On the Semantification of 5-Star Technical Documentation

Sebastian Furth[1] and Joachim Baumeister[1,2]

[1]denkbares GmbH, Friedrich-Bergius-Ring 15, 97076 Würzburg, Germany
[2]University of Würzburg, Institute of Computer Science,
Am Hubland, 97074 Würzburg, Germany
{sebastian.furth,joachim.baumeister}@denkbares.com

Abstract. Technical documentation is a special purpose content describing machines and plants with high complexity. The documentation covers operation, maintenance and repair of the technical artifacts. The high complexity of the machines yields a voluminous documentation, where it increasingly becomes difficult to find the relevant information for a given problem. The paper discusses the use of semantic technologies to organize the documentation on a syntactic and semantic level. Also, a scheme for the assessment of the maturity of existing documentation is proposed, that simplifies the application of semantic technologies.

Keywords: Semantic Publishing, Ontology Engineering, Information Extraction

1 Introduction

The complexity of machines has grown dramatically in the past years. As a consequence, the technical documentation became a fundamental source for service technicians in their daily work. Service technicians need fast and focused access methods to handle the massive volumes of technical documents. For this reason semantic search emerged as the new system paradigm for the presentation of technical documentation. However, the existing corpora are usually not semantically prepared. The best existing solutions may give access to dedicated sections, while the information relevant for the service technician remains concealed. In this paper we present a novel ontological representation for technical documents that combines structural and rhetorical elements to enable direct access to *Core Documentation Entities*. We additionally introduce a maturity schema that allows the assessment of existing technical documentation with respect to these Core Documentation Entities.

The remainder of this paper is structured as follows. In Section 2 we first give a general introduction to technical documentation and present a novel maturity

 In: R. Bergmann, S. Gorg, G. Muller (Eds.): Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB. Trier, Germany, 7.-9. October 2015, published at http://ceur-ws.org

scheme for the assessment of their quality. The maturity scheme relies on semantic technologies, hence we present ontologies for the ontological description of technical documents in Section 3. Section 4 shows the practical applicability of the presented ontologies. We conclude with a summary and a statement of future research directions.

2 5-Star Technical Documentation

In this section we introduce the domain of technical documentation as a special type of textual and multimedia resources. We motivate that the semantification enables reuse and integration of the resources for various applications.

2.1 Uses of Technical Documentation

Builders of machinery and plants provide technical documentation to support the service technician to ensure the save operation and maintenance of their products. Typically, the documentation is created to efficiently support the following tasks:

1. Operation of the machine
2. Maintenance of the machine
3. Localization of specific components
4. Diagnosis of problems
5. Repair of a localized damage

Historically, the documentation is partitioned into a number of books supporting the particular tasks by technical descriptions:

User Manual describes the operation of the machine, i.e., how to activate and perform the machine functions.

Repair Manual shows the replacement and maintenance of specific components of the machine.

Technical Functions and Diagnosis describe the logical connections and relations of components within the machine, in order to support the diagnosis of observed faults. Typical examples are electrical and hydraulic wiring diagrams.

Spare Parts provide a detailed view of parts located in particular components. Service technicians locate parts by using this documentation, but also to order new parts in exchange for faulty parts.

In the past, the documentation was printed on paper. With the increasing complexity of the machines many vendors switched to electronic versions of the books in recent years (PDF and HTML). For instance, the documentation for a full-featured harvesting machine or other special purpose vehicles comprise about 10,000 pages. With the electronic availability the metaphor of a single 'book' is not necessary anymore.

In recent years, semantic technologies emerged to (re-)organize the structuring of documents in corporate environments [3]. Hence, advanced methods are emerging for searching for relevant chapters and navigating between information units. Applications within the infrastructure of the technical documentation were improved, such as automated term extraction and general information extraction tasks. More importantly, interesting end-user applications become possible such as *semantic search* [7, 11] and *semantic assistants* [15].

However, existing documentation data does not necessarily fulfill all requirements for semantic applications. The quality state of existing documentation data can vary massively, ranging from scanned image PDF documents to products of XML content management systems. In practice, it is helpful to provide a classification schema to assess the maturity level of the existing documentation. This schema also gives advices for improving the current state of documentation.

2.2 Towards the quality of Technical Documentation

We introduce a maturity schema for the assessment of technical documentation data. The schema lists a number of quality criteria building on each other. For each criteria we give one star; that way the maturity of documentation data can range from one star to five stars. This schema is inspired by the idea of evaluating the quality of data in the linked open data cloud [1, 9], and was adapted to the needs of technical documentation. The aims of the schemes, however, are identical: First, users should obtain an intuitive impression about the maturity of their data; second, users should get motivated to increase the stars of their data by adding more semantics. The schema for *5-Star Technical Documentation* data is depicted in Figure 1.

Fig. 1. The levels of the 5-stars maturity schema for technical documentation.

The first star is given, when the documentation is accessible in an electronic format, for instance, as PDF or MS-Word. The documentation gets two stars, when it is accessible in a structured and non-proprietary format, e.g., XML,

SGML, or Markdown. Three stars are received for documentation that is accessible in a standardized format, e.g., DocBook XML or ASD S1000D[1]. Documentation with four stars provide URIs for all relevant elements of the content. That way, the book itself, the particular chapters, and paragraphs can be clearly named and thus can be linked by external applications. Five stars documentation adds semantics to the relevant elements by attaching meta-data to the elements that refers to concepts of an ontology. Using an ontology enables the automated interlinkage of document elements by using the same concepts of the ontology. Also external ontologies with similar semantics can by aligned to the used ontology. In the following, we discuss the use of open documentation standards and ontologies in order to receive the 5-stars level.

3 Ontologies for Technical Documentation

For the semantic representation of technical documentation we pick up the established idea from the semantic publishing community of the definition of OWL [8] or RDFS [14] vocabularies that describe certain aspects of the publishing domain. Such aspects typically comprise structural components (e.g. paragraphs, sections, sentences) and rhetorical elements (e.g. discourse elements / sections like "Motivation", "Problem Statement" or "Discussion"). Complementary ontologies often provide annotation vocabulary that allows the definition of additional meta data. In the following we first describe suitable vocabularies for the representation of structural and rhetorical aspects of a technical document. Building upon these vocabularies we introduce a novel ontology that exploits structural and rhetorical aspects to facilitate direct access to core documentation entities like component overviews or repair procedures. At this point the technical documentation already gets four out of five stars. The addition of annotation vocabularies completes the section with the achievement of 5-star technical documentation.

3.1 Structural Components

Considering only the pure structural composition of a document, the required vocabulary is rather independent of the underlying problem domain. The Document Ontology schema of the SALT ontology [6] or the pattern ontology [4] are popular examples for the description of (scientific) publications. However, for publications in the technical domain DocBook [12] is a de facto standard maintained by the Organization for the Advancement of Structured Information Standards (OASIS)[2]. Following the maturity schema introduced in Section 2.2 documents written according to this standard receive the 3-stars level. Thus we encourage the usage of a DocBook-like ontology for the structural description of technical documentation. Şah and Wade [13] proposed an ontology that covers a

[1] http://www.s1000d.org/
[2] https://www.oasis-open.org/

reasonable subset of the DocBook standard. Table 1 briefly introduces the most important elements of this ontology, e.g. `docbook:Book`, `docbook:Article`, `docbook:Chapter` or block elements like `docbook:Paragraph`, `docbook:Procedure` or `docbook:Figure`.

Element	Type	Description
`docbook:Book`	`Class`	Represents the top level element that has a number of sub-components like articles or chapters.
`docbook:Article` / `docbook:Chapter`	`Class`	Articles and chapters contain (sequences of) block elements.
`docbook:BlockElement`	`Class`	Block elements are typically used as atomic information units. Common examples that are available as subclasses are `docbook:Paragraph`, `docbook:Table`, `docbook:List`, `docbook:Procedure` or `docbook:Figure`
`dc:hasPart`	`Property`	Property from the Dublin Core ontology that connects instances of the DocBook classes

Table 1. Important elements of the DocBook ontology [13].

3.2 Rhetorical Components

In contrast to the structural organisation of a document the rhetorical ontology concentrates on modeling the rhetorical structures and elements of the document. A correspondence of structural components does not necessarily exist in the rhetorical organisation of the document. However, core rhetorical structures like safety instructions can often be linked explicity to particular structures like chapters, sections or paragraphs. For the representation of scientific articles the Rhetorical Ontology schema of the SALT ontology [6] or the Discourse Elements Ontology [2] provide appropriate vocabulary. Thus, rhetorical aspects like the motivation, background, methods etc. can be modeled as instances of respective classes. While the underlying idea also facilitates the rhetorical modeling of technical documentation the concrete classes do not fit the technical domain. For instance law requires technical documentation to follow a certain rhetorical organisation, e.g. safety notes need to preceed actual operation instructions. Thus it would be benefitial to semantically represent safety notes. Table 2 gives a non-exhaustive overview of common rhetorical elements in technical documents.

3.3 Core Documentation Entities = Structure + Rhetoric

The maturity schema introduced in Section 2.2 requires that relevant elements are identifiable by URIs. Representing the structural and rhetorical aspects of technical documentation is a considerable step in this direction. However, the most important aspects of technical documents are interweaved in these two

Element	Description
`rtc:Index`	Indices like table of contents, subject catalogs, list of abbreviations etc.
`rtc:GeneralInformation`	General aspects of the document or the machine in focus.
`rtc:SafetyInstruction`	Safety notes to be obtained while working with the machine.
`rtc:Description`	Information about specific components or functions.
`rtc:Operation`	Information about the usage of the machine, specific components or functions.
`rtc:Repair`	Repair procedures; important subclasses are `rtc:Assembly` and `rtc:Disassembly`
`rtc:Maintenance`	Information about maintenance works, schedules etc.
`rtc:Adjustment`	Information about necessary adjustments in specific situations.
`rtc:FaultIsolation`	Detailed troubleshooting information.
`rtc:Parts`	Spare part information.

Table 2. Common rhetorical components in technical documentation.

structures. The entropy of these aspectes is typically sufficient to satisfy an immediate information need. In the following we give excerpts of a novel ontology, that combines structural and rhetorical aspects in order to make these *Core Documentation Entities* easily accessible. A typical example for such an information need is a (dis-)assembly procedure. The corresponding information can be obtained by combining the rhetorical structure `rtc:Assembly` with the structual element `docbook:Procedure`:

$$\texttt{cde:AssemblyProcedure} \sqsubseteq \texttt{rtc:Assembly} \sqcap \texttt{docbook:Procedure}$$

Another example are component overviews that can typically be found in a section describing the machine or in the spare part information. Component overviews typically consist of an exploded-view drawing and an associated list of labels, product numbers etc.:

$$\begin{aligned}&\texttt{cde:ComponentOverview} \sqsubseteq \\ &\quad (\texttt{rtc:Description} \sqcup \texttt{rtc:Parts}) \sqcap \\ &\quad \exists\, (\texttt{dc:hasPart.docbook:Figure} \sqcap \texttt{dc:hasPart.docbook:List})\end{aligned}$$

3.4 Linked Documentation Data

The structural and rhetorical representation of technical documents and the subsequent identification of core documentation entities receives a publication four stars in the presented maturity schema. The maturity schema requires that documents have meta-data from an ontolgy attached to receive the fifth star. We recommend the usage of the `dc:subject` property from the Dublin Core [10] ontology for the annotation of structural, rhetorical or core documentation entities with concepts from (enterprise) ontologies. For instance, consider a document

that has been annotated with concepts describing relevant components or functions of a machine. Then a complete repair instruction (assembly + disassembly) for a concrete component (`ex:componentA`) can be identified as follows:

$$\texttt{ex:RepairComponentA} \sqsubseteq (\texttt{rtc:AssemblyProcedure} \sqcup \texttt{rtc:DisassemblyProcedure}) \sqcap \forall\, (\texttt{dc:subject.ex:componentA})$$

4 Extended Example

The following Turtle excerpt, is an example of how the ontologies described in Section 3 may be used to represent a technical document. The example gives an ontological description of a repair manual that contains detailed information (`docbook:Step`) about the assembly and disassembly of a concrete component.

```
:repair-manual a docbook:Book ;
  dc:hasPart :index , :general , :safety , :repair .

:repair a docbook:Chapter , rtc:Repair ;
  dc:hasPart :repair-a , :repair-b .

:repair-b a docbook:Chapter ;
  dc:hasPart :disassembly-b, :assembly-b ;
  dc:subject :component-b .

:disassembly-b a docbook:Chapter, rtc:Disassembly ;
  dc:hasPart :safety-note; :some-text; :some-procedure .

:some-procedure a docbook:Procedure ;
  dc:hasPart
    [ a docbook:Step ;
      dc:description "Insert stem into the fork." ],
    [ a docbook:Step ;
      dc:description "Point stem towards the front." ] .
...
```

Listing 1. Example ontology representing a repair manual.

5 Summary and Future Work

This paper introduced a maturity schema that allows the assesment of existing technical documents according to certain quality criterias. The schema is inspired by the 5-star Linked Open Data idea but consideres important aspects of the Technical Documentation and Publishing domain. The maturity schema requires the usage of documentation standards and ontologies. Thus we proposed

the representation of technical publications in a DocBook-like ontology. This representation is accompanied by a novel ontology that covers the rhetorical aspects of a technical document. Combining both ontologies in complex OWL [8] classes reveals core documentation entities. These high entropy elements can immediatly satisfy an information need. Hence, effective access to these elements yields huge time savings. The completion of rhetorical elements for technical documentation as well as the definition of supplementary core documentation entities will be subject of future work. We additionally plan to implement methods for the automatic conversion of 1-star legacy data to 4-star ontological data. These methods shall also be combined with our existing semantification approaches [5].

References

1. Berners-Lee, T.: Linked data (http://www.w3.org/designissues/linkeddata.html)
2. Constantin, A., Peroni, S., Pettifer, S., Shotton, D., Vitali, F.: The Document Components Ontology (DoCO). Semantic Web Preprint(Preprint) (2015)
3. Coskun, G., Streibel, O., Paschke, A., Schäfermeier, R., Heese, R., Luczak-Rösch, M., Oldakowski, R.: Towards a corporate semantic web. In: International Conference on Semantic Systems (I-SEMANTICS '09). pp. 602–610. Graz, Austria (2009)
4. Di Iorio, A., Peroni, S., Poggi, F., Vitali, F.: Dealing with structural patterns of xml documents. Journal of the Association for Information Science and Technology 65(9), 1884–1900 (2014)
5. Furth, S., Baumeister, J.: Towards the semantification of technical documents. In: FGIR'13: Proceedings of German Workshop of Information Retrieval (at LWA'2013) (2013)
6. Groza, T., Handschuh, S., Möller, K., Decker, S.: SALT-Semantically Annotated LaTeX for Scientific Publications. In: The Semantic Web: Research and Applications, pp. 518–532. Springer (2007)
7. Guha, R., McCool, R., Miller, E.: Semantic search. In: Twelfth International World Wide Web Conference (WWW 2003) (2003)
8. Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S. (eds.): OWL 2 Web Ontology Language: Primer. W3C Recommendation (27 October 2009), available at `http://www.w3.org/TR/owl2-primer/`
9. Janowicz, K., Hitzler, P., Adams, B., Kolas, D., II, C.V.: Five stars of linked data vocabulary use. Semantic Web 5(3) (2014)
10. Kokkelink, S., Schwänzl, R.: Expressing qualified dublin core in RDF/XML (2001)
11. Mäkelä, E.: Survey of semantic search research. In: Proceedings of the seminar on knowledge management on the semantic web (2005)
12. Norman, W., Hamilton, R.L.: DocBook 5: The Definitive Guide. O'Reilly Media, Inc. (2010)
13. Şah, M., Wade, V.: Automatic metadata extraction from multilingual enterprise content. In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 1665–1668. ACM (2010)
14. W3C: RDF Schema 1.1 – W3C Recommendation. `http://www.w3.org/TR/rdf-schema` (February 2014)
15. Witte, R., Gitzinger, T.: Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients. In: 3rd Asian Semantic Web Conference (ASWC 2008). LNCS, vol. 5367, pp. 360–374. Springer, Bangkok, Thailand (February 2–5 2009), `http://rene-witte.net/semantic-assistants-aswc08`