Contents lists available at SciVerse ScienceDirect

J. of Acc. Ed.

journal homepage: www.elsevier.com/locate/jaccedu

Main article

Ontology-based e-assessment for accounting: Outcomes of a pilot study and future prospects

Kate Litherland [a,*], Patrick Carmichael [b,1], Agustina Martínez-García [a]

[a] *Faculty of Education, Community and Leisure, Liverpool John Moores University, Barkhill Building, IM Marsh Campus, Barkhill Road, Liverpool L17 6BD, UK*
[b] *Faculty of Education and Sport, University of Bedford, Polhill Road, Bedford MK41 9AE, UK*

ARTICLE INFO

Article history:
Available online 23 April 2013

Keywords:
Semantic technologies
Subject ontologies
Online assessment
Empirical study

ABSTRACT

This article reports on a pilot of a novel ontology-based e-assessment system in accounting that draws on the potential of emerging semantic technologies to produce an online assessment environment capable of marking students' free-text answers to questions of a conceptual nature. It does this by matching their response with a "concept map" or "ontology" of domain knowledge expressed by subject specialists. The system used, OeLe, allows not only for marking, but also for feedback to individual students and teachers about student strengths and weaknesses, as well as to whole cohorts, thus providing both a formative and a summative assessment function. This article reports on the results of a "proof of concept" trial of OeLe, in which the system was implemented and evaluated outside its original development environment (an online course in education being used instead in an undergraduate course in financial accounting. It describes the potential affordances and demands of implementing ontology-based assessment in accounting, together with suggestions of what needs to be done if such approaches are to be more widely implemented.

1. Introduction

In this paper, we describe the implementation and initial testing of a novel approach to online assessment or "e-assessment" of student understanding that draws on emerging semantic web tech-

* Corresponding author. Tel.: +44 (0)151 231 4608.
E-mail addresses: k.litherland@ljmu.ac.uk (K. Litherland), patrick.carmichael@beds.ac.uk (P. Carmichael), a.martinez-garcia@ljmu.ac.uk (A. Martínez-García).
[1] Tel.: +44 (0)1234 793100.

0748-5751/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jaccedu.2013.03.003

nologies. Specifically, it explores how an e-assessment built around an 'ontology' or conceptual map of a particular domain – in this case, introductory financial accounting, can be used to provide both valid and reliable marking of short free-text answers that typically involved students engaging with between two and five key concepts. The aim was to offer teachers and students formative feedback as to which concepts were well understood and which required further attention and revision.

This trial project was funded by ACCA (Association of Chartered Certified Accountants) and the International Association for Accounting Education and Research (IAAER) under a programme of research to support the work of IFAC's International Accounting Education Standards Board (IAESB). Specifically, it was a response to a call to explore alternatives to well-established approaches in e-assessment (such as multiple-choice tests) and to support "question types ...which provide a more valid and realistic assessment of competency than has previously been possible" (ACCA, 2010, p. 3).

Given the cutting-edge nature of the technologies involved, the project's main objective was to implement and evaluate a small-scale instance of a newly developed and innovative e-assessment platform rather than to develop a new one or to carry out large-scale deployment. It built on the work of two other projects involved with developing teaching, learning and assessment approaches in higher education (HE) settings other than education. First is the 'Ensemble' project based in the UK, which explored the educational role of semantic web technologies in general (Carmichael, 2008). A summary review of this project's empirical work in diverse educational settings is set out in Martinez-Garcia, Morris, Tscholl, Tracy, and Carmichael (2012), which also sets out a research and development agenda for work on semantic and linked data in higher education, including semantic archives, rapid prototyping environments and support for multiple ontologies in pedagogical settings. The second project that contributed to the research described in this paper is the 'Ontology eLearning' project based at the University of Murcia in Spain, which was specifically concerned with online assessment approaches drawing on specific semantic web technologies. "OeLe," will be described in more detail in Section 2. Further background to the project rationale and more detail on the algorithms the platform uses to mark (i.e., assess) student work and generate feedback are described in Sánchez-Vera, Fernández-Breis, Castellanos-Nieves, Frutos-Morales, and Prendes-Espinosa (2012).

The e-assessment approaches enabled by OeLe were applied to the existing short-answer questions of an undergraduate exam in financial accounting where, as yet, e-assessment of any kind was little used and only limited formative feedback was provided to students in the form of written teacher comments addressed to the whole group. Our new work aimed to understand, but not necessarily to faithfully replicate, the marking practices that we observed being employed by examiners on student scripts, and to explore the extent to which machine marking (and the results it produces) compares with markings produced by existing 'manual' practices. This was the first time that the OeLe system had been implemented beyond its original development environment and in a context other than the distance learning course in education for which it was developed, and also the first time it had been used with examinations in English (necessitating translations of both the documentation and interfaces from the original Spanish). Consequently, this new project is to be viewed as exploratory in nature and as it sought primarily to establish the usefulness and limitations of the system in these new contexts. In describing the parameters of an ontology-based system's usefulness in accounting as they appeared in this study, this paper seeks to contribute to the future research agenda in this area.

The remainder of this article will describe the characteristics of ontology-based e-assessment in relation to other developments in e-assessment, before describing how trials of the OeLe system were conducted, presenting results of these and discussing implications of this work for wider applications of ontology-based systems and for assessment practice in accounting more widely.

2. Theoretical background

Multiple areas of theory and practice in HE, as well as specific developments in the area of semantic technologies, inform the research and development described in this paper. While work on formative assessment and 'assessment for learning' has informed the development and application of a range of e-assessment technologies, it is with the emergence of semantic web technologies, ontologies, and ‘linked data’ approaches that more significant opportunities for e-assessment have begun to be addressed.

2.1. E-assessment and formative assessment

For the most part, e-assessment remains dominated by automated objective (i.e., multiple-choice) testing and technology-supported assessment taking place in online environments (e.g., in virtual learning environments). There are exceptions such as in mathematics education, where well-understood misconceptions and the appearance of common errors have provided the basis for more sophisticated intelligent and adaptive assessment systems. Reviews by Bescherer, Kortenkamp, Müller, and Spannagel (2009), Bescherer, Herding, Kortenkamp, Müller, and Zimmermann (2012) of e-assessment systems in mathematics distinguish between systems that are “automated,” offering students generic responses, “intelligent” in that they identify common misconceptions or patterns of errors, and those that are “adaptive,” offering students tailored content and activities according to patterns in their responses to questions and problems. Semantic web technologies could be integrated into all of these categories of e-assessment software, but offer particular advantages in those that are considered “intelligent” or “adaptive.” Despite these developments, partially or fully automated e-assessment of free-text written answers, and support for other, novel kinds of assessment remains limited (Jordan & Mitchell, 2009). It is exactly this under-investigated assessment effort that the OeLe project sought to explore.

At the same time, formative assessment has become a central aspect of educational practice in schools and in post-compulsory, vocational and professional learning, with wide-ranging claims being made for its role in improving student engagement and achievement (Black & Wiliam, 1998; Sadler, 1989). Most formative assessment initiatives involve the development of practice in four areas: (i) sharing and discussing learning objectives; (ii) open and generative questioning strategies; (iii) supporting peer and self-assessment; and (iv) the provision of timely and appropriate feedback offered in such a way that learners can see how to apply this feedback to their own circumstances and learning trajectories (Black & Wiliam, 1998; James et al., 2007; Nicol & MacFarlane-Dick, 2006).

In HE in particular, it is the last of these—formative feedback—that has been a particular concern, not only because of the impact that effective feedback can have on learning across subjects, settings and institutional contexts (Hattie & Timperley, 2007), but also because enduring concerns about the quality, volume, and timeliness of feedback are reflected in comparatively lower levels of student satisfaction—so much so that international “league tables” of university performance now include “satisfaction with feedback” as one of the criteria used in calculating institutional rankings. This concern is heightened as institutions increasingly employ online and blended learning approaches. Much research on formative practice is premised on its employment in face-to-face classrooms in schools and universities (Nicol, 2009), and the question of how best to provide rich, useful feedback (on which students can act) through online environments remains unresolved (Nicol & Milligan, 2006). In the UK, initiatives such as the *Reengineering Assessment Practice* (REAP) project (see JISC, 2007; Nicol, 2009) have made progress in encouraging teachers in HE to adopt online environments in which they can offer written or audio feedback to individuals or groups of students; use conferencing tools to provide tutorials; or restructure courses to allow for more self-paced learning. Rather less headway has been made in actually transforming the nature of assessment activities and the feedback that students receive using technology.

These general patterns are reflected in accounting education in particular. Broad commitments to the adoption of formative assessment practice that include more effective feedback are evident, as is a concern to broaden the scope of e-assessment in accounting education. Marriott and Lau (2008) discuss the opportunities to use summative e-assessment more “formatively” and Lau and Blackey (2011) provide a useful overview of current practice and the opportunities offered by online environments to enable and support formative practice. Other promising work has been carried out in enhancing accounting self-study materials using artificial intelligence approaches (Johnson, Phillips, & Chase, 2009). However, “semantic web” technologies and their potential applications in HE have brought with them new possibilities of hybrid systems, in which teacher–student interactions are supplemented by intelligent assessment environments with a focus on explicating, assessing, and supporting understanding, and which use concept maps and ontologies as the basis of online assessment systems (Wang & Tsu, 2006).

2.2. Semantic web technologies in education

The "Semantic Web" has been defined variously as an extension, a reworking, or a next generation of the existing World Wide Web, and, after over a decade of development, a consensus view seems to have been reached that sees the semantic web as being about:

> ... common formats for integration and combination of data drawn from diverse sources, where on the original Web mainly concentrated on the interchange of documents. It is also about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing (World Wide Web Consortium, 2011, p. XXX).

The application of semantic web technologies and approaches, it has been argued, has considerable potential to contribute to the administration of routine educational tasks such as scheduling, marking and managing learning resources (Anderson & Whitelock, 2004). Koper (2004) suggests that the main role of semantic web technologies is to enable teachers and others to "perform tasks more effectively and efficiently in large, distributed, problem-based, multi-actor, multi-resource learning spaces" (Koper, 2004, p. 5). It is important to distinguish between this broad vision of the "Semantic Web" as envisaged by Berners-Lee, Hendler, and Lassila (2001), which proposed a new iteration of the World Wide Web characterized by seamless integration and personalization; and specific "semantic web technologies", which enable the enhancement of existing web technologies, educational platforms or, in this case, assessment systems. Semantic web technologies include metadata standards, data conversion utilities, visualization tools and, most importantly in the context of this article, *ontologies*. These structured representations of domain knowledge underpin description of objects and concepts, data exchange and linkage, and while they are an essential element of the machine reasoning across the linked databases of the semantic web described by Berners-Lee et al. (2001), they are also useful in "standalone" applications (see Carmichael & Jordan, 2012, for a more extensive discussion of these issues).

Ontology-based e-assessment has at its heart the idea that, in capturing the conceptual map of a particular domain, an ontology may be used to help structure and implement assessment activities in which students are presented with questions that demand that they exhibit higher-order thinking skills and argumentation. Assessment based on an ontology is therefore a promising approach for the assessment of students who are beginning to engage with the conceptual rather than procedural aspects of accounting, but who are not yet ready to undertake sustained and complex case studies. The aim at this stage of their learning is to assess students' work on the basis of their understanding and application of concepts, rather than on their performing calculations accurately or simply reproducing verbatim answers: an important stepping stone on the road to the kind of analytical and evaluative competences that are required for professional practice, and, for that matter, required to engage with professional (P) rather than foundational (F) level assessment activities. This conceptual basis of assessment also offers the possibility of offering formative feedback, couched in terms of conceptual understanding rather than being limited to how students approached particular questions in the context of a particular examination.

There are ontologies of accounting terms and concepts, but they are oriented towards the consistent and unambiguous description of standards (Gerber & Gerber, 2011); the design of accounting software systems (see for example Lupasc, Lupasc, & Negoescu, 2010); or the interchange of data between different software systems (e.g., Spohr, Cimiano, & Hollink, 2011). While engaging students directly with these kinds of ontologies may form the basis of some learning activities Allert, Markannen, & Richter, 2006, teachers often base their educational activities and resources around more situated ontologies, which are oriented towards more immediate student learning outcomes.

While ontologies such as those described above map the conceptual structure of a particular knowledge domain and therefore are intelligible to professionals and experts, not all of this knowledge is necessarily relevant to students who are still coming to terms with relatively small areas of it. Situating relevant parts of a more comprehensive ontology within a pedagogical context may therefore involve excluding some concepts or expressing relationships between them in a way that may not be wholly correct in a professional context, whereas in a pedagogical context completion and correctness may cause confusion and hinder learning. As teachers have different and often highly individual pedagogical practices, our work recognized that, while any e-assessment system for accounting education would need to be rooted in the International Accounting Standards (the assessment activities representing, in the words of Gerber and Gerber (2011, p. 15), a particular "interpretation" of these), it would additionally need to be capable of representing and responding to the pedagogical practices of teachers and examiners, and it is to this combination that the OeLe E-Assessment Platform seeks to respond.

2.3. The OeLe e-assessment platform

The OeLe (Ontology eLEarning) E-Assessment platform developed by a team at the University of Murcia (Castellanos-Nieves, Tomás Fernández-Breis, Valencia-García, Martínez-Béjar, & Iniesta-Moreno, 2011; Frutos-Morales et al., 2010; Sánchez-Vera et al., 2012) is described as: "... us[ing] ontologies, semantic annotations, natural language processing techniques and semantic similarity functions in order to support assessment processes, in particular, providing marks to free text answers to open questions" (Sánchez-Vera et al., 2012, p. 154). OeLe builds on a legacy of automated assessment that predates the semantic web and this is reflected in an architecture in which users access a database through a 'client' programme. This has progressively been enhanced by the introduction of semantic web technologies and approaches and then by the development of web interfaces, first for student and then for teacher and administrator functions. As is the case in many applications, the 'semantic web' label belies prior work in related fields such as artificial intelligence, natural language processing, and data visualization, and it is more accurate to describe OeLe as "including [Semantic Web technologies] in the E-learning teaching–learning process" (Castellanos-Nieves, Tomás Fernández-Breis, Valencia-García, & Martínez-Béjar, 2007, p. 451).

At the core of any implementation of OeLe is a model of the domain knowledge to be assessed, which is expressed as an ontology using Web Ontology Language (OWL).[2] At present, the ontology is created externally using a free, open-source tool (Protégé, www.protege.stanford.edu). While Protégé is a robust and well-supported knowledge-representation tool, the initial process of creating the ontology can be lengthy and can require several drafts as users attempt to represent explicitly a concept map in which concepts and relationships are expressed in a machine-readable form.[3] This process can be daunting to non-computer scientists and was identified in our trial as one aspect of the process that needed improvement. However, as the concepts in the ontology depict the relationships between concepts, rather than being tied to specific answers, the ontologies can be re-purposed and re-used by educators in subsequent exams. Ontology construction should, ideally, be a one-off task. As the concepts in the ontology carry no inherent importance until they are associated with a model answer, an ontology that, for example, asserts that concept c2 is a part of concept c1, may be applied to exams that require students to discuss either one of those ideas or the relationship between them, regardless of the specific nature of the questions relating to those ideas.

This pedagogically oriented ontology is associated with, and may be designed alongside, an examination, comprising a set of questions and, most critically, the model answers that accompany them, which teachers need to develop. Whereas in manual marking procedures this kind of mapping of the assessment domain would be an optional activity, when using OeLe it is necessary step, thereby encouraging more detailed planning of the assessment and marking activity. Unlike the ontologies, the question and model answers are closely related and new model answers do have to be created for each new question. However, the OeLe system does separate questions from exams, allowing for the possibility that teachers may create exams from a bank of pre-existing questions (providing that

[2] The abbreviation to "OWL" rather than "WOL" is a deliberate inconsistency, retrospectively justified on the grounds that the wise Owl character in A.A. Milne's 'Winnie the Pooh' books spells his name 'WOL' (http://www.w3.org/2003/08/owlfaq).

[3] Recent accounting education papers that focus on concept mapping include Simon (2010) and Leauby, Szabat, and Maas (2010).

the knowledge required in each question is represented in the same ontology). Each question is assigned a number of marks, as in any examination, but in addition, the relative values of the different concepts that appear in the model answers are also assigned values. This allows teachers to assert that, in scoring up to n marks for a particular question, it is more important (for example) that students apply concept c1 than concept c2. As students are highly unlikely to express their answers solely in terms of the concepts that appear in the ontology, a range of linguistic expressions may be defined that map to the concepts in the ontology.[4] The initial source of these is the model answer, but OeLe can also be trained. As students' answers are assessed and annotated markers can highlight additional acceptable linguistic expressions and associate them with concepts in the ontology so that subsequent student answers can be assigned marks even though their responses may not exactly match the model answers. In the work described in this paper, OeLe employed exactly the same model answers as those originally employed by the tutor, with exactly the same marking and credit scheme. However, in our trial this did not necessarily yield the same results, even for top-scoring answers, and we explore the implications of this in Section 5, below.

Students submit their answers through a web interface; examinations can include both closed (multiple choice) and open (text response) answers and can be opened for a set time period during which students may either make a single attempt to answer the questions, or return to revise their answers at any time during the specified examination period. This latter option offers the possibility of students being presented with open-book style questions on which they work over a period of time until they are satisfied with their answers, or in more reflective assessments.

Once an examination is closed, the marking process takes place in a two-phase process, though in fact the second of these is optional. The first phase involves annotation of student answers by a marker, using the subject ontology as a marking scheme. Markers do not, however, have to calculate actual marks, but, rather, highlight elements of the student answers in an online editing environment and select the concepts of which they demonstrate understanding. On completion of this annotation process, a mark is assigned based on the weightings attached to the concepts and the maximum mark available for the question.

It would be entirely possible that a marker might use the OeLe platform solely in this role; but, this process of annotation and ontological mapping of student answers also offers the potential of using the now trained system not only to mark annotated scripts, but to automatically annotate and mark subsequent student answers on the basis of the annotations. This, of course, raises important questions about the extent and outcomes of this training: how many answers need to be marked for automatic annotation to be as accurate and reliable as a human marker? And, are certain kinds of questions easier to reliably annotate automatically: some might have only a limited range of acceptable answers with a clear structure, but what of those that ask students to make judgments, construct arguments, or express opinions supported by evidence?

As Sánchez-Vera et al. (2012) explain, initially, the OeLe system was conceived simply as a means of carrying out annotation and marking in this way; subsequent developments added an additional set of features in which students received not only marks, but also feedback derived from the same ontology that underpins the annotation and marking processes. The platform as whole can then be envisaged as in Fig. 1.

From a student perspective, this means that they then receive feedback in which they are presented with their mark, the model answer to compare with their own, and a summary of the concepts for which they received credit and a list of concepts which, had they drawn on them, would have resulted in a higher mark. This feature of the system has the potential then to be linked to suggestions of useful resources, revision activities or course content that might be revisited in order to develop their understanding. Student feedback is currently offered through a web interface shown in Fig. 2.

[4] So for example a "preferred term" such as "durable" (in relation to a particular process) might have acceptable alternatives such as 'established' and more colloquial phrases such as "tried and tested" and "has stood the test of time" – the latter being the actual expression used by the teacher. In some cases these alternatives involve no more than different word order, but other differences may be more substantial. As the system is trained, the processes of annotation may supplement the original list of acceptable alternatives with others.

This insight into conceptual understanding is clearly valuable for teachers. Not only can individual students' areas of conceptual understanding be gauged, the OeLe system also presents teachers with reports that highlight areas of common understanding and lack of it, and those concepts on which levels of student understanding are highly differentiated. Even if individual student feedback is not offered, teachers' general feedback to a student cohort can be couched in terms of understanding and application of concepts rather than success in answering specific questions. The system also provides useful feedback to teachers themselves: not only about their success in conveying the conceptual basis of their course content, but also how well the assessment exercises they have set are indeed testing conceptual understanding. One view of the teacher interface is shown in Fig. 3.

While the ultimate purpose of our implementation of OeLe is to provide valid, reliable assessment of student understanding of key concepts, the ontology that was implemented is, as we suggested in Section 2, one that reflects both the stage of education students are at and the boundaries of what they might know, understand and express in the context of a particular test. During training, the original ontology is elaborated with additional local data that not only captures domain knowledge but information about students' learning and behavior in the context of assessment activities.

3. Methods: Implementing OeLe in undergraduate financial accounting

The module selected for trial deployment of the system was a second-year undergraduate course in financial accounting, one of the first in which students encounter the conceptual basis of accounting. Discussions with teaching staff suggested that in the past students taking this course were particularly challenged by the element of the course that required them not only to calculate accurately, but to engage with the concepts that underpin those calculations, and to begin to make choices about the definition, classification, and treatment of the figures based on those concepts.

Marked examination papers from a cohort of 103 students formed the basis of our study. The examination comprised one section in which students were asked to carry out calculations and another made up of six short-answer written questions. Initial analysis of the marks awarded indicated the extent to which many students struggled with the latter section: while 37 of the 103 students (36%) achieved a high passing grade (70% or more on the paper as a whole), only 11 scored the same 70%+ mark on the six written short answers. There was, therefore, a concern about moving to a full

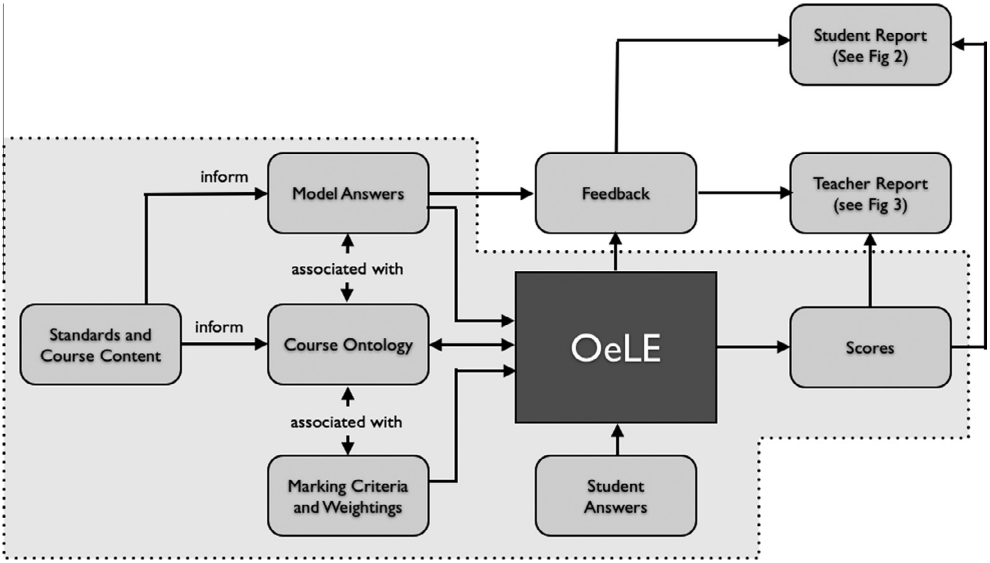
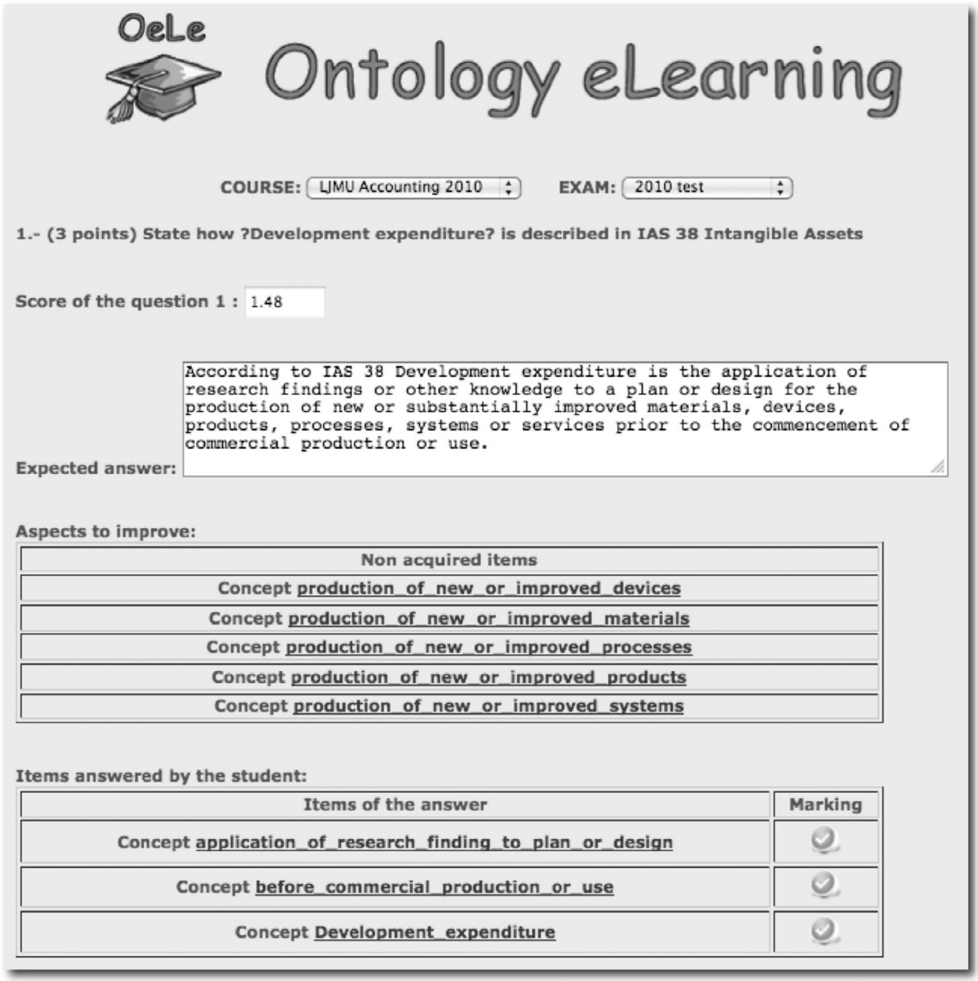


Fig. 1. Overview of the OeLe system showing relationships between different elements and interfaces.



COURSE: LJMU Accounting 2010 EXAM: 2010 test

1.- (3 points) State how ?Development expenditure? is described in IAS 38 Intangible Assets

Score of the question 1 : 1.48

Expected answer: According to IAS 38 Development expenditure is the application of research findings or other knowledge to a plan or design for the production of new or substantially improved materials, devices, products, processes, systems or services prior to the commencement of commercial production or use.

Aspects to improve:

Non acquired items
Concept production_of_new_or_improved_devices
Concept production_of_new_or_improved_materials
Concept production_of_new_or_improved_processes
Concept production_of_new_or_improved_products
Concept production_of_new_or_improved_systems

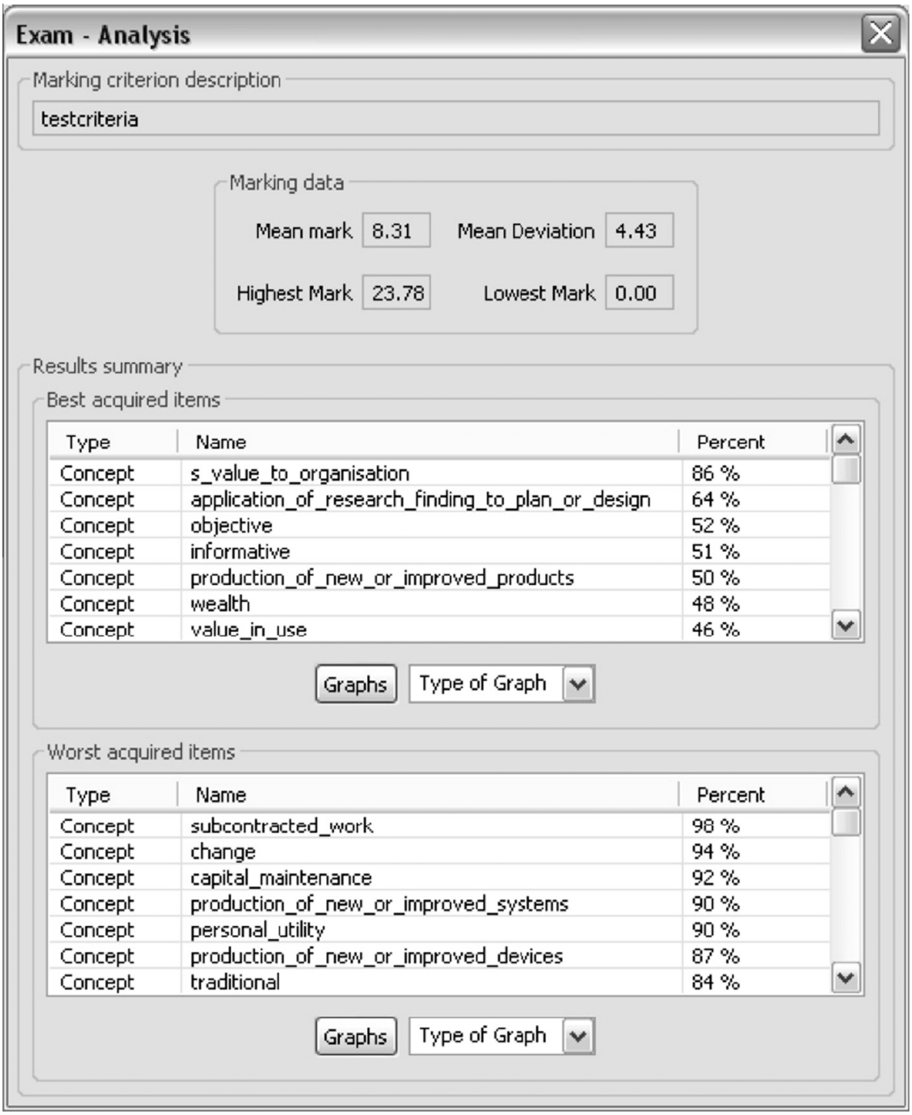
Items answered by the student:

Items of the answer	Marking
Concept application_of_research_finding_to_plan_or_design	✓
Concept before_commercial_production_or_use	✓
Concept Development_expenditure	✓

Fig. 2. OeLe student feedback interface.

implementation of any e-assessment system without first ascertaining whether it could reliably deal with a wide range of student answers—ranging from those who wrote extensive answers and achieved up to 21 marks out of the 23 available on the six questions to those who attempted few questions and scored very low marks. Students scored across the entire range from 0 – full marks on each of the six questions: a key issue for teachers was whether the OeLe system would be able to adequately to discriminate across the range of student answers.

While student responses to exam questions therefore represented authentic responses to a real exam situation, our work focused on a technical assessment of the system. We did not attempt an evaluation of users' experience taking the online test or of receiving the type of feedback that the online test might provide. Two linked trials were carried out: the first of these was based on a set of 30 papers which represented a representative sample but one that excluded those where students had written very little, as these would have given little basis for OeLe training. This sample was designed to ascertain how best to configure the system for accounting, implement the ontology and test the automatic marking of manually annotated scripts. This involved comparing the manual marks ("pencil-and-paper" style) with those achieved by a marker reading and annotating each



Marking criterion description: testcriteria

Marking data

Mean mark: 8.31
Mean Deviation: 4.43
Highest Mark: 23.78
Lowest Mark: 0.00

Results summary

Best acquired items

Type	Name	Percent
Concept	s_value_to_organisation	86 %
Concept	application_of_research_finding_to_plan_or_design	64 %
Concept	objective	52 %
Concept	informative	51 %
Concept	production_of_new_or_improved_products	50 %
Concept	wealth	48 %
Concept	value_in_use	46 %

Worst acquired items

Type	Name	Percent
Concept	subcontracted_work	98 %
Concept	change	94 %
Concept	capital_maintenance	92 %
Concept	production_of_new_or_improved_systems	90 %
Concept	personal_utility	90 %
Concept	production_of_new_or_improved_devices	87 %
Concept	traditional	84 %

Fig. 3. OeLe teacher interface showing conceptual knowledge demonstrated across a student cohort.

of the scripts using the terms in the ontology, with OeLe then calculating the marks to be awarded. The second trial used the annotations made by the marker on the first set of 30 papers to provide a range of additional linguistic expressions as training sets against which the entire cohort's papers could be marked. In order to establish how many papers needed to be treated in this way before most (if not all) variations in student answers were exhausted, this phase involved three separate runs with training sets of 10, 20, and 30 papers. This second trial thus assessed the potential of the fully-automated system to annotate and assign marks based on different training sets; its ability to deal with different types of questions and responses; and its accuracy and predictability in comparison to a human marker.

Table 1
Comparison of manually marked vs manually annotated and automatically marked scores across Q1–6 on 30 'sample' papers.

1 Question (max mark)	2 Manual marker Mean (SD)	3 Manual annotation, automarked Mean (SD)
1 (3)	1.97 (1.10)	1.38 (1.04)
2 (2)	1.23 (0.97)	1.06 (0.87)
3 (3)	0.73 (0.98)	0.80 (0.75)
4 (2)	0.57 (0.73)	0.60 (0.67)
5 (5)	2.20 (1.35)	1.51 (0.96)
6 (8)	4.23 (2.51)	4.53 (2.52)
Totals	10.93 (4.98)	9.88 (4.09)

4. Results

4.1. From hand marking to auto-marking

The first trial focused on the impact of using OeLe for manual annotation of student scripts. In this scenario, the human marker reads the student answer on-screen, identifies the ideas present, and associates these with the relevant parts of the ontology. The answer needs to be precise enough for a specific part of it to be recognizable as the expression of a specific concept, but it does not need to be couched in exactly the same terms as the marker can recognize acceptable synonyms. The results of the first trial, which compared manual marking with manual annotation plus automatic marking, where the human judges the student's level of conceptual understanding but the system calculates the marks, are summarized in Table 1. The first column shows the question number and, in brackets, the maximum number of marks available; the second, the marks originally given by the human marker using the conventional paper-and-pencil method, with the standard deviation in brackets; and the third, the scores obtained by manually annotating the scripts with ontology terms and then allowing the system to assign marks automatically on the basis of these annotations.

Using the ontology provided by OeLe to guide annotation (column 3) led to a more focused marking process than the wholly manual marking (column 2), as it compelled the marker to highlight text and then assert relationships with concepts from the ontology, causing them to justify the award of marks rather than placing an indicative tick on the script. Manual annotation (column 3) led to lower scores for many students and this is reflected in lower mean scores for several questions.[5] We discuss the significance of these for both teachers and students in Section 5. The more explicit marking process of the marker highlighting text and then asserting relationships with concepts from the ontology also meant that when marks were calculated, a range of marks was achieved, rather than manual marks of integer values. If calculated scores were rounded to the nearest integer value, the results varied little from the manual marks—although there were exceptions, as tendencies for the marker to overly "round up" scores were not replicated by the automatic system (for example, students who had written partially correct answers but been generously awarded 2 out of a possible 3 might, on the basis of the annotations, be awarded a mark of 1.4, for example, which would round down to 1). The importance and influence of the model answers was also evident: in Question 2, 17 of the 30 students achieved the maximum 2 marks. However, the model answer included a small detail that only 1 student included in his/her response (thus achieving the maximum 2 marks when auto-marked) while the other 16 students scored 1.78 when auto-marked. Again, rounding would have led to their resulting reported mark being the same, but this highlighted the fact that, in questions where very specific responses were required, the ontology-based annotation had the potential to discriminate in detail between answers.

Furthermore, instances of inconsistency in manual marking of near-identical answers were more consistently marked when manual annotations were scored (column 3). In the manual marks (column

[5] At this exploratory stage in our work the small sample size precluded meaningful assessment of statistical significance of these differences.

2), there were examples where different marks were awarded for answers that were not only near-identical in their conceptual basis but also in their linguistic composition. With a human marker manually annotating but allowing the OeLe system to calculate the resulting marks, students who wrote virtually identical answers but received different manual marks were all awarded a consistent score.[6] However, there are potential disadvantages to OeLe's less subjective approach too, and we discuss these in relation to its treatment of 'partial' answers below.

These results suggest that, by being asked to explicitly indicate for which part of their answer students are being awarded marks, markers are compelled to focus on what the student answer actually means, rather than being swayed by style or expression. Markers are at once discouraged from giving marks to concepts that are vaguely expressed, or merely implied; and at the same time, encouraged to recognize and reward detail where it is present. It is therefore up to teachers and examiners to create a sufficiently detailed and accurate conceptual structure at the beginning of the process, one that reflects the various components that may be present, and independently credited, in student answers; we discuss the challenges of this in Section 5.

4.2. The results of training: OeLe in action

For OeLe to be able to annotate the student responses as well as to calculate the marks, the system first needs to be trained in recognizing the acceptable alternatives expressions to the terms it already has in the ontology. Training consists of annotating some of the students' answers manually: this enables the system to append them to the concepts already in the ontology. Again, this distinguishes OeLe's situated ontology from a formal, expert ontology, which might contain exact synonyms. In this case, however, the main purpose of the exercise is to build a database of ways in which the concepts may be expressed by students, which are good enough synonyms in the context of this particular test. Our second trial used the annotations made by the marker on the first set of 30 papers to provide a range of additional linguistic expressions as training sets against which the entire cohort's papers could be marked. The whole set of papers was then both automatically annotated and then automatically marked by the system; the results of this trial are set out in Table 2. The first column again shows the question number and maximum mark, with following columns giving the mean marks and standard deviation for the total cohort of exams ($n = 103$). The table compares the marks given by the manual marker (column 2), to those given by OeLe trained by the marker on the model answer only (column 3), and subsequently and on 10, 20 and 30 papers from the sample set used in the first trial (columns 4–6).

The scores determined through auto-annotation against the ontology and the model answer (column 3) are, again, much lower than those awarded by the human marker in the original examination (column 2). However, as columns 4–6 show, the patterns of scores became progressively closer to those of the original human marker as the number of training scripts increases, with 20 scripts apparently enough to ensure a good agreement both in terms of the spread of marks shown here and a generally good correlation of original examination scores and full automatic annotation and marking.[7]

When these outcomes are translated into the terms in which teachers and students couched many of their questions about the role of e-assessment, the following observations can be made:

- Of the 103 students, 42 students would have gained marks (after rounding) had a trained OeLe system been used; 41 would have lost marks and the remainder would have emerged with the same mark as the original marker awarded.

[6] For example, two answers to Q1 read: "Development expenditure is defined as the application of the plan or design undertaken in order to achieve a new or improve a current product or material" (awarded 1 mark), and "Development expenditure is the money used to apply researching knowledge to the plan or design of a new or substantially improved product" (awarded 2 marks). These answers received marks of 1.3 and 0.9 respectively (both of which rounded to 1) when auto-annotated and auto-marked.

[7] A Pearson R correlation of 0.84 was achieved across all students' total scores (significant at $p = 0.01$). Correlations of scores on specific questions varied from 0.86 (Question 1) and 0.83 (Question 6) to a low of 0.38 (Question 4), which can be attributed to generally low scores on this question.

Table 2
Fully automated annotation and marking (*n* = 103) with OeLe trained on the model answer only, and then on 10, 20 and 30 scripts.

1 Q. (max)	2 Manual marker Mean (SD)	3 Trained on model answer Mean (SD)	4 Trained on model answer plus 10 Mean (SD)	5 Trained on model answer plus 20 Mean (SD)	6 Trained on model answer plus 30 Mean (SD)
1 (3)	1.22 (1.22)	0.45 (0.62)	0.78 (0.91)	1.12 (1.06)	1.13 (1.06)
2 (2)	0.80 (0.96)	1.04 (0.79)	1.04 (0.79)	1.04 (0.79)	1.04 (0.79)
3 (3)	0.72 (0.90)	0.91 (0.91)	0.95 (0.89)	0.97 (0.90)	0.99 (0.90)
4 (2)	0.49 (0.62)	0.31 (0.50)	0.35 (0.52)	0.39 (0.56)	0.44 (0.62)
5 (5)	1.79 (1.42)	0.97 (1.10)	1.00 (1.10)	1.09 (1.12)	1.10 (1.12)
6 (8)	2.96 (2.50)	3.07 (2.81)	3.48 (2.92)	3.56 (2.96)	3.62 (2.99)
Totals	7.97 (5.39)	6.75 (4.66)	7.60 (4.98)	8.17 (5.27)	8.31 (5.31)

- If we assume a bare passing score is around 40% and that which is considered a high pass, distinction or "first class" is 70%, then 10 students who would have failed this part of the examination would have been awarded passing when graded by OeLe; while 6 would have dropped below the 40% threshold.
- At the upper "first class" grade boundary, five students who did not achieve this would have been awarded 70% or more by OeLe, while six students would have dropped below the 70% threshold as a result of the automatic marking.

5. Discussion

While the overall pattern is one of improved consistency and agreement with the manual marker when using a trained system, within this trend there are epistemological issues worthy of further exploration, which ultimately raise broader questions about the validity of an approach that primarily aims to replicate human markers' strategies. As we have indicated, disparities between OeLe and the human marker were most evident where students expressed partial understanding of key concepts. Where students can state key ideas clearly, OeLe rewards their answers, even if their reasoning is incomplete or poorly expressed. In contrast, the reverse is true of the human marker, who tends to reward students who can express the gist of a correct response, but in very general and imprecise terms – and, if they are both teacher and marker, where students have intentionally or unintentionally reproduced the teacher's own words or arguments.

These divergent interpretations of understanding are difficult to reconcile. While OeLe operates on the basis that using the correct terminology in the specified context implies understanding, the human marker's approach is more subtle. But, because the human marker's approach is potentially also more inconsistent and subject to even more subjectivity when a number of different markers are used, as the tacit process of judgment of a student's answer is difficult to render explicit, particularly where the criteria are qualitative or where the subject itself carries some inherent uncertainty (Brooks, 2012). In these circumstances, while marker training and the use of rigorous mark schemes can improve consistency, with some formats such as essays, "there has to be an acceptance that the marks or grades that candidates receive will not be perfectly reliable" (Meadows & Billington, 2005, p. 68). Once trained to recognize a range of acceptable answers, OeLe has the benefit of not being susceptible to these subjectivities: but this may, paradoxically, disadvantage those students whose understanding is still phrased in lay terms, and who might benefit from the judgments a human marker could bring to their work.

Despite these differences in interpretation, OeLe performed best with marks at the upper and lower end of the range, clearly identifying students with very low scores, and rewarding those with accurate and precise answers that were scored highly. Scores within ±20% of the pass mark were less consistent, but even so, the first point of divergence from the human marker between overall pass and fail marks was the student in 74th position. One potential application of an ontology-based assessment system, therefore, could be to identify broad 'categories' of papers, to allow markers to focus their efforts where they are most required. Indications from this trial suggest that OeLe could be used in accounting education with reasonable confidence to screen out the lowest quartile of papers, allowing markers to differentiate their treatment of the papers as appropriate, perhaps directing attention to scripts where some relevant content has already been recognized, or targeting the weaker papers for more extensive feedback. For those with large numbers of papers, or where time and budgetary constraints are important factors to consider when allocating marking, the automatic marking function's ability to identify and filter out papers significantly below the pass mark could prove to be a useful feature. The system's ability to support human markers through structured annotation plus automatic marking also has wider potential application, perhaps as a tool to assist new markers, or to ensure some consistency between markers.

Making the most of the potential affordances of an e-assessment system like OeLe implies new approaches to testing, and changes in assessment practice, including both question setting and marking. As we have shown, OeLe can be used to apply fine-grained analysis of the concepts articulated in students' answers in a way that human markers may not be able to do consistently, but a different approach to the assessment process may be needed in order to fully exploit this potential. In our findings, complete responses only got a few tenths of a point more than merely "good" ones, because the test had been designed with existing practices in mind. A transition to ontology-based e-assessment would involve further work with educators and assessors, firstly to articulate the ontologies and the model answers, but also to decide the relative importance and weighting of concepts in an answer in order to set a target score, rather than first deciding a maximum and then deciding what different levels of answers might include. This part of the process, however, proved particularly challenging in our project, highlighting a need for more work with educators and examiners to understand the processes, which can assist with these new practices.

6. Conclusions

While our work has highlighted the potential applications of an ontology-based system in accounting examinations, it has also shown that adopting this form of e-assessment would require changes in current teaching and assessment practices. These changes should, perhaps, be seen in the more general context of increasingly automated practices, both in professional activities and in education, many of which are both enabled and performed by the semantic web technologies described in Section 2.2. Education for this post-human context, where machine reading of texts is more normal, implies a shift in pedagogical practices: not specifically tailoring tasks and assessments around the capabilities of the system, but recognizing that working practices evolve as technologies develop and permeate both the profession and the education that prepares people for it. This, too, may be an area where e-assessment can usefully support the professional development of students.

However, OeLe has some way to go to be a fully developed, mature system, and there are a number of avenues to further explore how an ontology-based e-assessment system could work in practice. Our work in this regard has identified three areas for future research. The first relates to the technology and its capacity to scale up. The question of to what extent the overheads of ontology generation and training would diminish as the number of answers increases is of particular interest here.

The second issue relates to integration with other aspects of teaching and learning. This was a relatively small-scale study, conducted with existing assessment materials: as we have indicated, more extensive technical testing, with question papers designed for this type of assessment is needed in order to fully understand the ways in which OeLe can cope with different types of questions and a variety of student responses, and to investigate how ontology-based systems could be used formatively, to aid ongoing learning. An exploration of how conceptually-based e-assessment may be developed would therefore require more work with educators and examiners, to understand more fully the practices around testing conceptual knowledge in accounting assessments: not only to design or adapt papers for these purposes, but also to gauge how educators understand the relationship between conceptual models and teaching and learning materials, in order to inform development of OeLe's feedback functions. In this respect, the diversity of student answers is another issue to consider: in our trial, students had all been prepared for the exam by the same teacher, and responded in fairly predictable ways (for example, making similar mistakes), but some of this homogeneity may be lost with a bigger group consisting of students from diverse educational backgrounds; OeLe has yet to be tested in such a circumstance.

Both the potential developments described above point to a third, perhaps more challenging, issue identified in our study: a need for better understanding of assessment and marking practices. Our trials using OeLe raises broader questions about how knowledge should be articulated in assessments: whether it is enough for students at the pre-professional level to express their developing knowledge in lay terms, or whether more systematic engagement with the terminology and discourses of accounting should be a prerequisite for advancement to a more advanced level. No clear message emerged here from our examples, highlighting a need to make relationships between pedagogical and professional practice more explicit as this may determine the nature and role of e-assessment systems that are implemented. Systems like OeLe may have one use (as an automated assessment tool) in contexts where engagement with specific terminology is important, but another use (as a tool to aid marking consistency) in different circumstances where expression of key concepts in general terms is acceptable. Decisions on which systems are deployed, and for what purposes, are dependent on a more fully articulated understanding of the pedagogical, conceptual and professional requirements of the assessment, which in turn is dependent on a clearer understanding of how knowledge is constructed in accounting.

What is the trajectory and nature of knowledge transmission in the discipline; how can (and should) assessments measure this knowledge, and is the approach of semantic web technologies (i.e., "concepts as philosophical primitives") congruent with the types of knowledge that students are required to express in assessments? These are just a few of the issues raised by e-Assessment which are for accounting education professionals, rather than system developers, to consider. Further development of e-assessment will therefore demand a continued dialogue among educationalists, computer scientists, accounting educators, and professional bodies to develop both technological systems and the pedagogical practices that accompany them.

Acknowledgements

This research was funded by ACCA (Association of Chartered Certified Accountants) and the International Association for Accounting Education and Research (IAAER) under a programme of research to support the work of IFAC's International Accounting Education Standards Board (IAESB). The authors would like to acknowledge the advice and guidance they have received from the funders in the course of this work. In addition, we would like to thank our research partners at the University of Murcia, especially Dr. Jesualdo Tomás Fernández-Breis and Dr. Maria del Mar Sanchez Vera; Rob Crichton, who worked as a project research assistant at Liverpool John Moores University; and the staff and students of the School of Business and Law of Liverpool John Moores University. We are also grateful for the opportunity to contribute this exploratory, empirical work and for the formative feedback provided by Andy Rosman (LIU Post) and David Lavoie (University of Connecticut) on an early draft of this article.

References

ACCA (2010). *Call for research proposals: International accounting education standards (RES-CALL-IAES2)*. London: Association of Chartered Certified Accountants.

Allert, H., Markannen, H., & Richter, C. (2006). Rethinking the use of ontologies. In *Proceedings of the 2nd international workshop on learner-oriented knowledge management and KM-oriented learning LOKMOL 06, in conjunction with the first European conference on technology-enhanced learning ECTEL 06* (October 2, 2006) (pp. 115–125).

Anderson, T., & Whitelock, D. (2004). The educational semantic Web: Visioning and practicing the future of education. *Journal of Interactive Media in Education, 24*(1). <http://www-jime.open.ac.uk/2004/1>.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic Web. *Scientific American, 284*(5), 34–43.

Bescherer, C., Herding, D., Kortenkamp, U., Müller, W., & Zimmermann, M. (2012). E-learning tools with intelligent assessment and feedback for mathematics study. In S. Graf, F. Lin, A. Kinshuk, & R. McGreal (Eds.), *Intelligent and adaptive learning systems: Technology enhanced support for learners and teachers* (pp. 151–163). Hershey, PA: Information Science Reference.

Bescherer, C., Kortenkamp, U., Müller, W., & Spannagel, C. (2009). Intelligent computer-aided assessment in mathematics classrooms. In A. McDougall, J. Murnane, A. Jones, & N. Reynolds (Eds.), *Researching IT in education: Theory, practice and future directions* (pp. 200–205). NY: Routledge.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–71.

Brooks, V. (2012). Marking as judgment. *Research Papers in Education, 27*(1), 63–80.

Carmichael, P. (2008). The semantic Web and 'Web 3.0'. In Selwyn, N. (Ed.), *Education 2.0? Designing the web for teaching and learning*. London: ESRC Teaching and Learning Research Programme.

Carmichael, P., & Jordan, K. (2012). Semantic Web technologies for education – Time for a 'turn to practice'? *Technology, Pedagogy and Education, 21*(2), 153–169.

Castellanos-Nieves, D., Tomás Fernández-Breis, J., Valencia-García, R., & Martínez-Béjar, R. (2007). A semantic Web technologies-based system for student assessmentin e-learning environment. *Proceedings of IADIS International Conference e-Learning* (July 6–8), 451–458.

Castellanos-Nieves, D., Tomás Fernández-Breis, J., Valencia-García, R., Martínez-Béjar, R., & Iniesta-Moreno, M. (2011). Semantic Web technologies for supporting learning assessment. *Information Sciences, 181*(9), 1517–1537.

Frutos-Morales, F., Sánchez-Vera, M. M., Castellanos-Nieves, D., Esteban-Gil, D., Cruz-Corona, C., Prendes-Espinosa, M. P., et al (2010). An extension of the OeLe platform for generating semantic feedback for students and teachers. *Procedia – Social and Behavioral Sciences, 2*(2), 527–531.

Gerber, M., & Gerber, A. (2011). Towards the development of consistent and unambiguous financial accounting standards using ontology technologies. *Paper presented at IAEER conference 2011: 'Accounting Renaissance: Lessons from the Crisis and Looking into the Future', Venice, Italy*, 4–5 November 2011. <http://researchspace.csir.co.za/dspace/handle/10204/5797>.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

James, M., Black, P., Carmichael, P., Drummond, M.-J., Fox, A., & Honour, L. (2007). *Learning how to learn: In classrooms, schools and networks*. London: Routledge, TLRP Improving Learning Series.

JISC (2007). *Transformation story: Re-engineering assessment practices in Scottish higher education* (JISC/REAP).

Johnson, B. G., Phillips, F., & Chase, L. G. (2009). An intelligent tutoring system for the accounting cycle: Enhancing textbook homework with artificial intelligence. *Journal of Accounting Education, 27*(1), 30–39.

Jordan, S., & Mitchell, T. (2009). E-assessment for learning? The potential of short-answer free-text questions with tailored feedback. *British Journal of Educational Technology, 40*(2), 371–385.

Koper, R. (2004). Use of the semantic Web to solve some basic problems in education: Increase flexible, distributed lifelong learning, decrease teacher's workload. *Journal of Interactive Media in Education, 6*. <http://www-jime.open.ac.uk/2004/6>.

Lau, A., & Blackey, H. (2011). *Hybrid learning meets assessment for learning: Facing the misconceptions*. *Proceedings of 4th international conference on hybrid learning (ICHL'11)*. Berlin: Springer, pp. 105–115. <http://dl.acm.org/citation.cfm?id=2033007>.

Leauby, B. A., Szabat, K. A., & Maas, J. D. (2010). Concept mapping—An empirical study in introductory financial accounting. *Accounting Education: An International Journal, 19*(3), 279–300.

Lupasc, A., Lupasc, I., & Negoescu, G. (2010). The role of ontologies for designing accounting information systems. *Annals of "Dunarea de Jos" University of Galati: Economics and Applied Informatics, 16*(1), 101–108.

Marriott, P., & Lau, A. (2008). The use of on-line summative assessment in an undergraduate financial accounting course. *Journal of Accounting Education, 26*(2), 73–90.

Martinez-Garcia, A., Morris, S., Tscholl, M., Tracy, F., & Carmichael, P. (2012). Case based learning, pedagogical innovation and semantic web technologies. *IEEE Transactions on Learning Technologies, 5*(2), 104–116.

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability*. Report for the National Assessment Agency by AQA Centre for Education Research and Policy.

Nicol, D. (2009). Assessment for learner self-regulation: Enhancing achievement in the first year using learning technologies. *Assessment and Evaluation in Higher Education, 34*(3), 335–352.

Nicol, D., & MacFarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218.

Nicol, D. J., & Milligan, C. (2006). Rethinking technology-supported assessment in terms of the seven principles of good feedback practice. In C. Bryan & K. Clegg (Eds.), *Innovative Assessment in Higher Education*. London: Routledge.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.

Sánchez-Vera, M., Fernández-Breis, J., Castellanos-Nieves, D., Frutos-Morales, F., & Prendes-Espinosa, M. P. (2012). Semantic Web technologies for generating feedback in online assessment environments. *Knowledge-Based Systems, 33*(September), 152–165.

Simon, J. (2010). Curriculum changes using concept mapping. *Accounting Education: An International Journal, 19*(3), 301–307.

Spohr, D., Cimiano, P., & Hollink, L. (2011). Multilingual and cross-lingual ontology matching and its application to financial accounting standards. In *Proceedings of the 10th international semantic web conference (ISWC 2011)*, Bonn, Germany, October 23–27, 2011.

Wang, H. C., & Tsu, C. W. (2006). Teaching material design center: An ontology-based system for customizing reusable e-materials. *Computers and Education, 46*(4), 458–470.

World Wide Web Consortium (2011). W3C Semantic Web Activity. <http://www.w3.org/2011/sw>.