Chapter 9
Ontology (Network) Evaluation

Marta Sabou and Miriam Fernandez

Abstract Ontology evaluation refers to the activity of checking the technical quality of an ontology against a frame of reference. As such, it is of core importance for ontology engineering supporting scenarios such as ontology validation, knowledge selection, or the evaluation of knowledge extraction algorithms. In this chapter, we provide methodological guidelines for evaluating stand-alone ontologies as well as ontology networks. Our goal is not only to present the NeOn perspective on this issue but to also provide a practical outlook to the vast area of work in the area of ontology evaluation. Without performing an extensive state-of-the-art analysis of this research field, we aim to illustrate how various evaluation methods developed by the NeOn project, and not only, can be used at different stages of the evaluation process. We conclude the chapter with some concrete examples of performing ontology evaluation.

9.1 Motivation

Ontology (network) evaluation plays a key role in ensuring the quality of ontology networks, and it is employed within various ontology engineering scenarios. The main scenario is that of *ontology development*, namely the process during which the ontology is built. The goal in this case is to assess the quality and correctness of the obtained ontology. The process of ontology development can be achieved through different methods and the evaluation of the obtained ontology changes

M. Sabou (✉)
MODUL University Vienna, Am Kahlenberg 1, 1190 Vienna, Austria
e-mail: marta.sabou@modul.ac.at

M. Fernandez
Knowledge Media Institute (KMi), The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
e-mail: m.fernandez@open.ac.uk

M.C. Suárez-Figueroa et al. (eds.), *Ontology Engineering in a Networked World*,
DOI 10.1007/978-3-642-24794-1_9,

accordingly. For example, an ontology could be obtained through *automatic extraction* from representative data sources such as text (Cimiano and Völker 2005) or databases (Cerbah 2008). In this case, an important research question refers to evaluating ontology extraction algorithms with respect to the quality of the produced artifacts, as well as comparing the various algorithms to each other. Ontology evaluation can often be used as a means to automatically assess the quality of the output of such algorithms.

Alternatively, the ontology development phase could also involve an *ontology evolution* activity where a base ontology is extended, either manually or through automatic means, in order to cover new domain terminology or to correspond to new application requirements (Chap. 11). In this case, the goal of ontology evaluation is to assess whether the new additions have impacted on the quality of the base ontology.

Additionally to ontology development, another scenario where ontology evaluation plays an important role is that of *ontology selection*. With the recent advances in the area of the Semantic Web, in particular the proliferation of online available ontologies and semantic search engines such as Watson[1] or Sindice[2], an increased number of applications are built by reusing external knowledge rather than building it from scratch (d'Aquin, et al. 2008). Examples include cross-ontology question answering (Lopez et al. 2010), relation detection, ontology evolution (Zablith et al. 2010), or ontology matching (Sabou et al. 2008). For these applications, it is crucial to evaluate, often entirely automatically, the quality of the reused knowledge. Ontology evaluation here refers to the situation where existing ontologies are evaluated (and often ranked) in terms of selected criteria in order to select the most appropriate one for the task at hand.

A final usage scenario is during the *ontology modularization* process that leads to a network of interconnected ontology modules (Chap. 10), whose quality is iteratively assessed in order to decide whether the modularization has reached the expected results.

In this chapter, we further explore ontology (network) evaluation by providing a definition (Sect. 9.2), methodological guidelines (Sect. 9.3), and concrete examples (Sect. 9.4).

9.2 Definitions and Filling Card

Ontology evaluation is defined as *the activity of checking the technical quality of an ontology against a frame of reference* (Suárez-Figueroa and Gómez-Pérez 2008). Intuitively, whenever an evaluation is performed for a certain ontology

[1] http://kmi-web05.open.ac.uk/WatsonWUI/

[2] http://sindice.com/

(or alignment) aspect (e.g., modeling correctness), the process is always guided by the evaluator's understanding of what is best and what is worse. In some cases, these boundaries (which we refer to as *frame of reference*) are clearly defined and tangible (e.g., a reference ontology, a reference alignment), but in other cases, they are weakly defined and may be different from one person to another, or even across evaluation sessions. The NeOn Glossary distinguishes two types of ontology evaluations depending on the frame of reference used:

- *Ontology validation* is the ontology evaluation activity that compares the meaning of the ontology definitions against the intended model of the world that it

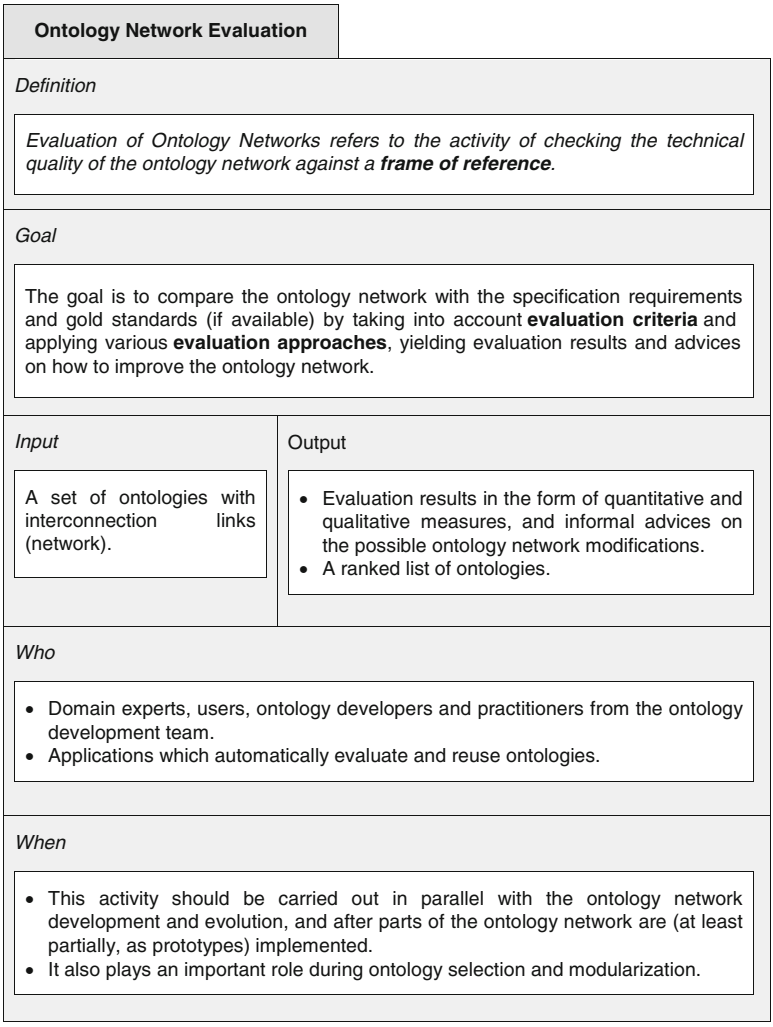

Fig. 9.1 Filling card for ontology (network) evaluation

aims to conceptualize (an intangible frame of reference). This activity answers the question: *are you producing the right ontology?*

- *Ontology verification* is the ontology evaluation activity which compares the ontology against the ontology specification document (ontology requirements and competency questions), thus ensuring that the ontology is built correctly (in compliance with the ontology specification). This activity answers the question: *Are you producing the ontology in the right way?*

The *filling card* shown in Fig. 9.1 provides a structured summary of the ontology (network) evaluation activity. Section 2.5 describes the main components of a filling card in more detail.

9.3 Ontology Network Evaluation Workflow and Guidelines

In this section, we describe the NeOn methodological guidelines for carrying out the ontology network evaluation activity. Besides prescribing a methodology, our aim is also to provide a brief overview of the various evaluation methods and techniques that can be used in each step of the methodology.

We propose a component-based evaluation approach where each element of the network (e.g., ontologies and alignments between ontology pairs) is evaluated as a stand-alone individual and then the findings of these evaluations are summed up (Fig. 9.2). An alternative to this approach would be the evaluation of the entire network from the point of view of the users or the organization that will use the ontology network. Methodologically, this approach is similar to evaluating a stand-alone component using, for example, a task-based evaluation, and therefore, it is covered by Tasks 2 and 3 of the proposed workflow. Figure 9.2 shows the *workflow* and the tasks for carrying out the ontology network evaluation.

Task 1. Selecting individual components of the ontology network. In a first instance, the ontology development team identifies the elements of the network that need to be evaluated including individual ontologies (Maedche and Staab 2002; Burton-Jones et al. 2005; Alani et al. 2006; Fernandez et al. 2006), alignments between ontology pairs (Euzenat and Shvaiko 2007), ontology statements (Lopez et al. 2009), ontology relations, etc. Their decision should be based on two criteria: (1) which ontology network elements are critical for the overall network and (2) which of these elements can actually be evaluated. The latter means that there must exist some frame of reference against which these individual components can be, at least in principle, evaluated. As we discussed before, the frame of reference is not necessarily tangible, but can be some idea of the perfect model, or canon, defined by the human evaluator for the particular evaluation task. Examples of frames of references will be given at Task 3.

Task 2. Selecting an evaluation goal and approach. For evaluating individual ontologies, the team needs to decide the goal of the evaluation and select an

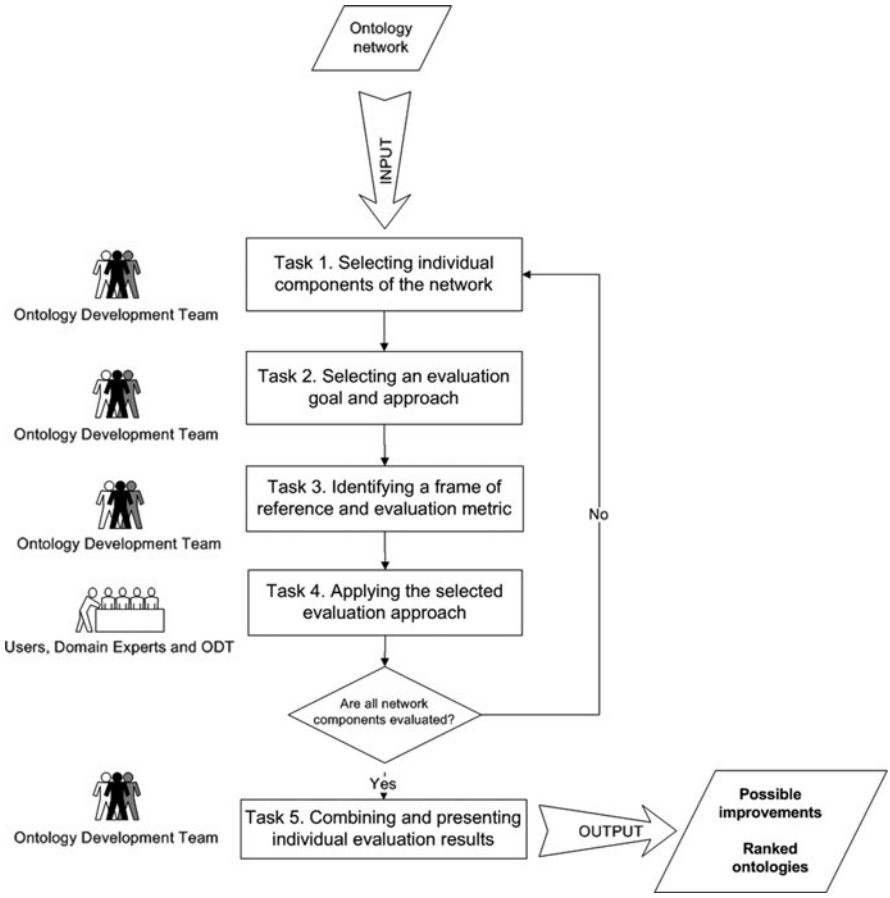

Fig. 9.2 Workflow and tasks for evaluating ontology networks

appropriate evaluation approach (as summarized in Table 9.1). We distinguish the following evaluation goals:

- *Domain coverage – Does the ontology cover a topic domain?* The extent to which an ontology covers a considered domain is an important factor to be considered both during the development and the selection of an ontology. The evaluation approaches employed to achieve this goal imply the comparison of the ontology to frames of references such as a gold standard ontology (Maedche and Staab 2002), or data sets that are representative for the domain (user-defined terms (Alani et al. 2006; Fernandez et al. 2006), tag sets (Cantador et al. 2007), document corpus (Brewster et al. 2004), etc.).
- *Quality of the modeling* in terms of the design and development process and in terms of the final result – *Does the ontology development process comply with*

Table 9.1 Evaluation goals, evaluation approaches, and relevant NeOn plugins

Evaluation goal	Evaluation approaches and relevant NeOn plugins
Domain coverage	Compare to a domain-specific gold standard ontology (Maedche and Staab 2002)
	Compare to unstructured or informal data (Brewster et al. 2004; Jones and Alani 2006)
	Compare to a user-defined set of terms – Sindice, Watson (Alani et al. 2006)
	Compare to an extended (using WordNet or other structured information sources) user-defined set of terms (Fernandez et al. 2006; Cantador et al. 2007)
Quality of modeling	Use human assessments to evaluate the syntactic, structural, and semantic quality of the ontology (Guarino and Welty 2004; Lozano-Tello and Gómez-Pérez 2004; Burton-Jones et al. 2005)
	Use reasoners to assess the logical correctness of the ontology (Horridge et al. 2009)
	Analyze the design and development process of the ontology to check its compliance with ontology modeling best practices/ODPs (Caracciolo and Heguiabehere 2009; Poveda-Villalón et al. 2009)
	Automatically compare to a reference alignment (Euzenat and Shvaiko 2007)
	Manually assess the quality of an alignment (Sabou et al. 2008)
	NeOn plugins:
	RaDON
	XDTools
	Alignment plugin
Suitability for an application/task	Use the ontology within an application/task and evaluate the task results and performance (Porzel and Malaka 2004; Strasunskas and Tomassen 2008; Fernandez et al. 2009)
	The work of Van Hage (Van Hage et al. 2007) presents two sampling-based evaluation approaches of ontology alignments
Adoption and use	Evaluation of the interlinking structure across ontologies – Sindice, Watson (Patel et al. 2003)
	Social rating systems (Lewen et al. 2006; Cantador et al. 2007)
	NeOn plugin:
	Watson for knowledge reuse

ontology modeling best practices/ODPs[3]*? Is the ontology modeled correctly?* Applicable both for the ontology development (Lozano-Tello and Gómez-Pérez 2004) and selection scenarios (Burton-Jones et al. 2005; Tartir et al. 2005), this evaluation goal focuses on the quality of the ontology which can be assessed using a wide range of approaches focusing on logical correctness or syntactic, structural, and semantic quality. Quality in terms or correctness, precision, and recall is an important goal when evaluating ontology alignments.

[3] ODP stands for Ontology Design Pattern.

- *Suitability for an application/task – Is the ontology suitable to use for a specific application/task?* (Porzel and Malaka 2004; Fernandez et al. 2009) *Will it produce the expected results?* (Strasunskas and Tomassen 2008) Different applications rely on different ontology (or alignment) characteristics. For example, for applications that use ontologies to support natural language processing tasks, domain coverage is often more important than logical correctness. As a result, measuring ontology (alignment) quality alone is not enough to predict how well the ontology (developed or selected) will support an application or a task. Task-based evaluations help assessing suitability for a task or application, rather than generic quality features.
- *Adoption and use – Has the ontology been reused (imported) as part of other ontologies?* (Sindice,[2] Watson[1]) *How did others rate the ontology?* (Cantador et al. 2007, Cupboard[4]) Understanding the extent of adoption of an ontology is of particular interest when selecting it, the assumption being that there is a direct correlation between the level of adoption and the quality of the ontology. Analyzing the degree of interlinking between an ontology and other ontologies (e.g., in terms of reused terms or ontology imports) as well as relying on social rating systems are two key approaches to achieve this goal.

Task 3. Identifying a frame of reference and evaluation metric. While in Task 2 the ontology development team decides on the key goal(s) of the evaluation and potential approaches, in Task 3, the team needs to select the concrete ingredients of the evaluation, consisting of:

- *A frame of reference – What are we comparing against?* The frame of reference denotes a set of representative resources that sets a baseline value against which the ontology should be compared.
- *Evaluation metric(s) – How to measure the features of the ontology that will be compared?* Example evaluation metrics are precision and recall, cost-based evaluation metrics, measures of similarity between an ontology or a mapping, and a corpus (domain knowledge), and lexical metrics. Table 9.2 summarizes the main evaluation metrics presented in the literature.

As exemplified in Table 9.2, evaluation metrics are generally specific for each frame of reference. There are however some generic metrics, such as precision and recall, which can be adapted for use with various frames of references.

Similarly to (Brank et al. 2005), we distinguish the following types of frames of references:

- *Gold standard*: The frame of reference is defined by a baseline ontology or some other kind of structured representation of the problem domain for which an appropriate ontology is needed. A gold standard is often used when the goal of

[4] http://cupboard.open.ac.uk:8081/cupboard-search/

Table 9.2 Evaluation metrics used for various evaluation frameworks

Frame of reference	Evaluation metric/approach
Gold standard	*Interpretability:* amount of terms of the ontology that have a WordNet[a] sense (Burton-Jones et al. 2005)
	Clarity: amount of WordNet senses of the ontological terms (Burton-Jones et al. 2005)
	Lexical similarity: average string matching between the set of gold standard terms and the set of ontology terms (Maedche and Staab 2002)
	Taxonomical similarity: maximum overlap between the concepts of the gold standard and the concepts of the ontology in terms of their "semantic cotopy" (their sets of super- and subconcepts) (Maedche and Staab 2002)
	Relation similarity: overlap between the relations of the gold standard and the relations of the ontology considering the geometric mean value of how similar their domain and range concepts are (Maedche and Staab 2002)
	Precision and recall of an alignment with respect to a reference alignment (gold standard): precision measures the ratio of correctly found correspondences (true positives) over the total number of returned correspondences (true and false positives). Recall measures the ratio of correctly found correspondences (true positives) over the total number of expected correspondences (true positives and true negatives) (Euzenat 2007)
	Semantic precision/semantic recall for alignment evaluation: This measure proposes an abstract generalization of precision and recall to discriminate among different degrees of alignment correctness (Euzenat 2007)
Application-based	*History:* number of times an ontology has been accessed
	Insertion, deletion, and substitution errors: errors according to the improvements in the task's output after fixing these errors in the employed ontology (Porzel and Malaka 2004)
	Search task fitness and *search enhancement capability:* these measures evaluate ontology quality in the context of an ontology-driven web search task (Strasunskas and Tomassen 2008)
	Watson's topology measures: these measures are used in the context of a relation correctness evaluation task (Fernandez et al. 2009)
Data-driven	*Class match:* coverage of an ontology with respect to a set of search terms (Alani et al. 2006)
	Best fit ontology: ontology that maximizes its conditional probability given a corpus. The probability is computed considering the terms and document clusters within the corpus (Brewster et al. 2004)
Assessment by humans	*Syntactic quality:* number of syntactical errors in the ontology (Burton-Jones et al. 2005)
	Accuracy: number of false statements in the ontology (Burton-Jones et al. 2005)
	Trust: correctness and usefulness of the information delivered by a certain reviewer with respect to the ontology (Lewen et al. 2006). This measure is defined and exploited in collaborative systems (d'Aquin et al. 2009)
	Collaborative evaluation: collaborative assessment of ontologies based on manual user evaluation (Cantador et al. 2007)
	Essence: assess if an entity is true in every possible world (Guarino and Welty 2004)

(continued)

Table 9.2 (continued)

Frame of reference	Evaluation metric/approach
	Identity: assess if individual entities of the world are the same or different (Guarino and Welty 2004)
	Unity: recognizes all the parts that form an individual entity (Guarino and Welty 2004)
Topology-based	*Topology of the graph:* set of topological evaluation measures including *number of classes, number of properties, number of individuals, ontology popularity* (number of ontologies importing a given ontology), and *ontology depth and breadth* (maximum, minimum, average, and variance); extracted from Watson
	Density: number of subclass, sibling, and domain relations of a given concept (Alani et al. 2006)
	Semantic similarity: closeness of the concepts of interest in the ontology structure (Alani et al. 2006)
	Betweenness: number of paths that pass through each node of the ontological graph (Alani et al. 2006)
	Comprehensiveness: number of classes and properties of an ontology (Burton-Jones et al. 2005)
	Authority: normalized value of times that an ontology is imported in the network (Burton-Jones et al. 2005)
	OntoRank: ranks ontologies based on the interlinking structure among ontologies in the network. Different versions of the similar evaluation principle are found in (Patel et al. 2003; Ding et al. 2005)
	Relationship richness: diversity of relations and placement of relations in the ontology (Tartir et al. 2005)
	Attribute richness: average number of properties per class (Tartir et al. 2005)
	Inheritance richness: average number of subclasses per class (Tartir et al. 2005)
	Class richness: ratio between the number of classes that contain instances divided by the total number of classes in the ontology (Tartir et al. 2005)
	Average population: ratio between the number of ontology instances and classes (Tartir et al. 2005)
	Cohesion: number of separated, connected components of the ontological graph (Tartir et al. 2005)
Language-based	Thirty-eight modeling language-specific criteria: if the *language allows axioms embedded in terms*, can *define disjoint decompositions*, etc. (Lozano-Tello and Gómez-Pérez 2004)
Methodology-based	Eleven methodology-based evaluation metrics: precision factors (e.g., the *delimitation of phases in the ontology construction)*, usability factors (e.g., the *quality of manuals*), and maturity factors (e.g., the *importance of the developed ontology*) (Lozano-Tello and Gómez-Pérez 2004)

[a]http://wordnet.princeton.edu/

the evaluation is *domain coverage*. For alignments, a reference alignment can play the role of a gold standard.

- *Application-based*: The frame of reference consists of the set of "ideal" results that an application should return when plugging the "perfect" ontology

(or alignment) into it. This frame of reference pertains to the assessment of the ontology's (alignment's) *suitability for an application/task*.
- *Data-driven*: The frame of reference is a collection of unstructured or informal data (e.g., text), which represents the problem domain. Similarly to structured representations used as gold standards, unstructured data collections are also mostly used to support the evaluation of *domain coverage*.
- *Assessment by humans*: The frame of reference is defined by human judgments that measure ontology features (or alignment characteristics) not recognizable by machines. Humans can (relatively) easily assess several *ontology quality* features which are not amenable to automatic processing. Human ratings also help to assess the level of *adoption and use* of the ontologies. Human-based ontology ratings are exploited to automatically select the most appropriate ontology according to previous users' experiences (Cantador et al. 2007).

Additionally, and based on the way in which human evaluators assess *ontology quality* features (by comparison with their mental idea of the perfect model or canon for these features), we have identified the next three nontangible frames of references as ideal models of topologies, languages, and ontology-construction methodologies, which constitute the boundaries within which comparisons are based when performing the evaluations: (a) the ontology with the optimal topology, (b) the potentially most powerful and expressive ontology language, and (c) the perfect set of steps to follow and requirements to fulfill in order to achieve the best modeled ontology. All these canons or ideal models of topologies, languages, and methodologies are weakly defined since they may vary across evaluations and across the evaluators who defined them.

- *Topology-based:* The frame of reference is defined by the minimum or maximum possible values of the topology evaluation metrics among ontologies within the network, or among ontology entities within the same ontology. Topology metrics automatically assess *ontology quality* features as well as *adoption and use* features, by measuring the interlinking structure of ontologies across the network (Ding et al. 2005).
- *Language-based:* The frame of reference is defined by the representational capabilities of the language used to construct the ontology.
- *Methodology-based:* The frame of reference is defined by the different quality factors of the selected ontology-development methodology.

Task 4. Applying the selected evaluation approach. Applying the selected evaluation approach requires a proper setup for the evaluation experiments and implementation of software tools to compute the evaluation metrics, and/or engage the human experts in stimulating sessions to collect their evaluations. We advise ontology developers to refer to the relevant scientific publications cited in this chapter for example evaluation setups and best practices. Evaluation approaches that rely on human judgment (Guarino and Welty 2004; Lozano-Tello and Gómez-Pérez 2004) are generally more time consuming and sophisticated than those which compare numeric values derived by automatic measures (Sindice, Watson),

although they often offer more valuable insight into the evaluation process. We advise using parallel evaluation with multiple human experts to account for cross-evaluator disagreements.

Task 5. Combining and presenting individual evaluation results. This task highlights the weakest spots in the ontology network by considering individual evaluation results and how they affect the rest of the network. The evaluation results derived for individual components are combined to reach a global understanding of the network's quality. The final task is to present the results of the evaluation in an appropriate form for possible repair (corrections, additions), improvements, and future evolution of the ontology network.

9.4 Examples of Ontology Evaluation

Since ontology network evaluation is not a widespread activity as yet, in this section, we present examples of various ontology evaluation studies and show how their stages map to the tasks prescribed by our guidelines. The examples cover all the key evaluation goals described in Task 2: domain coverage (Sect. 9.4.3), quality of modeling (Sects. 9.4.1 and 9.4.2), suitability for an application (Sects. 9.4.3 and 9.4.4), and adoption (Sect. 9.4.5).

9.4.1 Evaluation of an Individual Ontology

In this example, we describe the evaluation of YAGO (Suchanek et al. 2008), a large, lightweight, general-purpose ontology, automatically derived from Wikipedia and WordNet. YAGO has over 1.7 million entities (individuals and concepts) and 15 million facts (ground binary relations between entities). The relations include the taxonomic hierarchy as well as around 100 semantic relations between entities. YAGO's evaluation follows the main tasks of our methodology.

[*Task 2*] Since the evaluation was performed in an ontology development scenario, the authors' goal was to assess the *quality of modeling* of YAGO, namely its precision with respect to the data sets from where it has been derived. The approach was that of evaluating the precision by using human expert opinion.

[*Task 3*] To evaluate the *precision* of an ontology, its facts have to be compared to some ground truths. Since there is no computer-processable ground truth of suitable extent to be used as a *frame of reference*, the authors relied on manual evaluations against Wikipedia content, which was the frame of reference.

[*Task 4*] During the evaluation, human judges rated as "correct," "incorrect," or "don't know" facts that were randomly selected from YAGO. Since common sense often does not suffice to judge the correctness of the YAGO facts, a snippet of the corresponding Wikipedia page was also presented to the judges. Thus, the evaluation compared YAGO against the ground truth of Wikipedia (i.e., it does not deal

Table 9.3 Precision of some YAGO facts

Relation	No. of evaluation	Precision
1 hasExpenses	46	100.0% ± 0.0%
2 hasInflation	25	100.0% ± 0.0%
3 hasLaborForce	43	97.67441% ± 0.0%
4 during	232	97.48950% ± 1.838%
...		
88 hasGDPPPP	75	91.22189% ± 5.897%
89 hasGini	62	91.00750% ± 6.455%
90 discovered	84	90.98286% ± 5.702%

with the problem of Wikipedia containing some false information). Thirteen judges evaluated a total of 5,200 facts (ground relations between YAGO entities).

[*Task 5*] The authors use a tabular format (Table 9.3) to present the evaluation results in the decreasing order of the obtained precision (we only show the most and least precise relations). To make sure that the findings are significant, the Wilson confidence interval for $\alpha = 5\%$ was computed. A confidence interval of 0% means that the facts have been evaluated exhaustively. The evaluation shows very high quality results as 74 relations have a precision of over 95%.

This tabular presentation helps identifying the least precise relations and fosters the analysis of such cases. It can be concluded, for example, that a key source of error are inconsistencies of the underlying sources. For example, for the relation *bornOnDate*, most false facts stem from erroneous Wikipedia categories (e.g., persons born in 1802 are in the *1805 Births* Wikipedia category). For facts with literals (such as *hasHeight*), many errors stem from a nonstandard format of the numbers (e.g., height is considered 1.6 km, just because the infobox says 1,632 m instead of 1.632 m). Occasionally, the data in Wikipedia was updated between the time of extraction and the time of the evaluation. This explains many errors for frequently changing properties such as *hasGDPPPP* and *hasGini*.

9.4.2 Pattern-Based Ontology Evaluation

In this section, we show how ontology design patterns, specifically content design patterns (CPs), are used to evaluate an ontology. The example does not cover the complete evaluation of the ontology, but presents one specific case where a CP assisted in finding potential problems and additionally suggested a solution. The example is set within the fishery domain, and the evaluated ontology is version 0.3 of the "fishing areas" ontology, modeling the division of water areas into divisions and subdivisions. An example is the FAO major fishing area 51, Western Indian Ocean, and its subareas numbered from 1 to 8, where 1 corresponds to the Red Sea and 2 to the Persian Gulf, but where the subdivisions of these subareas are only numerically identified.

[*Task 2*] The goal of the evaluation was assessing the *quality of modeling*, and the chosen approach was manual evaluation by an ontology pattern expert.

[*Task 3*] The expert used the pattern catalog available in the ontology design pattern portal[5] as a "gold standard" of modeling to which the modeling solutions in the evaluated ontology were compared. CPs introduce best practices for solving particular modeling problems, but by introducing those solutions, the pattern catalog can also be seen as a catalog of modeling issues.

[*Task 4*] The ontology used a locally defined, transitive, "part-of" relation to model the division of subareas and further levels of divisions and subdivisions, thus using the same modeling approach as the "part-of" content pattern. This modeling solution, however, is not suitable for certain contexts, because, when using reasoning, it is not possible to distinguish between the direct and the indirect subparts of an area. For example, if the hierarchical structure of the partitioning of the areas should be reconstructed, for example, for browsing the ontology in a graphical interface, or when answering "what are the divisions of the Red Sea?," only the direct subareas of the Red Sea are of interest rather than all the inferable parts.

The "componency pattern" provides a modeling alternative using two inverse object properties: "hasComponent" and "isComponentOf." These are nontransitive properties that can be used in combination with the "part-of pattern" to both register general partitioning but also the nontransitive property of a "proper part," i.e., a direct component of something. When using these two patterns as "gold standards" for modeling, the ontology evaluator can discover the potential problem of a missing nontransitive property to distinguish the different "levels" of area decomposition and propose an appropriate solution.

9.4.3 Multiple Evaluations of an Ontology

An example of how various types of evaluations shed light on different aspects of an ontology is provided in (Sabou et al. 2005). Similar to this, when evaluating ontology networks, one needs to combine evaluation results for various network components. The authors of (Sabou et al. 2005) report on the multifaceted evaluation of an ontology that was automatically extracted from a corpus of textual web service descriptions in the bioinformatics domain. The various stages of this evaluation are graphically depicted in Fig. 9.3. The aim of the *extracted ontology* is to support the semantic description of web services. The *myGrid*[6] project provided a good context to evaluate this ontology as a bioinformatics expert has previously built a *gold standard* ontology for describing the same set of web

[5] http://www.ontologydesignpatterns.org

[6] http://www.mygrid.org.uk

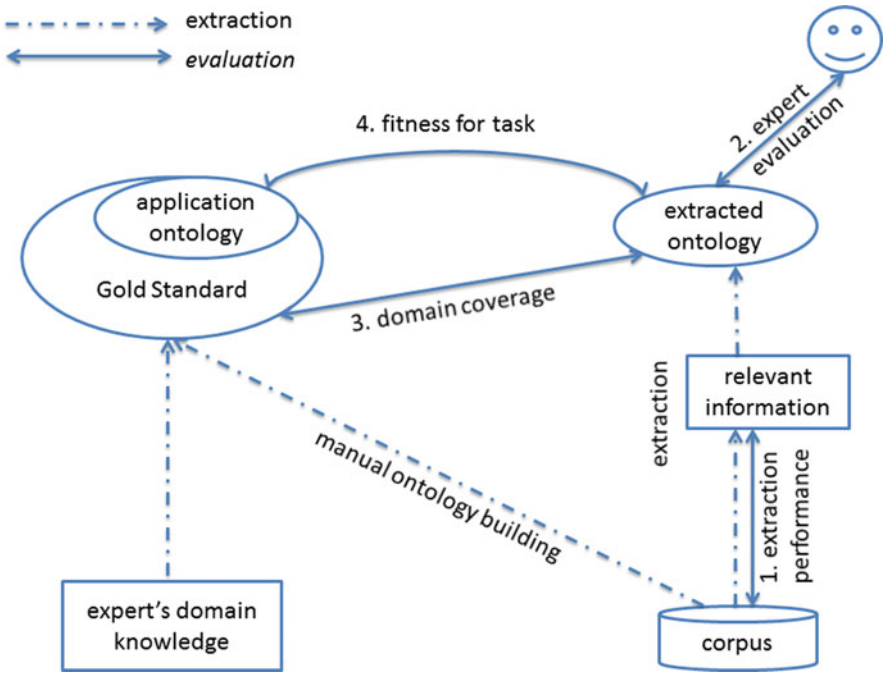

Fig. 9.3 Overview of various evaluations of an ontology (Sabou et al. 2005)

services. The domain expert has relied on his *domain knowledge* to build the ontology rather than on the description of web services (*corpus*), which were used as the main input for the automatic extraction algorithm. A part of the gold standard ontology, referred to as the *application ontology*, provides concepts for annotating web service descriptions in a form-based annotation tool and is subsequently used at web service discovery time to power the search.

[*Task 2*] In this ontology development scenario, the evaluations had several complementary goals. First, the authors aimed to assess whether the extracted ontology would be a good starting point for building an ontology and relied on an expert evaluation approach for this (shown as evaluation 2 in Fig. 9.3). Second, they wanted to evaluate *domain coverage* by comparison to the gold standard ontology (shown as evaluation 3 in Fig. 9.3). Third, the authors got an insight into how well the ontology would *support an application* by comparing it with the application ontology.

[*Task 3*] The authors made use of the following frames of references and metrics. For evaluation 2, the frame of reference consisted in the expert's knowledge of the domain as he was asked to review and rate the extracted concepts as either *correct* or *spurious* or *new*. A precision value was then computed as a ratio of the correct and new concepts over all extracted concepts. For evaluation 3, the authors used the gold standard ontology as a frame of reference and computed metrics such as *lexical overlap* (LO – the ratio of overlapping concepts), *ontological improvement*

Table 9.4 Results for domain coverage and task fitness from (Sabou et al. 2005)

Concepts	Gold standard	Application ontology
All	549	125
Correct	39	25
All_{missed}	510	100
$Missed_{corpus}$	3 (0.6%)	0 (0%)
$Missed_{external}$	360 (70.6%)	88 (88%)
$Missed_{abstract}$	101 (19.8%)	6 (6%)
$Missed_{composed}$	46 (9%)	6 (6%)
New	306	27
LO	7%	20%
OL	93%	80%
OI	56%	21.5%

(OI – the ratio of new concepts that were not in the gold standard but were domain relevant), and *ontological loss* (OL – the ratio of gold standard concepts which were not extracted). For evaluation 4, the application ontology was used as a frame of reference and compared to the extracted ontology using the metrics defined for evaluation 3.

[*Task 4*] Task 4 consisted in the evaluation performed by the domain expert as well as the computation of the various ontology comparison metrics.

[*Task 5*] The authors sum up the results of the various evaluations in tabular form and perform a subsequent analysis of these results. For example, Table 9.4 sums up the results when assessing domain coverage and suitability for a task by comparing the extracted ontology to the gold standard and application ontologies. The results show that although the overlap with the gold standard is low (7%), the extracted ontology contains a significant number of new, domain-relevant concepts (56%) that were identified in the automatically analyzed corpus but missed by the domain expert, which relied exclusively on his domain knowledge. A detailed analysis of all the missed concepts when comparing to the gold standard ontology shows that 70.6% of these terms did actually not appear in the corpus (but could be acquired if the corpus would be enlarged) and 19.8% referred to abstract concepts introduced by the domain expert to structure the ontology and which again were not in the corpus. It turns out that extraction algorithm–related issues only account for only 10% of the missed concepts.

9.4.4 Task-Based Ontology Evaluation

The authors of (Strasunskas and Tomassen 2008) investigate which ontology features influence the web search task. In their study, they consider different types of search tasks (fact-finding, exploratory search, comprehensive search), identify ontology features important for each task, and then introduce new evaluation metrics that measure these features respectively (e.g., fact-finding fitness

(FFF), exploratory search task fitness (EXF)). Such metrics can support ontology selection for search. Their theoretical considerations are experimentally verified, by correlating the values of the metrics for different ontology versions with the search performance obtained in the context of the WebOdIR web search application (Strasunskas and Tomassen 2008). Core to their study is therefore a task-based evaluation of ontologies.

[*Task 2*] The goal is to understand the *suitability for a task*, and the approach consists in exploiting ontologies to support web search and measuring the improvement in terms of search precision obtained in an experimental setting.

[*Task 3*] The frame of reference is defined by the *performance scores obtained in a web search task* with an original version of the ontology. The metrics used measure ontology features important for certain search tasks (e.g., FFF, EXF).

[*Task 4*] The experimental setup consists of relying on two groups of users to perform web search using WebOdIR within four different domains (two search tasks per domain, i.e., eight tasks in total). WebOdIR exploited a set of ontologies for one group and the extended version of the same ontologies for the second group. The performance score of the search task is computed and compared across the two versions of the ontologies as well as correlated with the computed values of the newly introduced metrics.

[*Task 5*] The authors present these correlations in both tabular and graphical form and conclude on the influence of ontology features on various search tasks. For example, they found that more instances and object properties improve fact finding, while the addition of disjoint and equivalent concepts is beneficial for explanatory and comprehensive search tasks.

9.4.5 Evaluating Ontology Adoption and Use

The work of Cantador and colleagues (Cantador et al. 2007) presents a tool for collaborative ontology evaluation and reuse (WebCORE) focused on evaluating *domain coverage* and *adoption and usage*. The goal of this tool is to help experts and practitioners to select the most appropriate ontologies from a repository. The tool has three main components. The first one helps the user to semiautomatically generate a gold standard representing the domain of interest. The second component evaluates the domain coverage of the ontologies by comparing them against the previously generated gold standard by means of lexical and taxonomical evaluation measures. The third component exploits previous users' judgments of those ontologies to automatically recommend the best ones.

[*Task 2*] Two main evaluation goals are considered when selecting the optimal ontology: (a) the *domain coverage* and (b) the *adoption and use* of the ontology.

[*Task 3*] To evaluate *domain coverage*, authors select a gold standard as a frame of reference. This gold standard is a representation of the domain of interest and is semiautomatically generated by the user with the support of the tool. To generate it, the user (a) introduces an initial set of terms or selects a textual source from which a

set of terms representing the domain of interest can be extracted, (b) complements this set of terms by selecting additional terms from a ranked list, automatically generated by the system by considering previous user-generated gold standards, and (c) extends this set of terms by selecting suggested hypernym, hyponym, and synonym relations from WordNet. To evaluate the *adoption and use* of the ontologies, this work relies on an *assessment by humans*' frame of reference. Users share their own experiences by evaluating the used ontologies according to five criteria: correctness, readability, flexibility, level of formality (highly informal, semi-informal, semiformal, and rigorously formal), and type of model (upper-level, core-ontology, domain-ontology, task-ontology, and application-ontology).

[*Task 4*] The tool evaluates the ontologies in two phases. First, the ontologies are evaluated according to their domain coverage by comparing them against the semiautomatically generated gold standard using lexical and taxonomical similarity measures. Second, the ontologies with sufficient domain coverage are assessed on their level of adoption and use with the help of a collaborative filtering algorithm (Adomavicius and Tuzhilin 2005) that explores the manual evaluations of the ontologies stored into the system. This algorithm takes into account not only previous users' experiences (usage) but also the number of times the ontologies were selected (adoption).

[*Task 5*] The representation of the results differs for the two types of evaluations. For domain coverage, the tool presents a ranked list of ontologies including their individual scores for the lexical and taxonomical evaluation measures, as well as a combined evaluation score. After the adoption and usage evaluation, the list of ontologies is reranked, and the collaborative ontology evaluation score is added to the previous scores. In addition, the system allows the user to provide her own judgment of the ontology so that her assessment can be exploited for future ontology evaluations and selections.

9.5 Relevant NeOn Toolkit Plugins

Given the complexity of the ontology evaluation task in terms of the variety of approaches and metrics, the NeOn Toolkit does not provide an evaluation plugin per se. However, various plugins exist that can support different evaluation approaches. We provide a brief description of these plugins here.

The *RaDON* plugin[7] supports the automatic detection of logical inconsistency and incoherence in an ontology or an ontology network. The plugin does not only detect these modeling errors but can also repair them automatically or support the user to manually solve these issues. As such, RaDON can support users whose goal is to assess the *quality of modeling* in their ontology.

[7] http://www.neon-toolkit.org/wiki/2.3.1/RaDON

The *XDTools* plugin[8] contains a suite of tools that support design pattern–based ontology development. One of the tools, XD Analyzer, provides suggestions and feedback to the user with respect to how good practices in ontology design have been followed, according to the eXtreme Design (XD) method (for instance, missing labels and comments, isolated entities, unused imported ontologies). Chapter 3 provides more information about the XD method. Similarly to RaDON, this plugin can also be used when checking the *quality of modeling*; however, the focus here is the quality of the domain conceptualization rather than logical correctness.

The *Watson for knowledge reuse*[9] plugin primarily supports knowledge reuse by allowing an ontology developer to search the Watson ontology search engine for relevant knowledge statements directly from within the NeOn Toolkit and then reuse those statements. The plugin also interfaces with the Cupboard ontology publication environment that allows users to rate various characteristics of the ontologies that they reused (e.g., reusability, correctness, completeness, domain coverage, modeling style). Individual ratings are aggregated into an overall score and can support other people when reusing ontologies. This plugin supports the evaluation of ontologies in terms of their *adoption and use* providing also reviews written by previous adopters.

9.6 Summary

Ontology evaluation is an important and complex ontology engineering activity. Its complexity stems both by its applicability in a variety of scenarios (Sect. 9.1) as well as the abundant number of existing approaches and metrics. In this chapter, we aimed at providing practitioners with the right balance of generic guidelines and specific techniques that they could use from the wide landscape of works in this area (Sect. 9.2). We hope that the five diverse evaluation examples in Sect. 9.3 will serve as useful material for exemplifying the proposed guidelines.

Although ontology networks contain both ontologies and their links in terms of alignments, we have mostly focused on ontology evaluation. Readers interested in ontology alignment evaluation should also consult Chap. 12. Finally, Chaps. 10 and 11 describe other ontology engineering activities that can benefit from ontology evaluation, namely ontology modularization and evolution.

[8] http://www.neon-toolkit.org/wiki/2.3.1/XDTools

[9] http://www.neon-toolkit.org/wiki/2.3.1/Watson_for_Knowledge_Reuse

References

Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17(6):734–749

Alani H, Brewster C, Shadbolt N (2006) Ranking ontologies with AKTiveRank. In: 5th international Semantic Web Conference (ISWC 2006), Athens, GA, USA, pp 1–15

Brank J, Grobelnik M, Mladenić D (2005) A survey of ontology evaluation techniques. In: Conference on Data Mining and Data Warehouses (SiKDD 2005), Ljubljana, Slovenia, pp 166–170

Brewster C, Alani H, Dasmahapatra S, Wilks Y (2004) Data driven ontology evaluation. In: 4th international conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, pp 164–169

Burton-Jones A, Storey VC, Sugumaran V, Ahluwalia P (2005) A semiotic metrics suite for assessing the quality of ontologies. Data & knowledge engineering – Special issue: Natural Language and Database and Information Systems: NLDB 2003, pp 84–102

Cantador I, Fernandez M, Castells P (2007) Improving ontology recommendation and reuse in WebCORE by Collaborative Assessments. In: Workshop on social and collaborative construction of structured knowledge at the 16th international World Wide Web conference (WWW 2007), Banff, Canada

Caracciolo C, Heguiabehere J (2009) NeOn deliverable D7.2.3. Initial network of fisheries ontologies. NeOn project

Cerbah F (2008) Learning highly structured semantic repositories from relational databases – RDBtoOnto tool. In: 5th European Semantic Web Conference (ESWC 2008), Tenerife, Spain, pp 777–781

Cimiano P, Völker J (2005) Text2Onto – a framework for ontology learning and data-driven change discovery. In: 10th international conference on applications of Natural Language to Information Systems (NLDB-2005), Alicante, Spain, pp 227–238

d'Aquin M, Motta E, Sabou M, Angeletou S, Gridinoc L, Lopez V, Guidi D (2008) Towards a new generation of semantic web applications. IEEE Intell Syst 23(3):20–28

d'Aquin M, Euzenat J, Duc C, Lewen H (2009) Sharing and reusing aligned ontologies with cupboard. Demo at international conference on Knowledge Capture (K-CAP 2009), Redondo Beach, CA, USA

Ding L, Pan R, Finin T, Joshi A, Peng Y, Kolari P (2005) Finding and ranking knowledge on the semantic web. In: 4th international Semantic Web Conference (ISWC 2005), Galway, Ireland, pp 156–170

Euzenat J (2007). Semantic precision and recall for ontology alignment evaluation. In: 20th international Joint Conference on Artificial Intelligence (IJCAI-2007), Hyderabad, India, pp 348–353

Euzenat J, Shvaiko P (2007) Ontology matching. Springer, Heidelberg

Fernandez M, Cantador I, Castells P (2006) CORE: a tool for collaborative ontology reuse and evaluation. In: 4th international workshop on evaluation of ontologies for the web at the 15th international World Wide Web conference (WWW 2006), Edinburgh, Scotland

Fernandez M, Overbeeke C, Sabou M, Motta E (2009) What makes a good ontology? A case-study in fine-grained knowledge reuse. In: 4th Asian Semantic Web Conference (ASWC 2009), Shanghai, China, pp 61–75

Guarino N, Welty C (2004) An overview of OntoClean. In: Handbook on ontologies. Springer, Berlin, pp 151–172

Horridge M, Parsia B, Sattler U (2009) Explaining inconsistencies in OWL ontologies. In: Scalable uncertainty management. Springer, Berlin/Heidelberg, pp 124–137

Jones M, Alani H (2006) Content-based ontology ranking. In: 9th international protege conference, Stanford, CA

Lewen H, Supekar K, Noy N, Musen M (2006) Topic-specific trust and open rating systems: an approach for ontology evaluation. In: 4th international workshop on Evaluation of Ontologies

for the Web (EON2006) at the 15th international World Wide Web conference (WWW 2006), Edinburgh, Scotland

Lopez V, Nikolov A, Fernandez M, Sabou M, Uren V, Motta E (2009) Merging and ranking answers in the semantic web: the wisdom of crowds. In: 4th Asian Semantic Web Conference (ASWC 2009), Shanghai, China, pp 135–152

Lopez V, Nikolov A, Sabou M, Uren V, Motta E (2010) Scaling up question-answering to linked data. In: Knowledge Engineering and Knowledge Management by the Masses (EKAW-2010), Lisbon, Portugal, pp 193–210

Lozano-Tello A, Gómez-Pérez A (2004) Ontometric: a method to choose the appropriate ontology. J Database Manage 15(2):1–18

Maedche M, Staab S (2002) Measuring similarity between ontologies. In: 13th international conference on Knowledge Engineering and Knowledge Management (EKAW 2002), Siguenza, Spain, pp 251–263

Patel C, Supekar K, Lee Y, Park E (2003) OntoKhoj: a semantic web Portal for ontology searching, ranking, and classification. In: 5th international workshop on Web Information and Data Management (WIDM 2003). In conjunction with the 12th international conference on Information and Knowledge Management (CIKM 2003), New Orleans, LA, USA

Porzel R, Malaka R (2004) A task-based approach for ontology evaluation. In: Proceeding of ECAI 2004 workshop on ontology learning and population, Valencia, Spain

Poveda-Villalón M, Suárez-Figueroa MC, Gómez-Pérez A (2009) Common pitfalls in ontology development. In: 13th Conference of the Spanish Association for Artificial Intelligence (CAEPIA 2009), Sevilla, Spain, pp 91–100

Sabou M, Wroe C, Goble C, Mishne G (2005) Learning domain ontologies for web service descriptions: an experiment in bioinformatics. In: 14th international World Wide Web conference (WWW 2005), Chiba, Japan, pp 190–198

Sabou M, d'Aquin M, Motta E (2008) Exploring the semantic web as background knowledge for ontology matching. J Data Semant 11:156–190

Strasunskas D, Tomassen S (2008) Empirical insights on a value of ontology quality in ontology-driven web search. OnTheMove 2008 confederated international conferences (OTM 2008), Monterrey, Mexico, pp 1319–1337

Suárez-Figueroa MC, Gómez-Pérez A (2008) First attempt towards a standard glossary of ontology engineering terminology. In: 8th international conference on Terminology and Knowledge Engineering (TKE 2008), Copenhagen, Demark, pp 1–15

Suchanek FM, Kasneci G, Weikum G (2008) YAGO: a large ontology from Wikipedia and WordNet. J Web Semant 6(3):203–217

Tartir S, Arpinar I, Moore M, Sheth A, 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, Houston, TX

Van Hage W, Isaac A, Aleksovski Z (2007). Sample evaluation of ontology matching systems. In: 5th international workshop on Evaluation of Ontologies and Ontology-based tools (EON 2007) Located at the 6th international Semantic Web Conference (ISWC 2007), Busan, Korea

Zablith F, d'Aquin M, Sabou M, Motta E (2010) Using ontological contexts to assess the relevance of statements in ontology evolution. In: 17th conference on Knowledge Engineering and Knowledge Management by the Masses (EKAW 2010), Lisbon, Portugal, pp 226–240