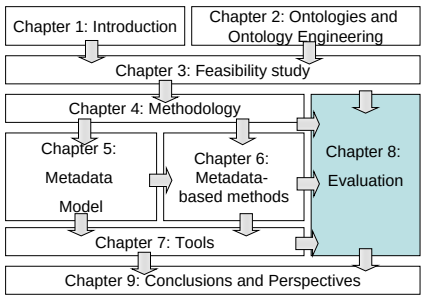
8 Evaluation

"Is not all honorable work also useful and good?"
–Plato, Protagoras, 358b



This chapter is dedicated to the validation of our research. We start by introducing the evaluation methodology, which provides the theoretical grounding of the evaluation results, in Section 8.1. Sections 8.2 to 8.4 describe the three building blocks of the evaluation procedure, which mainly consisted of professional reviews, case studies and goal-free analyses. Section 8.5 closes the chapter with a short summary.

8.1 Evaluation Methodology

In this section we give an overview of the most important evaluation approaches which can contribute to the validation of our research. Our goal is to identify from the multitude of existing evaluation methods and models those which are appropriate to our purposes and can reliably demonstrate the soundness of our solution in the field of ontology reuse.

8.1.1 Evaluation as a Field

According to the Merriam-Webster Online Dictionary the term *evaluation* can be defined as the process of determining *"the significance, worth, or condition of [something] usually by careful appraisal and study"*.[1] An evaluation should always start with the clear definition of a *testable hypothesis*. Depending on the type of the hypothesis to be tested (or the artifact being evaluated) one can apply multiple evaluation methodologies/models that are in turn supported by different methods. An evaluation model is determined by the following dimensions:

1. **Evaluation process**: it describes how to practically test the hypothesis. The process description includes activities, their order of execution, evaluators, target audience, inputs and outputs. It also refers to the requirements which should be fulfilled as a pre-requisite for a valid evaluation (i.e. the assumptions which the evaluators and the target audience should agree upon, prior to starting the evaluation process).

[1] `http://www.m-w.com` last visited in July, 2006

2. **Evaluation methods**: they describe concrete procedures for data collection and are associated to specific evaluation criteria.

3. **Evaluation criteria**: are the dimensions of the solution space which are relevant for determining the utility of the artifact.

4. **Interpretation of the evaluation results**: in order to decide whether or not the artifact fulfills the pre-defined set of requirements one needs means to objectively interpret the evaluation results. This includes quantifiable metrics for comparing among result sets, but also qualitative approaches based on observations and their unbiased interpretation (see below).

Evaluation methods can be classified in various ways. One of the most common distinctions is between *quantitative* and *qualitative* methods. Quantitative approaches are characteristic for natural sciences and concentrate on observable facts, quantitative techniques, systematic procedures and reproducible results. Representative methods for this category are *laboratory experiments*, *survey methods*, *econometrics* or *mathematical modelling*. Qualitative evaluation has its origins in social sciences. It focuses on human-driven hypothesis testing performed by conducting structured expert interviews and empirical studies. In this category we can mention established methods such as *professional reviews* and *case study research*. Every evaluation procedure (whether quantitative or qualitative) builds upon a certain notion of what constitutes *valid research* and which research methods are appropriate to achieve trustworthy results. These assumptions are primarily philosophically grounded and relate to the epistemology underlying the research endeavor. A well-acknowledged classification differentiates among *objective* and *subjective* epistemology in this respect. The former is concerned with the discovery of general laws which intend to predict or control a particular class of phenomena or artifacts, a goal which is often embraced by natural scientists. By contrast the latter aims at explaining and understanding empirical or artificially created situations. Independently of its evaluation approach and its epistemological principles every valid evaluation endeavor should satisfy two core requirements:

- **Credibility and trust**: the audience of the evaluation (and implicitly of the subject being evaluated) should trust the procedures of the evaluator. The audience may range from academia to practitioners and users. Each of them is interested—or can be convinced—by different evaluation methods and results. Managers, for instance, might trust quantified facts to choose the most efficient procedure, while practitioners could also be interested in empirical findings. Academic parties are definitely interested in the research question itself. The way the evaluators claim credibility is partially correlated to the evaluation method: in the objectivist case the audience should agree with the facts, while in the subjectivist case it should accept the experiences made in the case studies, the way they were interpreted and their implications. Nevertheless, independently of the method applied, the participants in an evaluation process should be unbiased; they should be interested in the evaluated findings without favoring any of the alternative solutions being analyzed.

- **Externalization and replication**: ideally, the evaluation should be externalized and explicit. For example, a case study evaluation should be based upon a dedicated case

study that was set up with the explicit purpose of validating a particular hypothesis. It should further report on those experiences and lessons learned during the case study operation, which are significant for the tested hypothesis and are not standardized or known in advance, and on the credentials of the participants. A second criterion for an evaluation to be credible and valid is the possibility to replicate it. This is a strong indicator for the significance of the achieved results. The replication degree highly varies across evaluation methods. Objectivist approaches tend to satisfy this requirement as they relate on quantitative measures (though they also depend on the statistical significance of the data sample utilized for the evaluation). Qualitative research is characterized by a comparably lower reproducibility, as insights might vary considerably between two externalizations. This can be however counterbalanced by the selection of a representative evaluation setting and of unbiased case study participants.

Design sciences—including computer science or ontology engineering—can make use of both types of evaluation approaches, quantitative or qualitative [99, 134, 136, 198]. The remaining of this section concentrates on presenting the basic notions of some of the most widely used evaluation models and methods, before deciding upon the evaluation framework applied in the context of our research.

Goal-free Evaluation

The *goal-free* approach is performed with the purpose of comparing different solutions of the same problem against a pre-defined set of criteria [195]. It is widely used across research disciplines, in qualitative, as well as quantitative research.

The process is initiated with the selection of the relevant quality criteria—ideally these should take into account the expected evaluation audience, be unambiguously defined and able to uniformly describe a preferably wide range of similar problem solutions. Each approach to be evaluated is assigned a set of scores corresponding to the pre-defined criteria and their associated quantifications. This task is accomplished with the help of a variety of data collection methods, the most common being interviews, questionnaires, laboratory experiments, field studies as well as action research.

The result of the goal-free evaluation is an objective analysis of the behavior of the analyzed solutions as regards the pre-defined quality dimensions, and the recordings of their measurements. The results are targeted at potential consumers of the tested artifacts (e.g., the applicants of a technology).

ADVANTAGES: The evaluation is objective and its outcomes can be used for different purposes and audience groups. The identification of evaluation criteria provides the target audience with an overview of the relevant aspects of the solution space and with a comparison of the available artifacts, their strong and weak points, while identifying directions for further research and development.

DISADVANTAGES: The evaluation outcomes depend on the quality of the evaluation framework. Furthermore, the approach assumes that the target audience agrees with the relevance and completeness of the quality criteria in the corresponding research field.

Goal-based Evaluation

In a goal-free evaluation a set of artifacts targeted at the same problem statement are evaluated against a set of quality criteria in order to give an account of their strengths and weaknesses. By contrast, the *goal-based* (or *objective-oriented*) evaluation focuses on the features of an artifact as stated by its developers and collects evidence as to which extent these have been truly achieved [220].

The process starts with the evaluator team collecting the data necessary to assess the quality of the artifact. Its goals are ideally quantifiable in order to enable an objective comparison between them and the outcomes of the artifact. However, this activity can also resort to human judgement: experts might be interviewed in order to express their opinion on the quality of the tested solution regarding its pre-defined aim. Similar to the previous approach, the goal-based evaluation can be operated with the help of both qualitative and quantitative methods, from expert-based ones to action research, field and case studies, as well as laboratory experiments.

The result of the evaluation is an account of the degree to which the tested artifact fulfills the initial requirements and is typically targeted at managers, analysts and psychologists.

ADVANTAGES: There is no need for defining an additional evaluation framework—provided the target audience agrees on the goals stated by the developers. Using these relieves the evaluator from having to presume certain functions or features of the artifact. Further on, parts of the evaluation process can be accomplished in advance by the developers themselves, who perform preliminary tests prior to submitting their work to an external reviewing panel.

DISADVANTAGES: The credibility of the approach depends on the appropriateness and the relevance of the artifact goals for the problem space. Further on, possessing too deep knowledge on the features and the objectives of the studied subject is considered to bias the evaluators [195].

Professional Review

The *professional review* (also called *accreditation* or *connoisseurship*) model relies exclusively on the judgement of people possessing deep knowledge in an area of interest [57]. A number of professionals study an artifact with respect to the plausibility of its account. The reviewing process can be structured with the help of pre-defined dimensions; this is considered to be independent of the quality of the results, though it might help systematizing the evaluation and distilling the review results. Experts recommend a careful selection of the reviewing team, the optimal size of the team being estimated between 5 to 10 [101].

This model produces expert opinions regarding the suitability of the evaluated artifact in regard to the given problem space (or its associated evaluation criteria). It is typically applied prior to other evaluation methods in order to eliminate obvious errors and rapidly increase the quality of the artifact. The results are targeted to a wide audience of professionals in the corresponding field of activity.

ADVANTAGES: The approach reduces bias, since many aspects of the proposed solution are evaluated by the experts. Obvious errors, as well as hidden consequences and implications of

the solution can be efficiently detected due to the involvement of several professionals holding various views upon the area of interest.

DISADVANTAGES: The evaluation depends on the availability of experts and can not be repeated regularly. The target audience should agree to the competence of the reviewing panel and the methods used for collecting the data from the experts.

Case Study

The main objective of the *case study* method is to understand how a particular process can be optimally carried out. The hypothesis being evaluated refers to an existing process description. The case study can provide insights on the validity of this process description with respect to the associated class of situations. Its outcomes are targeted primarily at practitioners.

The case study participants accomplish their tasks as recommended by the proposed process-driven model, or suggest modifications of the original solution if appropriate. A case study can have a pre-defined duration. The evaluators should start the empirical analysis with the formulation of the research hypothesis [231]. They should consult additional information sources and perform a literature review in order to increase the precision of research question definition. They should further agree upon the data collection and analysis techniques e.g., interviews, surveys, action research. Another important aspect is related to the group of participants: besides having a high number of individuals involved in the study, it is important that they have some prior knowledge on the process being evaluated and that they are unbiased. In the course of the case study the evaluator collects the data from the different participants in the process as required by the aforementioned methods. Many participants should provide data, in order to capture multiple views on the process and thus increase the representativeness of the observations. The analysis of the data should point out the important issues and relevant findings. After the completion of the study the evaluators describe the results, taking into account the organizational setting, the participants, the collected data and its analysis. They also produce experience reports for the tested process model, outlining best practices and guidelines.

ADVANTAGES: The main advantage of the case study method is its emphasis on the experiences of the practitioners in a given situation. It contributes to the empirically driven refinement and revision of the artifact being tested, thus increasing its real-world usability.

DISADVANTAGES: The value of the case study description depends on the capabilities and the experience of the team of evaluators. Further on, it is difficult to compare different case studies as they depend on the underlying setting and on the characteristics of the participants. An important pre-condition for the acceptance of the study findings by the target audience is an agreement between them and the evaluators with respect to the process model applied.

Other Approaches

Further widely used models are the *system analysis*, the *decision making*, the *art criticism* and the *quasi-legal* or *adversary* evaluation [101]. Many other approaches are derivations of the enumerated ones or combinations of these major types. They are not included in this overview as they are not directly relevant to our research, due to divergences in the objectives, the expected audience, or both [101].

8.1.2 Selection of Appropriate Evaluation Approaches

The selection of an appropriate evaluation method depends on the type of research conducted and on its desired outcomes. In computer science we can primarily resort to methods or combinations of methods which are acknowledged to obtain reliable results in design sciences [99, 134, 136].[2]. In this thesis we rely on the framework for design science research proposed by Hevner and March [99]. They differentiate among four types of research outcomes:

1. **Constructs**: provide the language, the terminology in which a problem/solution space is defined and explained.

2. **Models**: cover the most important facts and concepts within a domain of interest or class of situations. They use constructs as a description language for the problem/solution space.

3. **Methods**: describe processes and guide their users in how to identify solutions to a given research question. They can range from rigorous mathematical algorithms to descriptions how to perform a process, best practices, guidelines etc. From a terminological perspective, methods—as understood by [99]—can be considered synonymous to "methodology", "technique" or "algorithm" in computer science.

4. **Instantiations**: implement constructs, models and methods, thus demonstrating their feasibility.

For each of the four types Hevner and March summarize the *recommended content and structure*, as well as *proved and tested evaluation methods* and *quality criteria*. The structure of each research artifact includes the information sources which are required to give full particulars on the actual solution, and thus enable a feasible evaluation procedure. The main components of each artifact and the suitable evaluation methods and criteria are listed in Table 8.1.

In terms of the aforementioned framework we can be localize the outcomes of our research in three of the four categories as subsequently explained:

1. **Ontology reuse methodology**: is a *method artifact*, as it provides a detailed description of ontology reuse processes from an application-oriented perspective.

[2] Design sciences are concerned with the study of *"artificial objects or phenomena designed to meet certain goals"* [198]. Typical examples are, besides computer science, engineering or product design.

CONSTRUCTS		
STRUCTURE	EVALUATION METHOD	CRITERIA
vocabulary meta-model	ontological analysis	construct deficit construct overload construct redundancy construct excess
MODELS		
STRUCTURE	EVALUATION METHOD	CRITERIA
domain and terminology scope and purpose syntax and semantics intended applications and use cases ref. to constructs and methods	syntactic validation semantic consistency matching to real phenomena using sample data user surveys integration tests	correctness completeness clarity and simplicity flexibility extendability applicability implementability
METHODS		
STRUCTURE	EVALUATION METHOD	CRITERIA
process-based meta-model intended applications and use cases outcomes of the method application ref. to constructs and models	user surveys case studies action research laboratory experiments	appropriateness completeness consistency implementability
INSTANTIATIONS		
STRUCTURE	EVALUATION METHOD	CRITERIA
implementation ref. to design model ref. to requirements specification ref. to documentation ref. to quality assurance documents ref. to management documents ref. to user guide	code inspection testing code analysis verification	functionality usability performance reliability maintainability

Table 8.1: Design Research Outputs and Their Evaluation

2. **Ontology metadata model**: primarily comprises a *model* of ontology metadata, as it identifies the information types required to perform particular reuse activities and organizes these constructs in an ontology. The model was implemented in OWL (i.e. *instantiation*).

3. **Support methods for ontology reuse**: can be classified as *method artifacts* associated with the reuse methodology and the ontology metadata model.

4. **Support tools for ontology reuse**:are an *instantiation* of our theoretical work in terms of implemented methodology and methods.

Consequently we derive the following evaluation framework for this thesis:

1. **Ontology reuse methodology**: the methodology is evaluated using professional reviews, goal-based evaluation on the basis of empirical studies and goal-free evaluation.

2. **Ontology metadata model**: the conceptual model was evaluated as using professional reviews and goal-free evaluation. Its instantiation was evaluated goal-based as regards

	METHODOLOGY	METHODS	METADATA MODEL
Professional reviews	reviews using pre-defined quality dimensions		
Goal-free evaluation	structured interviews	-	structured interviews
Goal-based evaluation	empirical studies	empirical studies	-

Table 8.2: Evaluation Framework of Our Research

its usability in relation to the support methods and tools.

3. **Support methods for ontology reuse**: the methods have been evaluated using professional reviews and goal-based evaluation conducted using the case study method. An in-depth goal-free analysis was estimated as unappropriate due to the specificity of the context of our work, which is application-oriented ontology reuse.

4. **Support tools for ontology reuse**: the prototypical implementation has been tested using common software engineering techniques.

The choice upon an ideal evaluation approach is not trivial, due to the advantages and disadvantages of each approach and their partially divergent objectives and foci. Restricting to a single approach might thus be misleading, since this fails capturing all aspects and implications of the studied artifact. For validating the results of our research we fundamentally rely on three types of evaluation: goal-free evaluations for the methodology and the ontology metadata model, empirical studies for the methodology and its associated methods, software testing techniques for the prototypical tools, as well as professional reviews for the overall approach.

The *goal-free* approach aims at situating our solution in the domain of research and is targeted at its potential users, expected to choose among similar approaches to solve a specific problem. With the help of the *empirical experiments* our evaluation is also targeted at practitioners. In this way they are provided with insights over the diversity of the solution space and with practical experiences on the application of the artifact in real-world scenarios. Software tests provide evidence on the implementability of our ideas. Finally *professional reviews* test the general quality of our approach according to their level of experience of the reviewing panel.

In the following we describe the results of the evaluation. Firstly we present the results obtained in the professional reviewing procedure. This evaluated the overall approach to ontology reuse at various maturity stages. We continue with the experimental studies, in which the applicability of the methodology and its associated support methods was investigated. We also describe the results of testing the implementation of the metadata generation method. We finalize the evaluation by comparing our proposal (the methodology and the ontology metadata model) with other approaches primarily situated in the area of ontology engineering.

8.2 Professional Reviews

The research reported in this thesis has been subject to various professional reviews, which allowed us to identify the limitations of its initial ideas. In the following we summarize the results of these reviews and their implications on the final outcomes of our work.

8.2.1 Ontology Reuse Methodology and Associated Methods

The ontology reuse process model was evaluated at various development stages by thirteen experts as part of different conference reviewing sessions and by three researchers in the ontology engineering field. The most significant evaluation criteria involved in this process were *originality*, *technical quality* and *impact*. The originality judges the novelty of the proposed solution. The technical quality measures the validity and the rigourousness of the conducted research. Finally, impact refers to the influence the artifact is estimated to have on the community.

The first version of our methodology comprised a description of the three main stages of the process and sketched the role of contextual information in the reusability of ontologies. At this point we had performed the requirements analysis and initiated the eRecruitment case study subsequently presented. Five reviewers appreciated the process model and its application-narrow orientation while criticizing the generality of the process description and the lack of automatic methods clearly demonstrating the advantages of a context-sensitive approach in empirical settings.

The ontology reuse methodology evolved towards a more in-depth description of the ontology merging and integration task. On the basis of the ontology metadata model, which was in parallel implemented, we refined the description of this core reuse step and designed a method for automatically selecting the appropriate strategy for comparing, merging and integrating ontological sources on the basis of contextual dependencies and metadata. The PROMI environment was realized as a result of these ideas. These developments were peer-reviewed by five experts with positive results. The main critique point at this stage was the absence of a powerful reasoning service by means of which the feasibility of the rule-based approach could be arbitrarily demonstrated. This can be considered an open issue of our work. However, our research was not focused at the question of ontology matching, merging and integration, but merely exemplified the utility of contextual information within ontology reuse by means of the ontology matching task.

A last phase of the presented methodology elaborated on ontology evaluation aspects. The approach was evaluated by conducting structured interviews with a group of three experts. One of the participants was were IT practitioners affiliated to academia (one) and industry (two) possessing in-depth knowledge on semantic technologies. They were given a detailed overview of our overall work in the field and were asked to estimate its plausibility. The experts appreciated the context orientation of the method, the possibility to choose among different evaluation dimensions and the technological support (cf. Section 8.3.3). They had several suggestions for the improvement of the presented tools and expressed their concerns in respect to the feasibility of ontology reuse in the absence of high quality ontological content and of powerful ontology search engines. These aspects are still marginally explored in

the scope of this thesis. Further on we acknowledge the need for further functional and usability refinements at tool level and the utility of a wide scale user study for the implemented prototypes.

8.2.2 Ontology Reuse Metadata Model

The evaluation of the metadata model was conducted in two parallel phases: on one hand, the content of the model was subject to human-driven evaluation with respect to the inventory of the included metadata elements, their meaning and labeling. On the other hand the ontology was incorporated to OMV (**O**ntology **M**etadata **V**ocabulary), which is a metadata standard for ontologies developed within the European Network of Excellence Knowledge Web.[3] Within this context several applications using this ontology have already been developed (cf. Section 5.3).

The content-based evaluation was performed by conducting interviews with a group of experts in the area of ontology engineering. Considering that the people best placed to give a comprehensive assessment of the ontology metadata vocabulary are currently researchers being directly involved in theoretical or practical issues of ontology engineering, we organized an expert group of six academics affiliated in this community and in the EU Network of Excellence KnowledgeWeb, which agreed on evaluating the model against the set of criteria defined in [79]:

- **Consistency**: this criterion refers to the existence of explicit or implicit contradictions in the represented ontological content.
- **Completeness**: according to [79], an ontology is complete if it (explicitly or implicitly) covers the intended domain.
- **Conciseness**: complementary to the previous feature, conciseness states for the redundancy-free representation of the application domain of an ontology and for the avoidance of useless definitions.
- **Extendability/sensitiveness**: the criterion refers to the possibility of adding new definitions to the ontology without altering the existent content.

The aforementioned evaluation framework was extended with a fifth dimension, the **readability**, which accounts for the usage of intuitive labels to denominate metadata entities.

The evaluation resulted in changes on both conceptual and implementation levels of the ontology:

- **Completeness**: during the evaluation the participants identified several aspects which were missing in the initial draft of the metadata model. For instance, information about the representation language of an ontology (syntax, representation paradigm etc.) was found to be insufficiently covered. As a result we introduced concepts such as `RepresentationParadigm` and `RepresentationLanguage` to account for these aspects. Two of the evaluators

[3] `http://omv.ontoware.org` last visited in February, 2006

argued for a separation of the generic `Ontology` concept in two concepts stating for the conceptualization and the implementations of the same conceptual model in various KR languages. However, we decided against this suggestion, as the remaining interviewees were against an academically justified refinement of the ontology, which can not be easily understood in practice without deep theoretical knowledge on ontologies. A second point against this distinction was related to the difficulties related to assigning particular ontology properties to the conceptual or the implementation level of the ontology. For example, it was not clear whether an entry such as knowledge representation paradigm should be defined at the conceptualization or at the implementation counterpart of an ontology.

- **Conciseness**: parallel to extending the ontology, the experts expressed their concerns regarding a series of concepts which were too specific for a core metadata vocabulary. Most of these concepts related to particular aspects of the application settings the ontology was previously used and to descriptions of persons and organizations. The engineering team decided to reduce the description of application systems to a simple taxonomy and a classification of industrial sectors based on NAICS, without linking these concepts to information about the developers and users of the corresponding applications.

- **Readability**: the naming of particular metadata entities was one of most challenging parts of the evaluation process. The experts proposed alternative names for several fundamental concepts, such as `OntologyLanguage`, `Domain` and `View`. They were modified to `RepresentationLanguage` (by contrast to the natural language used to label ontological primitives), `OntologyDomain` and `ViewUponTheDomain`, respectively. Further on, the experts indicated the poor readability of abbreviated concept labels, which were changed respectively.

- **Consistency**: according to human judgement and to the automatic consistency checking no inconsistencies were found. However, several evaluators emphasized the challenges associated with modelling metadata with the help of OWL and with the representation of tasks, roles, processes and situations. While the ontology was positively evaluated by the experts in respect to the correctness of its modelling, it was not clear whether some modelling decisions would be appropriate for automatizing ontology reuse. This question was, however, positively answered a posteriori in relation with the implemented support methods for ontology reuse.

- **Extendability**: the model was estimated to provide easy to use means for refinements and extensions. The interviewees proposed several extension modules, which would refine aspects of ontology engineering processes which are not properly modelled in the context of a generic metadata schema like ours. These will be developed in relation to the OMV ontology as an activity within the Knowledge Web project.

In summary, the results of the expert-driven evaluation significantly contributed to the quality of our approach, confirming our expectations towards the realization of useful metadata schema for Semantic Web ontologies and their reuse. However, the evaluation process has

already pointed out two challenges of our approach. The first is related to the usability of the model (conciseness, extendability): the achievement of a common agreement in a large community of ontology users with respect to their requirements and perceptions about ontology metadata has a direct impact on its usability. This issue was addressed by aligning the model to the OMV initiative (cf. Chapter 5). A second challenge questions the usage of languages like OWL DL for modelling metadata and the need for a more complex representation of processes, tasks, roles and situations. As aforementioned, these questions were partially approached in relation to the designed ontology reuse support methods, which demonstrate the utility of the current metadata ontology within this scope.

8.3 Goal-Based Evaluation

The research we report on in this thesis is concerned with the question of methodologies, methods and tools which enable an efficient and effective operation of ontology reuse processes. The hypothesis derived after completing the analysis of the state of the art in the field was that the reusability of currently available ontologies and ontology-like knowledge sources would benefit from explicitly having regard to the context-sensitive nature of the process. This hypothesis was validated using the goal-based evaluation model applied on several empirical case and user studies. The objectives underlying the approach were in our case the *user-perceived process operation efficiency* and the *fitness of use of the reuse outcomes in the target application setting*.

A case study in the domain of eRecruitment was performed in parallel with the first professional reviews reported in Section 8.2. It covered all main process steps, including preliminary considerations on the application-oriented ontology evaluation method. Orthogonal to this experiment, we conducted at a later date two in situ user studies aiming at validating the ontology evaluation and the automatic metadata acquisition methods.

8.3.1 Case Study eRecruitment

Organizational Setting

The case study was performed in collaboration with an international eRecruitment solution provider and used the technical infrastructure of the case study partner as starting point for the investigations. Details on the architecture of the system as well as on exact heuristics are in possession of the case study partner. The scenario shared several commonalities with the human resources case study investigated in the "KnowledgeNets" project (cf. Section 3.4) with respect to the requirements for the ontology to be developed. A team of two practitioners with experience in eRecruitment applications assisted by an ontology engineer aimed at building an ontology for (automatically or manually) classifying English documents such as job postings or job seeker profiles in the *life sciences* field. The domain experts were given an introduction to the ontology reuse methodology and the associated methods. The study was carried out over a period of approximately 3 months. The data was collected by interviewing the participants and by own observations. The implemented tools were not part of the study.

Reusing Job Classifications

Ontology Discovery The discovery step was performed as described in the ontology reuse methodology. Domain experts carried out manual searches on pre-selected keywords in repositories or using general-purpose search engines. They also collected information from the Web sites of international and national organizations in the areas of life sciences and employment. The ontology engineering expert was concerned with the seek for relevant resources on existing Semantic Web initiatives. While this attempt did not produced any notable positive hits, we obtained better results by resorting to collections of ontology-related resources such as taxonomies and controlled vocabularies. Extending the range of our search to this kind of lightweight ontologies contributed to the discovery of a substantially larger set of potential reuse candidates, a fact which was positively appreciated by the case study participants.

The result was a list of approximately 60 resources, most of them in form of informal or semi-formal, freely available classifications. Their characteristics were compatible with the context-specific recommendations of the methodology. In the same time, the industrial partners were reluctant to eliminating any of the reuse candidates at this point. Despite our expectations, neither availability, nor provenance information seemed to play a crucial role in this decision. Contrariwise, the domain experts possessing a wide body of knowledge in the human resources area, initially showed a slight preference for commercial artifacts, while standard classifications were generally estimated to be too comprehensive for single job portal applications.

Ontology Evaluation The 61 lightweight ontologies covered the following domains:

- **Occupations**: in this category the search resulted in approximately 30 standard occupational classifications, most of them being developed by established governmental and international organizations and employment agencies.
- **Skills**: the approximately 20 reuse candidates in this category provided descriptions of competency sectors, qualifications and skills related to particular educational levels or occupational profiles. In contrast to the occupational classifications, they were mainly proprietary products, which significantly varied with respect to the represented content and its quality (see below).
- **Life sciences**: in this category we took into consideration 10 reuse candidates. The majority described the medical domain, originated from academic projects, and showed the same irregularities of content and quality as the skill resources. However, due to their preponderantly research provenance, the number of Semantic Web resources was higher than in the former categories.

For each of the sub-domains the engineering team evaluated the usability of the discovered resources as described in our methodology. Firstly we agreed on the evaluation dimensions, which were in this case concerned with the content, the applicability and the technical suitability of the analyzed artifacts. For each of the evaluation dimensions we assigned a percentage score reflecting the perceived degree to which a specific ontology fulfilled a set of pre-defined criteria.

While the evaluation of the life sciences ontologies was relatively straightforward because of the low number of candidates and the unsuitability of their content, the parallel activities on the occupational and skill classifications were tedious due to the heterogeneity of the sources. As some of the resources were not available—or were too comprehensive to be "read through" by humans—the evaluation was carried out solely on the basis of the associated metadata. The metadata was acquired manually by the ontology engineering team and was compliant to the proposed schema. An automatic generation of the metadata was not feasible, due to the proprietary formats in which the ontologies were formalized and the lack of resources to develop adequate computer-aided parsers. More research is needed in order to handle this significant class of ontological knowledge sources programmatically.

The content-related evaluation was not very productive in the absence of a precise and comprehensive description of the employment domain. However, many of the life sciences ontologies proved to be inadequate for our purposes. While the domain of the these ontologies was compatible to the case study setting, the view upon the domain of life sciences (or some fragment of it) was not relevant; the majority of the resources described scientific facts of the domain of interest, while we were interested in a simple classification of the life sciences field in sub-fields.

For the remaining two evaluation dimensions we created an evaluation framework on the basis of the considerations preliminarily elaborated in the methodology.[4] The catalogue of evaluation criteria was created along two face-to-face meetings of the engineering team: in the first brainstorming workshop the participants conceived a preliminary list of ideas regarding potential evaluation criteria and dimensions. A document summarizing the agreed results was distributed to the team, who further refined the evaluation framework with respect to their particular interests. In the second workshop, the change requests were discussed and integrated into the evaluation framework, whose final version consisted of the following dimensions:

- **Application-related aspects**: is the ontology appropriate for the tasks it should be involved in at application level? Since the final ontology was targeted at manual and automatic indexing the application-related requirements were as follows:
  - are the concepts denominated in English?
  - are they labeled using naming conventions and in a linguistically predictable way?
  - is the ontology large enough to be representative for the domain corpus to be annotated?
  - is the ontology "thesaurus-like"? does it contain synonyms or alternative spelling information about the contained concepts?
- **Technical aspects**:
  - is the information available in a (semi-)structured form? is the classification available in a standard representation language? can the classification be easily converted to new formats?
  - is the source subject of frequent updates which have to be propagated to the application scenario?

[4] At that point, the ontology evaluation method had not been described in detail in our work.

  - can one easily use fragments of the ontology?
  - can the ontology be merged with or integrated into other ontologies without substantial effort?

The case study partners acknowledged the benefits of a controlled approach to evaluation and participated actively at the elaboration of the criteria catalogue. Nevertheless they expressed concerns with respect to several of the proposed criteria, which were perceived to be difficult to estimate. These were not utilized in the study.

The results of the evaluation can be summarized as follows:

- **Occupations**: despite of the complexity of the evaluated sources, the process was performed in a straight forward manner due to the high number of additional documents and support tools available for most of the reuse candidates. However, several occupational classifications covering the same areas obtained very high evaluation scores; this state of affairs made the selection of a single ontology difficult.

- **Skills**: in contrast to the previous category, the evaluated sources were assessed comparatively low relevance rankings. Due to the associated licence costs and contrary to their initial opinion, the engineering team decided against the re-usage of any of the resources and for the adoption of the skill information available in occupational classifications such as O*NET and SOC.

- **Life sciences**: the evaluation was accomplished with manageable effort; most of the ontologies were ranked with low quality scores and the engineering team decided not to use them as basis for a new topic hierarchy of the application domain.

The occupational classification O*NET and the Occupational Thesaurus were selected to be reused in the eRecruitment scenario. The Occupational Thesaurus was intended to linguistically complement the vocabulary of the occupational classification O*NET. The decision to use a single occupational classification might be surprising if we recall that several of the analyzed resources were assigned a high relevance score in respect to the agreed criteria. The main objective against using multiple classifications—which were likely to provide an increased domain coverage—was the difficulties associated with merging these resources in correlation with the absence of uniform naming conventions. Such conventions were available at individual resource level, but their benefits would have been diminished in case of merging several ones. Further on, the O*NET classification already included skills, thus reducing once more the need for integrating external ontologies dedicated at modelling this type of information. The domain experts were highly aware of the challenges associated with such aspects, probably due to the relevance of data and application integration in real-world scenarios. They expressed their concerns with respect to their ability to develop a feasible merging tool, or to deploy an existing one without additional consultancy, confirming our assumptions with respect to the high expertise requirements related to the choice upon an adequate merging strategy.

Nevertheless, due to their size and lack of focus on the domain of life sciences, the two sources to be reused had to be customized in accordance to the requirements of the application scenario.

Ontology customization, merging and integration In this step we identified fragments of the occupational classification which are relevant for life science disciplines. Due to the tree-like structure of the O*NET classification, most of the relevant concepts could be easily localized in the categories *"Biology"*, *"Chemistry"*, *"Medicine and Dentistry"* and *"Physics"*. The availability of tools for navigating the hierarchical structure of the O*NET greatly contributed to the efficient accomplishment of this process step. The list of relevant concept contained approximately 400 elements, including job identifiers, skills, abilities, work activities and industrial sectors.

After identifying the relevant knowledge sources and the list of concepts which can be used as input for the application, the next steps which have to be carried out are the translation of the O*NET data model to OWL and the transformation of the relevant data from one format to another. The OWL implementation of the eRecruitment ontology has not been performed in this case study. However, in order to accomplish this task, one has to implement a program, which reads and parses the occupational profiles using the O*NET Internet front-end and generates the corresponding OWL constructs (e.g., using Jena2). The dimensions of the multi-facet taxonomy can be modelled as OWL properties of a main class `OccupationalProfile`, while alternative names for individual concept labels have to be extracted from the Occupational Thesaurus—possibly using linguistic matching algorithms—and integrated into the O*NET-resulting ontology in form of RDFS labels. Every occupational profile consists of several building blocks: disciplines in which the occupation is relevant, skills, abilities, interests and work activities. These can be modelled as classes and organized in hierarchies. The creation of the RDFS labels can be automatized using for instance the merging component in PROMI.

Lessons learned

The main challenge of the reuse case study described in this section were the *discovery of the reuse candidates* and the *evaluation of existing sources*. The first has definitely proved to be an art and not a science due to the absence of fully-fledged component repositories. While sites such as Taxonomy Warehouse provide a large inventory of useful knowledge sources and a comprehensive classification index, the way these indexes are used for search purposes is still restricted to keyword-based techniques. The result of the heuristics-based ontology discovery task was a considerable list of classification systems, which differed in the formalized domain, quality and appropriateness for the ontology scope i.e. indexing and semantic annotation of the job descriptions and job seeker profiles. The overload caused by this heterogeneity was avoided with the help of the ontology metadata schema. The evaluation step required additional efforts to be invested in the elaboration of a commonly agreed quality framework, due to the lack of applicable methodological background in the field at that point. On the other hand, once the evaluation criteria have been agreed by the case study partners, the evaluation and customization of the sources was performed in due time in a systematic manner. The availability of systematic metadata information as that provided by Taxonomy Warehouse, which descriptively eased the access of the domain experts to the ontologies to be evaluated, and the taxonomical structure of the surveyed ontologies contributed to this success.

In similar situations, existing reuse methodologies proved to be related to considerable

post-processing efforts, mainly because of their high level of generality. In contrast, instantiating these process models on the application context of the corresponding scenario—in terms of the scope/task/role of the ontology to be built—provided a feasible basis for the evaluation procedure to be performed efficiently and effectively.

The development costs averaged 1.25 person months. The effort distribution during the case study was as follows: the efforts related to the discovery of the source ontologies required over 20% of the time necessary to build the target ontology. Further 50% of the engineering time were spent on setting up the evaluation framework, while the remaining 30% were invested to customize the reuse candidates.[5] Compared to the Knowledge Nets scenario the ontology engineering task was performed more efficiently and effectively (1.25 vs. 1.8 PM). The development costs were relatively lower, though the resulting ontology necessitated additional customization efforts in order to prune the general-purpose eRecruitment classification systems to the domain of life sciences. Aligning the particularities of the overlapping tasks in the two studies confirms the role of the descriptive metadata information about ontologies, which simplified the understandability of the sources to be evaluated, and the importance of a strongly task-focused approach to ontology reuse. This led to a simplification of the reuse candidate selection while guaranteeing the usability of the outcomes in the target application setting.

In the same time the application of the methodology in the eRecruitment setting confirmed the need for an application-focused approach to ontology evaluation. The general-purpose dimensions of this problem, typically addressed by current ontology engineering research, proved to be insufficient to allow an effective selection of the relevant reuse candidates. The content-related evaluation led to the elimination of several of the original set of ontologies only after considering the view upon the modelled domain of interest in addition to the latter. The provenance and availability information played a minor role in this experiment. Nevertheless, the domain experts agreed on their importance in arbitrary situations. In respect to the topic of merging/integration the case study participants preferred discarding resources modelling overlapping facts to the costly generation of a possibly broader ontology.

The main result of this case study was a first exhaustive outline of our ontology evaluation method.

[5] The 50% also included the generation of the task-relevant metadata about the surveyed ontologies.

8.3.2 OntoMeta User Study

The goal of this study was to demonstrate the feasibility of an automatic approach to metadata creation as compared to human estimations. The heuristics implemented in OntoMeta were tested against the opinions of three computer science students with prior basic knowledge on ontologies. The study was supervised by a team of two evaluators, one of them being directly involved in the design and development of OntoMeta. The test set contained 52 OWL and RDFS ontologies crawled from the Web.

The quality of the automatically computer information was evaluated using two types of similarity functions:

- a boolean function, which returns 1 in case the actual test result coincided with the reference value, and 0 otherwise.
- a multi-valued function, which measures the resemblance between the actual and the reference values as estimated by the three study participants using some heuristics.

Many of the metadata elements can be evaluated using a two-valued function. This particularly applies for the information which is modelled using datatype properties, in numerical or string form (cf. Chapter 5) in the metadata schema. In this case the specification of the reference values was carried out with the help of external ontology management tools (see below). The metadata elements *ratio of inconsistent/invalid statements* as well as *OWL sub-language* were computed using the reasoning engine Pellet. The correct values of these elements depend, however, on the quality of the underlying tools and can not be computed by humans without investing considerable efforts correlated with deep knowledge in the field of formal logics. The metadata element *documentation* could not be subject to a systematic evaluation because of the lack of an appropriate instrument to identify the associated reference result reliably.

Some of the semantic metadata elements—notably those modelled using object properties—could not be evaluated using two-valued similarity measures. They capture information such as the *domain of the ontology*, the *view upon the modelled domain*, the *readability* of the ontology, which can not be estimated by humans in a purely objective manner. In order to comply to this ambiguity, the evaluators applied a five-values resemblance function instead of the more rigid boolean approach. The function is defined as follows:

- value 1: if at least two of the human results coincide with the machine delivered result
- value 2: if at least one human result equals the program output or if all human results are close to it.
- value 3: if no human results are exactly equal to tool output, but at least two of the results are close to it.
- value 4: if at least one human result approximates the machine result.
- value 5: otherwise.

Tested feature	Gold standard	Similarity function	Test result
number of classes	Protégé/DAML Statistics	equals	100%
number of properties	Protégé/DAML Statistics	equals	100%
number of instances	Protégé/DAML Statistics	equals	100%
number of asserted axioms	Protégé	equals	100%
depth of inheritance tree	Protégé	equals	100%
ratio of invalid statements	Online OWL Validator	equals	100%
ratio of used syntax	Human	equals	100%
imported ontologies	Protégé	equals	100%
OWL sub-language	Pellet	equals	100%
ratio of inconsistent statements	Pellet	equals	100%
number of comments	Protégé	equals	100%
label unambiguity	Human	approx	1
label readability	Human	approx	1
type of ontology	Human	approx	2
ontology domain	Human	approx	2
ontology view upon the domain	Human	approx	2
label natural language	Protégé/Human	equals	100%
creation date	Human	equals	92%
versioning	Protégé/Human	equals	100%
author	Protégé/Human	equals	96%
creation tool	Human	equals	94%
documentation	-	-	-

Table 8.3: Evaluation Overview OntoMeta

The three participants were required to study each of the 52 ontologies (using the Protégé ontology editor) and to assign a reference value for the corresponding metadata elements from the range of values defined by the ontology metadata schema. For example, in case of the entry "type of ontology" they were asked to estimate whether a given ontology is an "upper-level", "core", "domain", "application" or "task" ontology. These options are defined in the metadata ontology, as individuals of the class `OntologyType`. The human estimations have been then compared to the output of the tool using the similarity function introduced above.

The test results are summarized in Table 8.3. As expected the tool is able to automatically compute those metadata elements which are easily quantifiable. Moreover, it delivered promising results to more sophisticated issues, such as the *label unambiguity* and *readability*, the *domain of the ontology* and the *view hold by the model* and the *type of ontology* (as regards the generality of the modelled domain, cf. Chapter 5). Pragmatic metadata proved to be difficult to be acquired automatically due to the lack of a predictable structure. Nevertheless the elements considered by OntoMeta can be reliably generated in many situations. This is also due to the detailed analysis of the characteristics of current ontologies, which underlies the implemented heuristics. For many of the considered metadata the decision upon a specific algorithm was justified by statistically studying the particularities of Web available ontologies, as those stored in common ontology repositories or indexed by Semantic Web search engines [38]. In the remaining of this section we focus on these semantic metadata elements.

Table 8.4 summarizes the test results according to the aforementioned similarity function

Metadata	Similarity 1	Similarity 2	Similarity 3	Similarity 4	Similarity 5
Ontology domain	51.9%	19.2%	15.4%	11.5%	2%
View upon the modelled domain	53.8%	17.3%	13.5%	13.5%	1.9%

Table 8.4: Evaluation Results for Domain and View upon the Modelled Domain

[38] for content-related metadata elements. As aforementioned the reference values were specified by the study participants, who were asked to choose the in their opinion most representative `dmoz:Topic` describing the analyzed ontologies.

The type of ontology (with respect to the generality of the modelled domain) could be reliably estimated to a satisfactory degree as well. Though the boundaries between different classes of ontologies are still blurred even for humans, the distinction between application, upper-level, domain and task ontologies proved to be understandable for the study participants, who agreed in the majority of the assessed reference values for this test unit. We obtained a similarity value of 1 in 65% of the cases. The readability and unambiguity of the labels was correctly predicted (i.e. similarity value 1) to a degree of over 75%. Here the participants were given an explanation of the way we understand the two concepts in relation to ontologies and were asked to specify whether the surveyed ontologies fulfill the corresponding requirements or not. The reference values (in this case 1 or 0) were compared to the output of the tool, while this alignment was based on the five point similarity scale. The reason for using it instead of the boolean function was the partial divergent reference values specified by the study participants.

The performed user study demonstrated that many aspects about Web ontologies can be acquired automatically at a viable quality level. A pre-requisite for the realization of these heuristics is the availability of a formal metadata model using standardized taxonomies and classifications. The metadata provision issue could benefit from the standardization of such a model and from its deployment in a wide range of ontology management tools; this would allow many of the pragmatic metadata elements, which can not be ex post generated programmatically, to be created and stored during the operation of the tools. Further on, the realization of fully-fledged ontology discovery solutions could trigger the refinement of the proposed model towards a more structured definition of currently fuzzy entries such as ontology documentation or ontology application systems. Such aspects play a decisive role in improving the real-world reusability of Web ontologies, as indicated many times in this thesis.

8.3.3 PROMI User Study

In an in situ experiment we tested the prototypical implementation of the two main components of PROMI, which are concerned with the evaluation, and the customization, merging and integration of the reuse candidates, respectively. The organization of the study was adjusted to the particularities of these tools: in case of the ontology evaluation tool we needed to compare the results delivered by the automatic procedure with those estimated in the same conjuncture by human experts; this approach is also adequate for testing the relevance of the results delivered by the context-dependent selection of the merging strategy, while not being optimal for the overall merging/integration component. The reason for this divergence fundamentally relies in the intrinsic nature of the latter—the novelty of the merging tool is primarily methodological nature. It does not propose new matching or merging technologies, but integrates existing elementary solutions into a new workflow. Hence, the way ontologies are customized, merged and integrated within PROMI can be meaningfully evaluated only from a usability point of view; the quality of individual matchers is expected to be tested ex ante by their developers (refer, for example, to [42] for a discussion on this topic).

Both the ontology evaluation and the matching selection method can be tested in a similar way by comparing the results of the provided automatic support with the way human experts accomplish the corresponding tasks. Hence the objective of the study is to demonstrate to which extent these methods correspond to expert judgement and, thus, to which extent they are able to provide technological and methodological support to ontology developers.

The standard procedure to evaluate such search problems is with the help of metrics such as precision, recall and f-measure [6]. The latter is defined as

$$F = \frac{2 * P * R}{P + R} \tag{8.1}$$

$$\tag{8.2}$$

We now turn to a description of the test setting of the evaluation task. The testing of the matching selection tool is not discussed in this work and will be elaborated in the context of the MOMA framework [148, 150] in the future.

In relation to the ontology evaluation task precision and recall are interpreted as follows

$$P = \frac{numberofrelevantretrievedontologies}{numberofretrievedontologies} \tag{8.3}$$

$$R = \frac{numberofrelevantretrievedontologies}{numberofrelevantontologies} \tag{8.4}$$

The total number of relevant ontologies is determined by humans, a task which is acknowledged to be challenging in classical information retrieval [6]. In relation to ontology reuse, estimating the total number of relevant hits is hampered by the fact that the evaluators are not familiar with the surveyed ontologies. Provided these ontologies are available to the evaluators, the efforts invested in studying them is directly proportional with the size of the ontologies, the complexity of the modelled domain and the quality of the documentation—all of them are well-known to be time-consuming tasks even for experienced individuals.

In order to overcome such problems we asked three ontology engineering experts to estimate the best five reuse candidate from a set of 21 human resources ontologies with percentage scores between 1 and 100. The participants possessed some knowledge in the domain of interest of the ontologies, though not being well-versed in the field. The operation of the study was supervised by the evaluator, who was also the developer of the evaluated artifact. The domain experts were provided with a detailed description of the context of the ontology reuse process, in terms of the prospected application system, and the tasks in which the ontologies to be built will be involved in. The evaluator opted for two typical application scenarios for systematically collecting and interpreting the test results:

- **Semantic annotation/indexing**: the ontology should be used to classify job postings and job seeker profiles in the healthcare provision domain. This task was supposed to be realized manually, while the information sources to be indexed were assumed to be formalized in English. This scenario shows major similarities with the eRecruitment case study we described in the previous sections.
- **Semantic search**: the ontology should form the basis for a reasoning-supported semantic retrieval component in the aforementioned job portal. Again the documents are formalized in English and cover the domain of healthcare human resources.

The organizers did not impose any particular evaluation workflow or quality criteria to support or influence the decision making process—the goal of the study was to investigate to which extent the designed method simulates the way humans approach the ontology usability assessment issue, no matter what heuristics they follow in this endeavor.

The first participant was asked to carry out the task based on the metadata descriptions of the 21 ontologies. The second participant was provided with a list of the reuse candidates and links to additional documentation. The third expert used PROMI for this purpose. Additionally to the list of potential reuse candidates, the evaluator collected data about the effort invested in each of the experiments. For the tool-supported task, the efforts included several iterations of the evaluation workflow, caused by an inappropriate selection of the evaluation criteria, and by the additional resources required to get familiar with the proposed approach. The costs for the metadata-driven evaluation also covered the time invested in gathering the associated information and formalizing it in a schema-conform manner.

The collected data showed that the effort invested in the evaluation task can be reduced provided ontology metadata information. Further on, the approach implemented in PROMI, which is based upon the usage of contextual dependencies in deciding upon the usability of a specific ontology in a new application scenario, appeared to cause further cost savings.

The content-driven evaluation experiment requested 0.62 person months. The study hypothesis was confirmed by the efforts invested in the parallel experiments: using the metadata about ontologies implied savings of approximately 33%, while PROMI was responsible for a further decrease of almost 30% . The relatively high costs arisen in the first experiment can be elucidated if we recall that the experts were required to collect the evaluation-relevant information about the surveyed ontologies manually. Due to the heterogeneity of the reuse candidates, this operation could not be carried out efficiently. Further on, the study participant involved in this activity expressed his concerns that the information he used in the decision making process is representative for this purpose. This overload was avoided in the

SCENARIO	ONTOLOGY	SCORE
SEMANTIC SEARCH	Occupational Net (O*NET)	60%
	SOC/NOC	50%
	Factiva	30%
	Healthcare provider taxonomy	30%
	MeSH	20%
SEMANTIC ANNOTATION	Occupational Net (O*NET)	80%
	SOC/NOC	70%
	Healthcare provider taxonomy	60%
	Factiva	60%
	Occupational Thesaurus	40%

Table 8.5: Results of the Expert-based Evaluation on the Basis of the Ontological Content

SCENARIO	ONTOLOGY	SCORE
SEMANTIC SEARCH	Occupational Net (O*NET)	60%
	Standard Occupational Classification (SOC)	30%
	National Occupational Classification (NOC)	30%
	Industry and Occupations 2000	20%
	Factiva	20%
SEMANTIC ANNOTATION	Occupational Net (O*NET)	70%
	Standard Occupational Classification (SOC)	60%
	National Occupational Classification (NOC)	60%

Table 8.6: Results of the Expert-based Evaluation on the Basis of the Ontology Metadata

SCENARIO	ONTOLOGY	SCORE
SEMANTIC SEARCH	Occupational Net (O*NET)	72%
	National Occupational Classification (NOC)	64%
	Occupation Thesaurus	61%
	National Generic Competency Standards	60%
	Industry and Occupations 2000	55%
SEMANTIC ANNOTATION	Occupational Net (O*NET)	75%
	Healthcare provider taxonomy	71%
	Dictionary of Occupational Titles	60%
	Factiva	44%
	Occupation Thesaurus	40%

Table 8.7: Results of the Tool-based Evaluation

metadata-supported approach at the costs of a partially insufficient number of ontological hits with relatively lower relevance scores. The expert justified this state of affairs by the absence of several metadata elements which could have fortified his decisions. Nevertheless, the presence of a structured set of ontology features—with a proven and tested relevance to the reusability question—contributed to the human-perceived acceptance of the evaluation results. Besides substantial cost savings, the main advantage of the tool-based experiment was

the provision of a systematic, yet flexible workflow for conducting the usability assessment at comparable results.

The rankings generated in each of the three endeavors are surely comparable. In particular we mention the difficulties encountered to differentiate between classifications such as O*NET, SOC and NOC. The experts motivated the unanimous decision for O*NET by the level of detail of the covered information and the availability of tools supporting its usage. Nevertheless this fact suggests the need for a refinement of the evaluation heuristics and for additional tests on other domains using detailed metadata descriptions.

The lower relevance scores assigned to the second application scenario become clear if we recall that one of the key requirements for implementing ontology-based retrieval systems is the formality of the represented knowledge. This requirement was not fulfilled by any of the 21 surveyed reuse candidates.

Despite the limited scope of the experiments we believe that the collected data complements and reinforces the results obtained in the parallel evaluation tracks presented in this chapter. The application scenarios chosen as test cases are compatible with the case study in which we analyzed the practicability of our methodological approach. Further on, the test results indicate that a partial automatization of the evaluation step within ontology reuse is possible, though further tests are required to rigorously validate the algorithm. They also confirm the benefits of our context-oriented approach with respect to the user-perceived quality of the performed process and the associated costs.

8.4 Goal-free Evaluation

Additionally to the professional reviews the core of our solution was compared to other approaches sharing the goal of enabling or improving the reuse of Web ontologies. This comparison was performed in compliance with existing quality frameworks in the field, which are expected to be widely accepted by the Semantic Web community and the potential users of our research.

8.4.1 Ontology Reuse Methodology

The goal-free evaluation of the methodology builds upon the quality framework for ontology engineering methodologies introduced in Chapter 3.5. This framework, originally conceived by Fernández-López and Gómez-Pérez [63], covered the following criteria:

1. **Level of detail**: the methodology should provide a fine-grained and precise description of the process and its main steps, decomposing them into manageable activities and assigning these activities to roles. It should clearly define the inputs, outputs and decisions of each process phase.

2. **Relation to application scenarios**: the process model should suggest means to optimize its application for particular real-world scenarios.

3. **Recommended life cycle**: the methodology should make a clear statement about the order of execution of the main process steps and the underlying activities and should allow users to directly control intermediary process results.

4. **Support methods and tools**: in direct relation with the focus on the application scenarios, the usability of the process-driven methodology should be increased by methods and tools.

5. **Methodology validation**: the process description should be subject to a careful evaluation.

Table 8.8 compares our methodology to related approaches to ontology reuse. The selection of the latter was performed on the basis of the literature survey associated with our ontology reuse feasibility study. Pinto and Martins were among the first who took a look at the question of ontology reuse from a process perspective [180]. The ONION methodology elaborated the subject for semi-structured ontological sources sharing the same domain [73]. Finally, Goméz-Pérez and Rojas-Amaya [81] concentrated on the re-engineering aspects of ontology reuse. OntoMetric aims at increasing the reusability of Web ontologies by aiding ontology developers in assessing their relevance in consideration of a set of (application-narrow) requirements [128]. The process is based on decision trees and quantified ontology features.

The comparison reveals that our methodology is well suited for ontology engineering tasks where reuse is a major concern. Further on, it is the first ontology reuse methodology explicitly giving an account on the application-oriented aspects of reusability, an issue which is essential for the industrial take up of semantic technologies. Its usability in real-world scenarios is enhanced by the precision of its descriptions, the coverage of all main reuse stages and an integrated inventory of methods and tools.

Our methodology is currently less adequate for use cases in which the customization, merging and integration of the reuse candidates is essential. Though we provide support at process level for this kind of activities, the implemented methods and tools require further refinements and extensions.

8.4.2 Ontology Reuse Metadata Model

Metadata and metadata standards have a long-tradition in a variety of computer sciences areas, such as digital libraries or data management and maintenance systems. We briefly mention related metadata standards, eminently including those ones relevant for the Semantic Web. The comparison relies on the survey in Section 1.4.2 from the introductory chapter.

By contrast to metadata standards used for the description of electronic information sources [153], be that documents or multimedia, our metadata model covers elements which are specific to the ontology engineering field. Hence it enables a more precise description of ontologies and the ontology life cycle. The importance of this specific information was demonstrated by the methods and tools using the metadata schema to automatize ontology reuse activities such as the evaluation of ontologies.

Further metadata schemas emerged in relation to ontology search engines. The Semantic Web search engine Swoogle [49] implicitly uses a simple metadata schema, which covers sta-

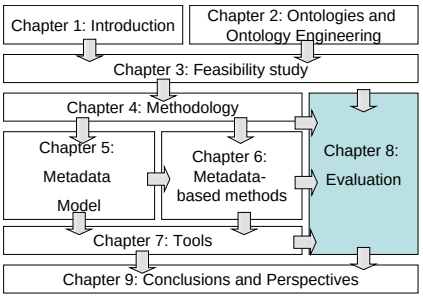
	Pinto & Martins	Gangemi et al	Gómez-Pérez & Rojas-Amaya	Lozano-Tello & Gómez-Pérez	Paslaru Bontas
LEVEL OF DETAIL					
Ontology discovery	marginally discussed	-	-	-	discussed in detail
Ontology evaluation	discussed in detail	discussed in detail from an ontological perspective	marginally discussed	discussed in detail	discussed in detail
Ontology customization	discussed in detail	marginally discussed	discussed in terms of re-engineering	-	marginally discussed
Ontology merging & integration	discussed in detail in terms of integration operations	marginally discussed	marginally discussed	-	discussed in detail at process level
Process participants	marginally discussed	-	-	-	discussed in detail
Low-level activities	discussed in detail	-	-	discussed in detail for ontology evaluation	discussed in detail
APPLICATION SCENARIOS					
Applicability across scenarios	-	-	-	-	discussed in detail
Application-narrow guidelines	-	-	-	marginally discussed	discussed in detail
LIFE CYCLE					
Recommended life cycle	sequential, incremental	-	sequential	sequential, incremental	sequential, incremental
SUPPORT METHODS AND TOOLS					
Methods	Reference Ontology, metrics for evaluation	-	re-engineering method	Reference ontology, method for ontology evaluation	ontology metadata, methods for evaluation and matching
Tools	-	-	WebOde plug in	OntoMetric	PROMI, OntoMeta
Integration	-	-	integrated solution	integrated to WebOde	integrated solution
METHODOLOGY VALIDATION					
Approach	case study	case study	case study	case study	several case studies and professional reviews

Table 8.8: Results of the Goal-free Evaluation of the Methodology

tistical information automatically captured from ontology implementations. Its range is thus much limited than our approach. Metadata is also present in ontology repositories. However, the majority of these repositories rely on a restricted, implicitly declared vocabulary, whose meaning is not machine-understandable (cf. Section 1.4.2). Some core pragmatic metadata elements for ontologies have been included as so-called "annotation properties" in ontology representation languages such as OWL [176] and RDFS [22].

The need for a dedicated model is acknowledged in the Semantic Web community; [112] performed a detailed requirements analysis for ontology metadata. We used the results of this survey in our work.

8.5 Summary



This chapter rounds off the contents of this thesis by presenting the way we validated its main outcomes. After introducing the evaluation framework and justifying its relevance in the context of our research, we elaborated the results we obtained by evaluating the proposed solution with respect to several dimensions. Professional reviewers positively appreciated the technical quality of our proposal according to their level of experience in the field. Further on, we discussed its applicability in real-world scenarios by means of a goal-based evaluation supported by empirical studies. Finally, a goal-free evaluation revealed how our work can be situated in the current state of the art in ontology reuse and in which situations it can be optimally applied.