009, p. 272).

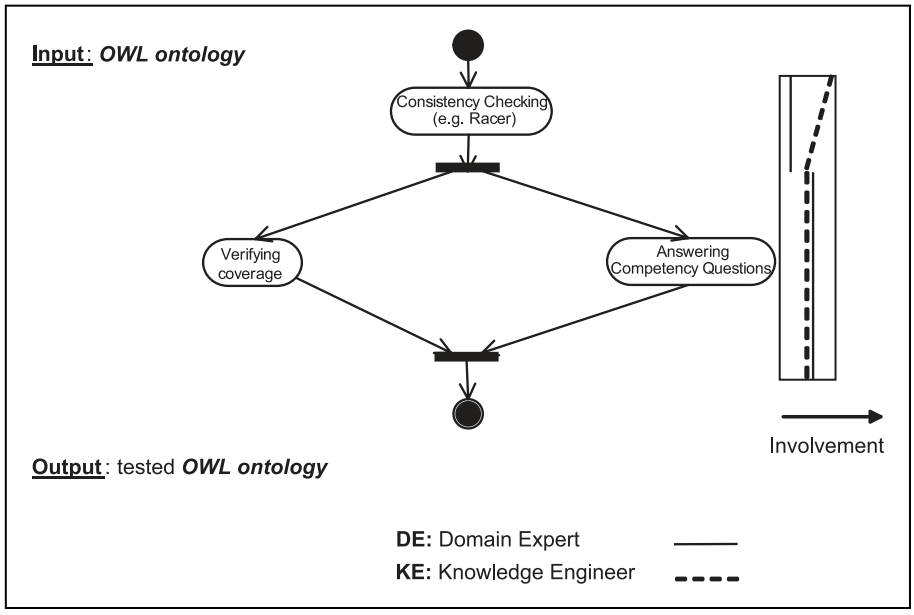

Figure 130: An overview of the test workflow (Nicola et al., 2009, p. 271)

Appendix D: Alternative documentary reference units

In the following, two alternative approaches to documentation are briefly outlined, which are less suitable than *concept-oriented* documentation.

a) Using an *individual database table* as documentary reference unit instead of an individual concept. For a given database table, it is documented which rows, columns or values correspond to which concepts. This can save time if the bottom-up approach is applied. Such a documentation concept is not suitable for other approaches such as the top-down approach. Data is not the starting point there.

b) Using a *delimited subdomain* or *predefined set of concepts* as documentary reference unit instead of an individual concept. Concepts that are known from the outset to be directly related are documented as a set. This can save time if the top-down or middle-out approach is applied. This is not possible when using the bottom-up approach, as the data are usually not organized according to their meaning. For example, database tables often only seem to represent direct semantic relationships. In the worst case, data resulting from text mining or similar processes are arranged in such a way that subdomains cannot be identified at the beginning.

References

Bawden, David; Robinson, Lyn (2012) *Introduction to information science*, London, Facet.

Baxter, Kathy; Courage, Catherine; Caine, Kelly (2015) *Understanding your users: A practical guide to user research methods*, 2nd edn, Amsterdam, Elsevier/Morgan Kaufmann.

Bayer (2018) *Internal documents* (intranet).

Bayer AG (ed) (2017) *Names | Figures | Facts 2017/2018* [Online]. Available at https://www.bayer.com/en/names-figures-facts-2017-2018.pdfx (Accessed 2019-02-18).

Berners-Lee, Tim (2006) *Design Issues: Linked Data* [Online]. Available at https://www.w3.org/DesignIssues/LinkedData.html (Accessed 2019-02-18).

Borst, Pim; Akkermans, Hans; Top, Jan (1997) 'Engineering ontologies', *International Journal of Human-Computer Studies*, vol. 46, 2-3, pp. 365–406. DOI: 10.1006/ijhc.1996.0096.

Borst, Willem N. (1997) *Construction of engineering ontologies for knowledge sharing and reuse*, PhD Thesis, Enschede, NL, University of Twente.

Brickley, Dan; Guha, R. V. (2014) *RDF Schema 1.1* (W3C Recommendation 25 February 2014) [Online], W3C. Available at https://www.w3.org/TR/rdf-schema/ (Accessed 2019-02-18).

Burton-Jones, Andrew; Storey, Veda C.; Sugumaran, Vijayan; Ahluwalia, Punit (2005) 'A semiotic metrics suite for assessing the quality of ontologies', *Data & Knowledge Engineering*, vol. 55, no. 1, pp. 84–102. DOI: 10.1016/j.datak.2004.11.010.

Carnegie Bosch Institute (1995) *Knowledge in international corporations – Outline of research area*, Pittsburg, CBI.

Comité Européen de Normalisation (2004) *European Guide to good Practice in Knowledge Management: Part 5: KM Terminology* (CEN Workshop Agreement CWA 14924-1).

Cooper, Alan (1999) *The inmates are running the asylum: Why high-tech products drive us crazy and how to restore the sanity*, Indianapolis, Ind., SAMS.

Corcho, Oscar; Fernández-López, Mariano; Gómez-Pérez, Asunción (2003) 'Methodologies, tools and languages for building ontologies. Where is their meeting point?', *Data & Knowledge Engineering*, vol. 46, no. 1, pp. 41–64. DOI: 10.1016/S0169-023X(02)00195-7.

D'Antonio, F.; Missikoff, M.; Taglino, F. (2007) 'Formalizing the OPAL eBusiness ontology design patterns with OWL', in Gonçalves, Ricardo J.; Mertins, Kai; Müller, Jörg P.; Zelm, Martin (eds) *Enterprise Interoperability II: New Challenges and Approaches*, London, Springer-Verlag London Limited, pp. 345–356.

Directorate-General for Research and Innovation (2018a) *Prompting an EOSC in practice: Final report and recommendations of the Commission 2nd High Level Expert Group on the European Open Science Cloud (EOSC)* [Online], European Commission.

Directorate-General for Research and Innovation (2018b) *Turning FAIR data into reality: Final report and action plan from the European Commission expert group on FAIR data* [Online], European Commission.

ELIXIR (2017a) *About us* [Online]. Available at https://www.elixir-europe.org/about-us/ (Accessed 2019-02-18).

ELIXIR (2017b) *Interoperability Platform* [Online]. Available at https://www.elixir-europe.org/platforms/interoperability/ (Accessed 2019-02-18).

Fernández, Mariano; Gómez-Pérez, Asunción; Juristo, Natalia (1997) 'METHONTOLOGY: From Ontological Art Towards Ontological Engineering', in Farquhar, Adam (ed) *Ontological engineering: Papers from the 1997 AAAI Symposium, March 24-26, Stanford, California*, Menlo Park, Calif., AAAI Press, pp. 33–40.

Fernández-López, Mariano (1999) 'Overview of Methodologies for Building Ontologies', in *Proceedings of the IJCAI-99 Workshop on ontologies and problem-solving methods: Lessons learned and future trends*. IJCAI-99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends. Stockholm, Sweden, August 2, 1999.

Fernández-López, Mariano; Gómez-Pérez, Asunción (2002) 'Overview and analysis of methodologies for building ontologies', *The Knowledge Engineering Review*, vol. 17, no. 02. DOI: 10.1017/S0269888902000462.

Forbes, David E.; Wongthongtham, Pornpit; Terblanche, Chamonix; Pakdeetrakulwong, Udsanee (2018) *Ontology Engineering Applications in Healthcare and Workforce Management Systems*, Cham, Springer International Publishing.

Frické, Martin (2009) 'The knowledge pyramid: a critique of the DIKW hierarchy', *Journal of Information Science*, vol. 35, no. 2, pp. 131–142. DOI: 10.1177/0165551508094050.

Gangemi, Aldo; Guarino, Nicola; Masolo, Claudio; Oltramari, Alessandro; Schneider, Luc (2002) 'Sweetening Ontologies with DOLCE', in *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*. 13th International Conference, EKAW 2002. Sigüenza, Spain, October 1-4, 2002. Berlin, Heidelberg, Springer, pp. 166–181. DOI: 10.1007/3-540-45810-7_18.

Gaus, Wilhelm (2005) *Dokumentations- und Ordnungslehre: Theorie und Praxis des Information Retrieval*, 5th edn, Berlin, Heidelberg, Springer-Verlag Berlin Heidelberg.

GO FAIR (2018) *Strategy* [Online]. Available at https://www.go-fair.org/go-fair-initiative/strategy/ (Accessed 2019-02-18).

Gómez-Pérez, Asunción; Fernández-López, Mariano; Corcho, Oscar (2004) *Ontological Engineering: With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web*, London, Springer-Verlag London Limited.

Goodman, Elizabeth; Kuniavsky, Mike; Moed, Andrea (2012) *Observing the user experience: A practitioner's guide to user research*, 2nd edn, Amsterdam, Elsevier/Morgan Kaufmann.

Gruber, Thomas R. (1993) 'A translation approach to portable ontology specifications', *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220. DOI: 10.1006/knac.1993.1008.

Guarino, Nicola (1997) 'Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration', in *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*. International Summer School, SCIE-97. Frascati, Italy, July 14-18, 1997. Berlin, Heidelberg, Springer, pp. 139–170.

Guarino, Nicola (1999) 'The Role of Identity Conditions in Ontology Design', in *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science*. International Conference COSIT'99. Stade, Germany, August 25-29, 1999. Berlin, Heidelberg, Springer, pp. 221–234. DOI: 10.1007/3-540-48384-5_15.

Guizzardi, Giancarlo; Herre, Heinrich; Wagner, Gerd (2002) 'Towards Ontological Foundations for UML Conceptual Models', in Meersman, Robert (ed) *On the move to meaningful internet systems 2002: Confederated international conferences CoopIS*, DOA, and ODBASE 2002, Berlin, Springer, pp. 1100–1117.

Heinrich, Lutz J.; Stelzer, Dirk; Riedl, René (2014) *Informationsmanagement: Grundlagen, Aufgaben, Methoden*, 11th edn, Berlin/Boston, De Gruyter.

HL7.org (2017) *FHIR Release 3* (STU): FHIR Linked Data Module [Online]. Available at https://www.hl7.org/fhir/linked-data-module.html (Accessed 2019-02-18).

Hobohm, Hans-Christoph (2016) 'DIKW-Hierarchie', in Umlauf, Konrad (ed) *Grundwissen Medien, Information, Bibliothek*, Stuttgart, Anton Hiersemann Verlag, pp. 9–10.

Ingold, M. (2016) 'Information', in Umlauf, Konrad (ed) *Grundwissen Medien, Information, Bibliothek, Stuttgart*, Anton Hiersemann Verlag, pp. 5–7.

Iqbal, Rizwan; Murad, Masrah A. A.; Mustapha, Aida; Sharef, Nurfadhlina M. (2013) 'An Analysis of Ontology Engineering Methodologies: A Literature Review', *Research Journal of Applied Sciences, Engineering and Technology*, vol. 6, no. 16, pp. 2993–3000. DOI: 10.19026/rjaset.6.3684.

Jacobson, Ivar; Booch, Grady; Rumbaugh, James (1999) *The Unified Software Development Process*, Reading, Mass, Addison-Wesley.

Kent Innovation Centre (2017) *PhUSE Working Groups: Linked Data & Graph Databases* [Online]. Available at https://www.phuse.eu/linked-data-graph-databases/ (Accessed 2019-02-18).

Landes, Dieter; Schneider, Kurt; Houdek, Frank (1999) 'Organizational learning and experience documentation in industrial software projects', *International Journal of Human-Computer Studies*, vol. 51, no. 3, pp. 643–661. DOI: 10.1006/ijhc.1999.0280.

Lehmann, Alexander (2018) 'A Documentation Approach for Higher Education', in *Proceedings of 2018 IEEE Global Engineering Education Conference (EDUCON)*. 2018 IEEE Global Engineering Education Conference (EDUCON). Santa Cruz de Tenerife, Canary Islands, Spain, April 17-20, 2018. Piscataway, NJ, IEEE, pp. 43–50. DOI: 10.1109/EDUCON.2018.8363207.

Levesque, Hector J.; Brachman, Ronald J. (1985) 'A Fundamental Tradeoff in Knowledge Representation and Reasoning (Revised Version)', in Brachman, Ronald J.; Levesque, Hector J. (eds) *Readings in Knowledge Representation*, San Francisco, California, Morgan Kaufmann Publishers, pp. 41–70.

Mihindukulasooriya, Nandana; Poveda-Villalón, María; Garcí-Castro, Rául; Gómez-Pérez, Asunción (2017) 'Collaborative Ontology Evolution and Data Quality - An Empirical Analysis', in *OWL: Experiences and Directions – Reasoner Evaluation*. 13th International Workshop, OWLED 2016 and 5th International Workshop, ORE 2016. Bologna, Italy, November 20, 2016. Cham, Springer International Publishing, pp. 95–114. DOI: 10.1007/978-3-319-54627-8_8.

NeOn Project (n.y.a) *About NeOn* [Online]. Available at http://neon-project.org/nw/About_NeOn.html (Accessed 2019-02-18).

NeOn Project (n.y.b) *Partners* [Online]. Available at http://neon-project.org/nw/Partners.html (Accessed 2019-02-18).

Nicola, Antonio de; Missikoff, Michele; Navigli, Roberto (2005) 'A Proposal for a Unified Process for Ontology Building: UPON', in *Database and Expert Systems Applications*. 16th International Conference, DEXA 2005. Copenhagen, Denmark, August 22-26, 2005. Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 655–664. DOI: 10.1007/11546924_64.

Nicola, Antonio de; Missikoff, Michele; Navigli, Roberto (2009) 'A software engineering approach to ontology building', *Information Systems*, vol. 34, no. 2, pp. 258–275. DOI: 10.1016/j.is.2008.07.002.

Niklas Blomberg (2017) *ELIXIR position paper on FAIR data management in the life sciences*, ELIXIR Consortium.

Nonaka, Ikujiro; Takeuchi, Hirotaka (1995) *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*, New York, Oxford University Press.

North, Klaus; Kumta, Gita (2018) *Knowledge Management: Value Creation Through Organizational Learning*, 2nd edn, Cham, Springer International Publishing.

Office of the Commissioner (2018) Study Data Standards: Janus [Online], U.S. Food and Drug Administration. Available at https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm155327.htm (Accessed 2019-02-18).

Paschke, Adrian; Burger, Albert; Romano, Paolo; Marshall, M. S.; Splendiani, Andrea (2018) *International SWAT4LS Workshop: Semantic Web Applications and Tools for Healthcare and Life Sciences: About* [Online]. Available at http://www.swat4ls.org/about/ (Accessed 2019-02-18).

Peters, Isabella; Weller, Katrin (2008) 'Paradigmatic and Syntagmatic Relations in Knowledge Organization Systems', *Information - Wissenschaft & Praxis*, vol. 59, no. 2, pp. 100–107.

Polanyi, Michael (1966) *The tacit dimension*, Chicago, University of Chicago Press.

Probst, Gilbert; Raub, Steffen; Romhardt, Kai (2010) *Wissen managen: Wie Unternehmen ihre wertvollste Ressource optimal nutzen*, 6th edn, Wiesbaden, Gabler Verlag / GWV Fachverlage GmbH Wiesbaden.

Pruitt, John S.; Adlin, Tamara, eds. (2006) *The persona lifecycle: Keeping people in mind throughout product design*, Amsterdam, Boston, Elsevier/Morgan Kaufmann.

Rowley, Jennifer (2007) 'The wisdom hierarchy: representations of the DIKW hierarchy', *Journal of Information Science*, vol. 33, no. 2, pp. 163–180. DOI: 10.1177/0165551506070706.

Schmidt, Daniela; Trojahn, Cassia; Vieira, Renata (2016) 'Analysing Top-level and Domain Ontology Alignments from Matching Systems', in *CEUR Workshop Proceedings: OM 2016*. 11th International Workshop on Ontology Matching. Kobe, Japan, 2016-10-18, pp. 13–24.

Schreiber, Guus; Raimond, Yves (2014) *RDF 1.1 Primer: W3C Working Group Note 24 June 2014* [Online], W3C. Available at https://www.w3.org/TR/rdf11-primer/ (Accessed 2019-02-18).

Schultz, Susan I.; Darrow, Jennifer J.; Kavanagh, Frank X.; Morse, Marjorie J. (1993) *The Digital Technical Documentation Handbook*, Burlington, MA, USA, Digital Press.

Semantic Web Health Care and Life Sciences (HCLS) Interest Group (2018) *Introduction* [Online]. Available at https://www.w3.org/blog/hcls/ (Accessed 2019-02-18).

Simperl, Elena; Luczak-Rösch, Markus (2014) 'Collaborative ontology engineering: a survey', *The Knowledge Engineering Review*, vol. 29, no. 01, pp. 101–131. DOI: 10.1017/S0269888913000192.

Software & Systems Engineering Standards Committee of the IEEE Computer Society (2010) *ISO/IEC/IEEE 24765:2010: Systems and software engineering — Vocabulary*.

Staab, Steffen; Studer, Rudi, eds. (2009) *Handbook on Ontologies*, 2nd edn, Berlin, Heidelberg, Springer Berlin Heidelberg.

Statista (2018) *Top 100 Companies: Pharma & Biotech* (study id 30864).

Stephens, Susie (2008) *New W3C Semantic Web for Health Care and Life Sciences Interest Group* [Online], Semantic Web Health Care and Life Sciences (HCLS) Interest Group. Available at https://www.w3.org/blog/hcls/2008/06/11/new_w3c_semantic_web_for_health_care_and_1/ (Accessed 2019-02-18).

Stuckenschmidt, Heiner (2011) *Ontologien: Konzepte, Technologien und Anwendungen*, 2nd edn, Berlin, Heidelberg, Springer Berlin Heidelberg.

Studer, Rudi; Benjamins, V. R.; Fensel, Dieter (1998) 'Knowledge Engineering: Principles and methods', *Data & Knowledge Engineering*, vol. 25, 1-2, pp. 161–197. DOI: 10.1016/S0169-023X(97)00056-6.

Suárez-Figueroa, Mari C. (2011) *NeOn Glossary* [Online]. Available at http://mayor2.dia.fi.upm.es/oeg-upm/files/pdf/NeOnGlossary.pdf (Accessed 2019-02-10).

Suárez-Figueroa, Mari C.; Brockmans, Saartje; Gangemi, Aldo; Gómez-Pérez, Asunción; Lehmann, Jos; Lewen, Holger; Presutti, Valentina; Sabou, Marta (2007) *D5.1.1 NeOn Modelling Components* (Class Deliverable: NEON EU-IST-2005-027595; Document Identifier: NEON/2007/D5.1.1/v1.1) [Online]. Available at http://neon-project.org/deliverables/WP5/NeOn_2007_D5.1.1v3.pdf (Accessed 2019-02-07).

Suárez-Figueroa, Mari C.; Gómez-Pérez, Asunción (2012) 'Ontology Requirements Specification', in Suárez-Figueroa, Mari C.; Gómez-Pérez, Asunción; Motta, Enrico; Gangemi, Aldo (eds) *Ontology Engineering in a Networked World*, Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 93–106.

Suárez-Figueroa, Mari C.; Gómez-Pérez, Asunción; Fernández-López, Mariano (2012) 'The NeOn Methodology for Ontology Engineering', in Suárez-Figueroa, Mari C.; Gómez-Pérez, Asunción; Motta, Enrico; Gangemi, Aldo (eds) *Ontology Engineering in a Networked World*, Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 9–34.

Suárez-Figueroa, Mari C.; Gómez-Pérez, Asunción; Fernández-López, Mariano (2015) 'The NeOn Methodology framework: A scenario-based methodology for ontology development', *Applied Ontology*, vol. 10, no. 2, pp. 107–145. DOI: 10.3233/AO-150145.

Suárez-Figueroa, Mari C.; Gómez-Pérez, Asunción; Motta, Enrico; Gangemi, Aldo, eds. (2012) *Ontology Engineering in a Networked World*, Berlin, Heidelberg, Springer Berlin Heidelberg.

Suárez-Figueroa, Mari C.; Gómez-Pérez, Asunción; Poveda, María; Ramos, Jose A.; Euzenat, Jérôme; Le Duc, Chan (2010) *D5.4.3. Revision and Extension of the NeOn Methodology for Building Contextualized Ontology Networks* (Class Deliverable: NEON EU-IST-2005-027595; Document Identifier: NEON/2010/D5.4.3/v1.0) [Online]. Available at http://neon-project.org/deliverables/WP5/NeOn_2010_D543.pdf (Accessed 2019-02-07).

Teymourian, Kia; Coskun, Gökhan; Paschke, Adrian (2010) 'Modular Upper-Level Ontologies for Semantic Complex Event Processing', in Kutz, Oliver; Hois, Joana (eds) *Modular ontologies: Proceedings of the fourth international workshop (WoMO 2010)*, Amsterdam, IOS Press, pp. 81–93.

U.S. Food and Drug Administration (2018) *Study Data Technical Conformance Guide* (Technical Specifications Document) [Online]. Available at https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf (Accessed 2019-02-18).

van Heijst, G.; Schreiber, A. T.; Wielinga, B. J. (1997) 'Using explicit ontologies in KBS development', *International Journal of Human-Computer Studies*, vol. 46, 2-3, pp. 183–292. DOI: 10.1006/ijhc.1996.0090.

Villazón-Terrazas, Boris; Gómez-Pérez, Asunción (2012) 'Reusing and Re-engineering Non-ontological Resources for Building Ontologies', in Suárez-Figueroa, Mari C.; Gómez-Pérez, Asunción; Motta, Enrico; Gangemi, Aldo (eds) *Ontology Engineering in a Networked World*, Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 107–145.

W3C (2015) *Semantic Web* [Online]. Available at https://www.w3.org/standards/semanticweb/ (Accessed 2019-02-18).

W3C (2018) *Semantic Web in Health Care and Life Sciences Community Group* [Online]. Available at https://www.w3.org/community/hclscg/ (Accessed 2019-02-18).

Weik, Martin H. (1977) *Standard dictionary of computers and information processing*, 2nd edn, Rochelle Park, N. J., Hayden.

Weller, Katrin (2014) 'B 6 Ontologien', in Kuhlen, Rainer; Semar, Wolfgang; Strauch, Dietmar (eds) *Grundlagen der praktischen Information und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und -praxis*, 6th edn, Berlin, Boston, DE GRUYTER SAUR.

Wersig, Gernot (1971) *Wersig: Ein Beitrag zur Orientierung der Informations- und Dokumentationswissenschaft*, München-Pullach und Berlin, Verlag Dokumenation.

Wersig, Gernot; Neveling, Ulrich (1976) *Terminology of documentation: A selection of 1,200 basic terms published in English, French, German, Russian and Spanish = Terminologie de la documentation = Terminologie der Dokumentation = Terminolgija v oblasti dokumentacii = Terminología de la documentación*, Paris, München, Unesco Press; Verl. Dokumentation.

Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, I. J. J.; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz B.; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J. G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; Hoen, Peter A. C. 't; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific data*, vol. 3. DOI: 10.1038/sdata.2016.18.

Wise, John; Barron, Alexandra G. de; Splendiani, Andrea; Balali-Mood, Beeta; Vasant, Drashtti; Little, Eric; Mellino, Gaspare; Harrow, Ian; Smith, Ian; Taubert, Jan; van Bochove, Kees; Romacker, Martin; Walgemoed, Peter; Jimenez, Rafael C.; Winnerberg, Rainer; Plasterer, Tom; Gupta, Vibhor; Hedley, Victoria (2019) 'Implementation and relevance of FAIR data principles in biopharmaceutical R&D', *Drug Discovery Today*. DOI: 10.1016/j.drudis.2019.01.008.

Yu, Liyang (2014) *A Developer's Guide to the Semantic Web*, 2nd edn, Berlin, Heidelberg, Springer Berlin Heidelberg.

Index

ABox, 23
Analysis workflow, 81
Application lexicon, 81
Application ontology, 27
Bottom-up approach, 29
Class, 23
Classification, 22
Codification strategy, 18
Combination, 18
Competency questions, 31
Concept, 23
Conversion of knowledge, 17
Data, 12
Design workflow, 82
DIKW hierarchy, 12
Dimension-specific module, 104
Documentary reference unit, 99
Documentary unit, 99
Documentation, 8
Documentation concept, 9
Documentation entry, 100
Documentation module, 102
Documentation object, 99
Domain expert, 30
Domain knowledge dimension, 103
Domain lexicon, 81
Domain ontology, 27
Engineering knowledge dimension, 103
Explicit knowledge, 16
Externalization, 18
Fact, 23
Formal, 11
Forms of knowledge representation, 17
General module, 104
Graph, 25
Implementation workflow, 82
Informal, 10
Information, 13
Instance, 23
Insufficient documentation, 4
Internalization, 19
Knowledge, 14
Knowledge engineer, 30
Knowledge spiral, 19
Language tag, 108
Life-cycle models, 94
Management knowledge dimension, 103
Methodology, 28
Middle-out approach, 29
Missing documentation, 4
Modes of knowledge conversion, 17
Module extension, 104
NeOn glossary, 93
N-Triples, 25
Ontology, 21
Ontology engineering methodology, 28
Ontology language, 24
Ontology libraries, 27
Ontology model, 23
Ontology modules, 27
Ontology specification, 32
Ontology types, 26
Personalization strategy, 18
Primary documentation unit, 100
Property, 23
RDF Schema, 26
Reference glossary, 81
Reference lexicon, 81
Requirements workflow, 81
Resource Description Framework, 25
Restriction, 23
Scenarios, 92
Semantic network, 82
Semi-formal, 10
Socialization, 17
Statement, 108
Tacit knowledge, 16
Task ontology, 27
Taxonomy, 22
TBox, 23
Test workflow, 82
Thesaurus, 22
Top-down approach, 29
Top-level ontology, 27
Triple, 25
Turtle, 25

The text of the following declaration of authorship (English and German version) is taken over in its entirety from:

Sven Kleinknecht (2013) *Ehrenwörtliche Erklärung deutsch und englisch* [Online], Technische Universität München. Available at https://www.ent.wi.tum.de/fileadmin/w00bcx/www/Ehrenwoertliche_Erklaerung_deutsch_und_englisch.pdf (Accessed 2019-02-18).

Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the thesis as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future theses submitted. Further rights of reproduction and usage, however, are not granted here.

This paper was not previously presented to another examination board and has not been published.

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit selbständig angefertigt habe. Die aus fremden Quellen direkt und indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Ich weiß, dass die Arbeit in digitalisierter Form daraufhin überprüft werden kann, ob unerlaubte Hilfsmittel verwendet wurden und ob es sich – insgesamt oder in Teilen – um ein Plagiat handelt. Zum Vergleich meiner Arbeit mit existierenden Quellen darf sie in eine Datenbank eingestellt werden und nach der Überprüfung zum Vergleich mit künftig eingehenden Arbeiten dort verbleiben. Weitere Vervielfältigungs- und Verwertungsrechte werden dadurch nicht eingeräumt.

Die Arbeit wurde weder einer anderen Prüfungsbehörde vorgelegt noch veröffentlicht.

Marius Michaelis — first and last name

Potsdam, 18.02.2019 — city, date and signature