INTELLIGENT SYSTEMS IN ACCOUNTING, FINANCE AND MANAGEMENT
Intell. Sys. Acc. Fin. Mgmt. **22**, 81–99 (2015)
Published online 19 January 2015 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/isaf.1360

AN ANALYSIS OF APPLICABILITY USING QUALITY METRICS FOR ONTOLOGIES ON ONTOLOGY DESIGN PATTERNS

BIRGER LANTOW* AND KURT SANDKUHL
University of Rostock, Chair of Business Information Systems, Rostock, Germany

SUMMARY

Ontology design patterns (ODPs) provide best-practice solutions for common or recurring ontology design problems. This work focuses on content ODPs, which form small ontologies themselves and thus can be subject to ontology quality metrics in general. We investigate the use of such metrics for content ODP evaluation in terms of metrics applicability and validity. The quality metrics used for this investigation are taken from existing work in the area of ontology quality evaluation. We discuss the general applicability to content ODP of each metric considering its definition, ODP characteristics, and the defined goals of ODPs. The research process presented in this paper has two phases. In the first phase, we conducted a literature research in the area of metrics for assessing ontology quality. The second phase consisted of a two-step evaluation of the ontology metrics identified in the literature analysis. During the first step, we investigated whether the metrics are appropriate to differentiate between content ODPs of different quality. Metrics that proved to be applicable were calculated for a random set of 14 content ODPs. In the second step, a controlled experiment, the quality indicated by the metric value was contrasted with the perception of ontology engineers; that is, do 'measured quality' and 'perceived quality' match?. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: Ontology Design Patterns; Quality Metrics; Semantic Web; Ontology Engineering

1. INTRODUCTION

In many engineering disciplines, quality is considered an essential factor for acceptance of technologies and solutions, for efficiency of the processes and for robustness and usability of products. Quality is often considered from different perspectives, including the product-based and the user-based perspectives (Garvin, 1984). The product-based perspective usually views quality as a precise and measurable variable; that is, differences in quality are reflected in differences of some attribute possessed by a product (Hallak and Schott, 2011). The user-based perspective assumes that quality shows when using a product or service and to some extent 'lies in the eyes of the user' (Bevan, 1995). International standardization attempts of quality include these two perspectives and often add other aspects like, for example, efficiency and reliability (Jung *et al.*, 2004). Independently of perspective and definition, how to determine and measure quality continues to be an issue that is subject of research and industrial development; for example, see Garrido *et al.* (2014) or Jozsef and Blaga (2014). With an increasing use of ontologies in industrial applications, standards, procedures and metrics for quality assessment of ontology construction processes and the artefacts produced during these processes also gain in importance. Although considerable efforts have been spent on developing ontology assessment and evaluation approaches, including metrics and ways to measure quality (see Section 2.3), generally accepted practices for industrial use are still missing.

The objective of this paper is to contribute to quality ontologies by focusing on ontology design patterns (ODPs) and ways to determine their quality. ODPs have been proposed as encodings of best

* Correspondence to: Birger Lantow, University of Rostock, Chair of Business Information Systems, Rostock, Germany. E-mail: birger.lantow@uni-rostock.de

practices (see Section 2.2) supporting ontology construction by facilitating reuse of proven solution principles. Different kinds of ODPs have been proposed, like logical, transformation or content ODPs, which represent different aspects of best practices. This paper focuses specifically on content ODPs and on investigating the transferability of ontology quality metrics to content ODP. The long-term objective is to create an instrument for quality assurance in practice; that is, the main intention is not to develop new fundamental knowledge about ODP characteristics and measurement options, but rather to evaluate how to transfer metrics from the ontology area and what metrics to transfer.

Research results presented in this paper are based on a research process with two phases. In the first phase we conducted a literature research in the area of metrics for assessing ontology quality. The results of this step are summarized in Section 2.3 and Section 3 respectively. The second phase consisted of a two-step evaluation of the ontology metrics identified in the literature analysis. During the first step, we investigated whether the metrics are appropriate to differentiate between content ODPs of proposedly different quality. For large ontologies a metric value may well characterize an ontology, but for small ODPs the same metric may always show very similar or identical values, which are unlikely to help differentiating quality. We used the measurement procedures defined for a metric and determined the actual value for a given set of patterns. The set of patterns used consisted of 14 randomly selected patterns from one of the major sources for ODPs, the ODP portal at http://www.ontologydesignpatterns.org/. The results are discussed in Section 4.1. In the second step, we only considered those metrics that passed the differentiation test during the first step. In a controlled experiment, the quality indicated by the metric value was contrasted with the perception of ontology engineers; that is, do ‘measured quality’ and ‘perceived quality’ match (see Section 4.2)?

The contributions of this paper are (1) the evaluation of a selected set of ontology metrics regarding their applicability for content ODPs, (2) an assessment of applicable metrics regarding their ability to differentiate between ODPs of different quality and (3) an investigation on how the proposed metrics correlate with perceived quality.

The remainder of this paper is structured as follows. Section 2 gives an overview to related work, which includes the area of content ODPs and ontology evaluation. We discuss possible quality metrics and their calculation and applicability in Section 3. The metrics that qualify for content ODPs are validated by a set of experiments that we describe in Section 4. Section 5 aggregates our findings and gives an outlook on future research needs.

2. BACKGROUND AND RELATED WORK

Relevant background for this paper includes knowledge patterns (Section 2.1), ODPs (Section 2.2) and approaches for quality assurance of ontologies and ODPs (Section 2.3).

2.1. Knowledge Patterns

For more than 20 years patterns have been popular in computer science, and they were introduced for numerous areas, like software design, information modelling or business processes. Although there is no generally accepted definition of the term *pattern*, most publications in the field get some inspiration from Christopher Alexander’s definition (Alexander *et al.*, 1977):

> Each pattern describes a problem which occurs over and over again in our environment, and then describes the core of the solution to that problem, in such a way that you can use this solution a million times over, without ever doing it the same way twice.

Whilst Alexander's focus is on the solution, many pattern approaches in computer science focus more on capturing proven practices on how to approach certain problems. The seminal book on patterns was published by the 'Gang of Four' (Gamma *et al.*, 1995) and focuses on software design patterns. Many other books followed and proposed patterns for most phases of the software development process, including analysis patterns (Fowler, 1997), data model patterns (Hay, 1995) or software architecture patterns (Buschmann *et al.*, 1996). The pattern idea was adapted in other areas of computer science, like workflow patterns (van der Aalst *et al.*, 2003) and knowledge patterns.

The term *knowledge pattern* has been explicitly defined by Clark *et al.* (2000) in the context of knowledge representation as 'a pattern as a first-order theory whose axioms are not part of the target knowledge-base, but can be incorporated via a renaming of the non-logical symbols'. The intention is to help construct formal ontologies by explicitly representing recurring patterns of knowledge, so-called theory schemata, and by mapping these patterns on domain-specific concepts. Staab *et al.* (2001) investigated the use of so-called 'semantic patterns' for enabling reuse across languages when engineering machine-processable knowledge. Semantic patterns consist in this approach of one description of the core elements independent from the actual implementation and for each target language a description that allows for translating the core elements into the target language. The structure of the informal description consists of eight elements, which resemble the elements of design patterns (e.g. name, intent, motivation, structure); the translation into a language includes translation mapping, samples, applicability and comments. Compared with knowledge patterns, semantic patterns try to separate engineering knowledge from language-specific implementations instead of theories from domains they are applied in.

Knowledge formalization patterns have been proposed by Puppe (2000) as rather simple templates proven in practice for the (mass) formalization of knowledge. Puppe puts a lot of emphasis on proven problem-solving methods, which uncover implicit knowledge of experts. Knowledge formalization patterns consist of well-defined problem-solving methods, a graphical notation and a simple-to-understand mental model.

2.2. Ontology Design Patterns

In a computer science context, ontologies usually are defined as explicit specifications of a shared conceptualization (Gruber, 1993). Owing to the increasing use of ontologies in industrial applications, ontology design, ontology engineering and ontology evaluation have become a major concern. The aim is to efficiently produce high-quality ontologies as a basis for Semantic Web applications or enterprise knowledge management. Despite quite a few well-defined ontology construction methods and a number of reusable ontologies offered on the Internet, efficient ontology development continues to be a challenge, since this still requires a lot of experience and knowledge of the underlying logical theory.

ODPs are considered a promising contribution to this challenge. In 2005, the term *ontology design pattern* in its current interpretation was mentioned by Gangemi (2005) and introduced by Blomqvist and Sandkuhl (2005). Blomqvist (2009) defines the term as 'a set of ontological elements, structures or construction principles that solve a clearly defined particular modeling problem'. ODPs are considered as encodings of best practices, which help to reduce the need for extensive experience when developing ontologies; that is, the well-defined solutions encoded in the patterns can be exploited by less-experienced engineers when creating ontologies.

The area of ODP research is closely related to reusable problem-solving methods (Puppe, 2000) and knowledge patterns (Clark *et al.*, 2000) (Section 2.1). Different types of ODPs are under investigation, which are discussed in Gangemi and Presutti (2009) regarding their differences and the terminology used. The two types of ODPs probably receiving most attention are logical and content ODPs. Logical ODPs focus only on the logical structure of the representation; that is, this pattern type is targeting aspects of language expressivity, common problems and misconceptions. Content ODPs are often instantiations of logical ODPs offering actual modelling solutions. Owing to the fact that these solutions contain actual classes, properties and axioms, content ODPs are considered by many researchers as domain dependent, even though the domain might be considering general issues like 'events' or 'situations'. Platforms offering ODPs currently include the ODP wiki portal initiated by the NeOn-project and the logical ODPs maintained by the University of Manchester.

2.3. Quality Assurance of Ontologies and Ontology Design Patterns

Work in the area of quality assurance for ontologies and ODPs includes different perspectives, such as the quality of the ontology or ODP as such, the quality of the process of ontology construction, and tools supporting the ontology engineer in achieving high quality.

From the tool perspective, there are tools for the identification of the origin of inconsistencies or unexpected entailments (Horridge *et al.*, 2009) using reasoners. Such logical errors are clear-cut and easily identifiable. However, content errors are often harder to detect, and their consequences often show only in the usage situation. A line of work attempting to detect content errors has focused on rendering ontology axioms by translating them into natural language. Examples are the GALEN project (Baud *et al.*, 1997) and the generation of natural language sentences by Duque-Ramos *et al.* (2011) which encompasses class definitions and entailments.

The quality assessment of the ontology construction process has received less attention than the assessment of tools and ontologies as such (Gorovoy and Gavrilova, 2007). From a process perspective, there is an approach of using workflow diagrams for formalizing the ontology construction process. The workflow support translating upper-level axioms and meta-properties (Guarino and Welty, 2009) into decision trees that interactively guide an incremental ontology construction process (Seyed, 2012a,b).

The quality assessment of ontologies as such has been the subject of many research activities (Vrandečić and Sure, 2007), but the quality criteria vary considerably between different approaches and often address structural, logical and computational aspects of ontologies. Furthermore, metrics originating from software quality evaluation have been investigated (Duque-Ramos *et al.*, 2011). Many of the metrics that have been proposed during recent years lack an empirical validation in a large number of cases; that is, what metrics value can be considered as 'good' or as 'bad' often has not been defined owing to an insufficient number of reported applications.

Besides approaches like that of Maedche and Staab (2002) that suggest a gold standard for reference, an ontology content evaluation is poorly feasible for tool support. Thus, we focus on structural metrics and their validation. The work of Gangemi *et al.* (2005) has been chosen as a starting point. Among others things, they suggest structural and usability metrics that can be calculated automatically. Additionally, a rough idea of 'bad' and 'good' values is given.

3. ONTOLOGY METRICS

In order to evaluate and to compare the quality of ontologies, formally defined metrics are an instrument of choice. They allow for automated or semi-automated metrics calculation. This section starts

DOI: 10.1002/isaf

by giving an overview of the different kinds of metrics (Section 3.1) and discusses their use for ODPs (Section 3.2).

3.1. Overview

Gangemi *et al.* (2005) define ontology metrics based on a metaontology O^2. This leads to three metric types for ontology evaluation (Gangemi *et al.*, 2005): structural, functional and usability-profiling metrics.

Structural Metrics

Structural metrics focus on syntax and formal semantics. Structurally seen, is a graph whose nodes and arcs represent concepts. Structural metrics mainly refer to the syntax of the ontology graph. Sometimes, formal (abstract) semantics is in focus. However, formal semantics can also be considered as additional syntax. Intended meaning, semantics and context are not referred to by such metrics.

Concrete metrics measure topological and logical properties (Gangemi *et al.*, 2005: 8). In general, depth and breadth metrics count *isa*- or *subclass-of* relationships respectively. Density metrics, in contrast, count all other relationships. A common representation of a metric is given by

$$M = \langle D, S, \mathrm{mp}, c \rangle$$

D identifies the dimension to be measured. Hence, it is the graph property of interest. The set of graph elements is represented by S. The measuring procedure mp is the calculation rule for the respective metric. The coefficient of measurement c allows adjustments for different measurement contexts.

Measuring structural metrics is usually based on counting. Thus, it relates natural numbers to a set of graph elements (Gangemi *et al.*, 2005: 10). In order to make such measuring procedures applicable, common element sets are defined and identified by symbols (see Gangemi *et al.* (2005: 10) for reference).

Functional Metrics

These metrics focus on the relation between the ontology graph and its intended meaning.

Assessing functional metrics requires expert knowledge in the ontology domain. The ODPs that are used for our investigation may be understandable based on common knowledge. But, when it comes to questions regarding the completeness and accuracy of modelled concepts, more than common knowledge is necessary. Furthermore, content ODPs do not intend to be complete regarding a given domain. They are considered as ontology building blocks and modelling guidance. Owing to this dependence on the domain of the actual usage of ODPs and on expert knowledge, a measurement procedure for functional metrics also will have to depend on these two aspects and can hardly be defined in a generally applicable way. Thus, functional metrics are not discussed further.

Usability-Profiling Metrics

Usability-profiling metrics aim at the ontology profile. The ontology profile is a set of ontology annotations that contains metadata about the ontology and its elements with regard to ontology use and development. This includes structural, functional and user-oriented information. Gangemi *et al.* (2005: 36) distinguish three analytical levels of information:

DOI: 10.1002/isaf

- **Recognition annotation.** These describe objects, actions and options. The goal is a complete documentation that guarantees effective access. Ontology structure, function and life cycle can be described by annotations. We focus on life-cycle annotations, which contain information about provenance, employed methods, versioning and compatibility.
- **Efficiency annotations.** These support the cost–benefit calculation in the use of ontologies.
- **Interfacing annotations.** These describe the alignment of an ontology to a user interface. If there is a strong connection between ontology context and ontology representation, such annotations can be helpful.

Possible metrics of usability profiling are presence, completeness and reliability of all three kinds of annotations. Recognition annotations have been present in all investigated content ODPs, while the other two kinds of annotations were missing. Completeness and reliability have not been assessed during our investigations. First, completeness has to consider the context in which an ontology or an ODP is used. Second, the assessment of reliability requires insight into the ontology development process. This is not given regarding the published ODPs.

3.2. Application of Structural Metrics to Ontology Design Patterns

Following the discussion in Section 3.1, we concentrate on the application of structural metrics to content ODPs. Functional and usability-profiling metrics can only be evaluated in a given context. This means there would have to be a defined domain of interest and defined requirements for the ontology in use.

Gangemi *et al.* collected 31 structural metrics together with measuring procedures. Additionally, *density* and degree *distribution* are mentioned for completeness (Gangemi *et al.*, 2005: 17, 21–22). For triangulation and in order to assure generality of the findings, we considered further metrics proposed by Yang *et al.* (2006), Bolotnikova *et al.* (2011), Tartir *et al.* (2005, 2010) and Djedidi and Aufaure (2010).

All structural metrics are based on counting graph elements. Differences are in the measurement procedure (mp) and the way how the ontology structure is represented (S). The measurement procedure (mp) defines the elements of the ontology representation (S) selected for counting, the way how the counted values are aggregated to the metric value and the region of the ontology that is going to be assessed.

In order to make the metrics comparable, a common model of S must be defined. Based on the selected proposals for ontology metrics, it can be described as follows:

$$S = \langle C,\ I, R, \mathrm{At}, \mathrm{An},\ \mathrm{Ax} \rangle$$

where C is the set of classes, I is the set of individuals or instances, R is the set of relations, At is the set of attributes, An is the set of annotations (do not mix up with ontology profile) and Ax is the set of axioms.

There are three general types of relations. First, relations that describe inheritance between classes R_{hsc}. A relation $R_{\mathrm{hsc}}(i,j)$ means that class j is a subclass of i. Second, relations between classes that have a different semantic than inheritance R_{sem}. And third, relations that assign individuals to classes R_{ind}. A relation $R_{\mathrm{ind}}(i,j)$ means that j is an individual or instance of class i. Attributes At_i and annotations An_i are seen as parts of the specification of a class C_i. The axioms Ax are general assumptions regarding the

DOI: 10.1002/isaf

ontology. Thus, the classes C and the individuals I are the nodes of the ontology structure graph. R is the set of edges. The other subsets of the ontology structure are seen outside of the ontology graph.

There are more complex views on the structure of an ontology. The OWL 2 language model, for example, allows inheritance between relations. However, the structure as shown here is sufficient for the evaluation of selected ontology metrics.

A number of ontology metrics (see below) only considers the inheritance structure of an ontology $S_{\mathrm{inh}} = \langle C, R_{\mathrm{hsc}} \rangle$. According to OWL 2, there is a single root C_{Thing} within S_{inh}. Ontologies that are not represented in OWL may not contain the Thing-class. Since most of the ODPs are formalized using OWL, we base all metric calculations on the existence of the Thing-class. For example, metrics regarding breadth, width, sibling sets and leaf sets are based on S_{inh}.

Based on the given ontology structure model and a first investigation of existing ODPs, some ontology metrics can be excluded from consideration. In general, ODPs do not contain individuals. Metrics regarding individuals can be excluded. Additionally, annotations are usually restricted to one language tag per class. Thus, no differences regarding annotation count are expected between ODPs. Owing to the small size of the ODPs, metrics that differentiate between regions of an ontology do not seem to be appropriate. This includes metrics for modularity, density and degree distribution. Some metrics treat attributes as graph nodes connected to classes using special relations. This is not compatible with the OWL 2 language model and is not considered further.

The class of metrics for logical adequacy as defined by Gangemi *et al.* (2005: 18) also contains some metrics that are not appropriate in the given context. First there are metrics checking the ontology graph for consistency with formal requirements. Except for the language syntax (of OWL 2), there are no formal requirements defined for ODPs. Second, there are metrics that aim at computational complexity for reasoners working on the ontology. Phases of the ontology lifecycle where reasoners are used are not the focus of our investigations.

Based on the above reasons of exclusion, a set of metrics for further evaluations has been selected from the proposals (see also Table I).

This includes 17 out of the metrics by Gangemi *et al.* (2005). M1 to M3 measure the depth of S_{inh}, while M4 to M6 deal with its breadth. According to Gangemi *et al.* (2005: 12), the tangledness metric (M7) counts the multi-hierarchical nodes of S_{inh}. Hence, it aims at concepts that have more than one superclass. Looking, for example, into the work of Duque-Ramos *et al.* (2011), tangledness is defined as the mean number of parents per class. However, the metrics description in Gangemi *et al.*'s (2005) report refers to nodes that have more than one outgoing $R_{\mathrm{hsc}}(i,j)$. This holds for all parent nodes (including the root-node) that have more than one child. We assume that there is just a mistake in Gangemi *et al.*'s (2005) metric description. The following formula has been applied for M7 calculation in our investigations:

$$\mathrm{M7} = \frac{\left|\left\{C_j; \exists i,h : R_{\mathrm{hsc}}(h,j) \wedge R_{\mathrm{hsc}}(i,j),\ h \neq i\right\}\right|}{|C|}$$

It counts the ratio of nodes having more than one parent.

The metrics of fan-outness and sibling fan-outness (M8–M18) are generally applicable to ODPs. However, M17 and M18 are exceptions. They do not conform to the ontology model as discussed above. This is also the case for the following metrics up to M21, which refer to shared attributes. Metrics of modularity (M22, M23) have already been excluded from consideration. Looking at metrics of logical adequacy which are not restricted to S_{inh}, only M29 will be evaluated further. M24 and M32 refer to formal specifications. M25 to M28 aim at computational complexity. M30 counts the number of

DOI: 10.1002/isaf

Table I. Metric values of selected ODPs.

Name	Source	Symbol	Componency	Species Conditions	Time-indexed Person Role	Time-indexed Participation	Object Role	Vertical Distribution	Classification	Role Task	Species Conservation	Agent Role	Constituency	Description	Gear Water Area	Tagging	No. of values
Absolute depth	Gangemi *et al.* (2005)	M1	2	11	24	12	5	17	2	4	2	7	1	4	4	28	10
Average depth	Gangemi *et al.* (2005), Djedidi and Aufaure (2010)	M2	2.00	2.20	2.67	2.40	2.50	2.43	2.00	2.00	2.00	2.33	0.00	2.00	2.00	3.11	9
Maximal depth	Gangemi *et al.* (2005), Yang *et al.* (2006)	M3	2	3	5	4	3	4	2	2	2	3	1	2	2	4	5
Absolute breadth	Gangemi *et al.* (2005)	M4	2	7	12	7	4	10	2	3	2	5	1	3	5	15	9
Average breadth	Gangemi *et al.* (2005)	M5	1.00	2.33	4.00	1.75	1.33	2.50	1.00	1.50	1.00	1.67	1.00	1.50	2.50	3.75	9
Maximal breadth	Gangemi *et al.* (2005)	M6	1	5	4	4	2	6	1	2	1	2	1	2	4	5	5
Tangledness	Gangemi *et al.* (2005)	M7	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	3
Absolute leaf cardinality	Gangemi *et al.* (2005)	M8	1	5	4	4	2	6	1	2	1	2	1	2	4	6	5
Ratio of leaf fan-outness	Gangemi *et al.* (2005)	M9	0.50	0.71	0.33	0.57	0.50	0.60	0.50	0.67	0.50	0.40	1.00	0.67	0.80	0.43	10
Weighted ratio of leaf fan-outness	Gangemi *et al.* (2005)	M10	0.50	0.45	0.21	0.40	0.40	0.40	0.50	0.50	0.50	0.33	1.00	0.50	0.50	0.25	7
Maximal leaf fan-outness	Gangemi *et al.* (2005)	M11	1	4	1	4	2	4	1	2	1	2	1	2	4	4	3
Total sibling fan-outness	Gangemi *et al.* (2005)	M12	1	2	8	3	2	4	1	1	1	3	0	1	1	8	6
Ratio of sibling fan-outness	Gangemi *et al.* (2005)	M13	0.50	0.29	0.67	0.43	0.50	0.40	0.50	0.33	0.50	0.60	0.00	0.33	0.20	0.57	10
Weighted ratio of sibling fan-outness	Gangemi *et al.* (2005)	M14	0.50	0.18	0.33	0.25	0.40	0.24	0.50	0.25	0.50	0.43	0.00	0.25	0.25	0.29	9
Average sibling fan-outness	Gangemi *et al.* (2005)	M15	1	2	4	2	1	3	1	2	1	2	1	2	3	4	9

Maximal sibling fan-outness	Gangemi *et al.* (2005)	M16	1	5	4	4	2	6	1	2	1	2	1	2	4	5	5
Class/relation ratio	Gangemi *et al.* (2005)	M29	0.67	0.50	0.92	0.50	1.00	0.77	1.00	1.00	2.00	1.25	1.00	0.60	1.00	0.61	9
Total number of concepts	Yang *et al.* (2006), Bolotnikova *et al.* (2011)	M33	2	7	12	7	4	10	2	3	2	5	1	3	5	14	9
Total number of paths	Yang *et al.* (2006)	M34	2	7	14	7	4	10	2	3	2	5	1	3	5	15	9
Average number of paths	Yang *et al.* (2006), Djedidi and Aufaure (2010)	M35	1.00	1.00	1.17	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.07	3
Total number of relations	Yang *et al.* (2006)	M36	3	14	13	14	4	13	2	3	1	4	1	5	5	23	8
Average path length	Yang *et al.* (2006)	M37	1.50	2.00	2.43	2.29	2.00	2.30	1.50	1.67	1.50	2.20	1.00	1.67	1.80	2.80	10
Average number of relations	Yang *et al.* (2006)	M38	1.50	2.00	1.08	2.00	1.00	1.30	1.00	1.00	0.50	0.80	1.00	1.67	1.00	1.64	9
Ratio of nodes containing leaf and non-leaf nodes	Bolotnikova *et al.* (2011)	M40	1.00	2.00	1.33	2.00	2.00	1.50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.20	5
Ratio of nodes with normal a degree	Bolotnikova *et al.* (2011)	M41	1.00	0.86	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.93	3
Ratio of nodes with different types of outgoing relations	Bolotnikova *et al.* (2011)	M42	1.00	0.14	0.00	0.14	0.00	0.10	0.50	0.33	0.00	0.00	1.00	0.67	0.20	0.21	9
Ratio of nodes with different types of ingoing relations	Bolotnikova *et al.* (2011)	M43	1.00	0.71	0.08	0.57	0.50	0.30	0.50	0.33	0.00	0.00	1.00	0.67	0.20	0.29	11
Maximum ratio of the widths of adjacent levels	Bolotnikova *et al.* (2011)	M44	1	5	4	4	2	6	1	2	0	2	1	2	4	5	6

(*Continues*)

DOI: 10.1002/isaf

Table I. (Continued)

Name	Source	Symbol	Metric values														No. of values
			Componency	Species Conditions	Time-indexed Person Role	Time-indexed Participation	Object Role	Vertical Distribution	Classification	Role Task	Species Conservation	Agent Role	Constituency	Description	Gear Water Area	Tagging	
Semantic Richness	Tartir *et al.* (2010), Djedidi and Aufaure (2010)	M45	0.80	0.67	0.08	0.50	0.40	0.40	0.67	0.67	0.00	0.00	1.00	0.71	0.33	0.48	10
Attribute Richness	Tartir *et al.* (2010), Djedidi and Aufaure (2010)	M46	0.00	0.00	0.25	0.43	0.00	0.20	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.29	6
Inheritance Richness	Tartir *et al.* (2010), Djedidi and Aufaure (2010)	M47	0.50	0.86	1.00	0.86	0.75	0.90	0.50	0.67	0.50	0.80	0.00	0.67	0.80	1.00	8
Coherence	Yang *et al.* (2006)	M48	1.00	1.36	1.88	1.67	1.20	1.65	1.00	1.00	0.50	1.29	0.00	1.00	1.00	1.29	9
Average number of semantic relations per class	Djedidi and Aufaure (2010)	M49	2.00	1.71	0.08	0.86	0.50	0.60	1.00	1.33	1.00	0.00	2.00	1.67	0.40	0.93	12
Mean-square deviation of the degree of a graph node	Bolotnikova *et al.* (2011)	M50	0.00	18.33	0.70	7.95	0.33	4.54	0.00	3.00	0.50	0.30	0.00	2.33	1.80	9.30	12
Ratio of the mean-square deviation of the depth to the mean depth	Bolotnikova *et al.* (2011)	M51	0.33	0.17	0.51	0.40	0.33	0.29	0.33	0.20	0.33	0.32	0.00	1.67	0.11	0.77	11
Mean-square deviation of descendant leaf nodes	Bolotnikova *et al.* (2011)	M52	0.50	2.24	0.24	1.29	0.33	1.60	0.50	1.33	0.50	0.30	0.00	1.33	3.58	0.42	11
Total number of different semantic relationships	Bolotnikova *et al.* (2011)	M53	4	11	1	6	2	6	2	4	0	0	2	5	2	13	8

DOI: 10.1002/isaf

axioms. But axioms have not been available for evaluation. There are a lot different ways for the formulation of axioms (graphical, textual). Thus, the influence of axiom representation could not be controlled within the experiments. M31 counts individuals. The latter are not present in ODPs.

The work of Yang *et al.* (2006) defines some 'primitive' graph metrics. The metrics are based on counting classes and relations. There is no differentiation between relation types. However, the approach seems to imply the restriction to S_{inh}. The authors refer to a most general class or concept as root and they use the notion of paths. The latter would need a more precise definition if other relation types than inheritance were to be allowed. For reference, the numbering schema for ontology metrics by Gangemi *et al.* (2005) is continued with M33 and following. The primitive metrics by Yang *et al.* are total number of classes (TNOC, M33), total number of paths (TNOP, M34), maximum path length (Λ, M3), average path length ($\overline{\Lambda}$, M37) and total number of relations (TNOR, M36). There is one correspondence with the metrics by Gangemi *et al.* (2005). The maximum path length equals the maximum depth (M3) regarding S_{inh}. Yang *et al.* (2006) also define some 'complexity metrics': average relations per class ($\mu = \mathrm{TNOR}/\mathrm{TNOC}$, M38), average paths per class ($\varrho = \mathrm{TNOP}/\mathrm{TNOC}$ M35) and ratio maximal path length to average path length or coherence ($\sigma = \Lambda/\overline{\Lambda}$, M48). The average path length for a class has been excluded from our investigations. It does not refer to the whole ontology or ODPs as such.

Bolotnikova *et al.* (2011) describe an eight-stage process for ontology evaluation. Metrics are defined for most of the stages. Stage 1 deals with the ontology size. The metrics suggested here are maximal depth (M3) and total number of classes (TNOC, M33). At stage 2, ontology errors are analysed. Bolotnikova *et al.* (2011) see cycles as such errors. Cycles in S_{inh} are generally not present in ODPs. Cycles regarding other relation types are common and do not impose errors in ODPs. Thus, we did not consider the suggested cycle metrics. Further metrics discussed for stage 2 are tangledness (M7) and the ratio of classes having leaf and non-leaf nodes as descendants to those having only leaf nodes as descendants (M40) in S_{inh}. Values higher than 1 for both metrics indicate possible errors. Stage 3 analyses Yngve–Miller metrics. This aims directly at cognitive ergonomics. The human short-term memory can only store 7 ± 2 elements according to Yngve–Miller. Thus, a graph node having a degree ≤ 9 is said to have a normal degree. The ratio of classes having a normal degree (M41) is defined based on this assumption. Another metric used at this stage according to Bolotnikova *et al.* (2011) is the mean-square deviation of the degree of a graph node (M50). All relation types (except for R_{ind}) are considered regarding this metrics, which is also true for the metrics of stage 4, the analysis of types of ontology relations. These are the total number of different relations types including types in R_{sem} (M53), the ratio of classes with different outgoing relations (M42), and the ratio of classes with different ingoing relations (M43). Breadth and width of the ontology graph are analysed at stages 5 and 6. Gangemi *et al.* (2005) already introduced metrics here (M1–M6). An addition by Bolotnikova *et al.* (2011) to the considered metrics is the maximum ratio of the widths of adjacent levels (M44) in S_{inh}. Branching metrics in S_{inh} are analysed at stage 7. Namely, there are the ratio of the mean-square deviation of the depth to the mean depth (M51) and the mean-square deviation of descendant-leaf nodes (M52). Stage 8 does not define additional metrics. It is the stage of taking actions and making recommendations based on the metric values.

OntoQA presented by Tartir *et al.* (2005, 2010) contains two sets of metrics regarding the ontology schema and the knowledge base respectively. Knowledge base metrics cannot be applied to ODPs. Generally, there are no instances/individuals in ODPs. The set of schema metrics refers to C, R (except for R_{ind}) and At. Relationship richness (RR, M45) is the ratio of relations in R_{sem} to the total number of

DOI: 10.1002/isaf

relations. Attribute richness (AR, M46) is the average number of attributes per class. And last, inheritance richness (IR, M47) is the average number of subclasses per class.

Djedidi and Aufaure (2010) name some ontology metrics as a part of their ontology evolution framework. The metrics are assigned to six evaluation criteria. Metrics for cohesion, completeness and comprehension are not further investigated here. The suggested cohesion metric refers to instances that are generally not present in ODPs. Completeness metrics correspond to the functional metrics by Gangemi *et al.* (2005). These have already been excluded. Comprehension metrics refer to annotations. As stated earlier, generally no differences between ODPs have been found regarding annotations. Djedidi and Aufaure (2010) suggest three metrics for the complexity criteria. Average depth (D, M2) has already been proposed by Gangemi *et al.* (2005). The average number of semantic relations R_{sem} per class (SRC, M49) and average number of paths to reach a node from root in S_{inh} (CP, TNOP/TNOC, M35) are the other two. The metrics regarding the conceptualization criteria have already been discussed (M45–M47). For the last criterion by Djedidi and Aufaure (2010), abstraction, again the average depth (M3) is suggested as a metric.

Altogether, 37 ontology metrics have been selected for further evaluation in Section 4. This selection does not have the ambition to be exhaustive. However, with extending the number of metrics considered compared with our earlier studies in Alm *et al.* (2013), the selection process showed a number of identical or near-identical metrics. The average depth (M2) is an example here (see Table I). These duplicates indicate that we covered a majority of ontology metrics.

4. VALIDATION OF SELECTED ONTOLOGY METRICS

In order to further investigate the validity of the selected ontology metrics to ODPs, we define two requirements. First, the metrics must be useful for differentiating between different ODPs of different quality. This is discussed in Section 4.1. Second, the metrics must correlate with the perceived quality of the ODPs. The steps undertaken to check this are described in Section 4.2.

4.1. Differentiation Criteria

The following 14 content ODPs from the ODPs wiki portal (http://ontologydesignpatterns.org) that was initiated by the NeOn-project are the base for our further discussion: (1) Componency, (2) SpeciesConditions, (3) TimeIndexedPersonRole, (4) TimeIndexedParticipation, (5) ObjectRole, (6) VerticalDistribution, (7) Classification, (8) RoleTask, (9) SpeciesConservation, (10) AgentRole, (11) Constituency, (12) Description, (13) GearWaterArea and (14) Tagging.

The patterns have been chosen by applying a pseudorandom number generator. A 15th selected ODP, Airline, had to be excluded because of an improper OWL source.

Intuitively, the ODPs have different qualities and different applications. The graphs provided by the OntoGraf-plugin of Protege (http://protege.stanford.edu/) have been used for metrics calculation. For reference, Figure 1 shows the structure of the VerticalDistribution ODP as an example of ODP visualization. Inheritance relations are marked with continuous lines. Semantic relations are symbolized by dashed lines. A different color stands for a different relation type.

Table I shows the calculated values for all selected structural metrics and ODPs. Significant differences between the calculated values of appropriate metrics can be seen, because of the diversity of the ODPs in structure and size. If ODPs are constructed similarly there are small or simply no differences in the values.

DOI: 10.1002/isaf

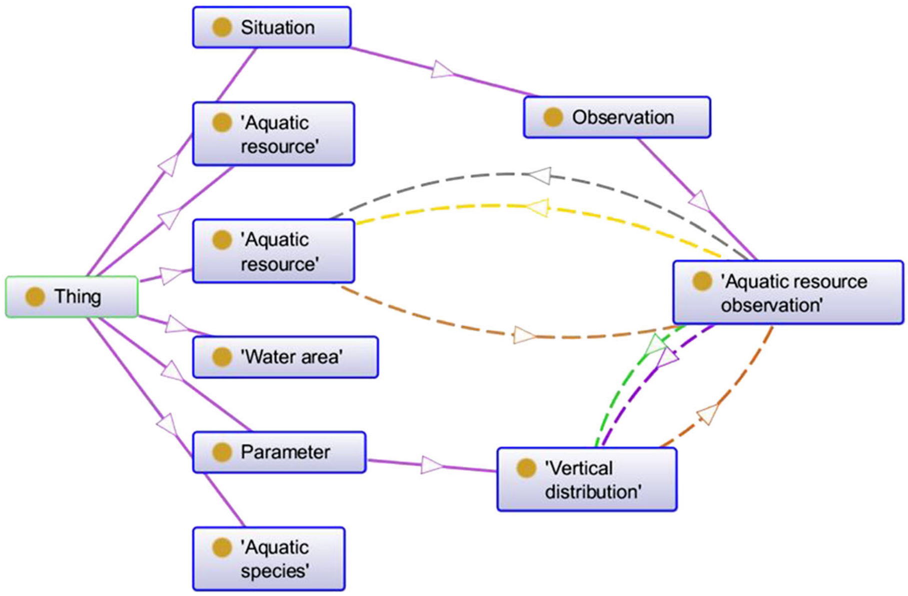

Figure 1. Graphical representation of "Vertical Distribution" ODP

Thus, we put a threshold of at least five different values for the metrics. This means – assuming a correlation between the metrics and perceived quality – that the metric is able to distinguish at least five qualities of ODPs. There are a few metrics that do not meet the criteria. First, tangledness (M7) and average number of paths (M35). Both are based on S_{inh}. They only show values different from 1 if multiple inheritance occurs. This is true for two of the selected ODPs. In accordance with Bolotnikova *et al.* (2011), a recommendation would be to avoid multiple inheritance. However, multiple inheritance is a basic feature of OWL 2. In the case of ODPs, those constructs seem to be seldom. We conclude that the existence of multiple inheritance should be assessed when evaluating ODPs. If there are such inheritance relations, a redesign should be considered or at least special attention should be paid to the quality of such ODPs. In the case of ODPs, metrics like M7 and M35 are not applicable to a correlation test as done in Section 4.2. Thus, a different approach will be necessary to validate their importance for quality assessments. The same assumptions can be made for the ratio of nodes with a normal degree (M41). Two ODPs out of the selected ones had one node each that did not comply with the Yngve–YMiller metrics. Such nodes have poor performance with respect to cognitive ergonomics. Thus, they should not be part of ODPs. The last metric that did not meet the five different values threshold is maximal leaf fan-outness (M11). The largest leaf set of a node in the selected ODPs was 4. This limits the number of possible different values to four. There were only three different values for M11 among the selected ODPs. However, we decided to further investigate the validity of M11 for quality evaluation in the next section.

4.2. Correlation Criteria

While the selected metrics allow one to describe the characteristics of content ODPs and to distinguish content ODPs with respect to these characteristics, in order to evaluate content ODPs the desired

DOI: 10.1002/isaf

characteristics or a preferential order for metrics values has to be determined. Gangemi *et al.* (2005: 39) provide principles that may be important in a project context for ontology evaluation. Each principle is based on a set of metrics that have impact on the fulfilment of the respective principle. Furthermore, the kind of impact is roughly expressed.

Gangemi *et al.* (2005) look into ontology use in general. In our case, the intention behind the idea of ODPs is the starting point. There is a strong focus on reuse and adaptability. ODPs should present best practices and should be accessible by a large number of nonexpert users. Thus, user-centred aspects like clarity and understandability are important. For example, Gangemi *et al.*'s (2005) principle of 'cognitive ergonomics' aims in the same direction.

In order to investigate how the defined metrics correlate with the fulfilment of desired principles, a number of experiments have been done. During these experiments, users evaluated selected content ODPs in a controlled situation with respect to

1. **Clarity.** Recognition of all concepts, relationships and their correspondences.
2. **Understandability.** Comprehension of all concepts, relationships, their correspondences and their meaning.
3. **Adaptability** to a given use case (the users got the task to adapt the respective pattern prior to evaluation).
4. **Reusability.** For example, as a part of a larger pattern

Setting

We had a total of 27 participants within the experiment who were divided into four groups. All of them were students in the BSc/MSc 'Business Information Systems' study program at Rostock University. The participants were familiar with the purpose, the syntax and semantics of ontologies and ontology graphs. However, the concept of ODPs had been introduced to them only briefly in conjunction with the experiment. The prior experience with ontologies and ODPs had been assessed by a survey in conjunction with the experiments. Biographical data and a self-assessment of skills and experience regarding ontologies and ontology engineering were part of that survey. It revealed that all groups were homogeneous regarding the level of skills and experiences.

Each group had to evaluate a set of three ODPs from the originally selected 14 ODPs. This resulted in 78 evaluations of ODPs. In total, nine different ODPs were included in the experiments (see Figure 2).

During the experiments, the participants were provided with the ontology graph (see Figure 1), the labels of semantic relations used and the ontology annotations (Recognition annotations) on paper. In consequence, the role of attributes and attribute-based metrics could not be evaluated in the experiments. Furthermore, the participants had to design and draw an ontology containing the evaluated ODPs. This was done in order to foster the examination of ODPs by the participants. The resulting ontologies have not been assessed.

The ODPs were evaluated according to the four criteria (Clarity, Understandability, Adaptability and Reusability) based on a Likert-scale containing the values 1 (very good), 2 (good), 3 (satisfactory), 4 (fair) and 5 (unsatisfactory). The rating was done by the participants according to their perception. Figure 2 shows the average rating of each ODP with regard to the criteria.

Results

A statistical evaluation has been done in order to assess the correlation of the ontology metrics regarding perceived Clarity, Understandability, Adaptability and Reusability of ODPs. First, the Pearson correlation coefficient was calculated for each metric X and each criteria Y with $n = 78$:

DOI: 10.1002/isaf

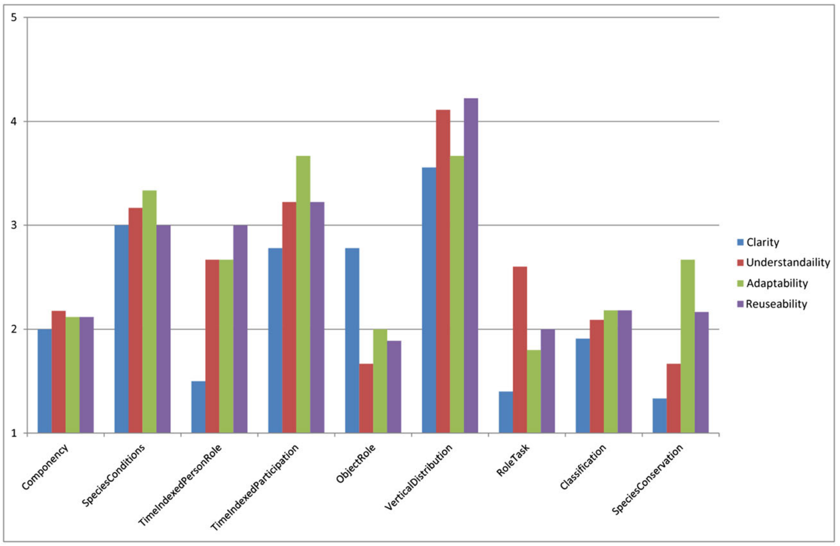

Figure 2. Average user ratings of the ODPs under investigation

$$r_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{s_X}\right)\left(\frac{Y_i - \overline{Y}}{s_Y}\right)$$

This measures the strength of the linear correlation of two variables. In a second step the significance of the correlation has been evaluated by a *t*-test with $t_{XY} = r_{XY}/\sqrt{(1-r_{XY}^2)/(n-2)}$. A threshold of $\alpha \leq 0.05$ for the error probability of rejecting $H_0 : r_{XY} = 0$ based on the Student's *t*-distribution was applied. Hence, the probability that the experienced correlation of the metric and the perceived quality stems from noise in the data is statistically below 5%. A significant correlation between metric and perceived quality does not prove causality, but the construction of causality is part of the metric definition by the different originators of the selected metrics.

For the majority of the metrics a significant correlation with at least one of the quality criteria could be proven. The exceptions are ratio of nodes with different types of ingoing relations (M43), semantic richness (M45) and ratio of mean-square deviation of depth to the mean depth (M51). Interestingly, the ratio of nodes with different outgoing relations (M42) shows some significant correlations in contrast to M43. Looking into the differences of metric values for M42 and M43, the cause is 'core' classes of the ODPs that have a lot of different outgoing relations connecting to a number of other classes (see, for example, VerticalDistribution ODP). Thus, M43 is very sensitive to the existence of such nodes. This may not result in linear or near-linear correlation. However, this is a possible explanation that needs further proof in the future. The poor validity of M45 and M51 may lay in the small size of the ODPs. Again, this is a future task to investigate possible reasons.

Besides significance, the strength of the correlation between metric and quality perception plays an important role. We speak of a strong correlation if $|r_{XY}| \geq 0.5$. Metrics that show a strong correlation are considered the best applicable metrics. Absolute leaf cardinality (M8), maximal sibling fan-outness

DOI: 10.1002/isaf

(M16), total number of relations (M36) and maximum ratio of the width of adjacent levels (M44) show a strong positive correlation for all three criteria 'Understandability', 'Adaptability', and 'Reusability'.

An analysis of the metrics revealed that M8, M16 and M44 resulted in identical metric values for the ODPs examined except for a small difference regarding Constituency and Tagging ODPs. The reason lies in the structure of the ODP graphs. Having *Thing* as a single root class, the maximum ratio of the widths of adjacent levels (M44) is found between the second level and first level (root) in all ODPs examined. This is also true for the maximal sibling fan-outness (M16), where the *Thing* node has the maximal number of descendant siblings. Furthermore, it seems that branching generally occurs only at these first ODP levels. In consequence, the leaf cardinality (M8) is almost identical to the number of branches at the first level. Summing up the discussion, the number of root concepts as direct descendants of *Thing* (similar values as M8, M16 and M44) is supposed to be a good metric for the assessment of 'Understandability', 'Adaptability', and 'Reusability'. This metric definition describes the semantics within the calculated values for ODPs better than those for M8, M16 and M44. The identifier M54 has been assigned to that metric.

Maximal leaf fan-outness (M11) is the only metric that showed a strong linear correlation to the 'Clarity' criterion. It seems that there are different ontology characteristics influencing perceived 'Clarity' on the one hand and perceived 'Understandability', 'Adaptability', and 'Reusability' on the other hand. However, a deep investigation of these interdependencies was not part of our investigations (see Section 4.3).

The overall results of the metrics evaluation regarding correlation can be found in Table II. The significance between metrics and quality criteria is calculated on the right-hand side of the table. Values meeting the desired error probability $\alpha \leq 0.05$ are in italics. The correlation strength is calculated on the left-hand side. Values indicating a strong correlation ($|r_{XY}| \geq 0.5$) are in italics.

Metrics (symbols in the middle) having no significant correlation with any of the criteria are marked underlined. Those having a significant correlation to some but not all criteria are in normal typeface. The remainder are in italics if there are no strong correlations and in bold in the case of at least one strong correlation.

4.3. Limitations

The participants in all groups form a homogeneous and specific set regarding their personal characteristics. All of them were in the same study programme at the same place. This limits the transferability of the results. However, the proposed metrics have a history of use and evaluation for the general case. Thus, showing a significant correlation in our special case of ODPs is a strong indicator for their applicability.

There is no single metric covering all characteristics of an ontology that influence perception regarding 'Clarity', 'Understandability', 'Adaptability' and 'Reusability'. There may also be interdependencies between these characteristics, limiting or fostering the influence of the other. Thus, the measured correlations may be influenced by those interdependencies.

The selection of patterns for the experiment may be seen critically. We selected the patterns based on our assumption of which of them are accessible and understandable by participants who are not domain experts.

Additionally, the sequence of patterns in the evaluation process has not been randomized. However, there was no evidence of a learning effect during the evaluation of patterns. The 'Componency' pattern,

DOI: 10.1002/isaf

Table II. Correlation between metrics and quality criteria.

Correlation strength				Metric symbol	Correlation significance (error probability)			
Clarity	Understandability	Adaptability	Reusability		Clarity	Understandability	Adaptability	Reusability
0.250776646	0.499779215	0.432058191	*0.504548665*	**M1**	*0.02679038*	*0.00000319*	*0.00007816*	*0.00000248*
0.313647082	0.279560051	0.297776833	0.335595026	*M2*	*0.00516970*	*0.01318314*	*0.00810199*	*0.00266746*
0.295559616	0.444303815	0.436415839	0.470153381	*M3*	*0.00861058*	*0.00004602*	*0.00006489*	*0.00001405*
0.307338226	*0.526964957*	0.45615236	*0.526579671*	**M4**	*0.00619838*	*0.00000072*	*0.00002703*	*0.00000073*
0.14655107	0.413395249	0.327527356	0.416521697	M5	0.20042816	*0.00016869*	*0.00342096*	*0.00014876*
0.461706214	*0.632487516*	*0.545385767*	*0.58171873*	**M6**	*0.00002092*	*0.00000000*	*0.00000024*	*0.00000002*
0.461706214	*0.632487516*	*0.545385767*	*0.58171873*	**M8**	*0.00002092*	*0.00000000*	*0.00000024*	*0.00000002*
0.312573259	0.337482338	0.239521256	0.17955178	M9	*0.00533332*	*0.00251424*	*0.03467869*	0.11572604
−0.118732113	−0.233634449	−0.235143417	−0.297121758	M10	0.30050481	*0.03952542*	*0.03823230*	*0.00824944*
0.528473097	*0.572411677*	*0.54004426*	0.484477968	(M11)	*0.00000066*	*0.00000004*	*0.00000033*	*0.00000697*
0.107943667	0.335092627	0.292051491	0.382262879	M12	0.34686202	*0.00270963*	*0.00947247*	*0.00055295*
−0.312573259	−0.337482338	−0.239521256	−0.17955178	M13	*0.00533332*	*0.00251424*	*0.03467869*	0.11572604
−0.377620779	*−0.560379804*	−0.45309536	−0.449604072	**M14**	*0.00065361*	*0.00000010*	*0.00003107*	*0.00003636*
0.14655107	0.413395249	0.327527356	0.416521697	M15	0.20042816	*0.00016869*	*0.00342096*	*0.00014876*
0.461706214	*0.632487516*	*0.545385767*	*0.58171873*	**M16**	*0.00002092*	*0.00000000*	*0.00000024*	*0.00000002*
−0.206622992	−0.252218728	−0.456792919	−0.357623249	M28	0.06952041	*0.02589917*	*0.00002625*	*0.00130688*
−0.331262619	−0.358788061	−0.187904086	−0.216423497	M29	*0.00305124*	*0.00125669*	0.09946741	0.05701926
0.307338226	*0.526964957*	0.45615236	*0.526579671*	**M33**	*0.00619838*	*0.00000072*	*0.00002703*	*0.00000073*
0.249903773	0.4835291	0.416258197	0.492709387	*M34*	*0.02734232*	*0.00000731*	*0.00015035*	*0.00000460*
0.396126258	*0.585093091*	*0.553977608*	*0.538296529*	**M36**	*0.00033061*	*0.00000002*	*0.00000014*	*0.00000037*
0.374374484	0.475948434	0.454468078	0.477041622	*M37*	*0.00073363*	*0.00001062*	*0.00002919*	*0.00001007*
0.3247366	0.371440524	0.335130563	0.247971892	*M38*	*0.00372275*	*0.00081353*	*0.00270642*	*0.02859817*
0.452521733	0.243803491	0.348495525	0.240607547	*M40*	*0.00003189*	*0.03147444*	*0.00176731*	*0.03384106*
−0.192812387	−0.18845125	−0.306572197	−0.259299333	M42	0.09078144	0.09846809	*0.00633484*	*0.02187928*
0.077224946	−0.093211314	−0.161101947	−0.18722819	M43	0.50156753	0.41696473	0.15881594	0.10071277
0.050490972	−0.016426143	−0.192497427	−0.169618281	M45	0.66065799	0.88649618	0.09132022	0.13764370
−0.117374106	0.010997711	0.26419837	0.12810475	M46	0.30610626	0.92386697	*0.01942045*	0.26367943
0.351440918	0.480948885	0.431240883	0.461056731	*M47*	*0.00160486*	*0.00000831*	0.00008092	0.00002156
0.355479345	*0.512497152*	0.41311013	0.474325424	**M48**	*0.00140401*	*0.00000161*	*0.00017063*	*0.00001149*
−0.123905597	−0.13255876	−0.188730287	−0.23759727	M49	0.27978240	0.24729697	0.09796150	*0.03620483*
0.332234163	0.406888869	0.425521358	0.307895286	*M50*	*0.00296112*	*0.00021830*	*0.00010286*	*0.00610079*
−0.165159925	−0.10279826	0.016844025	0.029585012	M51	0.14844456	0.37046244	0.88362854	0.79708189
0.388046804	*0.550351609*	0.448792434	0.409884529	**M52**	0.00044738	0.00000018	0.00003771	0.00019400
0.41438646	0.47681866	0.377358045	0.326883068	*M53*	*0.00016212*	*0.00001018*	*0.00065977*	*0.00348862*

for example, was evaluated in two groups: at first position in one group and at second position in the other group. It showed better evaluations in the first group. Considering learning effects, one would intuitively expect the opposite.

5. CONCLUSION AND FURTHER WORK

The goal of this work was to investigate the possibility of applying ontology quality metrics on content ODPs and to validate such metrics. Table II shows metrics that can be calculated for ODPs and that have a significant and strong correlation with the formulated engineering principles of ‘Clarity’, ‘Understandability’, ‘Adaptability’ and ‘Reusability’. With the number of root concepts as direct descendants of *Thing* (M54), a simple metric with a strong correlation to three of the four criteria has been proposed.

For future work, the points listed in Section 4.3 need to be addressed. This includes a deeper investigation of possible interdependencies regarding the influence of ODP characteristics on perceived quality. Furthermore, there are metrics that do not seem appropriate for application on ODPs; namely, M7, M35 and M41 (see Section 4.1) regarding a quality scale. However, the existence of phenomena addressed by these metrics has an influence on perceived quality. This should be investigated using a different approach. A deeper view into the reasons why metrics like M43, M45 and M51 have poor performance for ODPs may reveal interesting results.

The evaluation of perceived quality is bound to the situation of the user, including among others the work environment, the tool support and the step in the ontology lifecycle. Thus, the findings of our investigations are true for situations similar to the experimental setting. There might be some different results when varying the situations.

The validation of additional metrics may be worthwhile too. A tool support for the selected metrics seems to be desirable for both practice and further research.

REFERENCES

Alexander C, Ishikawa S, Silverstein M. 1977. A Pattern Language. Oxford University Press.

Alm R, Kiehl S, Lantow B, Sandkuhl K. 2013. Applicability of quality metrics for ontologies on ontology design patterns. In KEOD 2013 – Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Vilamoura, Algarve, Portugal, 19–22 September, Filipe J, Dietz JLG (eds). SciTePress: Philadelphia; 48–57.

Baud R, Rodrigues J-M, Wagner J, Rassinoux A-M, Lovis C, Rush P, Trombert-Paviot B. 1997. Validation of concept representation using natural language generation. *Proceedings of the AMIA Annual Fall Symposium* 1997: 841.

Bevan N. 1995. Measuring usability as quality of use. *Software Quality Journal* **4**(2): 115–130.

Blomqvist E. 2009. Semi-automatic ontology construction based on patterns. PhD thesis, Linköping University, Linköping.

Blomqvist E, Sandkuhl K. 2005. Patterns in ontology engineering: classification of ontology patterns. In *Proceedings of the 7th International Conference on Enterprise Information Systems*, Miami.

Bolotnikova ES, Gavrilova TA, Gorovoy VA. 2011. To a method of evaluating ontologies. *Journal of Computer and Systems Sciences International* **50**(3): 448–461. DOI: 10.1134/S1064230711010072.

Buschmann F, Meunier R, Rohnert H, Sommerlad P, Stal M. 1996. Pattern-oriented Software Architecture. John Wiley & Sons: Chichester.

Djedidi R, Aufaure M-A. 2010. *ONTO-EVO^{A} L* an ontology evolution approach guided by pattern modeling and quality evaluation. In Foundations of Information and Knowledge Systems, Link S, Prade H (eds). Lecture Notes in Computer Science, vol. **5956**. Springer: Berlin; 286–305. DOI:10.1007/978-3-642-11829-6_19 (accessed 18 December 2014).

Duque-Ramos A, Fernandez-Breis J, Stevens R, Aussenac-Gilles N. 2011. OQuaRE: A SQuaRE-based approach for evaluating the quality of ontologies. *Journal of Research and Practice in Information Technology* **43**(2): 159–176.

Fowler M. 1997. Analysis Patterns. Addison Wesley: Menlo Park, Calif., Harlow.

Gamma E, Helm R, Johnson R, Vlissides J. 1995. Design Patterns. Addison-Wesley: Reading, Mass., Wokingham.

Gangemi A. 2005. Ontology design patterns for Semantic Web content. In The Semantic Web – ISWC 2005, Gil Y, Motta E, Benjamins VR, Musen MA (eds). Lecture Notes in Computer Science, vol. **3729**. Springer: Berlin; 262–276.

Gangemi A, Catenacci C, Ciaramita M, Lehmann J. 2005. Ontology evaluation and validation: an integrated formal model for the quality diagnostic task. Technical report, Laboratory for Applied Ontology, CNR, Rome, Italy. http://www.loa.istc.cnr.it/old/Files/OntoEval4OntoDev_Final.pdf (accessed 18 December 2014).

Gangemi A, Presutti V. 2009. Ontology design patterns. In Handbook on Ontologies, Staab S, Studer D (eds). International Handbooks on Information Systems. Springer: Berlin; 221–243.

Garrido T, Kumar S, Lekas J, Lindberg M, Kadiyala D, Whippy A, Weissberg J. 2014. e-Measures: insight into the challenges and opportunities of automating publicly reported quality measures. *Journal of the American Medical Informatics Association* **21**(1): 181–184.

Garvin DA. 1984. What does 'product quality' really mean. *Sloan Management Review* **26**(1): 25–43. http://sloanreview.mit.edu/article/what-does-product-quality-really-mean/ (accessed 18 December 2014).

Gorovoy V, Gavrilova T. 2007. Technology for ontological engineering lifecycle support. *Information Theories & Applications* **14**(1): 19–25.

Gruber T. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* **5**: 199–220.

Guarino N, Welty C. 2009. An overview of OntoClean. In Handbook on Ontologies, Staab S, Studer D (eds). International Handbooks on Information Systems. Springer, Berlin; 201–220.

Hallak JC, Schott PK. 2011. Estimating cross-country differences in product quality. *The Quarterly Journal of Economics* **126**(1): 417–474.

Hay D. 1995. Data Model Patterns: Conventions of Thought. Dorset House: New York.

Horridge M, Parsia B, Sattler U. 2009. Explaining inconsistencies in OWL ontologies. In Scalable Uncertainty Management, Godo L, Pugliese A (eds). Lecture Notes in Computer Science, vol. **5785**. Springer: Berlin; 124–137.

Jozsef B, Blaga P. 2014. Production quality control in the process of coating in an electrostatic field. *Procedia Technology* **12**: 476–482.

Jung HW, Kim SG, Chung CS. 2004. Measuring software product quality: a survey of ISO/IEC 9126. *IEEE Software* **21**(5): 88–92.

Maedche A, Staab S. 2002. Measuring similarity between ontologies. In Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, Gómez-Pérez A, Benjamins V (eds). Lecture Notes in Computer Science, vol. **2473**. Springer: Berlin; 251–263.

Puppe F. 2000. Knowledge formalization patterns. In *Proceedings of PKAW 2000*, Sydney.

Seyed A. 2012a. A method for evaluating ontologies – introducing the BFO-rigidity decision tree wizard. In Formal Ontology in Information Systems: Proceedings of the Seventh International Conference (FOIS 2012), Donelly M, Guizzardi G (eds). IOS Press: Graz; 191–204.

Seyed A. 2012b. Integrating OntoClean's notion of unity and identity with a theory of classes and types – towards a method for evaluating ontologies. In Formal Ontology in Information Systems: Proceedings of the Seventh International Conference (FOIS 2012), Donelly M, Guizzardi G (eds). IOS Press: Graz; 205–218.

Staab S, Erdmann M, Maedche A. 2001. Engineering ontologies using semantic patterns. In *Proceedings of the IJCAI-01 Workshop on E-business & The Intelligent Web*, Seattle, Goméz-Pérez M, Gruninger M, Stuckenschmidt H, Uschold M (eds): 174–185. http://ceur-ws.org/Vol-47/ONTOL2-Proceedings.pdf (accessed 18 December 2014).

Tartir S, Arpinar IB, Moore M, Sheth AP, Aleman-Meza B. 2005. OntoQA: metric-based ontology quality analysis. In *IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources*, vol. 9.

Tartir S, Arpinar IB, Sheth AP. 2010. Ontological evaluation and validation. In Theory and Applications of Ontology: Computer Applications, Poli R, Healy M, Kameas A (eds). Springer: Dordrecht; 115–130. Available online at DOI:10.1007/978-90-481-8847-5_5 (accessed 18 December 2014).

Clark P, Thompson J, Porter B. 2000. Knowledge patterns. In KR2000: Principles of Knowledge Representation and Reasoning, Cohn AG, Giunchiglia J, Selman B (eds). Morgan Kaufman: San Francisco, CA; 591–600.

Van der Aalst W, ter Hofstede AHM, Kiepuszewski B, Barros AP. 2003: Workflow Patterns. *In Distributed and Parallel Databases* **14**(1): 5–51. DOI: 10.1023/A:1022883727209.

Vrandečić D, Sure Y. 2007. How to design better ontology metrics. In The Semantic Web: Research and Applications, Franconi E, Kifer M, May W (eds). Springer: Berlin; 311–325.

Yang Z, Zhang D, Ye C. 2006. Evaluation metrics for ontology complexity and evolution analysis. In ICEBE 2006: IEEE International Conference on e-Business Engineering, Tam W-T, Onmg J-Y, Younas M (eds). IEEE Computer Society Press: Los Alamitos, CA; 162–170.