Contributing Evidence to Data-driven Ontology Evaluation
Workflow Ontologies Perspective

Hlomani Hlomani and Deborah A. Stacey
School of Computer Science, University of Guelph, Guelph, Canada

Keywords: Workflow, Ontology, Ontology Evaluation, Corpus-driven Evaluation, Latent Semantic Analysis.

Abstract: Ontologies have established themselves as the single most important semantic web technology. They have attracted widespread interest from both academic and industrial domains. This has led to an increase in ontologies created. It has become apparent that more than one ontology may model the same domain yet they can be very different. The question then is, how do you determine which ontology best fits your purposes? This paper endeavours to answer this question by reviewing relevant literature and instantiating the data-driven ontology evaluation methodology in the context of workflow ontologies. This evaluation methodology is then evaluated through statistical means particularly the Kruskal-Wallis test and further post hoc testing using the Mann-Whiteny U test.

1 INTRODUCTION

Semantic web technologies, particularly ontologies, have seen increased interest from both academic and industrial domains. This is evident in the multitudes of academic publications, tool sets, methodologies and applications that either reference or are driven by ontologies. By definition, an ontology is a conceptualization of a target domain that explicitly specifies the concepts in a domain and the relations between them (Gruber, 1993). By using an ontology, you define a language for that domain thereby standardizing the use of concepts.

The increase of research interest in ontologies has led to a considerable increase in the number of ontologies(Vrande, 2009). An inevitable reality though, is that often there will be ontologies that model the same domain yet are very different in their modelling of the domain and the constructs used in the modelling of that domain. This is largely because, while an ontology creates a shared vocabulary of a domain, it is a conceptualization of the domain. This conceptualization is largely dependant upon the modeller's perception of the domain (Brank et al., 2005). The question then is: how does one determine which ontology best fits their purpose? This is the fundamental question that this paper investigates.

2 BACKGROUND

The need for ontology evaluation is a topic that needs no introduction. There are several factors about ontologies that heighten the need for ontology evaluation. These are discussed in Section 2.1. In an attempt to address this need, the research community has contributed solutions in the form of frameworks, methodologies and tools. Section 2.2 discusses some of these contributions.

2.1 Motivation for Ontology Evaluation

The following facts about ontologies heighten the need for their evaluation:

Play a Pivotal Role. Ontologies play a critical role in the semantic web and ontology-driven (enabled) applications. Proper representation of domain knowledge is therefore an obvious necessity.

Shared Conceptualization. By definition ontologies are an explicit specification of a shared conceptualization, in other words, they are a model of knowledge for a specific domain (Gruber, 1993). While the ontology should be a "shared conceptualization" of the domain, subjectivity is always a concern since it represents the time, place, and cultural environment in which it was created as well as the modeller's perception of the domain (Brank et al., 2005; Brewster et al., 2004).

Hlomani H. and Stacey D..
Contributing Evidence to Data-driven Ontology Evaluation - Workflow Ontologies Perspective.
DOI: 10.5220/0004543602070213
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2013), pages 207-213
ISBN: 978-989-8565-81-5

Potential of Reuse. Ontology reuse is perhaps the most obvious motivation to evaluate ontologies. Before an ontology can be reused, one has to evaluate the ontology's quality and most importantly, the ontology's fitness for a particular purpose.

2.2 Ontology Evaluation Methodologies

Several methods for ontology evaluation have been proposed over the years. The main methods have been surveyed by (Vrande, 2009; Brewster et al., 2004) and more recently by (Ouyang et al., 2011) to include the following:

Comparison against a "Gold Standard". The gold standard may itself be an ontology. The problem with this method is that it is difficult to establish the quality of the gold standard.

User-based Evaluation. This typically involves evaluating the ontology through users' experiences. The problem with this method is that it is difficult to establish objective standards pertaining to the criteria (metrics) for evaluation. In addition it is also hard to establish who the right users are.

Application-based Evaluation. This would typically involve evaluating how effective an ontology is in the context of an application. While this may be practical for the purposes of evaluating a single ontology, it may be challenging to evaluate a number of ontologies in an application area to determine which one is best fitted for the application especially in an automated fashion.

Congruence Evaluation. This involves evaluating the "fitness" or congruence between the ontology and a domain of knowledge. Several approaches have been pursued including comparison of the ontology to a "gold standard" as discussed above. Another approach is to evaluate the ontology or ontologies against knowledge from the domain the ontologies represent. More specifically, comparison can be made against a corpus or text extracted from the documents about the domain (*e.g.* (Brewster et al., 2004)).

Hybrid Evaluation: User-based Evaluation and Corpus-based. This method is exemplified by (Ouyang et al., 2011) which combines the corpus-based and user-based evaluations. The ontology here is evaluated against a set of metrics (coverage, coherence and coupling). Users are allowed the flexibility to weigh the influence of each of the metrics on the evaluation.

It is important to know that there is no "gold standard" evaluation; however, one should choose an evaluation technique based on the purposes (reasons) of the evaluation (Vrande, 2009).

3 PROPOSED ONTOLOGY EVALUATION

The ontology evaluation of this paper is an instantiation of the congruence ontology evaluation methodology initially proposed by (Brewster et al., 2004). The main motivation of this research is that, while there are some general methodologies proposed for ontology evaluation, there is a paucity of evidence in support of these methodologies. This is particularly true for the congruence evaluation or data-driven ontology evaluation methodology.

In defining and instantiating this methodology, (Brewster et al., 2004) considered the domain of arts. Our paper evaluates ontologies in the domain of workflow management. The general steps followed in this investigation are: corpus definition, similarity calculation and statistical evaluation.

3.1 Corpus Definition and Distance Measure

The ontologies considered in this paper pertain to the concept of workflow. A workflow is by definition:

> *"The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules"*
> (WFMC, 1999).

Workflow ontologies model the workflow domain based on concepts that have some relation to this definition. This is because an ontology is a formal conceptualization of a domain of interest (OMG, 2009; W3C, 2009; Gruber, 1993). A conceptualization is an abstraction of that which we wish to represent.The corpus for the ontology evaluation of this paper consists of text from documents about the workflow and process modelling domain. These documents consist of one hundred (100) peer reviewed academic articles. These documents were obtained through the assistance of three major search facilities: the IEEE eXplore, Google Scholar and Primo Central (via the university library). The key phrases that were used to search for content are: Workflow modelling, Business Process modelling, Workflow modelling languages, Business process modelling languages. We will refer to this corpus as the domain corpus.

In addition to the domain corpus we also define the ontology corpus which consists of the concepts extracted from the ontologies. This forms the documents to be compared to the domain corpus.

Following the provision of text which eventually forms the corpus, there is a need for some representa-

Table 1: Possible results.

Paper	Similarity
$Paper_1$	1
$Paper_2$	0.1455
$Paper_3$	0.9154
$Paper_4$	0.0463
$Paper_5$	0.8798
.	.
.	.
$Paper_n$	0.8798

tion model which in our case implies a form of automated term recognition. This reduces the documents to words that are representative of each document in the corpus thereby producing a matrix from which we can do calculations on. Latent Semantic Analysis (LSA) (Hofmann, 1999; Deerwester et al., 1990) was used for this purpose. We particularly used the Text Mining Library implementation (TML) (Villalon and Calvo, 2011) of LSA. Cosine distance between every document in the domain corpus and every document in the ontology corpus were calculated. This produced results in the form depicted in Table 1 for each ontology under investigation.

3.2 The Workflow Ontologies

1. The *IntelLEO* Workflow Ontology (Jovanovic et al., 2011). The *IntelLEO* Workflow Ontology has a sense of "traditional" workflows. This is because the ordering (either sequential or parallel) of activities to achieve some goal can be achieved. The ontology, however, lacks the expressive ability to capture even some of the most basic concepts of process models (*e.g.* the notions of routing beyond just "sequence" and "parallel").
2. The Protegé ontology: workflows for collaborative ontology development (Sebastian et al., 2008). This work focuses on a specific set of workflows viz.: those that describe collaboration during ontology development. These type of workflows are human-centred in that most, if not all, the activities require some form of human action. The activities are defined in terms of the steps (or states) that a proposed change goes through before it is published.
3. The business process ontology (BMO) (Jenz, 2003) is designed to be generic in its description of business processes. This is achieved through a static representation of a business process by focusing on *activities* or *tasks* as the "building blocks". The representation of the business process in an ontology then achieves two main goals viz.: provide a vendor-neutral and platform-independent description of the business process, and provide both human-understandable and machine-readable descriptions of the business process.
4. The process ontology (Martin et al., 2007) is an interesting addition to the list. This is because it has its origins from the context of web services (a process is a subset of the OWL-S description of web services). It is relevant in our context since it does define the workflow constructs that include control flow, input and output (pre-conditions and post-condition), categorization of the process concept (*i.e.* composite, atomic process etc.) and the like.
5. The workflow ontology by Tim Berners-Lee hereby dubbed the *Flow ontology* offers another perspective to workflow modelling. Unfortunately, there is not much documentation about the ontology. However, through inspecting the ontology, it was found that the ontology follows the trend of the other ontologies described in this paper.

3.3 Statistical Evaluation

Our experiment endeavours to investigate which of the hypotheses is true and therefore, answer the question, which of the ontologies is more representative of the workflow domain? The hypotheses are defined as thus:

1. Null Hypothesis (H_0): All ontologies have equal similarity on average or $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$, where $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$ are the mean similarity scores for the five ontologies.
2. Alternative Hypothesis (H_1): The mean similarity of at least one ontology is significantly different.

To accept or reject any of these hypotheses, a statistical procedure is followed. The sample data is the similarity scores for each ontology (roughly one hundred records for each). As will become obvious in later sections, the type of statistical test to perform will be depended upon the distribution the data follows (if normal then a parametric test like one-way ANOVA otherwise a non-parametric alternative such as the Kruskal-Wallis test). These are discussed in subsequent sections.

4 RESULTS AND ANALYSIS

This sections presents the results of the experiment set out in Section 3.3. These include results for the

normality test and statistical tests (Kruskal test and Mann-Whitney U test-pairwise comparisons).

4.1 Normality Test

An assessment of the normality of data is a prerequisite for many statistical tests because normal data is an underlying assumption in parametric testing. The normal Q-Q Plot was created for the individual samples (five plots) to determine their normality. The normality plot for the *Flow ontology* is depicted in Figure 1. The plot shows that this sample is not normally distributed since its data points deviate from the diagonal line (plot of the expected sample if the data is normally distributed) in a non-linear fashion. This observation is true for the other samples (similarity scores for the other ontologies) as can be see in Figures 2, 3, 4 and 5.

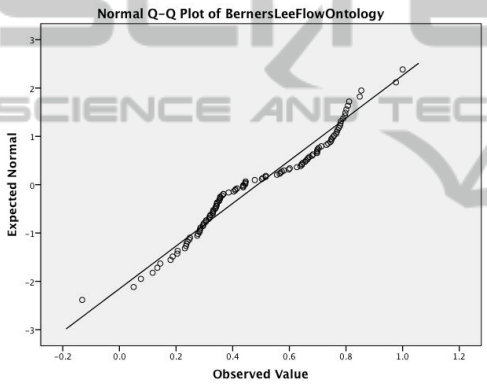

Figure 1: Normality plot for the *Flow ontology* by Tim Berners Lee.

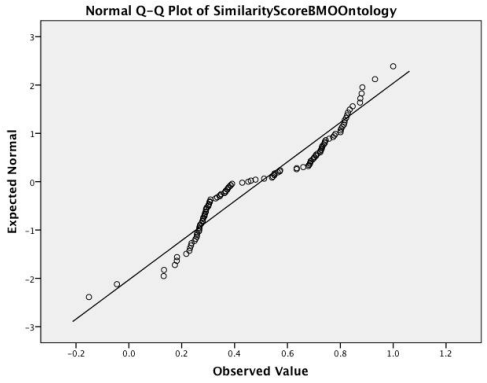

Figure 2: Normality plot for the *BMO ontology*.

4.2 Differences between the Different Ontologies

It is apparent that this study deals with a five level single factor type of scenario, where the single factor is

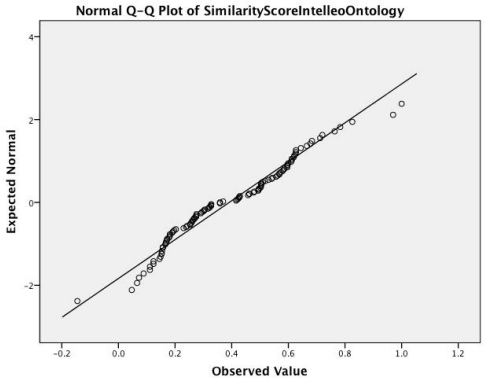

Figure 3: Normality plot for the *Intelleo ontology*.

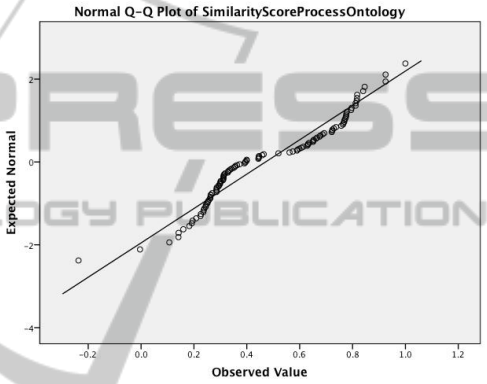

Figure 4: Normality plot for the *Process ontology*.

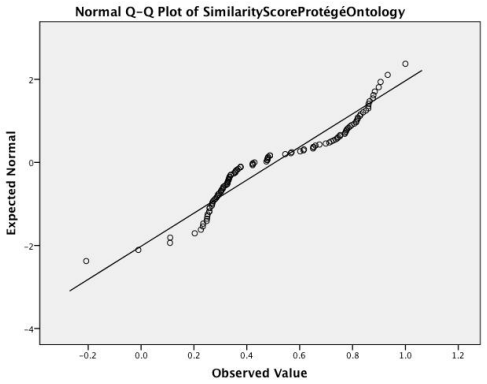

Figure 5: Normality plot for the *Protegé ontology*.

the variable of interest (the ontology similarity) and the different settings or levels are the different ontologies. Our main interest is to find out if the ontologies of interest have equal similarity to the corpus on average. The normality test reveals that all samples of interest are not normal distributed, hence ruling out the possibility of doing parametric statistics (*e.g.* *t* test, ANOVA, etc.). Since we have multiple levels (more than two comparisons), a non-parametric alternative to a one-way ANOVA (*i.e.* Kruskal-Wallis test) was

conducted. The hypotheses to be tested are stated in Section 3.3 A point of consideration in the analysis is the **significance level:** $\alpha = 0.05$ which thus defines a **rejection region**, *i.e.* reject H_0 if p_value $\leq \alpha$.

4.2.1 Kruskal-Wallis Test

The results of the Kruskal-Wallis test are summarized in Tables 2 and 3.

Table 2: Ranks.

Ontology	N	Mean Rank
BMO	108	455.19
Flow (Berners Lee)	107	190.48
Intelleo	106	245.50
Process	105	266.40
Workflow (Protegé)	104	165.17

Table 3: Test Statistics.

	Similarity Score
Chi-Square	237.835
df	4
Asymp. Sig.	.000

At the $\alpha = 0.05$ level of significance, there exists enough evidence to conclude that there is a difference in the median test scores (and, hence, the mean test scores) among the five ontologies (rather at least one of them is significantly different.)

4.2.2 Pairwise Comparisons

The Kruskal-Wallis test previously discussed only tells us if there is a significant difference between the ontologies. Pairwise comparisons of the ontologies will identify where the differences lie.

Since the study observes five different levels (*i.e.* ontologies) of a single factor (*i.e.* similarity), the number of pairwise comparisons is

$$k = (n*(n-1))/2, where\, n = number\, of\, levels \quad (1)$$

Hence, $k = (5*(5-1))/2 = 10$. To counteract the problem of multiple comparisons the results (p_values) have been subjected to the Bonferroni correction (corrected $p\_value = p\_value/k$. Tables 4, 5 and 6 depict the results from the Mann-Whitney comparisons of the ontologies' similarity (with Tables 5 and 6 showing both the original and corrected p_values).

The three Tables (4, 5 and 6) are very useful because they indicate which ontology had the highest similarity score (highest mean rank), the actual significance value of the test (U value and the asymptotic significance (2-tailed) p-value) and continuation of

Table 4: Ranks.

Ontology	N	Mean Rank	Sum of Ranks
BMO vs Flow			
BMO	108	156.13	16862.50
Flow (Berners Lee)	107	59.42	6357.50
BMO vs Intelleo			
BMO	108	155.27	16769.50
Intelleo	106	58.83	6235.50
BMO vs Process			
BMO	108	153.66	16595.50
Process	105	59.00	6195.50
BMO vs Protegé			
BMO	108	153.63	16591.50
Protegé	104	57.56	5986.50
Flow vs Intelleo			
Flow (Berners Lee)	107	94.17	10076.50
Intelleo	106	119.95	12714.50
Flow vs Process			
Flow (Berners Lee)	107	83.63	8948.50
Process	105	129.80	13629.50
Flow vs Protegé			
Flow (Berners Lee)	107	115.26	12332.50
Protege	104	96.48	10033.50
Intelleo vs Process			
Intelleo	106	106.29	11266.50
Process	105	105.71	11099.50
Intelleo vs Protegé			
Intelleo	106	120.94	12819.50
Protegé	104	89.76	9335.50
Process vs Protegé			
Process	105	130.88	13742.50
Protegé	104	78.87	8202.50

the significance value of the test, respectively. However, concluding that there is a higher congruence between a particular ontology and the corpus is only valid upon examination of the statistical significance of the difference between that ontology's similarity score and that of the other ontologies.

From this data, it can be concluded that there is a statistically significant difference between the similarity of the ontologies (pairwise comparisons) with the exception of the *Intelleo ontology* when compared to the *Process ontology*. The comparison of these two ontologies, reported a p_value of 0.0945 which is higher than the α value, hence, it falls outside the rejection region for the null hypothesis (the null hypothesis is accepted for this comparison). We can conclude that these two ontologies' coverage of the corpus is similar. Similarly, having ascertained statistical significance in the differences between the means, the original research question can now be addressed. The research question was stated as thus: "Given a pool of workflow ontologies, which of the ontologies has a higher congruence with the workflow corpus and hence, is more representative of the workflow domain?" The answer lies in Table 4. In each of the pairwise comparison, the *BMO ontology* had the highest similarity score (mean ranks) relative to the others it

Table 5: Test Statistics.

	Similarity Score
BMO vs Flow	
Mann-Whitney U	579.500
Wilcoxon W	6357.500
Z	-11.398
Asymp. Sig. (2-tailed)	.000
Corrected p_value	.000
BMO vs Intelleo	
Mann-Whitney U	564.500
Wilcoxon W	6235.500
Z	-11.392
Asymp. Sig. (2-tailed)	.000
Corrected p_value	.000
BMO vs Process	
Mann-Whitney U	630.500
Wilcoxon W	6195.500
Z	-11.206
Asymp. Sig. (2-tailed)	.000
Corrected p_value	.000
BMO vs Protegé	
Mann-Whitney U	526.500
Wilcoxon W	5986.500
Z	-11.399
Asymp. Sig. (2-tailed)	.000
Corrected p_value	.000
Flow vs Intelleo	
Mann-Whitney U	4298.500
Wilcoxon W	10076.500
Z	-3.052
Asymp. Sig. (2-tailed)	.002
Corrected p_value	.0002

Table 6: Test Statistics Continued.

	Similarity Score
Flow vs Process	
Mann-Whitney U	3170.500
Wilcoxon W	8948.500
Z	-5.480
Asymp. Sig. (2-tailed)	.000
Corrected p_value	.000
Flow vs Protegé	
Mann-Whitney U	4573.500
Wilcoxon W	10033.500
Z	-2.234
Asymp. Sig. (2-tailed)	.025
Corrected p_value	.0025
Intelleo vs Process	
Mann-Whitney U	5534.500
Wilcoxon W	11099.500
Z	-069
Asymp. Sig. (2-tailed)	.945
Corrected p_value	.0945
Intelleo vs Protegé	
Mann-Whitney U	3875.500
Wilcoxon W	9335.500
Z	-3.717
Asymp. Sig. (2-tailed)	.000
Corrected p_value	.000
Process vs Protegé	
Mann-Whitney U	2742.500
Wilcoxon W	8202.500
Z	-6.216
Asymp. Sig. (2-tailed)	.000
Corrected p_value	.000

was compared to. We can, therefore, conclude that there is a better congruence between the *BMO ontology* and the workflow domain than there is between the other ontologies. The ranking of the congruence between the workflow ontologies and the workflow corpus is depicted in Table 7. This is based on the pairwise comparison of the ontologies' mean similarity scores (refer to Table 4).

Table 7: The rank of the congruence between the ontologies and the corpus.

Ontology	Rank
BMO	1
Intelleo	2
Process	2
Flow	3
Protegé	4

5 CONCLUSIONS

This paper has been about a corpus-driven ontology evaluation. It considered the particular instance of workflow ontologies. The paper had set out to answer the question, how can we determine which ontology in a pool of ontologies best fits a particular domain? This has been answered through the motivation of the corpus-driven evaluation methodology and further validated through statistics.

From a set of five ontologies, it was demonstrated that there was significant statistical difference between the ontologies' similarity scores. It was particularly concluded that there was a better congruence between the *BMO ontology* and the workflow domain than there was with the other ontologies. This could be attributed to a better coverage of the domain by the ontology, hence, rendering the ontology more representative of the domain. The results of such a study could lead into the winning ontology being adopted, expanded and finally applied to the application it was initially evaluated for.

REFERENCES

Brank, J., Grobelnik, M., and Mladenić, D. (2005). A survey of ontology evaluation techniques. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005)*, pages 166–170.

Brewster, C., Alani, H., Dasmahapatra, S., and Wilks, Y. (2004). Data-driven ontology evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Gruber, T. R. (1993). Toward principles for the design of ontologies used for knowledge sharing. In *International Journal of Human-Computer Studies*, pages 907–928. Kluwer Academic Publishers.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. ACM.

Jenz, D. E. (2003). Simplifying the software development value chain through ontology-driven software artifact generation.

Jovanovic, J., Siadaty, M., Lages, B., and Spors, K. (2011). IntelLEO workflow ontology. http://www.intelleo.eu/ontologies/workflow/spec/#s31.

Martin, D., Burstein, M., Mcdermott, D., Mcilraith, S., Paolucci, M., Sycara, K., Mcguinness, D. L., Sirin, E., and Srinivasan, N. (2007). Bringing semantics to web services with owl-s. *World Wide Web*, 10(3):243–277.

OMG (2009). *Ontology Definition Metamodel - OMG Document Number: formal/2009-05-01*. Object Management Group.

Ouyang, L., Zou, B., Qu, M., and Zhang, C. (2011). A method of ontology evaluation based on coverage, cohesion and coupling. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, volume 4, pages 2451 –2455.

Sebastian, A., Noy, N., Tudorache, T., and Musen, M. (2008). A generic ontology for collaborative ontology-development workflows. In Gangemi, A. and Euzenat, J., editors, *Knowledge Engineering: Practice and Patterns*, volume 5268 of *Lecture Notes in Computer Science*, pages 318–328. Springer Berlin / Heidelberg.

Villalon, J. and Calvo, R. A. (2011). Concept maps as cognitive visualizations of writing assignments. *International Journal of Educational Technology and Society*, 14(3):16–27.

Vrande, D. (2009). Ontology Evaluation. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, pages 293–313. Springer Berlin Heidelberg, Berlin, Heidelberg.

W3C (2009). *OWL 2 Web Ontology Language, Document Overview: W3C Recommendation 27 October 2009*. World Wide Web Consortium. http://www.w3.org/TR/2009/REC-owl2-overview-20091027/.

WFMC (1999). *The Workflow Management Coalition Specification, Workflow Management Coalition Terminology and Glossary; Document Number: WFMC-TC-1011*. Workflow Management Coalition.