

Collaborative Authoring of Semantically Structured Multilingual Educational Content

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Darya Tarasova
aus
Novosibirsk

Bonn 2015

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Sören Auer
2. Gutachter(in): Prof. Dr. X Y

Tag der Promotion:
Erscheinungsjahr:

Acknowledgements

I would like to thank ...

You should probably use `\chapter*` for acknowledgements at the beginning of a thesis and `\chapter` for the end.

Publications

This thesis is based on the following journal and conference publications, in which I have been an author or a contributor:

Journal Publications, peer-reviewed

- Sören Auer, Ali Khalili, and Darya Tarasowa. Crowd-sourced open courseware authoring with slidewiki.org. *International Journal of Emerging Technologies in Learning (iJET)*, 8(1), 2013.
- Darya Tarasowa, Sören Auer, Ali Khalili, and Jörg Unbehauen. Crowd-sourcing (semantically) structured multilingual educational content (cosmec). *Open Praxis*, 6(2):159–170, 2014.
- Darya Tarasowa, Ali Khalili, and Sören Auer. Crowdlearn: Crowd-sourcing the creation of highly-structured e-learning content. *iJEP*, 5(4):47–54, 2015.

Conference Publications, peer-reviewed

- Ali Khalili, Sören Auer, Darya Tarasowa, and Ivan Ermilov. Slidewiki: Elicitation and sharing of corporate knowledge using presentations. In *Proceedings of the EKAW 2012*, Lecture Notes in Computer Science (LNCS) 7603. Springer, 2012.
- Darya Tarasowa and Sören Auer. Balanced scoring method for multiple-mark questions. In *5th International Conference on Computer Supported Education (CSEDU 2013)*, 2013.
- Darya Tarasowa, Ali Khalili, and Sören Auer. Crowdlearn: Collaborative engineering of (semi-)structured learning objects. In *Proceedings of the KESW 2012*, 2012.
- Darya Tarasowa, Ali Khalili, Sören Auer, and Jörg Unbehauen. Crowdlearn: Crowd-sourcing the creation of highly-structured e-learning content. In *5th International Conference on Computer Supported Education (CSEDU 2013)*, pages 33–42, 2013.
- Darya Tarasowa, Christoph Lange, and Sören Auer. Measuring the quality of relational-to-rdf mappings. In *Proceedings of the KESW 2015*, pages 210–224, 2015.

Book Chapters

- Darya Tarasowa and Sören Auer. Collaborative authoring of opencourseware: The best practices and complex solution. In *Open Data for Education*, pages 103–131. Springer International Publishing, 2016.

Abstract

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Problem Definition	5
1.3	Usage Scenario	6
1.4	Overview of our Approach	7
1.4.1	Content Reusability and Interoperability	7
1.4.2	Social Collaboration	8
1.4.3	Multilinguality	9
2	Collaborative authoring on the Web	11
2.1	Crowdsourcing	11
2.1.1	Definitions	11
2.1.2	Benefits, Limitations and Controversies	13
2.2	Web 2.0 and Social Networking	14
2.2.1	Definitions	15
2.2.2	Benefits, Limitations and Controversies	16
2.3	Wiki	18
2.3.1	Definitions	18
2.3.2	Benefits, Limitations and Controversies	19
3	State of the Art: Systematic Literature Study	21
3.1	Organization of the Study and Paper Selection	21
3.1.1	Research Questions	22
3.1.2	Search Strategy	22
3.1.3	Study Selection	23
3.2	Paper Categorization and Domain Overview	24
3.2.1	Analysis of Existing Surveys	25
3.2.2	Analysis of Paper Distribution	26
3.3	Terminology	29
3.3.1	Learning	30
3.3.2	Content	30
3.3.3	Authoring	30
3.3.4	Technical Concepts and Technologies	31
3.4	Educational Content Co-creation Studies	32
3.4.1	Collaborative Content Authoring	32
3.4.2	Content Categorization	33

3.4.3	Content Personalization and Recommendation	34
3.4.4	Content Localization	34
3.4.5	Content Quality Assurance	35
3.5	Educational Content Reuse and Re-purpose Studies	35
3.6	Social Collaboration on Educational Content Studies	36
3.6.1	User Coordination	37
3.6.2	User Engagement	37
3.7	Conclusions	38
4	Collaboration on Multilingual Structured Educational Content	41
4.1	CrowdLearn Concept	42
4.2	WikiApp Data Model	45
4.3	Co-evolution of Structured Multilingual Educational Content (CoSMEC)	48
4.4	Conclusions	50
5	Implementation and Evaluation	53
5.1	SlideWiki Architecture, Functionality and User Interface	54
5.1.1	Content Development.	55
5.1.2	Content Migration.	57
5.1.3	Multilinguality	68
5.1.4	User Engagement and Social networking	72
5.1.5	E-Learning and E-assessment.	73
5.2	Limitations of the Implementation and Description of the Next Generation	76
5.2.1	Usability Studies	76
5.2.2	Defined Issues of Current Implementation and Proposed Solutions	81
6	Discussion and Conclusion	85
A	Research Paper: Measuring the quality of relational to RDF mappings	89
A.1	Abstract	89
A.2	Introduction	89
A.3	Related work	90
A.4	Quality Requirements	91
A.4.1	Quality of the mapping implementation and representation	91
A.4.2	Faithfulness of the output	93
A.4.3	Quality of the output	94
A.4.4	Interoperability	96
A.5	Evaluation	98
A.6	Conclusion	100
B	Research Paper: Balanced scoring method for multiple-mark questions	101
B.1	Abstract	101
B.2	Introduction	101
B.3	Terminology	102
B.4	Related work	103
B.4.1	Dichotomous scoring	103
B.4.2	Morgan algorithm	104

B.4.3	MTF scoring	104
B.4.4	Ripkey algorithm	104
B.5	Balanced scoring method for MMQs	105
B.5.1	Concepts	105
B.5.2	Mathematical model	105
B.6	Evaluation	107
B.6.1	Synthetic experiments	107
B.7	Conclusions	110
C	RDB2RDF Mapping Documentation	113
C.1	R2RML mapping representation	113
C.2	Tabular mapping representation	127
C.3	SlideWiki knowledge graph	127
Bibliography		133
List of Figures		143
List of Tables		145

CHAPTER 1

Introduction

Knowledge exists to be imparted.

R. W. Emerson

The Web does not just connect machines, it connects people.

Tim Berners-Lee

1.1 Motivation

Knowledge expands people's possibilities. It promotes creativity and imagination. In addition to its intrinsic value, it has substantial instrumental value in expanding other freedoms. Being educated empowers people to advance their interests and resist exploitation. According to [89], the worldwide level of literacy has dramatically increased during the 20th century. Both males and females have got the right to be educated, number of schools worldwide grows constantly. However, while basic conventional education is accessible for the most people around the world, its quality often remains low. Often people do not have access to a high-quality education due to its extremely high costs, geographical, language or cultural barriers. Overcoming those barriers is a current challenge educators have to face.

Another challenge is the 21st century realities, which require people to continuously acquire new knowledge and skills. This fact led to the appearance of *life-long learning* concept. Ensuring the possibility of professionals to educate themselves is another important challenge of today's education programs. The both challenges are addressed by the concept of Open Educational Resources (OER).

According to the UNESCO definition¹, OERs are teaching, learning or research materials that are in the public domain or that can be used under an intellectual property license that allows re-use or adaptation (e.g Creative Commons²). The OER movement aims to break down

¹ <http://www.unesco.org/new/en/unesco/themes/icts/open-educational-resources/>

² <http://creativecommons.org>

barriers that have blocked free access to academic content. Until recently, most digital educational content was not freely available from outside its owning institutions. Moreover, content providers often ensured the materials are protected by passwords and proprietary formats. The OER movement aims to break down these strategies, proving that knowledge exchange is beneficial for all participants and society in whole. And the effort put in the promoting those ideas has been already accepted on the highest levels. The report from UNESCO forum on open content for higher education [56] stays: "OER represents a major step toward sharing teaching materials, methods and tools, just as academics have shared their work in scholarly papers for a long time. The result is to augment teaching resources while expanding knowledge opportunities for learners and faculty members".

Since the OER strategy has become widely accepted, meeting increasing and increasingly varied demand for quality higher education is an important consideration in the policy debate and institutional development in many countries. New developments in higher education – from virtual universities and e-learning to open source initiatives – speak to the efforts on the part of the traditional higher education community, as well as new providers, to address this increasing demand. The open source movement can be seen as reflecting the philosophy of academe, which is based upon a sharing of information and new discoveries through the peer-reviewed academic publication process to share knowledge. It is particularly important in the case of developing countries, for whom demand often greatly exceeds capacity in the existing higher education system.

Nowadays open initiatives in education have crystallized around three major areas of activity: the creation of open source software and development tools, the creation and provision of OpenCourseWare (OCW), and the development of standards and licensing tools. The success of OER movement depends on presence of efficient approaches in each one of them. While the idea of opening and sharing educational resources is nowadays widely accepted, the OER providers and consumers worldwide face significant challenges related to providing increased access, while containing or reducing costs. Perhaps the best-publicized and most copied institutional OER model is MIT OpenCourseWare³ project. However, being overall successful, the model is hard to follow due to the extremely high costs of development: the project costs approximately \$3.5 million every year, which means the release of each course took about 100 working hours in average. Apart from that, there is a number of additional challenges preventing the OCW movement from reaching its goals, as summarized in paragraphs below.

OpenCourseWare and MOOC Initiatives have inherent weaknesses. The OpenCourseWare concept arose from the success of open source software by expanding the concept of openness to the educational context. A vast amount of OCW is being published online to make educational content more attainable. The awareness of such courses among users and the popularity of systems providing such courses are increasing. However, from a subjective experience, OCW is frequently cursory, outdated or non-reusable. Many of the courses contain only a syllabus, are only available in one language or in formats, which are difficult to reuse. More than 250 institutions worldwide are openly publishing courses today. Some OCW initiatives by renowned universities are MIT OpenCourseWare, Stanford Engineering Everywhere, Carnegie Mellon Open Learning Initiative, Harvard University Extension School, Open Learning Initiative, Initiative, Open Yale Courses, Rice University's OpenStax, OpenLearn, and Webcast.Berkeley. Further OCW repositories have been made available by organizations such as OpenCourseWare Con-

³ <http://ocw.mit.edu/>

sortium, Open Education Resources Commons, and The Saylor Foundation's Free Education Initiative. The Open Education Consortium as the central community of institutions and organizations working on open education lists 26,611 courses from 80 providers. The systematic analysis [3] of 100 courses from major OCW repositories revealed the following weaknesses:

- Legal Reusability. Overall only 28 out of the 100 courses have a truly open license, the vast majority (i.e. 57 out of 100) are restricting reuse to non-commercial scenarios (i.e. CC-BY-NC-SA), which is not open licensing according to the Open Definition. Often, for example, if courses are offered with a fee or the training organization is a for profit-organization, the non-commercial restriction thus prevents reuse.
- Multilinguality. English is the original language of the vast majority of the courses, only two of them have been translated to other languages. 12 out of 100 courses were originally offered in languages other than English.
- Format repurposeability. Overall 68 courses are offered in formats supporting some form of repurposeability. However, 52 of the 68 courses are only available in PDF, thus preventing true reusability.
- Recency. Out of 100 courses assessed in 2014, only 32 courses have been updated in the last two years.
- Learning by Self-assessment. Self-assessment material is only available as part of the course material of 15 courses and separately for another 40 courses. Out of these 55 courses with some form of self-assessment, just 25 provide solutions.
- Engaging course material. Although, two thirds (i.e. 65) of the courses have at least one example and one illustration, just one quarter of the courses (i.e. 25) have more than 50 examples. Based on the ratio of illustrations per course unit, almost two thirds of the courses (i.e. 60) are objectively of low attractiveness.
- Community Involvement. The content of 61 courses has been created by a single author. Only 16 courses are the result of collaborative work.

Educators are reinventing the wheel every day. Despite the plethora of Learning Content Management Systems (LCMS) available and used today, educators are spending most of their time replicating what others have done before. In the educational domain, there is currently no systematic way of sharing and collaborating around content, as is the case in other online domains (e.g. Wikis for text authoring, GitHub⁴ for software development or OpenStreetMap⁵ for map creation). Learning resources are shared only in a cursory way and frequently do not allow repurposing by technical or licensing reasons. When resources are shared, there is no mechanism, to coordinate the synchronisation of various revisions of content (as Wikis, Git or OpenStreetMap allow). Hence, the maintenance and updating of learning resources is cumbersome and learning resources become quickly outdated.

There is no integrated support for multi-linguality and translation of learning material. There is currently a hype around Massive Open Online Courses (MOOCs). However, there are two key issues with MOOCs:

⁴ <http://github.com>

⁵ <http://openstreetmap.org>

- the development and creation of MOOCs is extremely expensive and can easily reach 250-500k Euro without even considering the MOOC delivery cost
- MOOCs are only available for relatively few courses covering popular topics and in very few languages

In order to improve this situation and increase the breadth and depth of OpenCourseWare, we have to substantially reduce the effort required for the creation, translation and maintenance of the content. More widely sharing the burden of content creation, translation and maintenance as well as employing crowd-sourcing techniques seems to be a natural candidate strategy for overcoming this obstacle.

Multiplatform learning delivery incorporating LCMS, MOOC. Learning Management Systems have dominated e-learning for several years. They have been widely used by academic institutions for delivering their distance learning programmes, as well as for supporting their students outside the classroom. They have also been established in the business sector as the mainstream platform for delivering training services to employees. A Learning Management System (LMS) is an online software application offering facilities for student registration, enrolment into courses, delivery of learning materials to students, student assessment and progress monitoring. Popular examples of LMSs used by the academic as well as the business world include *Blackboard*⁶, *Moodle*⁷, and *CLIX*⁸.

The emergence of OERs has greatly facilitated online education through the use and sharing of open and reusable learning resources on the Web and via social networks. Learners and trainers can now access, download, remix, and republish a wide variety of quality learning materials available through open services provided in the cloud. The emergence of affordable tablet technology has also enabled the mass production and delivery of rich interactive learning materials as eBooks, distributed via popular channels like *iTunes U*⁹.

The OER initiative has recently culminated in MOOCs delivered via providers such as Udacity¹⁰, Coursera¹¹ and edX¹². MOOCs have very quickly attracted large numbers of learners; for example over 400,000 students registered within four months for edX courses. More recently, the Open University established FutureLearn¹³ as the UK response to the emergence of MOOCs in collaboration with over 40 organizations (37 universities). Since its launch, FutureLearn has attracted more than 1 million registered learners worldwide with over 2.2 million course sign-ups.

These initiatives have led to widespread publicity and also strategic dialogue in the education sector. The consensus across different groups of stakeholders is that after the web-induced revolutions in communication, business, entertainment and media, amongst others, it is now the turn of universities. Exactly where this revolution will lead is not yet known, but some radical predictions have been made including the end of the need for university campuses, while milder future outlooks focus on ‘blended learning’ (traditional lectures with new digital

⁶ <http://blackboard.com>

⁷ <http://moodle.org>

⁸ <http://www.im-c.de/germany/en/solutions/learning-management/clix-learning-suite>

⁹ <https://www.apple.com/education/ipad/itunes-u/>

¹⁰ <http://www.udacity.com/>

¹¹ <http://coursera.org>

¹² <http://edx.org>

¹³ <http://futurelearn.com>

interactive activities). Whatever the future might bring, the landscape in higher education is about to change radically.

There is a plethora of systems covering other stages of educational value chains, such as Learning Management (e.g. ILIAS, Moodle, OLAT), Learning Content, Learning Delivery (e.g. OpenCast, FutureLearn), Learning Analytics Systems and social networks. Instead of trying to incorporate as much functionality as possible in a single system, we need to facilitate the exchange of learning content and educational data between different systems in order to establish sustainable educational value chains and learning eco-systems. Open (learning meta-data) standards such as SCORM [1] are first steps in this direction, but truly engaging, inclusive and multi-lingual value chains can be only realized if we take the content structuring to the next level and employ techniques such as interactive HTML5 (which can be presented meanwhile on almost all devices) and fine-grained semantic structuring and annotation of OpenCourseWare.

1.2 Problem Definition

Based on the discovered challenges discussed above, we define the problem addressed by our research as *producing reusable multilingual high-quality educational content with low resources consumption*. And while the high levels of quality and reusability of the produced content can be reached by using models similar to MIT OCW, the issue of decreasing production costs remains mostly unsolved. Even more resources are consumed if there is a need to offer the content in different languages.

An important step towards decreasing the costs while increasing the quality of the educational material is applying collaborative techniques to the production process. Such collaboration affects different aspects of content production such as content annotation, personalization and sharing. According to the study [117], already at that time most of the work in academia, business and industry was completed by groups of people collaboratively. This is why collaborative creation of education materials is natural. From psychological and sociological point of view it is proven, that people like to collaborate [138, 65]. Several psychologists demonstrated the effects of collective work on the process of cognitive development of people. Generally, people even *prefer* to work together. According to [138, 65], the reasons for this include:

- benefiting from partner's knowledge;
- building strong personal relations with others;
- increasing the work quality by critics and experience exchange;
- avoiding duplication and redundancy of tasks.

If we consider different domains of content creation, collaborative authoring has dramatically improved the efficiency, effectiveness, quality and timeliness of content creation. For text authoring, for example, Wiki's are meanwhile prevalent and used in corporate intranets of large companies as well as by online communities. In particular, for encyclopedic article creation Wikipedia has demonstrated how such a large-scale collaborative process can be superior to traditional content creation processes. However, as shown in Figure 1.1 the collaborative content authoring is not limited to textual content. For geospatial data, for example, OpenStreetMap has shown how a community of a few thousand collaborators can create a value, which required previously multi-billion Euro investments. In the domain of open-source software development

Text	Encyclopedic articles	Maps	Software	OpenCourseWare
				?

Figure 1.1: Collaborative authoring platforms in various domains

platforms such as SourceForge, Google Code or recently GitHub have dramatically improved the effectiveness and efficiency of community collaboration around software source code and facilitated the creation of widely-used software projects such as Linux, Firefox and LibreOffice.

These collaborative platforms have in common the fact that they were specifically created to support a particular content type. Wikis support text authoring and versioning, with OpenStreetMap editors, maps can be created and version control interfaces such as GitHub are tailored for software source-code. For OpenCourseWare, such technological and organizational support for collaborative authoring and translation does not exist. Namely, there is no authoring tool supporting true collaboration on learning [110], which means not only possibility to reuse (parts of) the content, but allowing authors to understand what other authors do (1), why a learning object should be created (2) and how the authoring process is proceeding (3).

In order to define the challenges preventing the development of such a system we have conducted an systematic literature review in the domain of collaborative authoring of educational content. The detailed results of the review are presented in chapter 3. According to the study, the main obstacles for implementing the support of large-scale collaboration on educational content are (1) content reusability and interoperability, (2) full support of social collaboration and (3) support of multilinguality. In scope of the thesis we define and discuss the obstacles, as well as propose and evaluate an example system design facing them.

1.3 Usage Scenario

Andrea Malupos is a trainer supporting young adults with disability and impairment for their professional life. At a training center in Madrid, she is training groups of 5-10 school graduates in technical and horizontal topics for improving their employment chances. A particular important aspect is auto-learning, where the trainees work with the course and accompanying material. Andrea was requested to prepare a new course module on career development, personal marketing as well as curricula vitae and reference portfolio presentation. Although she finds some material she could use for her course online, it is cumbersome to investigate licensing conditions, organize and translate the content (mostly only available in non-reusable PDF) and enrich it with specific material as well as self-learning and -assessment possibilities. However, Andrea is part of a national and international network of trainers providing similar training course. A collaborative web platform developed based on our approach is extensively use within the network. Using the platform she can quickly develop a structure of the envisions course material and integrate some existing modules she found. The colleagues in her network are interested to contribute specific parts related to their area of competence and using the semi-automatic and crowd-sourcing translation features the material is easily translated and kept-in-sync in five languages. The branching, merging and repurposing features enable every-

one to create his own version of the course material, while at the same time benefiting from updates others provide. Since the material is used by a number of collaborators, new additions and extensions are quickly incorporated into the course and available to everyone. Andrea's Italian colleague plans to produce a MOOC in Italian on the topic. Since the course material including self-assessment questions, tasks, project descriptions and illustrations is already available, the effort to produce the MOOC is reduced by more than 60%.

1.4 Overview of our Approach

The main goal of the research summarized in the thesis is providing a complex approach dealing with the challenges of collaborative authoring of OERs discussed above. Namely, our research answers the following questions:

1. Related to the content reusability:
 - How to make the content truly reusable?
 - How to make it searchable
 - How to publish the content effectively?
2. Related to the social collaboration:
 - How to organize the collaboration effectively?
 - How to deal with mistakes and vandalism?
 - How to collaborate on structured content?
3. Related to multilinguality:
 - How to decrease the costs of producing the multilingual content?
 - How to synchronize the content of different language versions?

In the section we briefly describe the strategies underlying the approach.

1.4.1 Content Reusability and Interoperability

The paradigm of *Open Educational Resources* supposes the creation and sharing of reusable, re-purposable learning objects, annotated by standardized metadata. Only learning objects that satisfy those requirements can be (re-)combined in maintainable high-quality OCW. In order to achieve the high levels of learning objects searchability, reusability and interoperability the content has to be structured, annotated by standard metadata and published on the web in an efficient manner. However, the state-of-art approaches in learning objects development and publishing often do not satisfy those requirements. Usage of different, often proprietary content formats and system architectures prevents effective migration of content between platforms; lack of structuring makes the repurposing of smaller content pieces impossible; plain-text metadata limits the searchability.

In order to deal with the issues we propose to use widely accepted content formats together with generous data models enabling fine-grained content structuring and efficient collaboration on it. Our approach deals with learning objects in a form of HTML-snippets organized in tree structure. Each branch of the tree as well as each leaf is an individual learning object annotated by metadata and it can be cut, copied or changed without updating the rest of the tree. The data

model we have developed in order to organize the content in such a way is called WikiApp. The model supports all operations necessary for collaboration on the structured content, including social networking and version control actions. In the [section 4.2](#) we introduce the model in details. We have chosen HTML format for the content not only due to its generosity, but as well due to the recently developed technologies of the HTML-content annotating and publishing. To be published on existing e-learning platforms, the HTML format can be transformed in a number of common e-learning formats automatically or with moderately low effort. For example, we have implemented the automatic content transformation into the SCORM packages, discussed in [item 5.1.2](#).

The content structured and organized in the proposed way can be transformed into machine-readable formats, such as Linked Data. The Semantic Web and Linked Data movements with the aim of creating, publishing and interconnecting machine readable information have gained traction in the last years (cf. *LODStats*¹⁴).

Making content machine-readable provides a wide list of advantages compared to not machine-readable formats [86]:

- For *search and retrieval* enriching documents with semantic representations helps to create more efficient and effective search interfaces, such as faceted search [151] or question answering [98].
- In *information presentation* semantically enriched documents can be used to create more sophisticated ways of flexibly visualizing information, such as by means of semantic overlays as described in [28].
- For *information integration* semantically enriched documents can be used to provide unified views on heterogeneous data stored in different applications by creating composite applications such as semantic mashups [7].
- To realize *personalization*, semantic documents provide customized and context-specific information which better fits user needs and will result in delivering customized applications such as personalized semantic portals [127].
- For *reusability* and *interoperability* enriching documents with semantic representations (e.g. using the Dublin Core vocabularies¹⁵) facilitates exchanging content between disparate systems and enables building applications such as executable papers [104].

In order to perform the transformation we use RDB2RDF¹⁶ approach, allowing to convert the relational data to linked data format directly from the database with ability to customize the output data model. We discuss semantic publishing and describe our approach for content conversion in [subsection 5.1.2](#).

1.4.2 Social Collaboration

The collaboration on structured educational content can be established within a limited group of experts (1) or involve the power of a crowd (2). The first approach has a number of issues:

¹⁴ <http://stats.lod2.eu>

¹⁵ <http://dublincore.org/>

¹⁶ <https://www.w3.org/standards/techs/rdb2rdf>

- it requires a lot of preparation, as the domain experts with the knowledge of certain languages have to be found first;
- each expert has her own background and point of view on the topic, thus, the negotiations can have a negative effect on the time costs;
- it results in production of content available in only a limited number of predefined languages;
- the content easily becomes outdated, as a limited number of experts being spread internationally might not be aware of all new findings in the domain;
- the content refinement is time-consuming and can cause de-synchronization of the content.

Since the emergence of Web2.0 concept the Internet has transformed from a large-scale source of information into a large-scale collaboration platform. Following the ideas of Tim Berners-Lee, it is now possible and even easy not only consume information, but to provide and share it through the web channels. Moreover, a number of technologies has been developed to support different aspects of the collaboration. Thus, the wiki-way of content authoring provides tools in order to make the content creation process accessible for larger audience; social networking channels support the communication and negotiation of collaborators, as well as allow sharing the content between users net. Those technologies allow decrease the costs of content development while keeping the content unbiased, up-to-date and open, which are especially important values for the educational content.

In our approach we apply the technologies supporting social collaboration to the structured educational content development. In order to do so, we needed to deal with versionning, merging and branching of the structured content, as well as with standard-compliance issues. To address the challenges we have developed the CrowdLearn concept, presented in [section 4.1](#). The proposed concept deals with learning objects of fine granularity, allowing their authoring and reuse for all users under CC-BY-SA ¹⁷ license. The license allows reuse, repurpose and update the content for anybody while the requirement of indicating a list of authors is satisfied. The structured content is built according to the WikiApp data model, that natively supports social networking actions to engage user negotiation and community building activities.

1.4.3 Multilinguality

Nowadays, the support of multilingual content becomes crucial for educational platforms because of the benefits it makes. One of the most important benefits of multilingual content supplying is an opportunity to educate without language barriers. This causes the number of learners to increase exponentially according to the number of languages content is available in. The learning quality for previously presented users increases as well, due to the possibility of studying in their mother language. Another benefit of the multilinguality is the ability to share and exchange knowledge within the multicultural environment. This opportunity is crucial for fostering the increase of education quality, especially in developing countries, where institutes and universities often base courses on outdated information, concepts and theories. Access to the international scientific knowledge through content synchronization would help experts and educational institutes to be aware of the state-of-art in their field.

¹⁷ <https://creativecommons.org/licenses/by-sa/4.0/legalcode>

In our approach we propose to use the power of a crowd to author the content available in a number of different languages. Apart from other benefits, using crowd-sourcing instead of expert-based translation ensures that the list of available languages is not fixed and the number of languages depends on the community size, diversity and activity. An example of the successful application of crowd-sourcing techniques to multilingual content authoring is Wikipedia¹⁸. However, the Wikipedia concept does not suppose an (semi-)automatic synchronization of the content presented in different languages. Although synchronization is sometimes being done manually by contributors, the versions in different languages often stay outdated or even contradictory to each other.

In order to deal with the synchronization issue we have developed the Crowd-sourcing (semantically) Structured Multilingual Educational Content (CoSMEC) concept. The concept is built on top of CrowdLearn concept and WikiApp data model discussed above. It adds additional relations and operations to model translating activities and connections between original and translated versions. The CoSMEC concept introduces the paradigm of *co-evolution* of multilingual content, that means the ability to update a translation to the current state of the source object and vice versa. In this manner the content versions in different languages can be (semi-)automatically synchronized, while staying individual learning objects and thus ensuring its abilities to reuse and repurpose.

¹⁸ <http://wikipedia.org>

CHAPTER 2

Collaborative authoring on the Web

(...) people need to be creative. They want to be able to record what they think. They want to be able to, if they see something wrong go and fix it.

Tim Berners-Lee

@TODO: generous introduction

In the chapter we give an overview of the technologies and paradigms underlying our research. Namely, we discuss crowdsourcing, social networking and wiki technology. We give the (various) formal definitions for the basic terms, as well as summarize the known benefits, limitations and controversies of the paradigms.

2.1 Crowdsourcing

The examples of what is now called crowdsourcing can be found far back in the humankind history. According to the timeline presented in [38] the first documented action that can be categorized as a crowdsourcing has taken place in 1711 in Great Britain, when the British government was trying to find a way to measure a ship's longitude. They offered the public a monetary prize to whomever came up with the best solution, which resulted in collaborative discussion and sharing ideas between previously unrelated people.

Nowadays the crowdsourcing projects are mostly initiated on web-based platforms. Although the crowdsourcing can be applied to a wide range of tasks, in the scope of our research we focus on the tasks (activities) related to all aspects of content development. In the section we summarized the formal term definitions we rely on and discuss the influence of crowdsourcing onto the content developing process.

2.1.1 Definitions

The specific definitions for the crowdsourcing are yet heavily debated. [51] The word *crowdsourcing* is a compound contraction of *crowd* and *outsourcing*. Thus crowdsourcing can be seen as outsourcing to the crowd, where a *crowd* is a large set of anonymous individuals [141].

Due to anonymity, individuals cannot be individually identified or recognized. Implicit in this definition is the idea that a firm cannot “build its own crowd”. Moreover, the crowd is generally composed of heterogeneous individuals. In particular, a crowd may be composed of scientists and experts in various fields, but also of novices.

Crowdsourcing as a business term was coined in 2005 by Jeff Howe¹, who later defined it in his book. The term is also defined by Merriam-Webster dictionary². We combined both definitions and the word meaning in order to create a more general term we will use later in the research (see Definition 3).

Definition 1 (By Jeff Howe) *Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.*

Definition 2 (Merriam-Webster dictionary) *Crowdsourcing is the process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from an online community, rather than from traditional employees or suppliers.*

Definition 3 (Combined definition) *Crowdsourcing is the process of obtaining needed services, ideas, or content by a form of outsourcing not directed to other companies but to the large set of anonymous individuals by means of an open call.*

Sometimes it might be challenging to decide if an application can be called a crowdsourcing platform. In scope of the current research we take into consideration only the systems dealing with the digital content. In order to define the class of systems we will discuss in the thesis, we adapted the definition [46] as following:

Definition 4 (Crowdsourcing system/platform) *A crowdsourcing system is a web-based system that enlists a crowd of humans to help solve a problem defined by the system owners, and in doing so, it addresses the following four fundamental challenges: (1) how to recruit and retain users; (2) what contributions can users make; (3) how to combine user contributions to solve the target problem and (4) how to evaluate users and their contributions.*

The crowdsourcing activities related to the web-based digital content delivery include but are not limited to:

- *Creative crowdsourcing*, that spans sourcing creative projects such as graphic design, crowdsourcing architecture, apparel design, movies, writing, illustration, etc. When applied to the educational domain, the creative crowdsourcing activity can be applied for collaborative authoring of educational content and its metadata.
- *Crowdsourcing language-related data collection*, that is functionally close to creative crowdsourcing, but requires the crowd to be multicultural. This special activity is used for creating and improving all kinds of dictionaries. In section 4.3 we discuss the adaptation of the activity for the creation of multilingual educational content and its metadata.

¹ <http://www.wired.com/2006/06/crowds/>

² <http://www.merriam-webster.com/dictionary/crowdsourcing>

- *Crowdvoting*, that allows applications to collect opinions and rankings on a certain topic from a large auditory. The crowdvoting may be applied to solve a variety of problems, such as quality assurance, product and services rankings, content recommendations and others. In the educational domain we will discuss the crowdvoting mostly as an approach for quality management.
- *Social tagging*, that collects tags for the content, usually in form of notes, questions, labels etc. Those tags are then used for content categorization, search improvements, personalization and recommendation. Social tagging allows to transform the plain text into learning objects with machine-readable metadata.

While those activities are examples of explicit crowdsourcing, a number of approaches to solve a wide range of problems are based on implicit crowdsourcing. According to [46], with explicit crowdsourcing, users can build artifacts by providing information and editing other people's work. Users can evaluate particular items like books or webpages, or share by posting products or items. Implicit crowdsourcing can take two forms: standalone and piggyback. Standalone allows people to solve problems as a side effect of the task they are actually doing, whereas piggyback takes users' information from a third-party website to gather information. Implicit crowdsourcing is most frequently used to collect user provided data in order to apply data-mining techniques to them. For example, a user's search history can be collected and analyzed in order to discover keywords for ads, spelling corrections, and finding synonyms. In this way, users are unintentionally helping to improve the performance of existing systems. Often gamification methods are used to motivate users to implicitly contribute. For example in the *ESP game* [155] users guess what is pictured on images and then these labels are used to tag Google images.

2.1.2 Benefits, Limitations and Controversies

Only based on the Definition 4, it is already clear that crowdsourcing as a task-solving approach has not only benefits when compare with expert-based approach, but a number of issues and challenges as well. When compared with a limited number of experts, a crowd has both strengths and weaknesses. As it was already mentioned in subsection 2.1.1, we focus our later discussion on the content development task. In the section we discuss the positive and negative influence of crowdsourcing on different aspects of content development process.

Content quality Allowing a community or even the whole population contribute to the content improves its quality due to continuous iterational development, improvement and update of the content. Crowdsourcing also positively influences the quality by producing unbiased content due to incorporating a large population of participants with a diverse background. However in order this to be true, the initiator of the crowdsourcing activity must be sure that the community has enough expertise to solve the task. The unbiasedness of the produced content is also not absolute. The research [71] proves that crowdworkers are a nonrandom sample of the population. For example many researchers use crowdsourcing in order to quickly and cheaply conduct studies with larger sample sizes than would be otherwise achievable. However, due to limited access to internet, participation in low developed countries is relatively low. Participation in highly developed countries is similarly low, because the low amount of pay is not a strong motivation for most users in these countries. These factors lead to a bias in the population pool towards

users in medium developed countries and cause the need to conduct a comprehensive user base study.

The need to study the contributing crowd is caused as well by the challenge to define a range of possible contributions the users are allowed to make. For example, when building a structured knowledge base, users might be allowed to supply attribute-value pairs, correct mistakes and discuss the content structure. But they can also supply inference rules, resolve controversial issues, and merge conflicting inputs [119]. Making this strategical decision has a huge impact on the content quality and is crucial for the overall project success. Moreover, the study of the crowd has to be conducted on a regular basis, as the quality of the crowd might significantly change over time.

Another issue is the the content consistency, that is the back side of its unbiasedness. Due to the crowd diversity, it is likely to have inconsistency in the used terminology or formatting due to absence of predefined rules. Moreover, the users background and cultural differences often lead to the semantic inconsistency, when different parts of the same artifact (or its versions in different languages) are based on the different assumptions and concepts. This is especially true for the multicultural crowds and in many cases might not be predicted even by the comprehensive user base analysis.

Economical efficiency During the iterations which will lead to the better quality content, many unusable and low quality contributions are made. This causes the necessity to sort and filter the contributions, that results in increased number of management tasks and stuff. However, most of the crowdsourcing works are done by people who are paid or directly benefit from the outcome. In many other cases, the end product is the outcome of a single person's endeavour, who creates the majority of the product, while the crowd only participates in minor details [160].

As the content quality strongly depends on the size and quality of the user base, it is required to spend more resources on the project dissemination and marketing. As well the challenging nature of the crowdsourcing requires the crowdsourcing systems to be more comprehensive (and more expensive) than the systems designed for the limited number of trusted experts collaboration.

End-user satisfaction End-user satisfaction is also influenced by choosing the crowdsourcing as a basic content development approach. From one side, users can follow the process of content creation from early stages and provide their opinions and edits. In that manner the responsibility for the final content quality is shared between content providers and consumers. Feeling themselves active participants, users most likely will not be disappointed if the process takes more time or resources as planned. From another side, the content development process might seem to the end users a never-ending story. The content pieces are often at different stages of development, making it hard to predict when the whole planned output will be ready to use.

2.2 Web 2.0 and Social Networking

Already from the time of the World Wide web invention, his inventor Tim Berners-Lee considered it to be the space for global collaboration. In one of his interviews ³, he says: “(...)

³ <http://www.ibm.com/developerworks/podcast/dwi/cm-int082206txt.html>

I really wanted it to be a collaborative authoring tool. And for some reason, it didn't really take off that way." Only around ten years later we first started to see the web sites, content of which can be changed by users. First, only small portion of administrators could edit the content of a page through the administration board. Later the systems allowed users to have their own pages, they have administrative rights to. This has made it possible for common users to write down and share their thoughts within blogs and bulletin boards. And finally the social interaction tools, allowing community building and instant communication between users within a group, has made it possible to authors the content collaboratively.

In the section we give an overview of the concepts behind these tools, which in combination have transformed the static read-only Web (Web 1.0) into dynamic writable Web (Web 2.0).

2.2.1 Definitions

The term *Web2.0* first appeared in [44], however the modern meaning was established and popularized due to the annual *Web 2.0 Summit*⁴ started in 2004. Since then, the concept of Web2.0 was growing and developing. New technologies and web-based applications added new meanings to the term, and this process is still going on. The Web2.0 can be defined as *writable* or *participative* web, which means that the Web2.0 applications and services allow users to use Internet in a more interactive and collaborative manner, emphasizing peers' social interaction and collective intelligence [108]. A conceptual definition of Web2.0 was given by [66]

Definition 5 *Conceptual definition Web 2.0 is a second generation, communicative form of the World Wide Web that emphasizes active participation, connectivity, collaboration and sharing of knowledge and ideas among its users.*

Although Web 2.0 suggests a "new version" of the World Wide Web, it does not refer to an update to any technical specification, but rather to cumulative changes in the way Web pages are made and used⁵. As for today, examples of Web 2.0 services include social networking sites, blogs, wikis, folksonomies, video sharing sites, hosted services, web applications, and mashups. Based on the research done in scope of the thesis we formally define Web2.0 as following:

Definition 6 *Web2.0 is the complex of Web sites and services allowing and supporting social collaboration on user-generated content.*

Below we explain the terms we used to construct the definition.

User-generated content. The concept of *participative web* supposes wide use of intelligent web services that empower users to contribute to developing, rating, collaborating and distributing content. The content created with the use of these services is called *user-generated content*. Although the term is self-explaining, it can be formally defined as following:

Definition 7 *User-generated content is the content made publicly available over the Internet, which reflects a certain amount of creative effort.*

Most user-generated content activity is undertaken with no expectation of remuneration or profit. Motivating factors include connecting with peers, self-expression, and achieving a certain level of fame, notoriety or prestige [154].

⁴ originally known as the Web 2.0 Conference

⁵ https://en.wikipedia.org/wiki/Web_2.0

Social collaboration. The emergence of Web2.0 concept brought to life a phenomena of social networking sites or platforms.

Definition 8 (By Educause Center for Applied Research [130]) *Social Networking site is a web-based service that allows to (1) construct a public or semi-public user profile within a bounded system, (2) articulate a list of other users with whom they share a connection and (3) view and traverse their list of connections and those made by others within the system.*

Those web-based applications first were created to enable large-scale online communication. The services like Facebook⁶, Google Plus⁷, Twitter⁸ etc. allow users to establish virtual connection with their real-life friends and users with similar interests. The distinguishing feature of social networking platforms when compared to other communication channels is that communication between users is mostly established through sharing digital materials in a form of textual content and media files. Thus, instead of sending private instant messages users often share their ideas and opinions publicly, creating and sharing their knowledge.

The developed culture of sharing allowed to apply social networking to support a variety of collaborative activities, including content production (see [Figure 2.1](#)). Social networking switched its role from providing a communication channel to supporting large-scale collaboration on the web and became a necessary tool in any Web2.0 application.

2.2.2 Benefits, Limitations and Controversies

Within the functions of Social Networking there are three most beneficial for the (educational) content development [21]: (1) support for conversational interaction; (2) support for social feedback; and (3) support for social networks and relationships between people

Support for conversational interaction. The conversational interaction of users during the content development is an essential feature of collaborative authoring tool. **@TODO: importance of content negotiation** In scope of our research we are focused mainly on the text-based communication channels, however the major principles can be applied to voice- and video- tools as well. There are two major types of tools enabling communication and discussions between users:

1. synchronous: text-, voice- and video-chats
2. asynchronous: bulletin boards, private and group (instant) messaging

The survey of learning instructors [23] showed that they prefer to use both types while preparing the course materials. Reasons for using synchronous communication included the possibility to implement brainstorming and other team decision-making techniques, and dealing with technical issues. On the other hand, the educators have found asynchronous communication to be helpful for encouraging in-depth, more thoughtful discussion and communicating with temporally diverse colleges as well as holding ongoing discussions where archiving is required. Both types of communication have their disadvantages, however. Disadvantages of synchronous communication include: getting colleges online at the same time, difficulty in moderating large-scale

⁶ <http://facebook.com>

⁷ <http://plus.google.com>

⁸ <http://twitter.com>



Figure 2.1: Social tools [122].

conversations, lack of reflection time for collaborators, and intimidation of poor typists. Educators also cited the limitations of asynchronous communication: lack of immediate feedback, collaborators not checking in often enough, length of time necessary for discussion to mature, and feeling a sense of social disconnection.

@TODO: attach conversations to the content

Support for social feedback. Ability of any reader to leave a feedback in easy and fast way is one of the distinguishing features of social networking sites. Moreover, the feedback itself (usually only positive) becomes a part of the user-generated content it refers to, letting other users judge the level of quality/actuality of the content. Sharing "likes" with the peers helps to spread the content, that users consider to be interesting and useful, while the content of low interest/quality will be filtered from the users attention. An important issue here is often low level of objectivity of such filtering. Thus, a piece of content can have a high user rating in one community and be out of interest in another based on very subjective, "marketing-related" factors, such as attractive illustration or title.

Support for social networks and relationships An important feature of Web2.0 is its support of social networks building.

Social network sites allow individuals to (1)construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. [22] Those features in the end allow user to build her own network and stay updated on what is happening within the net. Applied to content development, social networking helps to engage new users to the team through the relations of team-members. And opposite, it allows newcomers to find suitable teams and projects by following the activity of chosen users they have a common background and/or interests. Another application of the social networking to the content development is ability to follow not only users, but projects itself. This allows team-members to track the changes in an easy way, as well as facilitates the stepping-in process for newcomers.

2.3 Wiki

The fundamental class of tools for authoring of any content on the Web is *Content Management System* (CMS). These tools provide an environment where the creating, editing and managing of multimedia documents can be made on a friendly, reliable and secure way [121]. CMSs aim to facilitate the tasks related to the web administration of the content production, such as editors, user communication tools, styling etc.

The first generation of CMSs aimed to facilitated the content development mainly (only) in sense of allowing not-experienced IT users to develop, organize and publish content on the Web. While the asynchronous collaboration is enabled, these first systems like Blackboard of WebCT did not have sufficient support for truly large-scale collaboration. This limitation became especially important in educational domain with the emergency and wide-spreading of the OER movement. OERs are built on the belief that everyone should have the freedom to use, customize, improve and redistribute educational resources without constraint. This requirement is not satisfied with the conventional CMSs. In search of better solutions the educators started to research the application of wiki [94] technology to solve the issue of the large-scale collaborative content development.

2.3.1 Definitions

In 1995 Ward Cunningham created WikiWikiWeb, the first Wiki space, and he defined the Wiki as "the simplest online database that could possibly work". Later, in January 2001, the founders of the Nupedia project, Jimbo Wales and Larry Sanger, decided to use Wiki to develop the encyclopedia project, creating of this way Wikipedia⁹ - the biggest Wiki space of the world.

The Wiki technology is a combination of two paradigms: crowdsourcing and Web2.0. The crucial feature of a wiki environment is its focus on user collaboration that is enabled by adapting the concept of crowdsourcing for the web-based (textual) content authoring. The Wiki technology allows the site content to be edited in a collaborative way using a simple notation for contents formatting, indexing and interlinking. Additionally, wiki platforms incorporate tools for version control of the content permitting the pages restoration.

In order for a platform to be able to be called a wiki, it must satisfy the rules established by the community [62]:

⁹ wikipedia.org

- Not a single author is permitted: Wiki is a tool that forms its own community and provides democratic collaboration.
- It has a simplified formatting language (mark-up).
- Its focal point is its content, not appearance.
- It provides version follow-up open to all its users.
- Since the processes in wikis are carried out by a community created by a number of people, observations are fast and efficient. Therefore, no security procedure or no process of user acceptance is needed. However, often there is a user who is responsible for verifying and if necessary reverting of the changes made.

2.3.2 Benefits, Limitations and Controversies

Since its inception in the early 2000s [94], wiki technology became an ubiquitous pillar for enabling large-scale collaboration. Traditional, text-oriented wikis enabled the creation of the largest encyclopedia of human-mankind edited by tens of thousands of volunteer editors – Wikipedia. Also, for group collaboration, in corporate intranets or online communities wiki's meanwhile constitute a fundamental base technology.

Like the most of other crowdsourcing applications, Wiki platforms typically support two essential functions – open editing and edit preservation. Open editing refers to the ability for anyone to easily edit the content on a wiki. Edit preservation refers to the ability of wikis to retain all edits to and versions of content contained on the wiki. Taken together, these two features allow users to "roll back" any changes to the wiki and restore content to a previous version. These two relatively simple capabilities of wikis, individually and in combination, can create a robust and transparent collaborative environment [85].

Wikis may work best for knowledge building "over time" (through versions and groups). The specific of wikis is the iterative process of content (and knowledge) development. Even having the predefined plan of what has to be done, the scope of work may be extended due to the new collaboratively created knowledge. This specific allows wikis to become a perfect tool for collaborative knowledge engineering, especially in situations, where no one can have a full knowledge of the topic. Thus, the usage of wikis is not only limited to content development, but to collaborative problem solving. This is enabled by the fact, that wikis allow:

- progressive problem-solving (particularly open-ended problems) and even problem redefinition. For example, Wikis could work well for communities of practice whose goal is to develop solutions to common problems over time in order to improve practice
- explaining increasingly diverse and contrary ideas, as well as examining the relatedness of ideas from diverse contexts
- combining, synthesizing and evaluating definitions and terminology across disciplines
- questioning underlying causes and principles
- critically reading, and responding in a constructive and public way, to others' work
- adding both nuance and complexity to concepts in a given field, through systematic engagement and analysis with work produced by more advanced users

- observing deeply, stereotype less, and avoiding premature judgment

However, Ward Cunningham's wiki paradigm was mainly *only* applied to unstructured, textual content thus limiting the content structuring, repurposing and reuse. More recently with the appearance of semantic wiki's the concept was also applied and extended to semantic content (cf. [section 3.5](#)) In many potential usage scenarios, however, the content to be managed by a wiki is neither purely textual *nor* fully semantic. Often (semi-)structured content (e.g. presentations, educational content, laws, skill profiles etc.) should be managed and the collaboration of *large* user communities around such content should be effectively facilitated.

Another issue with the wiki paradigm is the content maintainability. Common wisdom has it that the Wikipedia has been created by "the crowd". However, the analysis of Wikipedia editing processes reveals that bots are 22 of the top 30 most prolific editors and collectively make about 16% of all edits to the English language version each month [59]. Thus, a large wiki-system is not able to perform efficiently without automatizing some of its functions, especially those related to content quality assurance.

One more issue that is illustrated by Wikipedia is synchronization of content presented in different languages.

In the scope of the current thesis we aim to overcome those limitations, by providing an innovative data model, allowing versionning and collaboration on the (semi-)structured multilingual content (see [chapter 4](#))

CHAPTER 3

State of the Art: Systematic Literature Study

The field of educational content authoring is well addressed in the research literature. In order to obtain the research papers related to our work as fully as possible and filter the unrelated work at the same time we needed to collect only the literature addressing the *collaborative* authoring of reusable educational content. Such filtering can not be done based only on titles or abstracts, as often from neither of them it is clear if the approach proposed is applicable to collaborative authoring. From another prospective the full-text reading of the whole scope of the research in the wide and multi-aspect field is not feasible. We needed a semi-automatic approach to filter the papers from the wide field.

In order to do so, we conducted a comprehensive systematic study of the state-of-art in collaborative authoring of reusable educational materials, following a formal systematic literature review process based on the guidelines proposed in [48, 88].

In the chapter we describe the organization of the study, including the research questions we have formulated. Then we categorize selected papers according to the scope of issues they address and solutions they propose. This process results in a mind map for the domain, as well as domain matrix where the major research problems are intersected with technologies and tools that can be applied for their solution. After that we collect the found terminology important for the field. And finally, we discuss the most influential approaches for solving the most complicated issues in the domain.

3.1 Organization of the Study and Paper Selection

A systematic literature review is an evidence-based approach to thoroughly search studies relevant to some predefined research questions and critically select, appraise, and synthesize findings for answering the research questions at hand. Systematic reviews maximize the chance to retrieve complete data sets and minimize the chance of bias. As a part of the review process, we developed a protocol (described in the sequel) that provides a plan for the review in terms of the method to be followed, including the research questions and the data to be extracted. In this chapter we summarize the results of the systematic literature review we conducted in January 2015.

3.1.1 Research Questions

In order to organize the survey we formulated four main research questions:

1. *What are the main challenging tasks in collaborative authoring of reusable OCW?*
2. *Which technologies are being used to solve these tasks?*
3. *How well covered are the challenging tasks in the scientific literature?*
4. *What are the main gaps in state-of-art research?*

3.1.2 Search Strategy

To cover as many relevant publications as possible, we used the following electronic libraries:

- ACM Digital Library
- IEEE Xplore Digital Library
- ScienceDirect
- Springerlink
- ISI Web of Sciences

Based on the research questions and pilot studies, we found the following basic terms to be most appropriate for the systematic review:

1. *crowd-sourcing OR crowdsourcing OR collaboration*
2. *collaborative OR collective*
3. *authoring OR creation OR edit OR editing*
4. *education OR educational OR learning OR e-learning*
5. *content OR resources OR material*

To construct the search string, all these search terms were combined using Boolean “AND” as follows:

$$(1 \text{ OR } (2 \text{ AND } 3)) \text{ AND } 4 \text{ AND } 5$$

The next decision was to find the suitable field (i.e. title, abstract and full-text) to apply the search string on. In our experience, searching in the ‘title’ alone does not always provide us with all relevant publications. Thus, ‘abstract’ or ‘full-text’ of publications should potentially be included. On the other hand, since the search on the full-text of studies results in many irrelevant publications, we chose to apply the search query additionally on the ‘abstract’ of the studies. This means a study is selected as a candidate study if its title or abstract contains the keywords defined in the search string. In addition, we limited our search to the publications that are written in English and are published after 1999, when, according to [45], "the first glimmerings of Web 2.0 were beginning to appear".

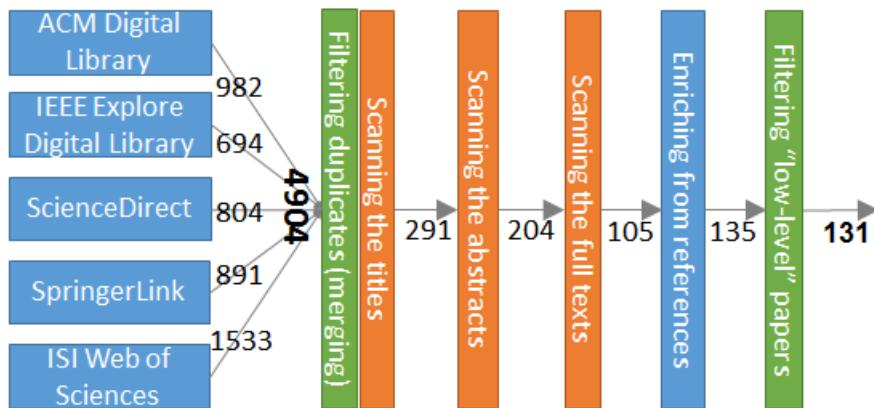


Figure 3.1: Steps followed to scope the search results.

3.1.3 Study Selection

Based on the search query discussed above, we have collected 4904 papers. The amount of papers found exceeded the expected one. We have imported the lists of papers received from the libraries into MicrosoftExcel in order to proceed filtering.

The filtering was proceeded as follows:

1. Remove duplicates by ordering the titles alphabetically and using in-built Excel functions.
2. Filter the titles, keeping only those which imply the conformity of the paper content to the research questions. Keeping in mind the large number of papers collected, we had to be more exclusive than was originally planned. As this might result in missing of important papers, we have added additional steps to ensure the presence of all significant papers in the field.
3. Abstract filtering.
4. Full-text filtering (see details below).
5. Enriching from the references in order to ensure the presence of all significant papers in the field. The references to be included were selected based on the rules for filtering the titles from the initial scope. before including the referenced paper, it was ensured that it is not already present in the scope.
6. Filtering "low-level" papers (see details below).

Details on full-text filtering criteria During the scanning of full-text content of the papers we often had to make decisions about including or not-including "borderline" cases. In the following list we collect the challenges and describe our decisions and their reasons.

- Some papers discuss the systems not allowing to author new content, but allowing to produce personalized courses out of content added in the repository or on the web. We have decided to consider such systems not to be OCW authoring tools and thus we do not included papers related to them.

- In the same manner we do not include papers about tools for collaborative annotation.
- Even harder was to decide either we should include papers related to collaborative writing tools. From one perspective, such tools can be used for writing learning text-books with reusable parts, but from another perspective, more often the tools are used for teaching writing or for writing artistic literature. Another argument to not include this kind of papers is that those tools are usually not complicated technically, their design is quite trivial. The papers related to collaborative writing mostly describe use cases and organizational issues. In result, it was decided not to include collaborative writing related papers as well.
- If initially we selected the papers written after 1999, during the filtering we have restricted the criteria and only kept the papers written after 2002. This is due to the significant difference in the terminology, points of view and challenges addressed. According to our analysis, we can conclude that Social Web started to influence the educational *research* area only in 2001. This can be implicitly proven by appearance of wiki paradigm and Wikipedia in 2001.
- We have filtered out the papers from the same authors and similar years addressing the same aspects, keeping only the latest or the most detailed one.
- We have filtered out the papers of the lowest quality. Here we did not want to use the quantitative analysis to evaluate the paper quality. Instead we took our decisions based on several factors important for the study.
 1. In order to be included, the paper should discuss at least one aspect of collaborative OCW authoring in details.
 2. If the paper describes an example implementation, the evaluation results have to be provided. However, the restriction is not applied to the novelty approaches not discussed in any other papers.
 3. The approaches proposed should be novel from the technological point of view rather than from pedagogical (however, we kept the papers discussing existing approaches from pedagogical point of view).

3.2 Paper Categorization and Domain Overview

In this section we describe our approach to the selected papers categorization. This process not only facilitates the analyze of the papers we have collected, but as well allows us to build the clear overview of the field itself.

We have processed the categorization in two steps. The first step of our state-of-the-art analysis was the identification of *aspects* (subdomains or major research problems) of collaborative OCW authoring. In order to do so we have researched the found surveys and essays. The analysis of the surveys revealed an absence of comprehensive general survey of the domain, however we reused the results of several surveys and essays for specific aspects, as described in [subsection 3.2.1](#). Based on the analysis of the surveys and our background knowledge of the field we have defined the first draft of the domain aspects as depicted in [Figure 3.2](#).

On the next step we classified the rest of the papers into one of aspects representing the main focus of the respective paper. We refine each aspect with additional sub-branches, thus

creating a domain mind map, described in [subsection 3.2.2](#). At the same time we have been filling in a matrix indicating the papers with regard to all aspects and technologies they address. The matrix discussed in [Figure 3.2.2](#) helps us to identify gaps and promising areas of further research.

3.2.1 Analysis of Existing Surveys

As was described above, we have started the papers categorization from selecting surveys and essays we found. Within the 131 selected papers there were in total 23 surveys, experiment studies and essays related to our topic of interest. A significant part of the surveys found is focused only on one particular aspect or technology (e.g "Social Networking" or "Collaborative adaptation authoring"). We incorporated the most important findings of the surveys into the related sections. The noticeable and influential essays and experiment studies results are discussed/cited mainly in [chapter 1](#) and [section 3.3](#).

Additionally we found several surveys aimed to answer research questions similar to the ones we defined for the literature study. However, the here presented analysis of the studies shows the absence of detailed and systematic surveys, and thereby proves the actuality of our research.

Thus, the study [58] in the related work section provides an analysis of the field considering five dimensions: (1) open hypermedia compliant systems, (2) design metaphors used to create the systems, (3) semantic characteristics, (4) collaborative characteristics and (5) adaptive and intelligence characteristics. The nature of the work however does not allow the authors to present a comprehensive analysis of existing systems (it is not a survey, but system description paper). Neither does the survey include the latest findings in the field due to its date.

Another review [12] studies the collaborative systems in the understanding of that time. This essential study provides detailed functional overview of the systems supporting collaborative work. The authors define main classes of collaborative systems and classify the studied applications accordingly. The survey however is not focused on educational materials authoring and is out-dated.

Several surveys [19, 140] discuss pedagogical and sociological aspects of e-collaboration rather than technical. Thus, in his paper [19] the author provides the classification of e-collaboration types, (e.g. synchronous versus asynchronous collaboration, continuous versus one-time contribution approaches etc.). Due to the nature of the survey, significant attention is paid to user motivation. The paper is important to understand social behavior of contributors, but addresses the e-collaboration from a different aspect than the current study.

The study [100] gives an overview of state-of-art in the collaborative authoring of OCW. Although the research questions are similar with those of the current study, the approach is different. The researchers interviewed the end users of the OERs (teachers) and summarized their experiences and the challenges they met. According to the approach, the study can not be considered systematic. It is nevertheless important from the practical point of view. The recommendations given by the authors in the conclusive part however lack technical depth and are too general to be directly incorporated by researches and developers on a technical level.

The rest of surveys and essays addresses specific aspects or technologies of collaborative OCW authoring. We have summarized all of them in [Table 3.1](#) and [Table 3.2](#), which were initially defined based on our a priori knowledge of the field, extended and polished with findings from the surveys and essays.

Concept	Surveys
Authoring paradigms	
crowdsourcing	[107, 116]
wiki	
workflow-based	
Semantic Web technologies	
semantic wiki	[112, 109]
ontologies	[31]
RSS-feed	
Social networking	
activity feed	
communication tools	[12, 125, 109]
social annotation	[118, 61, 92]
social voting	
social games	
Other	
AI	
tree-structure	
grids, LORs, mash-ups	[158]
Total	12

Table 3.1: Surveys and essays studying technologies in application to collaborative OCW authoring

Concept	Surveys
Co-creation	
Content authoring	
Metadata authoring	[111]
Assessment items authoring	
Quality assurance	
Categorization	
Personalization	[27, 31, 92]
Localization	
Reuse and re-purpose	
Search, aggregation, filtering	
Remixing	
Organization	
Social collaboration	
Negotiation	[84]
Awareness	[84]
Network building	[84, 19]
Engagement	[84, 19]
Total	6

Table 3.2: Surveys and essays addressing the issues of collaborative OCW authoring

3.2.2 Analysis of Paper Distribution

We have continued the paper categorization by building up a domain mind map and domain matrix as discussed below.

OCW Collaborative Authoring Mind Map

Our initial mind map constructed after analysis of selected surveys and essays included root branches for 11 concepts, some of which were further branched. However, as the detailed study of the selected papers has led us to significant changes, we do not illustrate the initial map here, limiting ourselves to a brief discussion of findings made already at this initial step. The final map is discussed further and presented at [Figure 3.3](#).

The mind map showed the main research trends in the field, as well as underrepresentation of some aspects, which we a priori assumed to be relevant and important to the field. The distribution of the papers between the concepts is presented at [Figure 3.2](#). There is an important assumption here, that each paper is related to just one concept (which is not true in the majority of cases). This assumption is only initially made at this point for simplifying the usage of the map.

As can be seen on the chart, more than one fifth of the selected papers are surveys essays or case studies (23 articles). We have separated these papers from the rest as they can not be assigned to any aspect. As well, we have separated research related to non-text-based content,

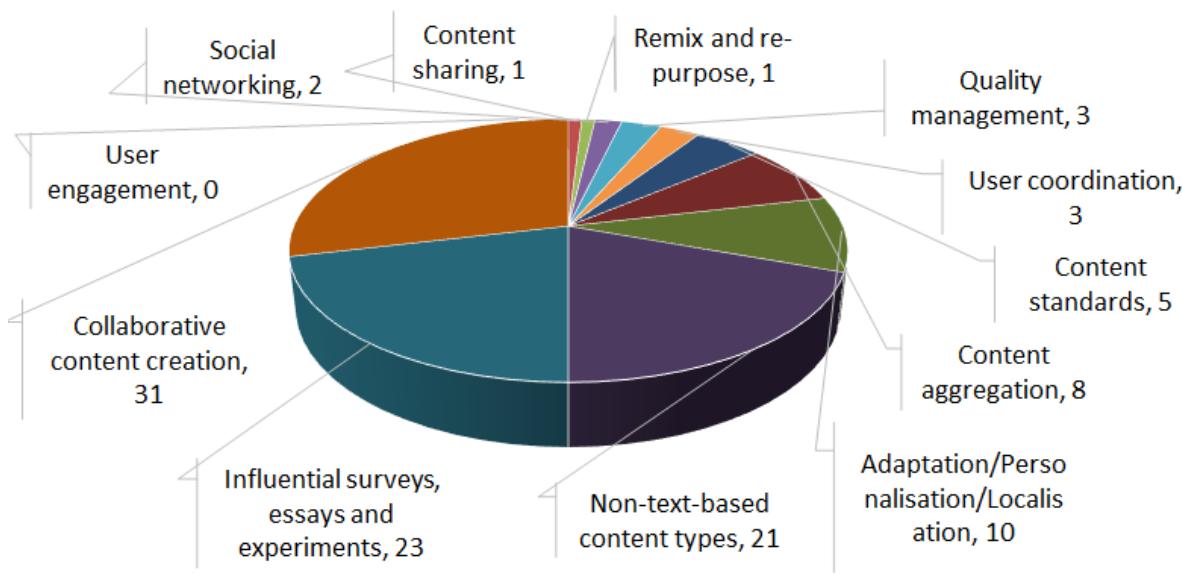


Figure 3.2: Distribution of selected papers between a priori identified concepts

such as video/audio records, graphics or three-dimensional models (21 articles). This is due to the significant difference in approaches used to deal with these types of content in comparison with text-based formats.

After separating these two kinds of papers, we observe that the majority of the remaining papers focuses on the aspect of content development using *crowdsourcing* or *wiki* paradigms (31 articles). The aspects of content adaptation/personalization and content aggregation receive decent attention as well (10 and 8 papers respectively). Less than 10 percent of the selected papers are focused on any of the other aspects and we could not identify a single paper making its main contribution in the user engagement aspect of the OCW collaborative authoring.

OCW Collaborative Authoring Matrix

During the detailed analysis of the collected papers we created a matrix indicating the distribution of papers between aspects versus employed technologies. Again, we do not consider the papers focusing on non-text-based content, as well as surveys and essays. The matrix shows the (proposed) use of technologies with regard to the OCW authoring aspects discussed in the selected papers. Each cell in the matrix includes the papers, in which the technology (indicated in the row) was applied to the aspect (indicated in the column). As articles usually cover multiple aspects and technologies, they might occur multiple times in the matrix. Due to the size of the matrix, in the paper we present a simplified version of the matrix in a [Table 3.3](#). Each cell of the table indicates the count of papers addressing or proposing the application of corresponding technology to a corresponding task. The total amount of papers considers every paper only once.

Technology	Aspect	Collaborative Authoring												Total		
		Content authoring	Metadata authoring	Assessment items auth.	Quality assurance	Categorization	Personalization	Localization	Search etc.	Remixing	Reuse and re-purpose	Negotiation	Social collaboration			
Authoring paradigms																
crowdsourcing		6		3										16		
wiki		16												16		
workflow-based		6												6		
Semantic Web technologies																
semantic wiki		5		1										5		
ontologies														6		
RSS-feed														7		
Social networking																
activity feed				1										9		
communication tools														15		
social annotation				17		2								20		
social voting							2							5		
social games								1						1		
Other																
AI		3		1										15		
tree-structure grids and mash-ups					6	4	4					5		5		
Total		30	18	5	13	9	10	5	26	4	10	28	10	3	3	65

Table 3.3: OCW collaborative authoring papers distribution

@TODO: discuss the matrix Such paper distribution shows us the lack of deep detailed research on the field. Only the major aspects are well covered, while the supporting but still crucial aspects such as user engagement of content repurpose are mostly not taken in consideration. In other words, the researches are re-inventing the wheel or polishing the existing approaches instead of focusing on solving particular crucial issues preventing the OER movement from full success.

The filling out of the matrix allowed us to build the final version of OCW mind map, presented in [Figure 3.3](#). The mind map represents the main concepts and technologies of the field. Together with terminology summarized in [section 3.3](#) it can be used for speedy entering the field for beginner researchers.

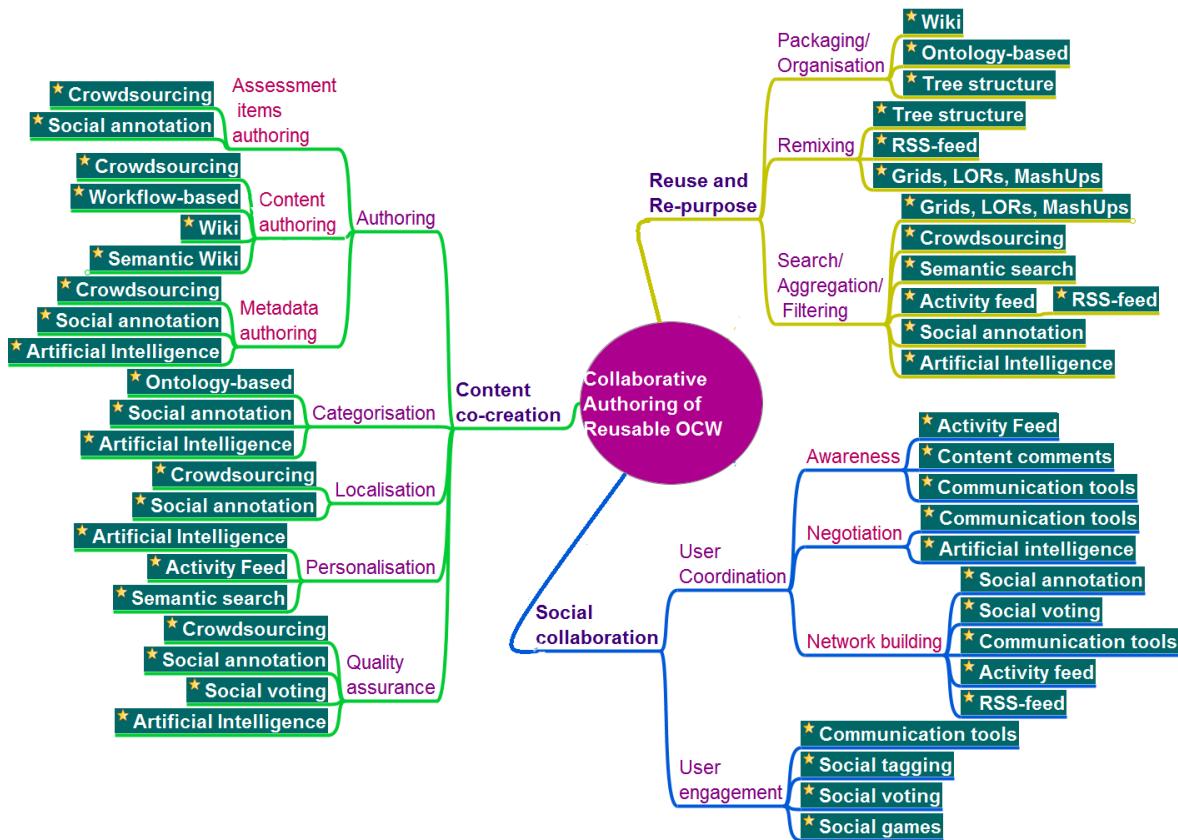


Figure 3.3: The mind map for Collaborative authoring of reusable OCW

3.3 Terminology

During our study we faced plenty of ambiguous terms and domain-based concepts that might be unclear for non-education specialists, for example, computer scientists aiming to implement an approach. In this section we focus on such terms, collecting their definitions and making

our own conclusions on their appropriate usage.

3.3.1 Learning

E-learning - a learning conducted on the Internet [36]. A wide set of applications and processes, which use available electronic media and tools to deliver vocational education and training [39].

SMLearning - type of e-learning that assumes the functions of a Social Media platform and extends its features to educational context [32].

Blended learning - the (organic) integration of online digital learning and face to face classroom learning[30].

Adaptive learning - learning that enables learners to customize their learning environments and dynamically adapts learning content to learners' learning needs [26]

Virtual attendance - the combination of synchronous and asynchronous ICT tools used to provide distance-education students with the same educational experience that conventional students receive in face-to-face (henceforth, F2F) taught classes.

3.3.2 Content

Learning object - a small, reusable digital component that can be selectively applied - alone or in combination - by computer software, learning facilitators or learners themselves, to meet individual needs for learning or performance support [137]. A later IEEE definition states: any entity digital or non-digital, which can be used, re-used, or referenced during technology-supported learning [99]. In the context of Semantic Wikis (**author?**) expands the term to include any real world objects such as people, places, organizations and events.

Learning Object Repository (LOR) - a general term for an online collection of learning objects.

OCW - a combination of learning objects, organized in a structured way according to pre-defined curriculum and serving as a unit for achieving a certain learning goal.

Adaptive OCW - OCW that is suitable to be used in adaptive learning environments.

Curriculum - structured plan that describes the educational program that is used in the educational resource [101].

Multimedia presentation - a digital slide presentation which includes diverse media objects such as graphics and videos [75].

Learning Design (LD) - a sequence of (collaborative) learning activities. It can incorporate single learner content, but also collaborative tasks such as discussion, voting, small group debate, etc. LD can be stored, re-used, customized, etc [123].

3.3.3 Authoring

Cooperation - a division of the labor among participants, into activities where each person is responsible for a portion of the problem. [124]

Collaboration - a mutual engagement of participants in a coordinated effort to solve a problem together. [123]

Communities of practice - the communities in which there exists "the sustained pursuit of shared enterprise [159].

The community-build system (CBS) - a system for content creation by a community operating on a dedicated engine (e.g. Wiki) [126]. Also, it is a system of virtual collaborations organized to provide an open resource development environment within a given community.

Wiki - a Website that allows visitors to add, remove, edit and change content [36]. A collection of web pages designed to enable anyone who accesses it to contribute or modify content using a simplified markup language ¹. A web application which allows people to add, modify, or delete content in collaboration with others ².

Semantic Wiki - a wiki that enables simple and quick collaborative text editing over the Web and Semantic Web. Semantic wiki extends a classical wiki by integrating it with the management capabilities for the formal knowledge representations ³.

Collaborative authoring - occurs in project like settings, where the project delegates authoring sub-tasks to a group of authors. This kind of authoring needs synchronization, dialogue support, and coordination of the whole project [42].

Cooperative authoring - mainly involves synchronous re-usage of authoring products, such as course materials, libraries, ontologies, etc [42].

Crowdsourcing - a problem-solving approach that outsources tasks to an undefined, often anonymous, population [76].

Workflow - a sequence of industrial, administrative, or other processes through which a piece of work passes from initiation to completion.

Workflow-based approach - an approach that uses workflow for solving a task.

Social annotation (Social tagging, folksonomy) - a system of classification derived from the practice and method of collaboratively creating and translating tags to annotate and categorize content [114].

Awareness - in the context of collaborative authoring, it is an understanding of activities of other collaborators, which provides a context for the own activity [65].

Knowledge sharing - an activity where agents - individuals, communities, or organizations - exchange their knowledge - information, skills, or expertise [79].

3.3.4 Technical Concepts and Technologies

Ontology - a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse ⁴.

Activity feed (Activity stream) - a list of recent activities performed by an individual(s), typically on a single website or on a single content piece.

Mashup - (in the domain of OCW authoring) it is a web page, or web application, that uses content from more than one source to create a single new learning object or OCW.

RSS - originally *RDF Site Summary*, but often dubbed as *Really Simple Syndication*. It is a mechanism to publish a feed of frequently updated information: blog entries, news headlines, audio, video etc. ⁵.

RSS aggregator - a tool that periodically checks for updates to the RSS feed and keeps the user informed of any changes [6].

¹ <http://wikipedia.com>

² <http://en.wikipedia.org/wiki/Wiki>

³ <https://goranzugic.wordpress.com/2010/09/09/semantic-wikis/>

⁴ [http://en.wikipedia.org/wiki/Ontology_\(information_science\)](http://en.wikipedia.org/wiki/Ontology_(information_science))

⁵ <http://en.wikipedia.org/wiki/RSS>

Grid - a collection of independently owned and administered resources which have been joined together by a software and hardware infrastructure that interacts with the resources and the users of the resources to provide coordinated dynamic resource sharing in a dependable and consistent way according to policies that have been agreed to by all parties [67].

Virtual Learning Environment (VLE) - tools that support e-learning through integrated provision of learning materials, and communication, administration, and assessment tools [35].

3.4 Educational Content Co-creation Studies

Changing reality poses new conditions and requirements for e-learning platforms. As it is mentioned in [113], “*the amount of knowledge that we deal with is much bigger than before, the interrelations between different forms of information are much more complex*”. Due to this, new approaches for learning process organization have to be created. Thus, a model in which a teacher is the monopolizing agent and the authorized representative of knowledge is no more adequate. Learners should be allowed to have their own knowledge and share it with teachers, in other words teachers and learners should be able to switch their roles during the learning process. The possibility of any user, including learners to edit the content together with modification history and discussion pages positively influences the reliability of the content. [81] In the section we summarize the known solutions and techniques to facilitate all aspects of such large-scale collaboration on open educational content.

3.4.1 Collaborative Content Authoring

During the literature study we have found three approaches of how the large-scale collaboration on educational content can be organized. The first solution is to use a complex of (interconnected) web2.0 services. For example, an approach described in [30] proposes the combination of Google Site, Google groups, microblogging and RSS technologies to collaboratively create OCW for blended learning. Course site includes curriculum, description, learning guide and teaching calendar. The updates on the site are sent to all subscribers through RSS. Site administrators can manually create a topic on the Google Groups page to discuss arguable changes before approving them. The similar approach proposed in [20] uses Google Drive to create self-assessment questions. The produced questions are then converted into digital flashcard by self-developed freely distributed Java-based program. The authors explain the choice of Google Drive by three factors: (1) Multiple people can simultaneously edit and view a document, spreadsheet, or PowerPoint, (2) it is user-friendly, and (3) just about all students are familiar with it.

The second solution is to add support of collaboration to existing non-collaborative OCW authoring tools. In the approach [63] authors integrate Joomla, Moodle and MediaWiki to support the collaborative authoring of educational content. The authors choose Joomla as the basis authoring tool due to the possibility to insert video records, Latex formulas, applets and highlighted code fragments. Moreover, during the content creation process, a publicly available knowledge base for each course is created through collaborative content annotation. The similar approach is described in [78]. The proposed web-based system allows students to share algorithmic representations with their peers. Students can create their algorithmic representation using any utility or software, and can upload those files using the system. These

representations are then available for other students to view, evaluate and discuss through discussion forums.

However, we have concluded that the most popular and promising solution is to use an integrated platform based on wiki technology.

Through the study of existing literature ([81, 62]), the list of benefits of the wiki technology when applied to educational context includes:

- users (teachers and students) are able to contribute to their private collections of various learning resources (documents, files, photos, videos, software etc.,) by uploading to given positions
- wiki-based collaborative learning environment provides a completely user-editable environment and thoroughly integrate the roles of author and reader.
- users can organize their resources in a thematic way, designing and creating different tags and share them with other users.
- authoring a wiki on a given topic produces a linked network of resources.

However, the specific of wikis arises a number of issues when applied to OCW authoring. The technical issues are indicated, for example in [156, 49]:

- All content is modifiable by any user. The instructor may want to restrict modifiability of certain pages.
- All content is open to everybody. A page, of which an important part is being developed, may be undesired to be issued, but wikis do not permit this.
- Simultaneous edits are allowed, but not successful. When simultaneous writings are being performed on a page, wikis are locked.
- Wikis can be evolved without end. An instructor may want to end the evaluation when the class ends; many wikis do not allow this.

Apart from the technical issues, there are two major obstacles preventing the usage of wikis in OCW development: (1) absence of possibility to structure the content as it is required by the learning object definition and (2) user engagement.

3.4.2 Content Categorization

The content categorization challenge is one of the key features of collaborative authoring systems. This is due to the fact that it enables (semi-)automatic combination of learning objects into OCW, strongly improves the searchability of the learning artifacts and allows to implement more advanced content quality control techniques. The conventional approach to content categorization is to choose one or more topics or concepts which the content piece is related to, usually from the pre-defined list. However the approach lacks flexibility and requires a strong pre-development planning, that is not always possible when the content is developed through collaborative iterative process.

One solution is to allow building the predefined list of interconnected concepts in the same manner as the content pieces: through the large-scale collaboration. In order to enrich such lists with the structure and semantic a number of authors propose to use ontologies. Even more

technologically advanced approach is to use the combination of social annotation techniques and artificial intelligence methods, allowing to (semi-)automatically cluster the content pieces based on the annotations. And finally, the most advanced, and due to this the most functional, approach is to combine social annotation and collaboratively developed ontologies.

This approach is implemented by a number of various authors [162, 5, 92, 106, 60, 4], within which we consider the approach described in [162] to be the most illustrative. The proposed system works with code snippets, allowing users to annotate them. Based on the annotations those content pieces are later added as individual learning objects to the courses based on the collaboratively developed curricula presented in a form of ontology. The ontology underlying the curricula can be updated and edited by the users as any other content piece. The topic matching is done using methods of artificial intelligence, such as NLP and machine learning techniques, however it can be later changed manually by any user. The byproduct of the approach is the ability to apply content recommendation and personalization approaches, as described in subsection 3.4.3

3.4.3 Content Personalization and Recommendation

As it was discussed above, the content personalization and recommendation approaches are based on the content categorization. Here by content recommendation systems we mean the ability of the platform to recommend to a learner a set of learning materials based on her background, interests or progress in other courses. The content personalization (or adaptation) system goes one step further and after selecting a set of suitable material organizes it in a complete course. Often the content personalization algorithms are able to adjust the initially provided course according to the learner performance.

In the scope of the thesis research we were not focused on the algorithms and approaches for the content recommendation and personalization. Instead, we were interested in what steps should be done on a stage of content development in order to ensure that the content can be (re)used in systems supporting content recommendation and personalization features. Based on the research done [24, 60, 29], we consider that a list of annotations attached to each learning object in combination with fine-grained content structuring make the OCW suitable for use in those systems.

3.4.4 Content Localization

The content localization issue is one of the less-discussed issues in state-of-art research. This is due to the fact, that the issue can not be effectively solved without solving related issues, such as community building, content structuring and synchronization, content quality assurance and other.

In the approach described in [93] the authors implement the possibility to translate the content by collaborators. Each new learning object created by the translation has its own metadata and can be further developed. However, neither the possibility to update the translation if the original object was changed nor vice versa are implemented. The synchronization is especially important, if versions in different languages are developed simultaneously, what often happens in open large-scale collaborative platforms. In the scope of our research we provide our own solution to this issue (see section 4.3).

3.4.5 Content Quality Assurance

Since the emergence of the OER movement, the quality of the open educational resources was a widely discussed issue. A number of authors were working on defining and summarizing the key factors for the OER's quality. For example, the following criteria are listed in [83, 93]:

- They should be relevant to the learner and thus easily modified to fit the learner's needs.
- They should be of good quality and contain no factual errors.
- They should disclose their point of view and in the case of science be free from bias.
- They should not have hidden costs or prohibiting limitations on use.
- A good learning resource should also be able to "travel well", to be easily translated and recontextualised

In [3] the authors formally define and discuss objective metrics which can be used for measuring an OER quality. They consider the following ten factors to determine the OER quality:

- Legal Reusability
- Multilinguality
- Format re-purposing
- Recency
- Sustainability
- Availability
- Learnability by Self-assessment
- Learnability by Examples and Illustrations
- Community Involvement
- Discoverability

The list of factors correlates well with the challenges of the collaborative OCW authoring we defined. Thus, solving the challenges will guarantee high quality of the resulting OERs.

3.5 Educational Content Reuse and Re-purpose Studies

The importance of creating reusable and re-purposable e-learning objects is widely accepted by the e-learning community [41]. However, most of the works address the learning object reuse problem rather by means of semantic meta-data annotations, content tagging and packaging than by creating richly structured, reusable learning objects from the ground. The importance of creating learning objects already with reuse in mind was, for example, stated by [113]: *Content ... should be represented not as an object of study but rather as necessary elements towards a series of objectives that will be discovered in the course of various tests.*

There are only few approaches for the direct collaborative authoring of reusable content, such as, for example, learning examples creation [91] or semantic structuring and annotation of video fragments [13]. This is due to the fact that the state-of-art collaborative authoring technologies do not deal with the (semi-)structured content.

As it was previously discussed in section 2.3, conventional wiki technology has not been yet fully applied to the structured content development. However, a couple of researches make steps in this direction. Thus, [64] introduces the concept of wiki templates that allow end-users to define the structure and appearance of a wiki page in order to facilitate the authoring of structured wiki pages. Similarly the Hybrid Wiki approach [102] aims to solve the problem of using (semi-)structured data in wikis by means of page attributes.

Another approach to combine wiki technology with structured representation are semantic wikis [131]. Keeping the ability of traditional Wikis to collaborative content building, Semantic Wikis allow to add meaningful relationships between resources. There are two types of semantic wikis. Semantic text wikis, such as Semantic MediaWiki [90] or KiWi [132] are based on semantic annotations of the textual content. Semantic data wikis, such as OntoWiki [69], are based on the RDF data model in the first place. Both types of semantic wikis, however, suffer from two disadvantages. Firstly, their *performance and scalability* is restricted by current triple store technology, which is still an order of magnitude slower when compared with relational data management, which is regularly confirmed by SPARQL benchmarks such as BSBM [18]. Secondly, semantic wikis are generic tools, which are not particularly adapted for a certain domain thus substantially increase the usage complexity for users. The latter problem was partially addressed by OntoWiki components such as Erfurt API, RDFauthor and Semantic Pingback, which evolved OntoWiki into a framework for Web Application development [69].

We should also mention the *Learning Objects Repositories* (LORs), that allow to produce structured reusable content. The first of them, *Connexions*⁶, presents the learning material as a combination of paragraphs, each of them could be easily edited or deleted. However, this structuring is done more for comfortable editing and does not have any functional benefits: the paragraphs cannot be reused or annotated independently. Thus, Connexions presents just an improved user interface for wiki-based system. The second example of structuring, is *LeMill*⁷. LeMill provides a way of collaborative editing of the presentations by implementing presentations as a group of images which can be edited collaboratively. However, to edit a slide, a user has to replace it with another one. Also, it is impossible to have several subgroups of the slides within a presentation. The search through the slides (not presentations) is also not implemented. Thus, slides cannot be manipulated as independent learning objects.

3.6 Social Collaboration on Educational Content Studies

In order to facilitate the social collaboration, two main features have to be enabled: user coordination and user engagement. In the section we discuss requirements for the features' implementation found in the selected papers.

⁶ <http://cnx.org/>

⁷ <http://lemill.net/>

3.6.1 User Coordination

One critical issue in collaborative authoring is the user coordination. In order the collaboration was efficient, all authors including novices must *be aware* of the state of the production process, current goals and emergent issues. State-of-art researchers [110] distinguish following forms of user awareness:

1. *Personal awareness* relates to authors' knowledge of what they have done in the collaborative authoring process, allowing users to review their past work and plan future actions based on that.
2. *Informal awareness* is related to the provision of information about who is in the collaborative work environment and what other authors have done during particular periods of time.
3. *Group awareness* relates to authors's shared awareness of what has to be done.

To enable support of the *personal awareness*, all user actions should be recorded and presented for example as user's activity feed in her profile. The activity feed should summarize information about what a certain author has done, which objects she has created or updated, when it has happened. It should as well allow to go directly to the edited object.

To support *informal awareness*, in addition to the personal activity feed, an object activity feed should summarize actions done to the object by all users. As well, user comments on the process should be stored and attached to the object in order to prevent discussions over the same question twice. Therefore, (novice) authors can easily find all annotations and comments that relate to a particular object and know what other authors have been doing recently.

To support the *group awareness*, some sort of an authoring plan should be available for all (prospective) collaborators. The planning can be done explicitly in a form of a separate document attached to the content object. As well, implicit planning can be chosen, when the scope of work is clear from the content structure. In case of implicit planning authors should be aware of empty and filled out (agreed) parts of the content, as well as about number of collaborators working on each part.

3.6.2 User Engagement

The issue of user motivation raises from the fact that individuals contribution is encouraged by the certain collaborative authoring platform popularity [25]. This makes it difficult to motivate users to contribute within a newly developed platform. Collaboration on projects like Wikipedia articles and open-source software is made easier by the fact that the goal is relatively well defined. Thus, the solution is to turn the development of each content piece into a mini-project. A number of authors define factors helping to proceed this turn and thus to increase authors participation [25, 55]:

- Easy first step towards participation that is legitimate. A newcomer should have something easy and satisfying to accomplish. From this initial task, there are a series of gradually more challenging tasks available. Users should be informed about what is expected to accomplish in each individual content piece and starting points for editing should be indicated.

- Social support for participation and learning. The presence of other collaborators who can potentially support her should be clear for a newcomer.
- Clear major objectives, understood by all its users gives collaborators feeling of the 'collaborative ownership' of the content. Working on a project the user cares about can motivate her to persevere when difficulties are encountered.
- Users are highly motivated if there exist definitive times for different 'stages' of use and definite end point.
- Clear indication of who can edit, which parts they can edit, acceptable and unacceptable use as well as clear navigation structure help users to feel confident while contributing.

Within other methods of user engagement the most promising are the ones based on gamification. For example, digital badges, rewarding users for certain achievements, can be shown not only inside the platform, but as well in the user's profile pages in popular social networks. This fact makes the successful work inside the platform more beneficial for a user, as she can use the badges earned for proving her skills for further online and offline projects.

3.7 Conclusions

We have built our state-of-art study around four major research questions:

1. *What are the main challenging tasks in collaborative authoring of reusable OCW?*
2. *Which technologies are being used to solve these tasks?*
3. *How well covered are the challenging tasks in the scientific literature?*
4. *What are the main gaps in state-of-art research?*

We have answered the first question by researching existing surveys. According to our study, the main challenging tasks of collaborative authoring of OCW can be divided in three dimensions: (1)content co-creation, including metadata and assessment items authoring, classification, localization, personalization and quality assurance; (2) content reuse and re-purpose, including content organization, remixing, search, aggregation and filtering; (3) social collaboration, including user coordination and user engagement.

The technologies which are currently commonly being used for solving the tasks are Web 2.0, artificial intelligence and Semantic Web technologies. We summarize the answers for these two questions in mind map presented in [Figure 3.3](#).

Having analyzed the paper distribution, we came to a conclusion that the issues and technologies of OCW collaborative authoring are covered in the literature with not enough depth, that means there is not enough *detailed* and *focused* studies on specific aspects and technologies.

The matrix we have filled out (cf. [Table 3.3](#)) gives additional details on the most and the least researched topics in the area. Especially, we consider the following topics not to be sufficiently addressed:

Content localization issue Although it is believed that the most important benefits OpenCourseWare would bring to the developing countries, the translation and localization approaches are almost ignored in the published research. This gap is from one perspective well-known, but from another perspective it is hard to solve. The power of crowdsourcing can be used to translate the content with reasonable quality, but this rises the problems of content synchronization and cultural barriers. The content synchronization is partly addressed by our proposed concepts (see [section 4.3](#)), while the actual localization which considers the cultural differences and barriers still needs to be additionally researched.

Content remixing The issue of content remixing is also often ignored in the state-of-art research. The educational content has to be re-designed from year to year, new topics often should be included, as well as parts of different courses often need to be combined. Although several approaches exist to address the challenge using mash-ups and LORs, these solutions do not involve content versioning and crowdsourcing. For example, an interesting approach proposed in [11] uses RSS-feed and machine-learning algorithms for filtering the content from the feed and assembling it according to the user preferences. A mash-up created in this way can be further edited by collaborators. In [section 4.2](#) we propose our own approach, assuming the content is authored in a wiki-way, has a tree-structure and due to that can be easily re-mixed. An important novelty here is the comprehensive algorithm for versioning the content in a tree-structure.

Network building and User engagement These two aspects are interconnected and are important for the success of any collaborative systems. Although several approaches are available from technological side (like digital badges or social gaming), only few approaches implement and evaluate them. These aspects lack especially the evaluation and comparison of the existing methods. In our approach, discussed in details in [section 4.1](#), we propose to use social networking functions to build highly-connected user groups, sharing the ownership of the content. As well, when developing the concept, we took into consideration the recommendations summarized in [subsection 3.6.2](#).

CHAPTER 4

Collaboration on Multilingual Structured Educational Content

The original idea of the web was that it should be a collaborative space where you can communicate through sharing information.

Tim Berners-Lee

I think in general it's clear that most bad things come from misunderstanding(...)

Tim Berners-Lee

While nowadays there is a plethora of Learning Content Management Systems (LCMS), the collaborative, community-based creation of rich e-learning content is still not sufficiently well supported.

When speaking about collaborative authoring, the crowdsourcing techniques and wiki paradigm are proven to be an effective solution, allowing continuous iterative increasing of the content quality while keeping the costs moderately low. Such techniques suppose the collaborative platform supports content versionning, branching and merging, that prevents unwanted changes while allows uncontrolled list of collaborators to contribute. The *wiki* paradigm allows every user to refine the published content by facilitating small changes and providing easy-to-use markup language. The best known example of successful crowdsourcing platform is Github, however there are adaptations of the crowdsourcing techniques to the domains other than classical software development, such as Wikipedia or OpenStreetMaps.

However, the application of the crowdsourcing techniques to the OCW production has not yet been done. This is due to the specific of the content type the concepts have to deal with. The paradigm of *Open Educational Resources* supposes the creation and sharing of reusable, re-purposable learning objects, annotated by standardized metadata. Thus, the content structuring is required, as discussed in [subsection 1.4.1](#). However, dealing with the structure when applying crowdsourcing techniques is challenging.

In the current chapter we present and discuss our *CrowdLearn concept*, aiming at application of crowdsourcing techniques to the structured content and supportive *WikiApp* data model we have developed to implement it. As well, we present the CoSMEC concept, which enables the support of multilingual content for this complex data model.

4.1 CrowdLearn Concept

In order to develop the concept which facilitates large-scale collaboration on OERs, we have first summarized the major challenges and possible solutions for that. The results of this work have been already discussed in the previous chapter. We have transformed the findings from our related work study in a [Table 4.1](#), which gives a brief overview on the challenges and approaches we have found to be the most promising.

Aspect	Approach chosen	Remarks
Content co-creation		
Content authoring	Crowdsourcing, wiki	
Metadata authoring	Crowdsourcing, wiki, standard compliance	
Assessmnt. items auth.	Crowdsourcing, wiki	
Quality assurance	Social networking, crowdsourcing, wiki, semantic structuring	Content objects, metadata, assessment items and styles are reusable and fully versioned Synergistic effect
Categorization	Crowdsourcing (social annotation)	
Personalization	Semantic structuring, crowdsourcing (social annotation)	
Localization	Crowdsourcing, wiki	Using CoSMEC concept. See section 4.3 for more details
Reuse and re-purpose		
Search, filtering, aggregation.	Standard compliance, crowdsourcing (social annotation), semantic structuring	
Remixing	Semantic structuring, standard compliance	Using WikiApp data model. See section 4.2 for more details
Organization	Semantic structuring	Using WikiApp data model. See section 4.2 for more details
Social Collaboration		
Negotiation	Social Networking	
Awareness	Social Networking	
Network building	Social Networking	
Engagement	Social Networking	

Table 4.1: Chosen approaches for solving OCW collaborative authoring challenges

As it can be seen from the table, most of the challenging tasks can be covered by five strategical approaches, as illustrated by [Figure 4.1](#). Moreover, the combination of the approaches gives a synergistic effect. However, the integration of the approaches requires additional concepts to be developed. We have called the resulting complex solution *CrowdLearn concept*. In the section we discuss the Crowdlearn concept and the synergistic effects it gives, while missing approaches we had to develop are discussed in the next sections.

The Crowdlearn concept aims to provide a framework for efficient collaborative authoring of reusable educational content. We see the CrowdLearn concept as an application of crowdsourcing techniques to the e-learning content creation, re-purpose and reuse. The sources of our inspiration were:

- Wiki paradigm and its implementation Wikipedia, the project which proved the effectiveness of applying crowdsourcing techniques to the textual content

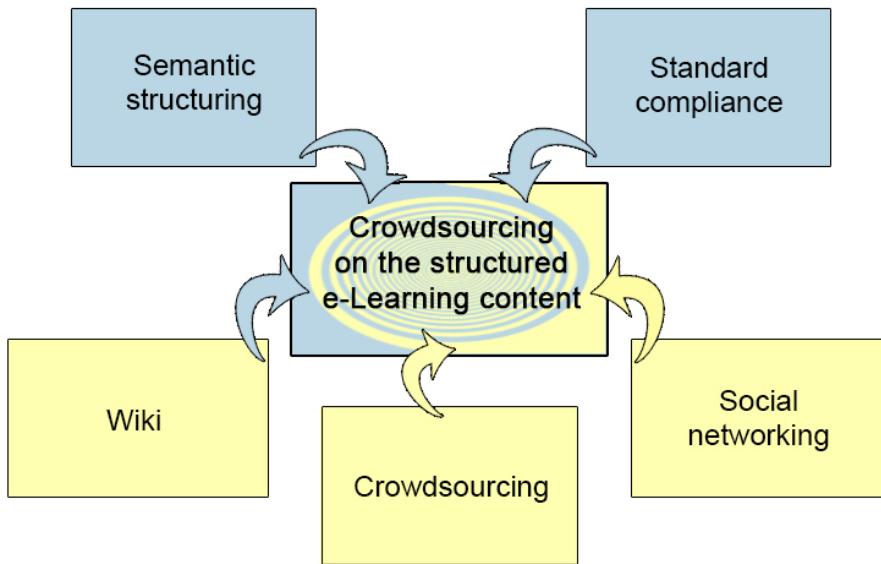


Figure 4.1: CrowdLearn concept.

- Content versioning systems and their online implementations (e.g. Github), which proved the reliability and effectiveness of opening large-scale projects for crowdsourcing with uncontrolled auditory.

In order the application to be effective, the concept considers the specific of the educational domain by combining five strategies, presented in [Figure 4.1](#). Below we discuss the five main components and their interrelations.

Crowdsourcing There are already vast amounts of amateur and expert users which are collaborating and contributing on the Social Web. Harnessing the power of such crowds can significantly enhance and widen the distribution of e-learning content. Crowd-sourcing as a distributed problem-solving and production model is defined to address this aspect of collective intelligence [77]. CrowdLearn as its main innovation combines the crowd-sourcing techniques with the creation of highly-structured e-learning content. E-learning material when combined with crowd-sourcing and collaborative social approaches can help to cultivate innovation by collecting and expressing (contradicting) individual's ideas. As Paulo Freire wrote in his 1968 book *Pedagogy of the Oppressed*, 'Education must begin with the solution of the teacher-student contradiction, by reconciling the poles of the contradiction so that both are simultaneously teachers and students...'. Therefore, crowd-sourcing in the domain of educational material not only increases the amount of e-learning content but also improves the quality of the content. Our concept assumes application of crowdsourcing techniques to all kinds of the content and its metadata, including self-assessment items. This, together with social networking and semantic structuring, completely solves the challenges from the *content co-creation* aspect group. The content quality assurance is then reached through facilitation of small contributions from the crowd and ability to discuss the individual learning artifacts, such as an individual slide or a question. Social annotation of the content pieces serves as a basis for content categorization and customization. An important task our concept allows to be done by the crowd is content

localization. Here the content structuring plays a crucial role in improving the quality of the translation, due to the possibility to translate and edit each learning artifact individually.

Wiki The wiki paradigm supposes the crowdsourcing of the content supported by facilitation of small contributions, formatting and version control. To be able to deal with structured content, the wiki paradigm needs a more comprehensive data model. Our CrowdLearn concept assumes the content to be presented according to our previously developed WikiApp data model, described in details in [section 4.2](#). The WikiApp data model is a refinement of the traditional entity-relationship data model. It adds some additional formalisms in order to make users as well as ownership, part-of and based-on relationships first-class citizens of the data model. A set of content objects connected by *part-of* relations can be arranged and manipulated in exactly the same manner, as an individual non-structured object. The model natively supports versioning and structuring of the different content objects.

Semantic structuring Instead of dealing with large learning objects (often whole presentations or tests), we decompose them into fine-grained learning artifacts. Thus, rather than a large presentation, user will be able to edit, discuss and reuse individual slides; instead of a whole test she/he will be able to work on the level of individual questions. This concept efficiently facilitates the reuse and re-purpose of the learning objects. Semantic structuring facilitates application of content personalization mechanisms, allowing recommendation systems to work with a finer tuned setup. Semantic structuring together with standard compliance and social annotation facilitates content publishing, aggregation and filtering due to the ability of algorithms to work with finer grained content. The semantic structuring of the content is implemented via using WikiApp data model discussed in the paragraph above.

Social networking The theoretical foundations for e-Learning 2.0 are drawn from social constructivism [157]. It is assumed that students learn as they work together to understand their experiences and create meaning. In this view, teachers are knowers who craft a curriculum to support a self-directed, collaborative search and discussion for meanings. Supporting social networking activities in CrowdLearn enables students to proactively interact with each other to acquire knowledge. With the CrowdLearn concept we address the following social networking activities:

- Users can follow individual learning objects as well as other users activities to receive notification messages about their updates.
- Users can discuss the content of learning objects in a forum-like manner.
- Users can share the learning objects within their social network websites such as Facebook, Google Plus, LinkedIn, etc.
- Users can rate the available questions in terms of their difficulty.

Besides increasing of the learning process quality, social activities improve the quality of the created learning material. Even when answering a quiz, users can contribute by analyzing the quality of the questions and making suggestions of how to improve them. Thus, the knowledge is being created not only explicitly by contributors, but also implicitly through discussions, answering the questions of assessment tests, or in other words through native learning activities.

Standard-compliance In order the CrowdLearn concept to be efficient in the real life, it has to support the content migration between different Learning Management Systems. However, often content migration is not completely adequate and can thus result in loss of valuable content, metadata or structure. Even if the transfer is possible, moving the content between systems can be more costly than just redeveloping that course in the new system. The strategy to overcome this challenge is the standard-compliance of both LMS and content. During our research on the available content and metadata formats we have decided to adopt the *SCORM standard* [1] and practical recommendations [2], expanding the standard to support the collaborative model of content creation (see item 5.1.2). @TODO: add about Linked Data

4.2 WikiApp Data Model

The implementation of the CrowdLearn concept requires a data model, that naturally supports required operations. However, to our knowledge, such model is not known to be described formulated at least in the open scientific research. To fulfill the gap we have developed and formalize an innovative data model called WikiApp.

The WikiApp data model is a refinement of the traditional entity-relationship data model. It adds some additional formalisms in order to make users as well as ownership, part-of and based-on relationships first-class citizens of the data model. A set of content objects connected by *part-of* relations can be arranged and manipulated in exactly the same manner, as an individual non-structured object. The model natively supports versioning and structuring of the different content objects.

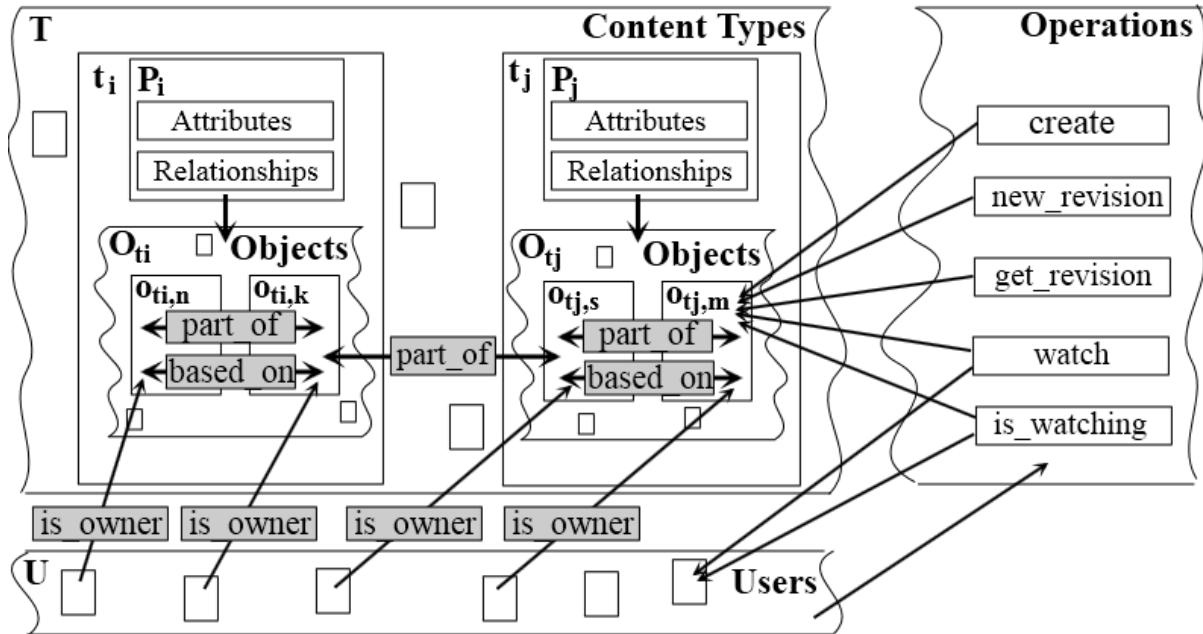


Figure 4.2: Conceptual view of the WikiApp data model.

We illustrate the WikiApp model in Figure 4.2 and formally define it as follows:

Definition 9 (WikiApp data model) *The WikiApp data model \mathcal{WA} can be formally described by a triple $\mathcal{M} = (U, T, O)$ with:*

- U - a set of users.
- T - a set of content types with associated property types P_t having this content type as their domain.
- $O = \{O_{t \in T}\}$ with O_t being sets of content objects for each content type $t \in T$.

Each O_t consists of content objects $o_{t,i}$ with $i \in I_T$ being a suitable identifier set for the content objects in O_t . Each $o_{t,i}$ comprises a set of properties $P_{t,i} = Attr_{t,i} \cup Rel_{t,i}$. $Attr_{t,i}$ is a set of literal, possibly typed attributes, and $Rel_{t,i}$ is a set of relationships with other content objects; The only necessary attribute for all content objects is $c_{t,i}$, which contains the creation timestamp of the object $o_{t,i}$. $Rel_{t,i}$ can particularly include the following designated relationships to related objects:

- $part_{t,i} \subset O$ refers to set of the content objects contained in $o_{t,i}$;
- $baset_{,i} \in O_t$ refers to base content object from which the object $o_{t,i}$ was derived;
- $user_{t,i} \in U$ refers to a user being the owner of the $o_{t,i}$;

The WikiApp model assumes that all content objects are versioned using the timestamp $c_{t,i}$ and the base content object relation $b_{t,i}$. In the spirit of the wiki paradigm, there is no deletion or updating of existing, versioned content objects. Instead new revisions of the content objects are created and linked to their base objects via the *base-content-object* relation. All operations have to be performed by a specific user and the newly created content objects will have this user being associated as their owner. In practice, however, usually only a subset of the content objects are required to be versioned. For auxiliary content (such as user profiles, preferences etc.) it is usually sufficient to omit a base content object relation. For reasons of simplicity of the presentation and space restrictions we have omitted a separate consideration of such content here. However, this is in fact just a special case of our general model, where the base content object relation $b_{t,i}$ is empty for a subset of the content objects.

The model is compatible with *both* the relational data model and the *Resource Description Framework* (RDF) data model (i.e. it is straightforward to map it to each one of these). When implemented as a relational data, content types correspond to tables and content objects to rows in these tables. Functional attributes and relationships as well as the *owner* and *base-content-object* relationships can be modeled as columns (the latter three representing foreign-key relationships) in these tables. The implementation of the WikiApp model in RDF is slightly more straightforward: content types resemble classes and content objects instances of these classes. Attributes and relationships can be attached to the classes via `rdfs:domain` and `rdfs:range` definitions and directly used as properties of the respective instances. For reasons of scalability we expect the WikiApp data model to be mainly used with relational backends.

Watching the users, as well as *following* the learning objects operations are natively supported by the model. This allows users to receive the information about changes of the followed content object or new objects created by the watched user. Also, these operations allow to easily find the followed object or user.

Example 1 (SlideWiki data model) For our SlideWiki example application (whose implementation is explained in detail in chapter 5) the data model consists of individual slides (consisting mainly of HTML snippets and some meta-data), decks (being ordered sequences of slides and sub-decks), media assets (which are used within slides) as well as themes (which are associated as default styles with decks and users):

- $T = \{deck, slide, media, theme\}$
- $Attr_{deck} = \{title \rightarrow text, abstract \rightarrow text, license \rightarrow \{CC - BY, CC - BY - SA\}\}, Rel_{deck} = \{default_theme \rightarrow theme\}, Part_{deck} = \{deck_content \rightarrow deck \cup slide\}$
- $Attr_{slide} = \{content \rightarrow text, speaker_note \rightarrow text, license \rightarrow \{CC - BY, CC - BY - SA\}\}, Rel_{slide} = \{uses \rightarrow media\}, Part_{slide} = \{\}$
- $Attr_{media} = \{type \rightarrow \{image, video, audio\}, uri \rightarrow string, license \rightarrow \{CC - BY, CC - BY - SA\}\}, Rel_{media} = \{\}, Part_{media} = \{\}$
- $Attr_{theme} = \{title \rightarrow string, css_definition \rightarrow text\}, Rel_{theme} = \{\}, Part_{theme} = \{\}$

After we introduced the WikiApp data model, we now describe the main operations on the data model. In the spirit of the Wiki paradigm, there is no deletion or updating of existing, versioned content objects. Instead new revisions of content objects are created and linked to their base objects via the $b_{t,i}$ relation.

Definition 10 (WikiApp operations) *Five base operations are defined on the WikiApp data model:*

- $create(u, t, p) : U \times T \times P_t \rightarrow O_t$ creates a new content object of type t with the owner u and properties p .
- $newRevision(u, t, i, p) : U \times T \times I_T \times P_t \rightarrow O_t$ creates a copy of an existing content object $o_{t,i}$ of type t potentially with a new owner u and overriding existing properties with p .
- $getRevision(t, i) : T \times I_T \rightarrow O_t \cup \text{false}$ returns the existing content object $o_{t,i}$ of type t including all its properties or false in case a content object of type t with identifier i does not exist.
- $isWatching(u, t, i) : U \times T \times I_T \rightarrow \{\text{true}, \text{false}\}$ returns true if the user u is watching the content object of type t with identifier i or false otherwise. Following users is a special case, where the content object type is set to user.
- $watch(u, t, i) : U \times T \times I_T \rightarrow \{\text{true}, \text{false}\}$ toggles user u watching the content object of type t with identifier i and returns the new watch status.

All operations have to be performed by a specific user and the newly created content objects will have this user being associated as their owner. In addition, when a new revision of an existing content object is created and the original content object is indicated to be part of another content object (by the distinguished part-of relations) the creation of a new revision of the containing content object has to be triggered as well. In our Example 1, this is, for example, triggered when a user creates a new revision of a slide being part of a deck. If the user is not the owner of the containing deck, a new deck revision is automatically created, so as to not implicitly modify other users' decks (cf. subsection 5.1.1).

4.3 Co-evolution of Structured Multilingual Educational Content (CoSMEC)

In order to enable the support of multilingual educational content we have expanded the CrowdLearn concept to support multilinguality. The proposed CoSMEC concept combines and expands several existing paradigms. The overview of the concept is presented in [Figure 4.3](#).

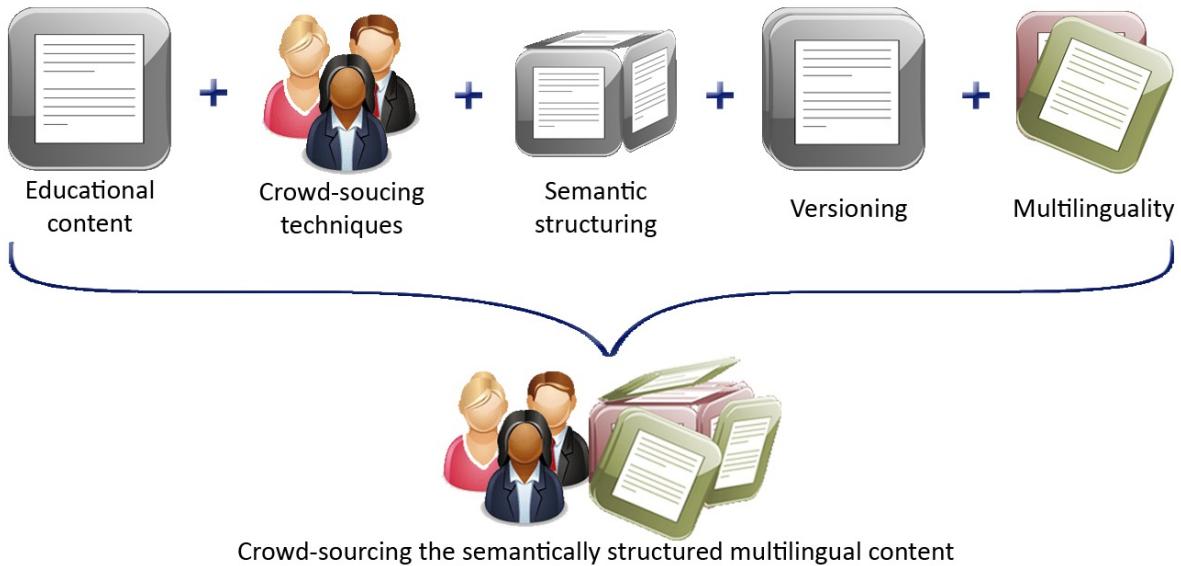


Figure 4.3: Overview of the CoSMEC concept

To enable collaborative authoring of semantically structured educational content, we use our CrowdLearn concept, as described in detail in [chapter 4](#). In order to enable the support of multilingual content, CoSMEC allows the content versions to be presented in different languages by adding the *translatedInto* and its inverse *translatedFrom* relations to the WikiApp data model. CoSMEC does not deal with the objects being originally created in different languages. Instead, our concept requires the presence of the *source* content object, as well as its *translations*. Enforcing this requirement allows us to:

- distinguish between users' authoring and translating contributions to the content,
- present the list of original content authors in all translations,
- propagate changes in order to synchronize the content of translations with the source.

The CoSMEC concept introduces the paradigm of *co-evolution* of multilingual content, that means the ability to update a translation to the current state of the source object and vice versa. However, the requirement above only allows users to update the content of a translation according to an original version, but not contrariwise. To overcome this limitation, we enable the translation of an object back to the original language, thus creating a revision of the source object. This *back-translation* requires a mechanism of merging the revisions, as we do not want to overwrite the original content with a repeatedly-translated one. Thus, in order to function efficiently, the mechanism of *co-evolution* requires three operations to be defined:

- initial translation (see [Figure 4.4](#), schema 1);

- synchronization of the translation with the source (see [Figure 4.4](#), schema 2);
- merging the revisions (see [Figure 4.4](#), schema 3)

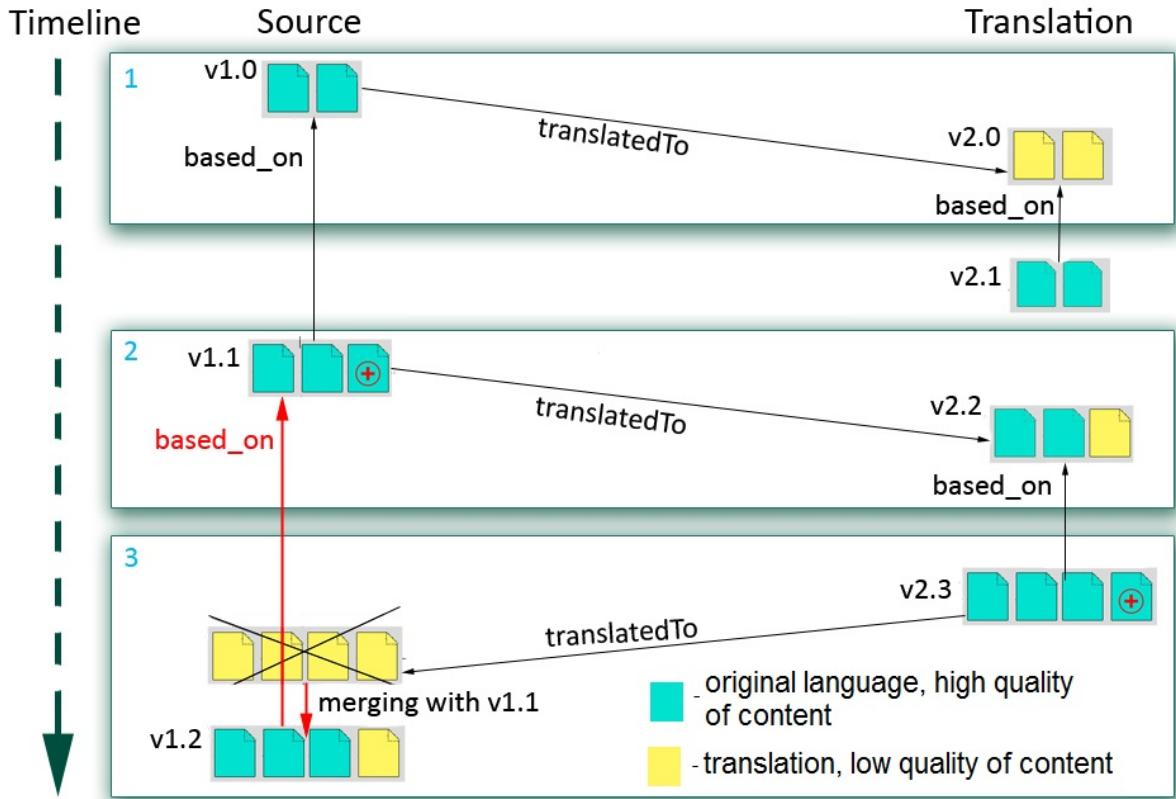


Figure 4.4: CoSMEC mechanism of the multilingual content co-evolution: 1 - initial translation, 2 - partial translation to synchronize the content of translation with the source, 3 - merging the revisions to avoid the repeatedly translated content

The issues of the mechanism implementation we have discovered and solved during example implementation are presented in [subsection 5.1.3](#), while the implementation in pseudocode is listed and discussed below.

@TODO: describe

```
function translate_slide(slide, target_language):
    if translation does not exist:
        create_empty_slide(new_slide)
        translate slide content(new_content) in target_language
        set new_slide author to the current user
        set new_slide language to the target_language
        set new_slide content to new_content
        set slide translatedTo to new_slide id
        return new_slide
    else:
```

```

        return existing slide translation

function initial_translation(source_deck , language):
    create_empty_deck(new_deck)
    for each slide of a source_deck:
        new_slide = translate_slide(slide , language)
        add new_slide to new_deck
    return new_deck

function update_translation(source_deck , translated_deck , language):
    create new revision of translated_deck
    for each added_slide of a source_deck:
        new_slide = translate_slide(added_slide , language)
        add new_slide to new revision of translated_deck
    return new revision of translated_deck

function find_source_deck(translated_deck):
    first_revision = get_first_revision of translated_deck
    find source_deck where translatedTo = first_revision
    return source_deck

function back_translation(translated_deck , language):
    source_deck = find_source_deck(translated_deck)
    source_deck = find last_revision of source deck
    for each slide of translated_deck:
        if slide does not exist in source_deck:
            add (translate_slide(slide , language)) to source_deck
    return source_deck

function translate_deck(source_deck , language):
    if translated_version of source_deck in language exists:
        if source_deck was translated from language:
            return back_translation(source_deck , language)
        else:
            return update_translation(source_deck , translated_version , language)
    else:
        return initial_translation(source_deck , language)

```

4.4 Conclusions

@TODO: create a picture of complex solution and aspects In the chapter we have presented the concepts we have developed in order to deal with the major challenges preventing in our opinion the OER movement to succeed. We have provided a complex solution supporting the whole life-cycle of a common OER, including reuse, re-purpose and localization of its content. In particular we aimed at finding solutions for the following issues:

Content production costs. Crowdsourcing on the content pieces, when supported by domain experts, enables the iterative production of quality content with lower resource production. As a byproduct, crowdsourcing involves not only specialists, but prospective learners as well, partly solving the learner engagement issue.

Content reusability and interoperability. This issue is addressed by WikiApp data model allowing collaboration on semantically structured content. The content structuring facilitates reuse and re-purpose of the content pieces, as well as their migration between educational platforms.

Multilinguality. Our developed CoSMEC concept allows not only to translated the content (pieces) into a number of languages, but to synchronize those translations and semi-automatically aggregate changes between them.

Moreover, being combined together the developed concepts give a synergistic effect on a quality of the produced resources. According to the [3], the content quality is affected by the following factors, and our solution positively influences each of them:

- *Legal Reusability.* The CrowdLearn concept assumes the content to be in shared ownership of all its contributors. At the same time, the Wiki paradigm allows every user to contribute, preventing vandalism by content versionning control.
- *Multilinguality.* The CoSMEC concept allows effective mechanism to manage the synchronization of the content presented in different languages, while WikiApp data model facilitates the translations by structuring the content into fine-grained pieces.
- *Format re-purposing.* The standard compliance component of the CrowdLearn concept together with content structuring paradigm ensures the content to be standard-compliant with
- *Recency.* Easily evolving content due to the power of crowd, reusability of content pieces, full version control and facilitation of minor edits.
- *Sustainability.* The community engagement ensures that useful and popular content is regularly updated and improved.
- *Learnability.* The WikiApp data model allows to deal with different types of OERs, including media objects and self-assessment items.
- *Community Involvement.* Social Networking features, enabled for each content piece, dramatically increase social interaction between the users.
- *Discoverability.* Crowdsourcing on resource metadata, enabled by CrowdLearn, dramatically increases its up-to-date completeness.

In order to evaluate the concepts in real life, we have implemented a web-based platform for collaborative OCW authoring. The details of the platform implementation and evaluation are discussed in the chapters below.

CHAPTER 5

Implementation and Evaluation

We implement and evaluate the developed concepts with SlideWiki¹ – a web-based OCW authoring and crowd-learning platform. Using SlideWiki large communities of teachers, lecturers and academics are empowered to create sophisticated educational content in a truly collaborative way. For newly emerging research fields, for example, a collaboration facility such as SlideWiki allows to disseminate content and educate PhD students and peer-researchers more rapidly, since the burden of creating and structuring the new field can be distributed among a large community. Specialists for individual aspects of the new field can focus on creating educational content in their particular area of expertise and still this content can be easily integrated with other content, re-structured and re-purposed.

With SlideWiki we have a holistic end-to-end view on the collaboration and engagement with OpenCourseWare targeting all relevant stakeholders, in particular:

- Primary and secondary education teachers who can use the platform to share, repurpose and present content.
- School students, who can access and comment the content and test their learning progress using the self-assessment functions.
- Parents, who can observe the learning progress of their kids.
- Professional and vocational trainers, who can easily collaborate in the courseware content creation.
- Professors and higher-education educators, who can use the platform for sharing, translating and maintaining their high-quality courseware content and building their reputation as OCW author.
- University students, who can obtain personalized and context specific learning resources.
- Companies, which can use the SlideWiki technology internally to train employees and partners.
- MOOC platforms, which can dramatically lower the content creation and maintenance costs.

¹ <http://slidewiki.org>

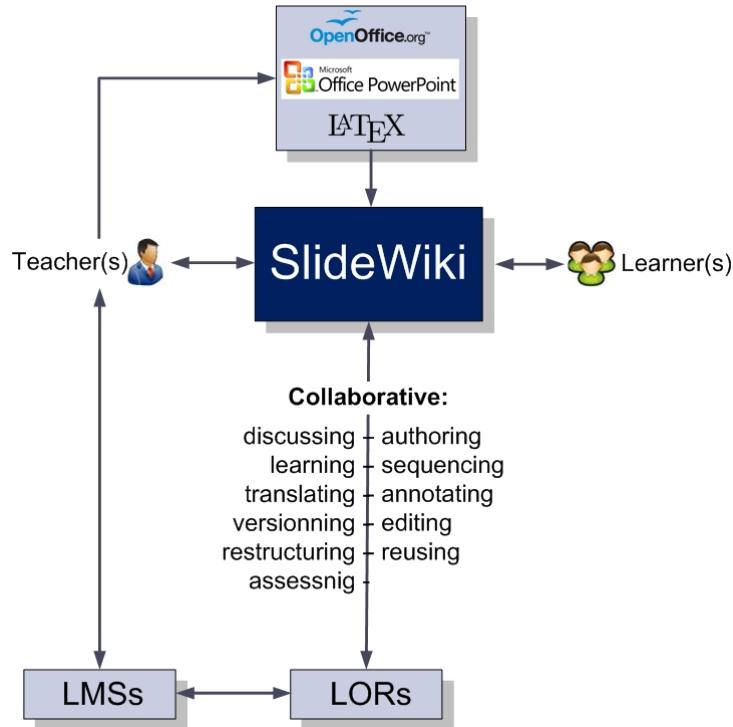


Figure 5.1: SlideWiki conceptual scheme.

The developed SlideWiki platform is described and evaluated in the sections below.

5.1 SlideWiki Architecture, Functionality and User Interface

SlideWiki deals with two types of (semi-)structured learning objects: slide presentations and assessment tests. The content is organized according to the *COSMEC* concept. Slide presentations are represented as a combination of individual slides (consisting mainly of HTML snippets, SVG images and meta-data), individual decks (ordered sequences of slides and sub-decks), themes (which are associated as default styles with decks and users) and media assets (which are used within slides). The self-assessment test sets combine individual tests (which can be organized manually by user or created automatically in accordance with the deck structure and content), questions for the slide material assigned to certain slides (which are part of tests) and answers (which are part of the questions).

The SlideWiki platform is built on top of the LAMP² technology stack where a relational approach (implemented by MySQL RDBMS) is used to implement the conceptual WikiApp data model of SlideWiki. The SlideWiki application makes extensive use of the model-view-controller (MVC) architecture pattern, as depicted at Figure 5.2. The MVC architecture enables the decoupling of the user interface, program logic and database controllers and thus allows developers to maintain each of these components separately.

SlideWiki makes extensive use of HTML5 features to provide users with intuitive and responsive interfaces. In addition to the overall MVC architecture, SlideWiki utilizes a *client-side MVC*

² Linux-Apache-MySQL-PHP

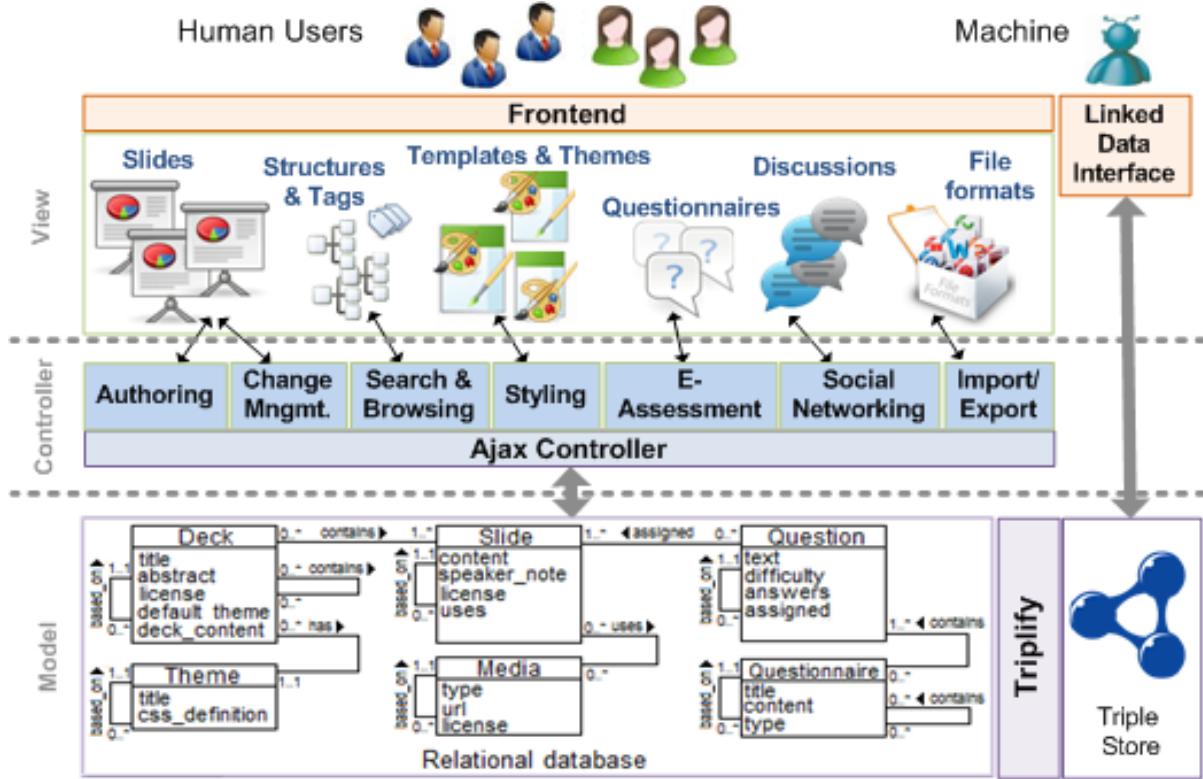


Figure 5.2: SlideWiki architecture overview.

approach (implemented in JavaScript and running inside the users Web browser). The client-side MVC handler as (singleton) controller listens to the hash fragment of the requested URLs and once a change has occurred the handler triggers the corresponding actions. Each action has a JavaScript template (implemented using *jQuery templates*) with the corresponding variable place holders. For each action an Ajax call is made and the results are returned to the controller in JSON format. Subsequently, the controller fills the templates with the results and renders them in the browser.

Additionally, SlideWiki supports *progressive loading* of presentations to guarantee the scalability when a large presentation is loaded. Progressive loading is a design pattern for web applications which adds content to a web page incrementally. It results in gradually increasing the workload over time when loading a large presentation thereby improving the performance of the system.

In the section we discuss the design solutions and implementation of main SlideWiki features.

5.1.1 Content Development.

Being a wiki, the SlideWiki platform supposes the prevalence of the content over format. Due to this, we were strongly focused on facilitation of the slide content editing process, trying to minimize the complications of slide content format. This means, that in oppose to desktop single-user slide editing applications (like Microsoft PowerPoint), we do not work with slide layouts. Instead, we operate on the slide content similarly to Microsoft Word, allowing embedding media within the textual blocks. Nevertheless, SlideWiki allows to individualize slide

formatting by applying style sheets. Moreover, the SlideWiki users are able to edit, share and reuse their own style sheets in the same manner as they share and reuse the content pieces.

In the section we discuss the tools supporting the authoring of slides and collaboration on them in detail.

Authoring SlideWiki employs an inline HTML5 based WYSIWYG (What-You-See-Is-What-You-Get) text editor for authoring the presentation slides (cf. [Figure 5.4](#), image 1). Using this approach, users will see the slideshow output at the same time as they are authoring their slides. The editor is implemented based on ALOHA editor³ extended with some additional features such as image manager, source manager, equation editor.

The inline editor uses SVG images for drawing shapes on slide canvas. Editing SVG images is supported by SVG-edit⁴ with some predefined shapes which are commonly used in presentations. We have chosen the SVG format due to the open XML-based encoding of the images it creates. This is especially important for the slide localization, as the textual content of the SVG images can be picked out the rest code blocks and translated.

For logical structuring of presentations, SlideWiki utilizes a tree structure in which users can append new or existing slides/decks and drag & drop items for positioning. When creating presentation decks, users can assign appropriate tags as well as footer text, default theme/transitions, abstract and additional meta-data to the deck.

SlideWiki follows the “anyone can edit” philosophy of Wiki for creating synergistic presentations. In order to manage the changes made by users, SlideWiki defines a revisioning system as described in the following.

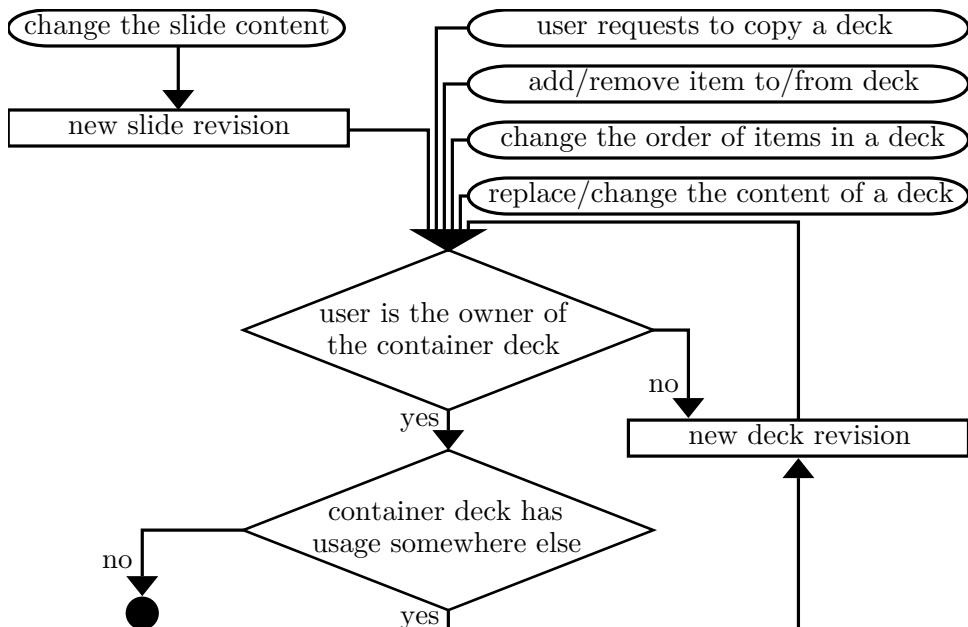


Figure 5.3: Decision flow during the creation of new slide and deck revisions.

³ <http://aloha-editor.org/>

⁴ <http://code.google.com/p/svg-edit/>

Change Management. Revision control is natively supported by WikiApp data model. We just define rules and restrictions to increase the performance. There are different circumstances in SlideWiki for which new slide or deck revisions have to be created. For decks, however, the situation is slightly more complicated, since we wanted to avoid an uncontrolled proliferation of deck revisions. This would, however, happen due to the fact, that every change of a slide would also trigger the creation of a new deck revision for all the decks the slide is a part of. Hence, we follow a more retentive strategy. We identified three situations which have to cause the creation of new revisions:

- The user specifically requests to create a new deck revision.
- The content of a deck is modified (e.g. slide order is changed, change in slides content, adding or deleting slides to/from the deck, replacing a deck content with new content, etc.) by a user which is neither the owner of a deck nor a member of the deck's editor group.
- The content of a deck is modified by the owner of a deck but the deck is used somewhere else.

The decision flow is presented in [Figure 5.3](#). In addition, when creating a new deck revision, we always need to recursively spread the change into the parent decks and create new revisions for them if necessary.

Styling. In order to create flexible and dynamic templates and styles for presentations, SlideWiki utilizes *Saas* (Syntactically Awesome Stylesheets) language⁵. Sass extends CSS by providing several mechanisms available in programming languages, particularly object-oriented languages, but not available in CSS3 itself. When Sass script is interpreted, it creates blocks of CSS rules for various selectors as defined by the Sass file. Using Saas, SlideWiki users can easily create and reuse presentation themes and transitions.

5.1.2 Content Migration.

Keeping up-to-date with these latest developments and trends in online education, the SlideWiki platform has been developed to target all the widely used and emerging educational channels, in order to maximize its outreach and impact to educators and learners worldwide. The flexible WikiApp data model allows to fully integrate SlideWiki with LMSs, MOOC platforms, as well as social networks and eBooks. This integration will ensure that the learning resources developed with the use of the SlideWiki platform are delivered via a variety of educational channels that address different learning contexts, pedagogical and technological requirements, and learning communities.

In the section we describe the implementation of WikiApp data model for both relational databases and RDF. We discuss as well bottlenecks and deficiencies we met during the implementation. We as well describe the conversion of SlideWiki content into SCOs, that allows to publish the content in a wide variety of existing educational platforms, such as Moodle.

⁵ <http://sass-lang.com/>

5 Implementation and Evaluation

The figure consists of four screenshots of the SlideWiki platform, each numbered 1 through 4:

- Screenshot 1:** Shows the tree structure of a presentation titled "Vorlesungsreihe Informationssysteme". A specific slide titled "E-Business E-Commerce" is selected, displaying its content. A red circle labeled "1" is positioned at the bottom right of this slide area.
- Screenshot 2:** Shows the "Edit Question for the slide" dialog. It includes fields for "Question" (a sample question about supported question types), "Difficulty" (set to 3 stars), and "Edit Answers" (a table for adding correct, rhetoric, stupid, and other types of answers). A red circle labeled "2" is positioned at the bottom right of the dialog.
- Screenshot 3:** Shows a question in "learning mode". The question is "What types of questions does SlideWiki support? by soeren". It has three options: "You are an educator. What are the benefits of SlideWiki for you?", "You are a student. What are the benefits of SlideWiki for you?", and "You are a teacher. What are the benefits of SlideWiki for you?". Below the options are "Correct?", "Explanation", and "Delete" buttons. A red circle labeled "3" is positioned at the bottom right of the question area.
- Screenshot 4:** Shows a "Test for 'Vorlesungsreihe Information'" page. It displays a list of questions from the "SOA-Technik" module, such as "Welche SOA Architekturstile existieren?", with multiple-choice options. A red circle labeled "4" is positioned at the bottom right of the test interface.

Figure 5.4: Four screenshots of SlideWiki features. 1 - Tree structure of the presentation, inline WYSIWYG editor; 2 - Editing of a question and manual assigning to a test using lists; 3 - Question in learning mode with correct answers displayed; 4 - Module-based scoring of an assessment test.

Data Model for Relational Database

Using the previously discussed WikiApp data model (see section 4.2) we have succeeded to implement the following principles on SlideWiki platform:

- Provenance: The origin and creation context of all information should be preserved and well documented.
- Transparency, openness and peer-review: Content should be visible and easily observable for the largest possible audience, thus facilitating review and quality improvements.
- Simplicity: The implementations should be simple to build and use.
- Social Collaboration: Following other users, watching the evolution of content as well as reusing and repurposing of content in social collaboration networks is at the heart of data model.
- Scalability: The implementations should be scalable and be implementable according to established Web application development practices.

SlideWiki example application uses two implementations of WikiApp data model. The first implementation is used for managing slides and presentations. It includes individual slides (consisting mainly of HTML snippets, SVG images and meta-data), decks (being ordered sequences of slides and sub-decks), themes (which are associated as default styles with decks and users) and media assets (which are used within slides). The second implementation was developed for managing questions and questionnaires (tests). It includes questions for the slide material (the question is assigned to all slide revisions), questionnaires (which could be organized manually by user or created automatically in accordance with the deck content), and answers (which are the part of the questions).

We implicitly connected these two WikiApp instances by adding two relations. Firstly, we assigned questions to slides. Thus, during the learning process users are able to try answer the questionnaires and have a look the assigned slide if necessary. The important issue here is that we assign question not to individual slide revision, but for the slide itself. This decision gives an opportunity to create new slide revision, that already has a list of questions, collected from other revisions. Secondly, we assigned questionnaires to concrete deck revisions. Thus automatically created test saves the structure of deck revision, which it is assigned to. This allows us to use module-based assessment to score the test results. The following two diagram [Figure 5.5](#)) illustrates the implementation of WikiApp in SlideWiki. These two special relations are indicated on the example object diagram by the dashed arrows.

The current database (see also [Figure 5.6](#)) consists of 30 tables interconnected by primary or foreign keys:

Data Model for RDF

In order the OER movement to fully success, the OERs have to be searchable and re-usable, in other words the interoperable. In order to ensure the interoperability, a number of competing approaches, metadata schemas and interface mechanisms was developed the last decade. For example, Dublin Core ⁶, IEEE Learning Object Metadata (LOM) [33] or ADL SCORM ⁷ standards and query interface mechanisms such as OAI-PMH ⁸ or SQI ⁹. However, so far Web-scale integration of resources is not facilitated, mainly due to the lack of commonly shared principles, datasets and schemas [43]. At the same time, the Linked Data approach has emerged as the de-facto standard for sharing data on the Web, offering a large potential to solve interoperability issues of any type of digital content. Thus, we consider the publishing content and its metadata as linked data to be the crucial requirement for the OER.

In order to publish the SlideWiki content in accessible way, we have adapted the developed WikiApp data model for linked data. We based the data model on three newly created ontologies: (i) WA ontology ¹⁰ models the WikiApp data model and can be reused in any type of WikiApp implementation; (ii) SW ontology ¹¹ models the base SlideWiki relations between content objects and users and (iii) SA ontology ¹² models the self-assessment items and can be reused for other e-assessment data mapping.

⁶ <http://dublincore.org/documents/dces/>

⁷ <http://www.adlnet.org/>

⁸ <http://www.openarchives.org/pmh/>

⁹ <http://www.cen-ltso.net/main.aspx?put=859>

¹⁰ <http://slidewiki.org/rdf/wa>

¹¹ <http://slidewiki.org/rdf/sw>

¹² [http://slidewiki.org/rdf\(sa](http://slidewiki.org/rdf(sa)

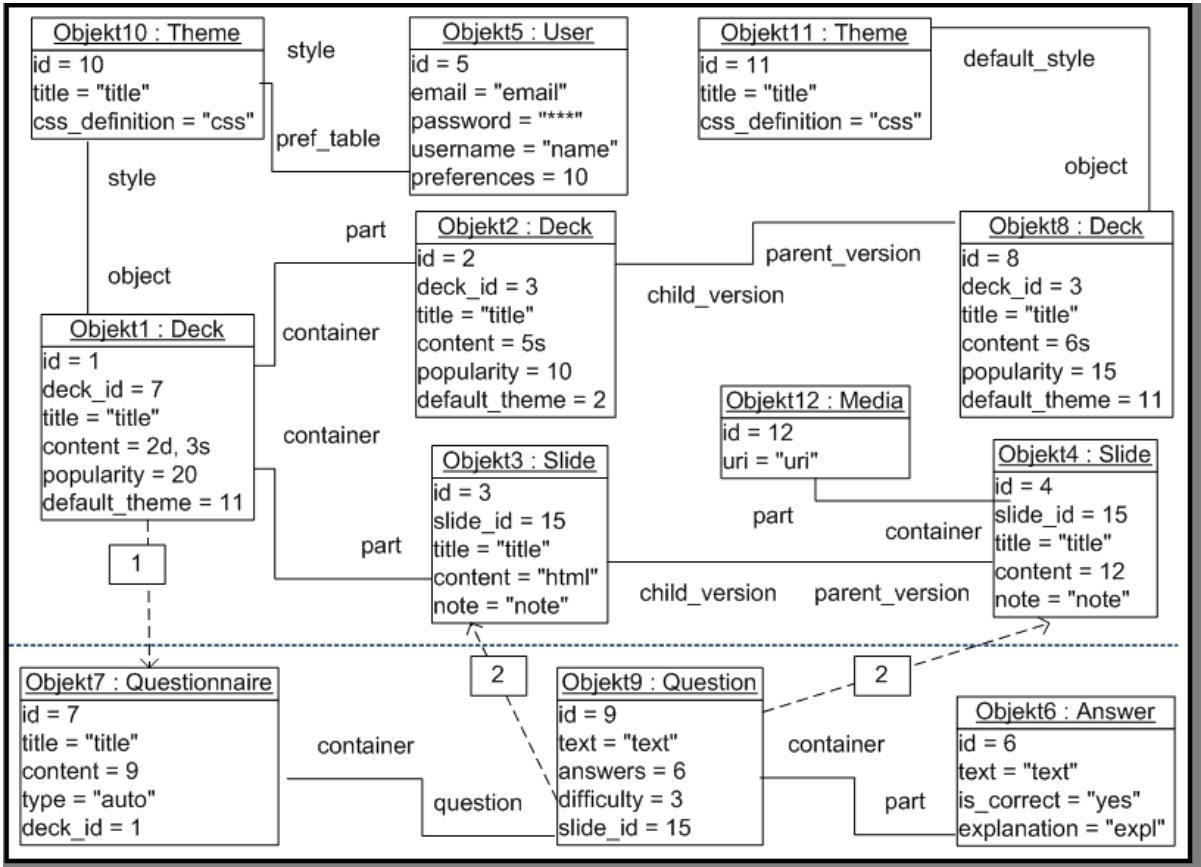


Figure 5.5: Example object diagram for Slidewiki. 1 - connection between questionnaires and deck revisions to enable module-based scoring; 2 - connection between questions and slides to allow users see the related material when answering the questionnaires

As there are no ontologies for modeling e-learning objects available, the number of reused properties and classes is not high. However, we have reused properties from three widely accepted ontologies:

- Dublin Core Metadata Initiative Terms (DCMI-TERMS)¹³ is a set of vocabularies which includes the elements from DublinCore, as well as sets of resource classes (including the DCMI Type Vocabulary [DCMI-TYPE]), vocabulary encoding schemes, and syntax encoding schemes. The terms in DCMI vocabularies are intended to be used in combination with terms from other, compatible vocabularies in the context of application profiles and on the basis of the DCMI Abstract Model [DCAM]. DublinCore (DC) is a vocabulary of fifteen properties for use in resource description. DC is a simple yet flexible metadata standard, intended to describe a wide variety of resources.
- Friend Of A Friend (FOAF)¹⁴ is an ontology describing persons, their activities and their relations to other people and objects. As used in most of the existing vocabularies, FOAF is a common way to add general information to persons and organizations. FOAF is one

¹³ <http://dublincore.org/documents/dc/terms/>

¹⁴ <http://www.foaf-project.org/>

Table	Function
Slide	describes the original slide
Slide_revision	describes the revisions of the original slide
Deck	describes the original deck containing a set of slides
Deck_revision	describes the revisions of the original deck
Deck_content	describes the organization of content in a deck
Questions	describes multiple-choice questions associated to a slide
Answers	describes answers to a question
Doubtful	describes the reasons for questions reported by users as doubtful
Users	describes the users of platform
User_group	describes different user groups on the platform
User_tests	describes the evaluation test created by a user
Test_content	describes the evaluation test as a set of questions
Test_results	describes the results of an evaluation test
Style	describes presentation themes
Transition	describes transition effects used in presentations
Impress_transition	describes a special type of transitions
Comment	describes user-added comments on a slide or deck
Tag	describes keywords assigned to an item (i.e. slide, deck, etc.)
Brand	describes branding ads for presentations
Message	describes messages communicated between platform users
Media	describes media items used in a presentation
Medial_relations	describes the relation between media and the associated item
Subscription	describes the user-subscription to a certain slide/deck
Badge	describes a user badge
Badge_import_details	describes the badge associated items
Notifications	describes user notifications
Preference	describes preferences for user profiling
Short_urls	describes the URLs for search engine optimization
Testing	describes performance tests on the platform
Translation_cronjobs	describes translation jobs to be sent to external translation services

Table 5.1: Description of SlideWiki Database Tables

of the key components of the WebID ¹⁵ specifications, and its support by browsers and web-services grows fast. For example, the Live Journal ¹⁶ and DeadJournal ¹⁷ blogging sites support FOAF profiles for all their members, My Opera community supports FOAF profiles for members as well as groups, FOAF support is present on Identi.ca, FriendFeed, WordPress and TypePad services. Yandex blog search platform supports search over FOAF profile information. Prominent client-side FOAF support is available in Safari, Firefox and Google Chrome web browsers.

¹⁵ <https://www.w3.org/2005/Incubator/webid/>

¹⁶ <http://www.livejournal.com/>

¹⁷ <http://www.deadjournal.com/>

5 Implementation and Evaluation

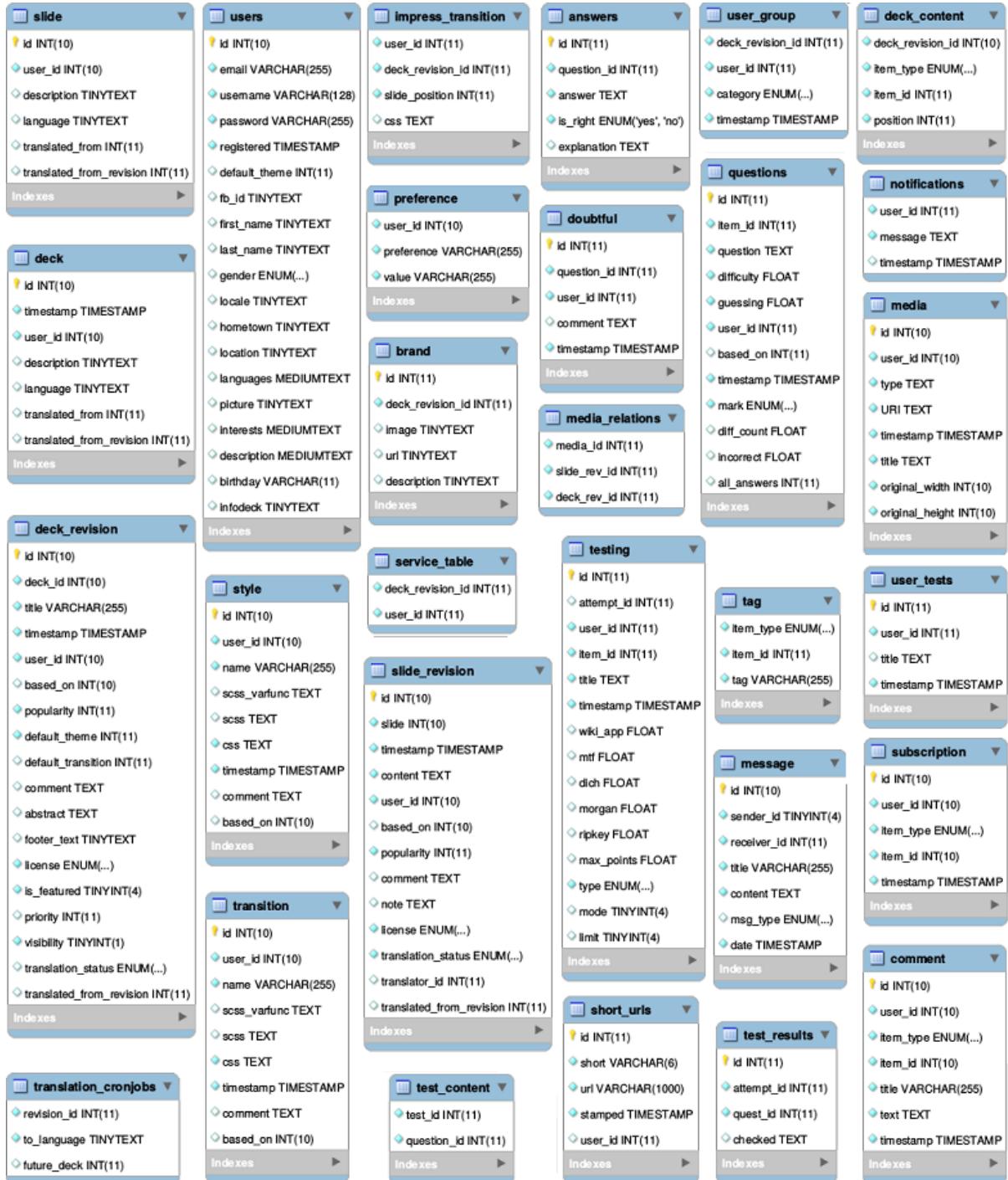


Figure 5.6: An overview of SlideWiki MySQL tables

- Provenance Ontology (PROV-O)¹⁸ provides a set of classes, properties, and restrictions that can be used to represent and interchange provenance information generated in different systems and under different contexts. Provenance is a critical ingredient for

¹⁸ <https://www.w3.org/TR/prov-o/>

establishing trust of published scientific content. This is true whether we are considering a data set, a computational workflow, a peer-reviewed publication or a simple scientific claim with supportive evidence. It is a domain-independent and general-purpose ontology and it allows and encourages for extensions to cover more specific needs. In particular, to track authoring and versioning information of content pieces, PROV-O provides a basic methodology but not any specific classes and properties for identifying or distinguishing between the various roles assumed by agents manipulating digital artifacts, such as author, contributor and curator.

Additionally, the W3C recommended ontologies are reused to model the CSS styles attached to the content objects ¹⁹ and the textual content of the objects ²⁰.

Following the WikiApp data model formalization, the core of the WA ontology is built on three main classes: (1) *wa:Container* which corresponds to the *Content Type* concept, (2) *wa:ContentObject* for storing the content objects and all kinds of relations between them and (3) *wa:User*, a subclass of *foaf:Person*. Being an implementation of WikiApp data model, the ontology requires each content object to have only the *dcterms:created* property specified. This ensures high flexibility and interoperability of the ontology, making it easy to be reused in a wide range of applications.

The core relations between content objects can be defined by using one out of four properties.

- ***prov:wasRevisionOf*** ²¹. The PROV ontology defines revision as a derivation for which the resulting entity is a revised version of some original. The implication here is that the resulting entity contains substantial content from the original. Revision is a particular case of derivation, that is reflected by the *prov:wasRevisionOf* property being a subclass of *prov:wasDerivedFrom* property. The suitable semantics of the property, its quality and popularity of the PROV ontology itself led us to the decision to reuse it instead of defining a new one. In order to simplify the usage of the future datasets, we define as well the inverted property *wa:hasRevisions*.
- ***wa:translationOf***. Despite the fact the integration of PROV vocabulary allowed us to use the *prov:hadPrimarySource* property for defining the relation between translation and its original, we have decided not to reuse the property. The reason for that decision is its potentially ambiguous meaning. For example, in the SlideWiki example implementation a slide presentation can have an assigned source, that is usually a textbook the slides are based on. In order to overcome the ambiguity we defined an own property, making it however a subproperty of *prov:wasDerivedFrom*. This allows us to use all the semantics defined for the PROV vocabulary and at the same time allows to add own specifications.
- ***dcterms:hasPart*** ²². We reuse the property as well as its inversion *dcterms:isPartOf* to define the *part-of* relationship between two content objects.
- ***wa:assignedTo*** The property defines the many to one relationship between two content objects. Its common usage includes the assigning comments, notes or questions to other content objects.

¹⁹ <http://www.w3.org/ns/oa>

²⁰ <http://www.w3.org/2011/content>

²¹ <http://www.w3.org/ns/prov#wasRevisionOf>

²² <http://dublincore.org/documents/dcmi-terms/#terms-hasPart>

@TODO: change the picture: wa:translationOf - subproperty of prov:wasDerivedFrom, wa:revisionOf = prov:wasRevisionOf

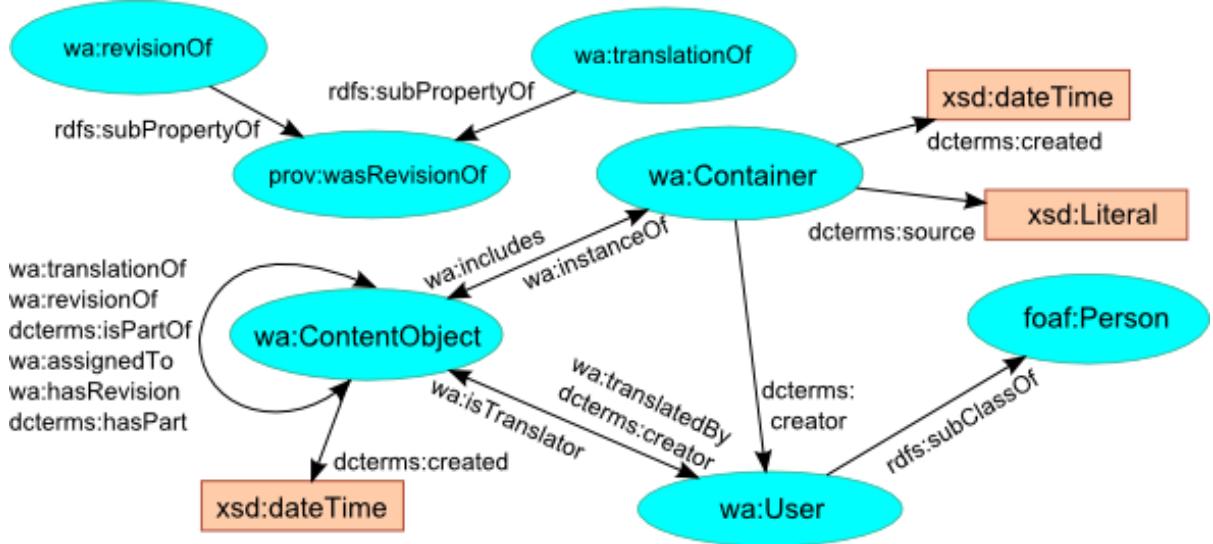


Figure 5.7: WikiApp data model implemented as an ontology

The complete SlideWiki knowledge graph is presented in [Appendix C](#).

RDB2RDF Conversion

An RDB2RDF mapping can either be represented as a set of customized rules or through a generic set of rules (as defined by the W3C Direct Mapping standard ²³). Often, the output of a Direct Mapping may not be useful, that is, it may not adequately take the structural and/or semantic complexity of the database schema into account. Due to the complexity of the SlideWiki data model, we have concluded that not using the customized rules would lead to a loss of resulting data quality.

A large number of tools for mapping relational data to RDF exist already²⁴, they implement different mapping approaches, often allowing to customize the mapping. Customized mappings in their turn can be represented in different material forms (proprietary formats, visual diagrams, tables, etc.). Languages for representing customized mappings have been surveyed in [139]. Until recently, no standard representation language for customized RDB2RDF mappings existed. However, in 2012 the W3C RDB2RDF Working Group has released recommendations for the Relational Database to RDF Mapping Language (R2RML) ²⁵. As it is the only mapping language that has been standardized to date, we do not consider usage of other (proprietary) formats reasonable.

The trial conversion of the SlideWiki relational data to RDF using R2RML has already been proceeded. We have used an open-source Java-based tool *SparqlMap*²⁶ that allows the execution of a SPARQL query fully inside a relational database and does not require dumping the data

²³ <https://www.w3.org/TR/rdb-direct-mapping/>

²⁴ One of the listings is available at <https://github.com/timrdf/csv2rdf4lod-automation/wiki/Alternative-Tabular-to-RDF-converters>.

²⁵ <https://www.w3.org/TR/r2rml/>

²⁶ <http://aksw.org/Projects/SparqlMap.html>

Requirement	Description	Measure
Quality of the mapping implementation and representation		
Data accessibility	Describes how the mapping result can be accessed.	ETL/on-demand/both
Standard compliance	Characterizes if the mapping representation is standard compliant.	boolean
Faithfulness of the output		
Coverage	Characterizes the mapping completeness	percentage of DB columns mapped
Accuracy of data representation	Characterizes the mapping correctness	percentage of correctly mapped DB columns
Incorporation of domain semantics	Shows level of domain semantics incorporation	percentage of properties that link to the results of SQL queries
Quality of the output		
Simplicity	Shows the simplicity of SPARQL queries returning the frequently demanded values	percentage of complex SQL queries results integrated into the mapping
Data quality	Characterizes the quality of output data	aggregation of linked data quality metrics
Data integration	Characterizes the interlinking degree of the output data	percentage of external instances integrated into the resultant dataset
Interoperability		
Reuse of existing ontologies	Shows the amount of reused vocabulary elements	percentage of reused properties and classes
Quality of reused vocabulary elements	Characterizes the quality of chosen for reuse properties and classes	accumulated quality and popularity of reused vocabulary elements
Accuracy of reused properties	Characterizes the accuracy of properties reuse	percentage of accurately reused properties
Accuracy of reused classes	Characterizes the accuracy of classes reuse	percentage of accurately reused classes
Quality of declared classes/properties	Shows the quality of ontology documentation	accumulated quality of declared classes/properties

Table 5.2: Summary table of the proposed metrics system

first. The tool takes as an input a list of settings (mainly related to the database connection) and the previously defined R2RML mapping representation. Based on those, SparqlMap outputs set of triples in Turtle notation ²⁷. The translation is a light-weight process with low requirements on processing power. The mapping of the relational database is expressed using R2RML [152].

SparqlMap allows to establish the access to the resulting triples using two main paradigms: (i) Extract Transform Load (ETL) and (ii) on-demand mapping. ETL means physically storing triples produced from relational data in an RDF store. The disadvantage of ETL is that, whenever the RDB is updated, you have to re-run the entire ETL process, even if just one RDB record has changed, carrying out an often redundant synchronization process. However, in the ETL case nothing more than the RDF store is needed to answer a query. On-demand mapping is realized by translating a SPARQL query into one or more SQL queries at query-time, evaluating these against (a set of) unmodified relational database(s) and constructing a SPARQL result set from the result sets of such SQL queries. In contrast to the ETL implementation, on-demand mapping requires more resources for processing each query. However, the on-demand implementation does not face the synchronization issue and does not replicate the data. In the light of these advantages and disadvantages, we have decided that the best solution is to implement the both data access approaches.

<i>W</i>	Requirement	<i>Potential</i> ₁	<i>Actual</i> ₁	<i>i</i> ₁	<i>Potential</i> ₂	<i>Actual</i> ₂	<i>i</i> ₂
0.1	Quality of mapping			1			1
0.5	- data accessibility	-	both	1	-	both	1
0.5	- representation	-	R2RML	1	-	R2RML	1
0.2	Faithfulness of output			0.7			0.76
0.4	- coverage	169	78	0.46	169	100	0.59
0.5	- accuracy	0	0	1	0	0	1
0.1	- domain semantic	39	7	0.18	43	9	0.21
0.3	Quality of output			0.51			0.913
0.3	- simplicity	18	14	0.78	18	14	0.78
0.4	- data quality	6	4.5	0.75	6	6	1
0.3	- data integration	1	0	0	1	1	1
0.4	Interoperability			0.9			0.91
0.15	- reuse of properties	39	24	0.615	44	28	0.64
0.05	- reuse of classes	10	2	0.2	11	3	0.27
0.15	- quality of reused properties	24	24	1	28	28	1
0.3	- accuracy of reused properties	24	24	1	28	28	1
0.15	- quality of reused classes	2	2	1	3	3	1
0.1	- accuracy of reused classes	2	2	1	3	3	1
0.1	- quality of newly created properties	15	15	1	16	16	1
		Total	0.778				0.95

Table 5.3: Mapping quality evaluation.

i - the index of evaluated requirement, *W* - weight of the index in the result mapping quality evaluation. The metrics with gray background were improved in the final mapping.

²⁷ <https://www.w3.org/TeamSubmission/turtle/>

We have created the R2RML file needed to customize the mapping in three steps. First, we have created the initial mapping, using common knowledge of linked data quality. On the second step, we have evaluated the mapping using our previously developed metric system (see [Appendix A](#)), summarized in [Table 5.2](#) with minor changes. And finally, we have improved the initial mapping based on the evaluation. The listing of the final R2RML file we created is presented in [section C.1](#)

In [Table 5.3](#) we summarize the results of the evaluation for both initially created and improved mappings. The metrics that we consider to be insufficiently good after the first iteration are shown with the gray background. Based on the discussion, we improved the quality of metrics from output faithfulness and quality dimensions, as discussed below:

1. *Quality of mapping implementation.* As we used R2RML file to represent the mapping and established both ETL and on-demand data access, the index of mapping implementation quality is maximal (thus, equal to 1).
2. *Faithfulness of output.* As SlideWiki contains a lot of usage statistics and service tables (private messages, data for cronjobs etc.), the index of coverage is quite low. However, during the assessment we took the decision to include two additional tables into the mapping. As we did not have any digital values mapped, the index of accuracy is 1. SlideWiki is based on a complicated data model. Thus, while implementing the platform, we tried to hide this complexity from the user as was done while creating the mapping. This resulted in the index value of domain semantics incorporation equal to 0.18, which we considered sufficient. According to our weighting, the index of faithfulness reaches 0.7 and should be increased by increasing the index of coverage.
3. *Quality of output.* To calculate the index of simplicity our domain specialists created a list of frequently used complicated queries (cf. [Table 5.4](#)). According to the evaluation, the index level of 0.78 was considered to be sufficient.

We have measured the data quality produced by the mapping using the metric system proposed by [\[161\]](#). After the evaluation, the dataset quality produced by the initial mapping was not sufficiently high. In order to increase the index, we have added machine-readable indication of a license (CC-BY-SA), metadata about publisher and dataset itself. These actions resulted in improvement of the data quality from the trustfulness aspect.

4. *Interoperability.* The initial mapping consisted of 10 classes (2 of which were reused) and 39 unique properties (24 reused). The properties were chosen based on LOV metrics; ranges, domains and comments were taken into account. For newly created properties dereferencable URIs were created. This activity resulted in a high level of interoperability.

SCORM

Sharable Content Object Reference Model (SCORM) [\[1\]](#) as one of the community-approved standards, requires the transformation of the learning material into the sequence of annotated *Sharable Content Objects* (SCOs). The granularity and sequencing of the SCOs should be determined by the content author depending on the audience needs and preferences [\[2\]](#). Each SCO can include several objects (images, diagrams, text blocks, questions to answer etc.), however, each SCO should be context neutral and should stand alone [\[103\]](#). Such a structuring

Query	Incorp.
- get content of a deck	1
- get usage of a deck	1
- get usage of a slide	1
- get questions assigned to a deck	0
- get questions assigned to a slide	1
- get contributors of a deck	1
- get contributors of a slide	0
- get language of a deck	1
- get language of a slide	1
- get translations of a deck	0
- get translations of a slide	0
- get all revisions of a deck	1
- get all revisions of a slide	1
- get tags of a deck	1
- get comments assigned to a deck	1
- get comments assigned to a slide	1
- get content of a user test	1
- get usage of a question in user tests	1
18	14

Table 5.4: Incorporation of frequently used complicated queries results into the mapping to simplify usage the linked data

optimizes the re-usability while still meets the needs of the learners for whom it was originally intended.

As the content of SlideWiki is semantically structured, it satisfies the requirements of the SCORM standard and the conversion is straightforward. For now we have implemented a beta-version of the conversion, which produces a structured list of slides followed by a plain list of self-assessment questions attached to each subdeck. We have tested the produced SCOs in Moodle. The resulting SCOs keep the content, styling, structure and metadata of the original SlideWiki slides, as shown at [Figure 5.8](#). Thus, a deck is converted to a SCO which can be used as a course module without additional work. The low efforts needed to implement the conversion prove the flexibility and interoperability of our data model.

5.1.3 Multilinguality

The implementation of multilingual content support in SlideWiki is based on the previously described CoSMEC concept and co-evolution paradigm (see [section 4.3](#)). According to the CoSMEC concept, the implementation of co-evolution of source object content and its translations supposes the implementation of three operations: initial translation, synchronization and merging of the revisions.

Our architecture allowed us to implement a *translation* operation backed by the Google Translate service. After translation into one of 71 currently supported languages, the presentation can be edited, re-structured and reused independently from its source. The interface for the translation mechanism is depicted at [Figure 5.9](#).

5.1 SlideWiki Architecture, Functionality and User Interface

Figure 5.8: SlideWiki deck imported in Moodle as a SCORM package

Afrikaans	Cebuano	Filipino	Hindi	Khmer
Albanian	Chinese (Simplified)	Finnish	Hmong	Korean
Arabic	Chinese (Traditional)	French	Hungarian	Lao
Azerbaijani	Croatian	Galician	Icelandic	Latin
Basque	Czech	Georgian	Indonesian	Latvian
Belarusian	Danish	German	Irish	Lithuanian
Bengali	Dutch	Greek	Italian	Macedonian
Bosnian	English	Gujarati	Japanese	Malay
Bulgarian	Esperanto	Haitian Creole	Javanese	Maltese
Catalan	Estonian	Hebrew	Kannada	Marathi

Figure 5.9: Interface of translations management: 1 - language of selected object; 2 - a drop-down list with links to languages available from the selected object; 3 - button for translation; 4 - dynamically updated list of the languages supported by Google Translate service

To enable *synchronization* of original and translated versions, every further revision of translated objects inherits the link to the source revision (see v2.1 at [Figure 5.10](#)). The changes in the original version of the object cause the creation of new revision v1.1. Additionally, users are notified of translations that have become out of sync with the source (exclamation marks

in v2.0 and v2.1).

In SlideWiki implementation of the synchronization is slightly different for slides and decks. We decided not to implement *manual* synchronization of decks, considering the process to be too complicated for users. Users can only trigger an *automatic* synchronization. However, our data model allows us to get all existing translations for all existing slide revisions. Then, during an automatic deck synchronization we do not repeatedly translate the slides which have already been translated to the target language before. Instead, we include in the translation the latest revision of the slide in the target language. Thus, the synchronization of decks only adds the slides and subdecks that were not included in the source deck at the moment of previous translation.

For the slides both automatic and manual synchronization are enabled. The users are able to compare v1.0 and v1.1 and decide, either they want to redo the translation (scenario 1 at [Figure 5.10](#)) or they want to update the content manually (scenario 2).

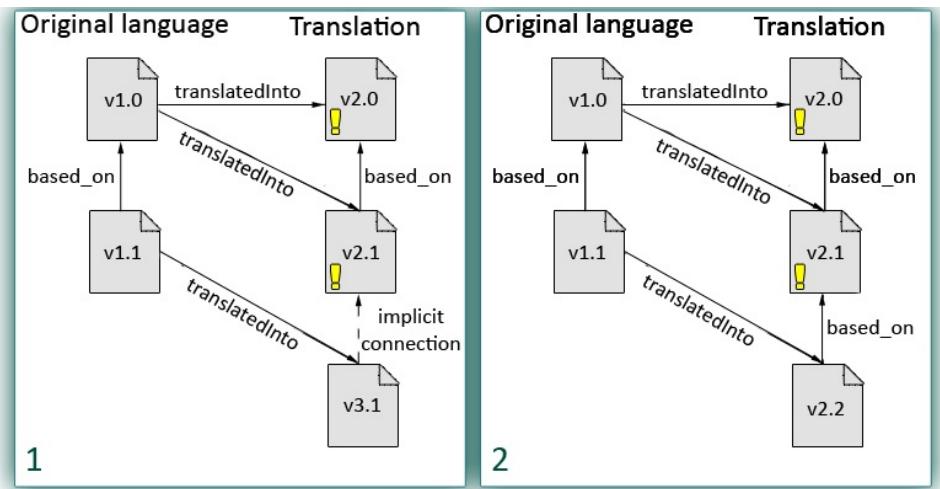


Figure 5.10: Two scenarios of content synchronization between translation and source: 1 - automatic; 2 - manual synchronization

SlideWiki implements the revision control in accordance with the WikiApp data model, where *merging the revisions* is supported as one of the core operations. However, as discussed above we defined rules and restrictions to increase the performance. Namely, we introduced the *content owner* and *member of editor group* roles. If the changes are made by a user belonging to one of these two roles, the creation of a new *deck revision* is not triggered (the new *slide revision* however is created). As we allow the owner of a deck revision to change it without the creation of a new revision, it was an important issue whether we should allow the multiple translation of the same revision into the same language or not. We decided to allow it, however, this led to the situation that we would get several identical presentations with content of bad quality, since it was translated automatically and not edited manually. However, we could not disable the multiple translations, because in that case it would be impossible for example to get translations of new slides if they were added by the owner. Thus, merging the revisions became the crucial operation, not only for merging back-translation with the source, but also for merging multiple translations in the same language. At [Figure 5.11](#) we illustrate the interface for comparing and merging deck revisions.

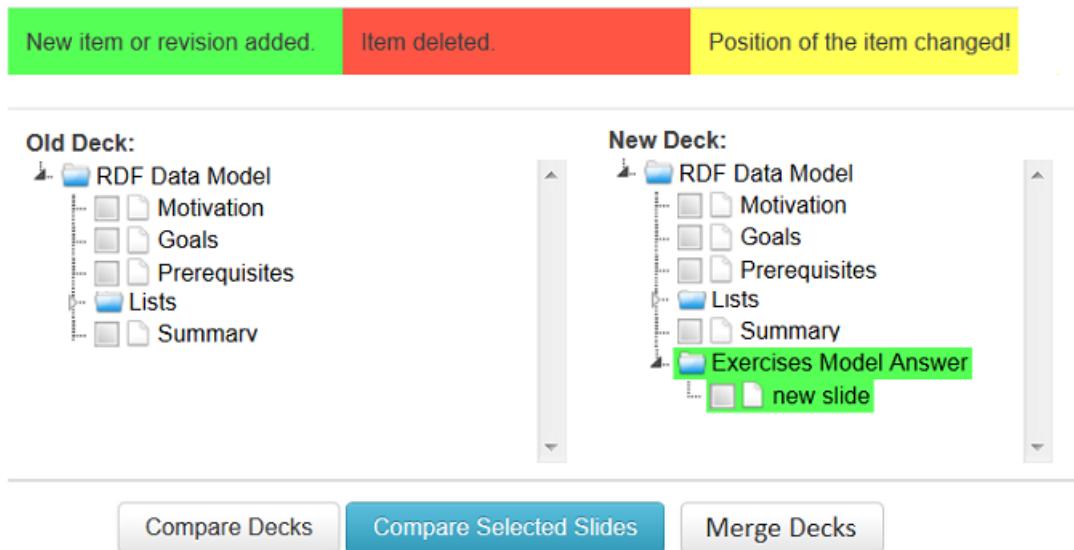


Figure 5.11: Interface for comparing and merging deck revisions

To evaluate the effectiveness of multilinguality support, we collected the statistics of usage the translation tool as well as statistics of the multilingual content it produced.

At [Figure 5.12](#) we present the distribution between original and translated versions in relation to the total number of content objects. As the WikiApp data model does not enable deletion or update of the content, the graphs can be viewed as time trends. The blue line shows an example moment in time, when the SlideWiki database consisted of 16321 slides overall. The graphic shows that 78% of the slides at the moment were in their original language, 22% were translations and 5% of total number of slides were revised after translation. At the same moment, about 35% of decks were translations. Thus, the percentage of translated decks increases faster than the same for slides. This means that the presentations consisting of less than average number of slides are being translated more often. This can be due to the fact that users want to try the feature before using it on large decks. However, the assumption needs further investigation.

According to the statistics, the percentage of content created by translation has a strongly increasing trend. We predict the percentage of translations will soon prevail over the percentage of source objects. From one perspective, the prevalence means decreasing the production costs and a large diversity of languages available. However, from another perspective, it causes the reduction of average content quality, as refining the translation needs time and human resources. This is illustrated by the decreasing percentage of revised slides in compartment with translated ones. The solution is in additional user motivation to put effort in refining the concrete translations according to their knowledge in both domain and languages.

The diagrams on the [Figure 5.13](#) show the language distributions for decks, slides and visitors (according to Google analytics). Only unique visitors are counted. Due to its large percentage, we excluded English language from the resulting diagrams to increase the informativeness. The statistics show the visible correlation between number of content objects available in a concrete language and number of visitors speaking the language (for the most of languages). The results look promising, as they prove the involving of not English-speakers into the global e-learning community activity. Especially promising looks the fact that more than 13% of overall visitors

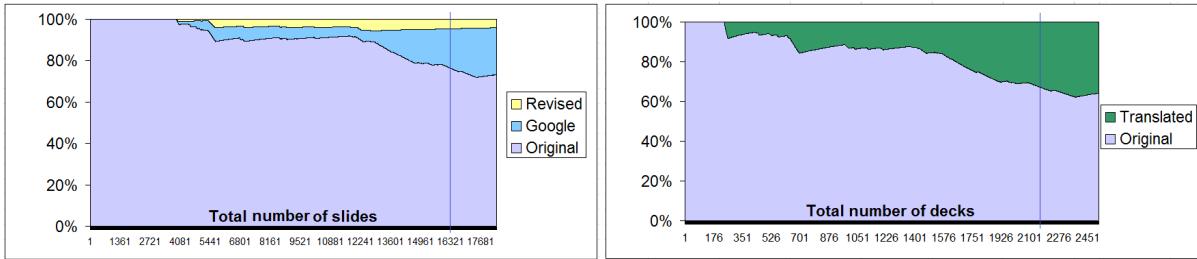


Figure 5.12: Percentage distribution between original and translated versions vs. total number of content objects for slides and decks. Blue line shows an example time moment, discussed in the paragraph above.

belong to developing countries and regions (mostly Eastern Europe and Russia, Turkey, Arabic-speaking countries, Thailand). We believe that this percentage will increase with increasing the popularity of the source.

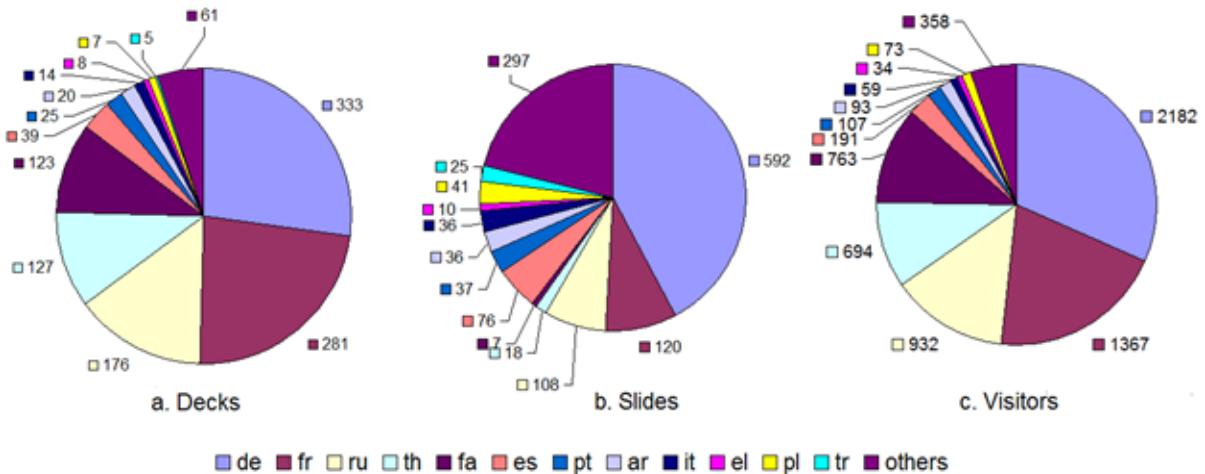


Figure 5.13: Distribution of languages for content objects and new visitors

5.1.4 User Engagement and Social networking

The theoretical foundations for e-Learning 2.0 are drawn from social constructivism. It is assumed that students learn as they work together to understand their experiences and create meaning. In this view, teachers are knowers who craft a curriculum to support a self-directed, collaborative search and discussion for meanings. Supporting social networking activities in SlideWiki enables students to proactively interact with each other to acquire knowledge. With the SlideWiki concept we address the following social networking activities:

- Users can follow individual learning objects as well as other users activities to receive notification messages about their updates.
- Users can discuss the content of learning objects in a forum-like manner.
- Users can share the learning objects within their social network websites such as Facebook, Google Plus, LinkedIn, etc.

- Users can rate the available self-assessment questions in terms of their difficulty and quality.

Besides increasing of the learning process quality, social activities improve the quality of the created learning material. Even when answering a quiz, users can contribute by analyzing the quality of the questions and making suggestions of how to improve them. Thus, the knowledge is being created not only explicitly by contributors, but also implicitly through discussions, answering the questions of assessment tests, or in other words through native learning activities.

5.1.5 E-Learning and E-assessment.

SlideWiki supports the creation of questionnaires and self-assessment tests from presentation slides. Questionnaires (like decks and slides) are supported by a revisioning system so that users can create, edit and share questionnaires according to the Wiki collaboration style. The created questionnaires can be employed for training organization members or to interview prospective members. Educators can manually or automatically generate evaluation tests from the questionnaires based on their preferences to assess the knowledge of employees to be trained.

Each question has to be assigned to at least one slide. Important note here, that the question is assigned not to the slide revision, but to slide itself. Thus, when a new slide revision appears, it continues to include all the list of previously assigned questions. Questions can be combined into tests. The *automatically created* tests include the last question revisions from all the slides within the current deck revision. Manually created tests present a collection of chosen questions and currently cannot be manipulated as objects (cf. [Figure 5.4](#), image 2). Thus, in our implementation only questions and answers have to be placed under the version control. However, their structure is trivial and the logic of creating their new revisions is intuitive. We just restricted the number of new revisions to be created similarly with the decks: changes made by the question owner do not trigger a new revision creation.

Students can start a chosen test (manually created or automatically collected) in one of two possible modes: “learning” or “examination”. In learning mode student can ask to show the slide, to which the question is assigned to remind the material, or simply show the correct answers (cf. [Figure 5.4](#), image 3). Thus, student should not spend time to find the material. However, after the user asked to show her/him either the slide or correct answers she/he will not get any points for the question. In examination mode these features are disabled.

After choosing the mode the user can also set up the amount of questions (all, all the difficult or concrete amount) and the order to show them (random or increase the difficulty). As the amount of questions can differ for the same test, we show the test results as a percentage of the maximum points for exactly this selection of question.

Our architecture allowed us to implement module-based scoring. Each module of the assessment test presents a sub-deck of the presentation and is scored individually. Then, all the “parent” modules are scored as a sum of “children” points and finally the whole test is scored as a sum of all the points for all the modules (cf. [Figure 5.4](#), image 4). To score the results the student (or the teacher) can choose one of five implemented algorithms. All five algorithms use the dynamically accumulated difficulty d of the question as the number of points for the fully correct response:

$$d = \frac{\text{incorr}}{\text{all}} \quad (5.1)$$

If the user prefers to use dichotomous scoring, the values of *incorr* and *all* mean, respectively, the accumulated number of *incorrect* answers and *all* answers of that question by any of users. In a case of partial scoring, *incorr* is determined as follows:

$$incorr = \sum_{i=1}^n \left(1 - \frac{p_i}{d_i}\right), \quad (5.2)$$

where n - number of attempts for the question, d_i - difficulty (or maximal points), that the question had at the moment, when the i^{th} attempt was made, p_i - points obtained in the i^{th} attempt

After the difficulty is determined, it's scaled to $(1, d_{max}]$, where d_{max} is the maximal weight, that a question can have. d_{max} is set up by the system administrator only for the users' comfort. In SlideWiki we set it up to 5.

Advantages and disadvantages of a learning assessment based on multiple-choice questions (MCQs) are a long and widely discussed issue in the scientific community. However, in practice this type of questions is very popular due to the possibility of automatic evaluation and scoring [53]. Consequently, an important research question is to exploit the strengths and mitigate the weaknesses of MCQs. Some systems (e.g. Moodle ²⁸) allow teachers to create MCQs with multiple correct options. This type of questions we will call multiple-mark questions (MMQs), to distinguish them from the conventional MCQs, where there is always only one correct option. Multiple-mark questions were already recommended by Cronbach [34]. Other research [120, 115, 74] considers MMQs to be more reliable, when compare them with conventional MCQs. However, even though the advantages of MMQs are meanwhile widely accepted, up to our knowledge there are no balanced methods for scoring multiple-mark questions available to date.

One possible approach to score the MMQs is to use dichotomous scoring system. The dichotomous scoring awards the constant amount of points, when the question is answered correctly and zero points in a case of *any* mistake. However, the partial scoring is preferable to the dichotomous, especially in case of MMQs. [120, 82, 14, 15]

After we have studied existing partial scoring algorithms we came to a conclusion about necessity of developing a new one (see [143]) We have called it balanced scoring method and give its formal definitions and results of synthetic experiments in Appendix B.

For evaluation of our algorithm we used a lecture series on "Business Information Systems". We chose this course since it comprises a large number of definitions and descriptions, which are well suited for the creation of MMQs. In total we have created 130 questions. A course of 30 students was offered to prepare for the final examination using SlideWiki. Overall, the students made 287 attempts to complete the questionnaire and we collected all their answers (also unfinished assessments) for the evaluation. After collecting the answers, we implemented all discussed algorithms to score and compare the results, in particular with regard to the ranking and the mean score. The results are summarized in Figure 5.14.

The study aimed to investigate three aspects of the proposed approach:

- How severe does the balanced scoring approach penalize?
- How does balanced scoring differ from Dichotomous scoring?
- How clear were the results scored by the proposed approach for the students?

²⁸ <https://moodle.org/>

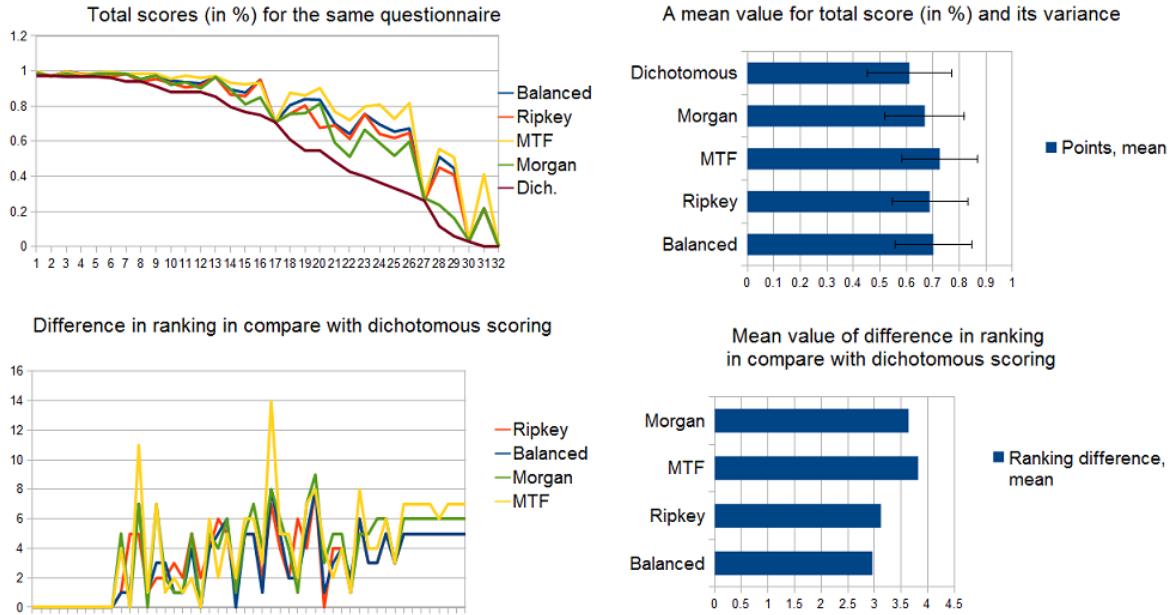


Figure 5.14: The statistics of the evaluation

We answer the first question by comparing the scores calculated using all discussed algorithms for the same questionnaire (see Figure 5.14, upper part). These two diagrams show, that on average the balanced scoring approach penalizes more severely than MTF scoring and less severely than other discussed approaches. Thus, the users study confirms the findings of our previous synthetic experiment.

We answer the second question by comparing the difference in student ranking. We rank all assessments based on the individual scores. That is, assessments with higher scores rank higher than assessments with lower scores and equal scores result in the same ranking. We compare the rankings of other approaches with the rankings calculated using the dichotomous scoring, since we consider the dichotomous scoring to be the ranking reference. The two lower diagrams in Figure 5.14 show the results of this evaluation. They show, that the ranking of the balanced scoring approach is the closest to the dichotomous ranking when compared to the other algorithms.

After the end of semester we asked the participants to answer the third question. They were offered to evaluate clarity of the results on a five-point scale from "very clear" to "very unclear". We have collected nine responses, two of them were "neutral", four – "clear" and three – "very clear". This confirms that the results obtained by the balanced scoring method are easy to understand for students.

The proposed approach has a list of restrictions, however it has advantages when compare with the discussed approaches. One of the main advantages is its clearness for the students, that was proven by the user evaluation presented above. Also, our approach is based on the mathematical model, it does not suffer from the skewness, as it has the same formula for all cases. At the same time, the proposed approach recognizes the attempts to guess the correct answer, for example choosing all the possible options. When compare with the existing approaches, the advantages of the proposed algorithm could be summarized as follows:

- The approach allows to score both multiple-mark and conventional multiple-choice questions.
- The approach is based on the partial scoring concept.
- The algorithm can be easily implemented, it is pure mathematical.
- The score does not highly depend on the amount of correct and incorrect options.
- The value of the penalty is in balance with the possibility, that the student is trying to guess.
- Due to the balance, the results are clear for the students.

5.2 Limitations of the Implementation and Description of the Next Generation

@TODO: Say that it is not all so bad as it looks =)

In order to evaluate the implementation we have conducted three usability studies. The first two in form of user survey aimed to evaluate the usability of the system at different stages of development. The third one in form of single-user observation aimed to define bugs and particular UI issues. Based on the study results, we have found that the SlideWiki concept itself is clear to the users, however performance and usability of the system have to be dramatically improved. In the section we describe the studies and their results, as well as propose the Next Generation architecture.

5.2.1 Usability Studies

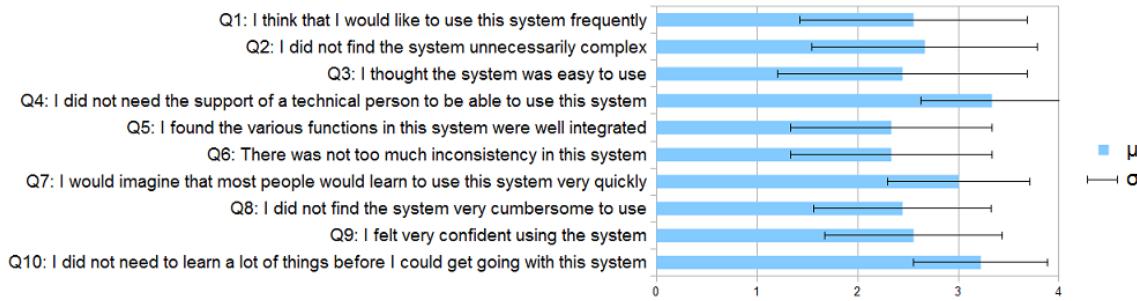
Study at Chemnitz Technical University

In order to evaluate the clearness of the concepts behind SlideWiki at early development stage, we used the platform for accompanying the Information Systems course at Chemnitz Technical University. We have structured the slides within the lecture series and added questions for student self-assessment before the final exam. We informed them about the different e-learning features of SlideWiki, in particular, how to prepare for the exam using SlideWiki. The experiment was not obligatory but students actively contributed by creating additional questions and fixing mistakes in slides. The experiment was announced to 30 students of the second year and 28 of them registered at SlideWiki.

The students were working with SlideWiki for several weeks, and we collected the statistics for that period. During that period, they created 252 new slide revisions which some of them were totally new slides, others were improved versions of the original lecture slides. Originally the whole course had 130 questions, and students changed 13 of them, fixing the typos or adding additional distractors to multiple choice questions. In total, students performed 287 self-assessment tests. The majority of these used the automatically and randomly created tests covering the whole course material. 20 tests included only difficult questions, 2 asked to show the questions with increasing difficulty. This showed us that the students liked the diversity of test organization. Students also liked the possibility to limit the number of questions – 80 attempts were made with such a setting. 8 students reached the 100% result for the whole course. On average, it took them 6 attempts before they succeeded.

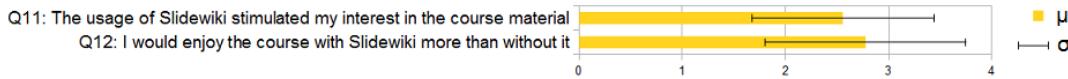
5.2 Limitations of the Implementation and Description of the Next Generation

Questions recommended by SUS technique



Questions for measuring the quality of learning

SlideWiki overall



Assessment tool



Figure 5.15: Results of SlideWiki evaluation survey: mean μ and standard deviation σ .

After the experiment we can claim, that more active SlideWiki users received better marks on the real examination. It shows that SlideWiki not only allows students to prepare for the examinations, but also engages them in active participation that helps to improve the quality of learning. After the end of semester, we have asked participants to fill out a questionnaire which consisted of three parts: usability experience questions, learning quality questions and open questions for collecting the qualitative feedbacks. We have collected 9 questionnaires that were filled out completely. They show us emergent problems and directions for the future.

In the first part of the questionnaire we included questions recommended by *System Usability Scale* (SUS) [95] system to grade the usability of SlideWiki. SUS is a standardized, simple, ten-item Likert scale-based questionnaire²⁹ giving a global view of subjective assessments of usability. It yields a single number in the range of 0 to 100 which represents a composite measure of the overall usability of the system. The results of our survey showed a mean usability score of 67.2 for SlideWiki which indicates a reasonable level of usability.

The second part of the questionnaire aimed to determine whether the SlideWiki helps to improve the quality of learning. It consisted of four questions with five options from “absolutely agree (1)” to “absolutely disagree (5)”. The evaluation results for these two parts are presented in Figure 5.15.

Although the positive answers prevailed, we were not satisfied by the fact that for many questions a third of participants chose the neutral value. The final part of the questionnaire helped us to understand the reasons. We included here four open questions:

1. What did you like most about Slidewiki?
2. What did you like least about Slidewiki?
3. What can we do to improve the Slidewiki’s usability?

²⁹ www.usabilitynet.org/trump/documents/Suschapt.doc

4. What features would you add to Slidewiki?

Within the answers we found repeated complaints about several bugs, that interfered the working process. We consider this fact to be the main reason of neutral and contradictory values. However, we collected also positive opinions, especially about features and possibilities that SlideWiki allows. Three of the recipients mentioned that they mostly liked that SlideWiki is easy to use, four of them noted, that they liked the idea of collaborative work and sharing the presentations itself. Within the collected answers we also got important suggestions, which could be roughly divided into two groups:

- Suggestions about desired improvements of existing features such as displaying the test results graphically, supporting more import formats, improving the SVG editor etc.
- Suggestions about totally new features, several of those were later implemented, e.g. translation, templates for presentation structure, etc.

The results of the initial evaluation were promising and we have continued the features development, trying to optimize the system performance and usability at the same time.

Study at the University of Bonn

After implementing the whole SlideWiki functionality we have conducted another study. This time the goal was to have a detailed view at the problematic aspects of the implementation discovered through the previous study. Namely, questions of this survey were designed to evaluate the following three aspects:

1. Responsiveness of the SlideWiki website
2. Design and usability of the SlideWiki website
3. Functionality and awareness of the features and usage of the SlideWiki website

The survey was divided into two parts. The first part consisted of 17 multiple choice questions and the remaining 5 questions were left for comments and ideas. We have received 23 anonymous responses out of 80 students registered for the “Semantic Data Web Technologies” course.

The results of the study, as presented in [Table 5.5](#), proved our previous concerns about usability of the trial SlideWiki implementation, especially on mobile devices. As well we have noticed even less satisfaction from SlideWiki performance than in the previous study, even though we have spent efforts to improve it. We consider there are two main reasons for even lower SlideWiki user satisfaction we have found out through the second evaluation:

- Actual decrease in SlideWiki performance due to substantial growth of its database. Although the architecture of SlideWiki was proven to be scalable with increasing amount of content, the increased amount of users simultaneously working on the same presentation was not taken into account.
- Increase of users expectations. During the development process new technological stacks and architectural patterns have appeared and spread widely, improving the possible performance of the systems with the similar scale. Thus, what have been considered to be relatively good performance at the beginning of development, has become not satisfying after the development has been finished.

However, the concepts underlying SlideWiki were still accepted by the users. This was proven by open answers we have collected in the second part of the survey:

If you were to review your overall SlideWiki experience, what score would you give it out of 10? 13 student gave it 5 and more, while 9 gave it less than 5.

What do you find most frustrating about SlideWiki? Most of the comments concerned navigation, especially on mobile devices and slowness of loading.

Question	Average
SlideWiki is easy to use.	2.39
I find it easy to navigate through SlideWiki pages.	2.3
SlideWiki provides a good user interface for learning slides.	2.35
I find SlideWiki clear and well organized.	2.57
SlideWiki loads quickly on my computer (3-4 sec).	2.39
SlideWiki loads quickly on my mobile (3-4 sec).	2.0
I can immediately find response to my actions (mouse click).	2.57
I can easily use SlideWiki on my mobile/tablet.	2.17
I can use Slidewiki with different browsers.	3.62
I can easily download the slides.	1.95
I found all the images, text, tables, etc. in the slides correctly rendered.	2.32
I can easily navigate through the slide pages using the navigation tree.	2.96
I had no problem finding the information I needed.	2.48
I could find some bugs while using SlideWiki.	3.78
I find all the links in the slides valid and active.	3.38
I tried the Questions section of the slides and I found it helpful	3.68
I tried the Discussion section of the slides and I found it helpful.	2.95

Table 5.5: The multiple choice questions and the average of the answers (1 = strongly disagree and 5 = strongly agree), categorized by the cell colors: dark-gray referring to “Responsiveness of the website”, (2) light-gray referring to “Design and usability” and (3) white referring to “Functionality and awareness of the website features and usage”

If you could change one thing about SlideWiki what would it be and why? Many students suggested improving the performance of loading slides.

What do you like best about SlideWiki? As it can be seen from the [Table 5.6](#), the students mentioned all main concepts we have based SlideWiki and the thesis except multilinguality:

Feature	# of students
Collaboration on the educational material	4
Different types of educational content, in particular, self-assessment items	4
Tree-structure of the content	2
Accessibility of the content from everywhere	2

Table 5.6: The most liked aspects of SlideWiki according to the final user evaluation

How can we improve SlideWiki? Please tell us your ideas and suggestions? Mostly students recommended changing layout (e.g. structure, minimalist design) and improving performance.

Based on the final evaluation results, we have concluded that the chosen architectural pattern and technological stack do not allow large-scale implementation of our concepts with satisfying performance. We have analyzed the technical issues and have found recently emerged technologies to solve them, as proposed in subsection 5.2.2. However, before the implementation of SlideWiki 2.0 we also needed to design the interface which considers all, even minor, UI issues indicated by the users.

Single-user Observation

In order to spot minor usability issues a single-user observation study [68] was conducted. Three users, whose demographical data is described in Table 5.7, were chosen.

#	Gender	Age	Occupation	Computer Skills
1	Female	19	Student Engineer	Good
2	Male	25	IT helpdesk	Excellent
3	Male	22	Economics Graduate	Average

Table 5.7: Demographical data of users participating in the single-user observation study.

In order to perform interaction with the Slidewiki, users were asked to complete the following sequence of actions:

1. Register on Slidewiki
2. Create a new empty presentation
3. Edit the presentation description
4. Add content to the first slide, using different text colors
5. Add an additional empty slide to your presentation
6. Add an image to the slide
7. Add an additional sub-deck to your presentation
8. Move the second slide you've created (with an image) to the new sub-deck
9. Change the title of the slide
10. Create a question to one of your slides
11. Download the presentation
12. Initiate the translation of the presentation into a different language
13. Comment on the translated version
14. Change the slides theme and transitions of the translated version
15. Edit the presentation description of the translated version
16. Play the translated version
17. Find and open the deck "Open Educational habdbook"
18. Open a user page of any of the collaborators
19. Message him/her
20. Contact the website owners
21. Search using the search field for the presentation "Semantic Web Lectures"
22. Exam yourself in there.

After finishing all the stages, the users were asked to give comments and suggestions. A screen recording tool was used in order to keep the video of the user interaction with the product for further exploit. The found issues and possible solutions are summarized in [68]. For example, while completing the task #5 (*Add an additional empty slide to your presentation*) all three users made at least 3 wrong actions and the fastest of them carried out the task within 47 seconds. In the current component implementation a user is supposed to complete the task by right-click on the deck title in the tree and

5.2 Limitations of the Implementation and Description of the Next Generation

choosing the necessary action from the drop-down list. However, this interface appeared to be not intuitive even for experienced users. In order to solve the issue, we propose to add two small buttons for slide and deck adding, as depicted at [Figure 5.16](#).



Figure 5.16: The deck tree navigation component: current state (left) and proposed improvements (right)

Another common issue found is the presentation of search results. The current state of the component is presented at [Figure 5.17](#). All three users called their created deck using the word "test" and it was difficult to find the exact presentation within dozens of others. As well, the tag cloud to filter the content is over-filled, and at the same time does not have to be always visible for the users. Our solution (see [Figure 5.18](#)) for these issues is to (1) enhance the search functionality by adding additional scope for search by authors and tags, and (2) hide those advanced options until user needs them.

Figure 5.17: The search component: current state

5.2.2 Defined Issues of Current Implementation and Proposed Solutions

Based on the evaluation results we can claim that the concepts behind SlideWiki are clear for users, they like the way of learning, storing and sharing of the presentations. However the trial implementation has a number of issues.

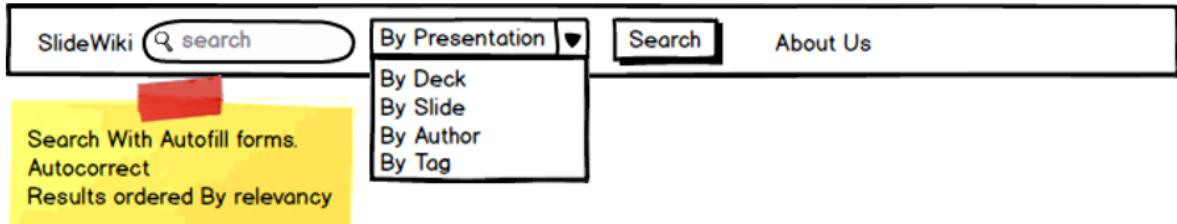


Figure 5.18: The search component: proposed improvements

The initial design decision for using a relational approach for storing the data was made mainly due to the centralized MVC architecture pattern for data flow in the SlideWiki system. As well, immaturity of triple stores and NoSQL database technologies at the design time was another incentive to adopt relational approach. The SlideWiki makes extensive use of the MVC where most of the complexity is handled on the backend-side (i.e. data layer) and views are mainly only serving content in different templates to users. In such a centralized model, a relational model with support for join and aggregate queries was crucial.

However, although we have made additional effort to increase the platform performance, users continue to complain about its slowness. Moreover, the emergence of mobile devices coupled with spreading wi-fi connection availability and quality led to the increased user expectations from web-based applications. In response to these increased expectations appeared new architectures and technological stacks which comes in place of less-flexible MVC architecture backed by LAMP.

Apart from the performance issue, as the number of developers, users and data on the platform started to grow, we faced new challenges with the relational data model and MVC architecture:

- Evolving the data schema with emerging of the new requirements. Using MySQL relational database, our application is constrained to a set of fixed schemata designed for the underlying tables. When more developers joined the project and new requirements were identified, handling the changes and extending the underlying schemata without affecting the related schemata became a cumbersome task. In this situation, a schema-less design approach (e.g. existing document-based data models) to provide a dynamic schema vs. a predefined schema seemed to be crucial and more efficient.
- Horizontal scalability of platform as the number of data and users increase. As the number of data and users on our system increased, we noticed some performance issues in loading the pages. Even though we implemented some caching mechanisms on the server, due to the dynamicity of SlideWiki environment, we still had to go for vertical scalability by increasing the CPU, RAM, SSD, etc, on our server which were quite expensive to sustain. On the other hand, with the emergence of NoSQL databases which allow horizontal scalability, we could potentially be able to handle large traffic by just adding few more servers easily in our infrastructure. In addition to that, MySQL database scaling was hard. A single MySQL table performance would degrade when crossing the 5-10GB per table in SlideWiki. In this case would need to partition and shard our database in a distributed way.
- Interoperability between the data model and front-end components. In the existing architecture, we have been using a complex object-relational mapping (ORM) layer that translates objects in code to relational tables. Since JSON format is mainly used in our front-end to interact with data, translation to SQL and building the abstract data layer seems to be another burden to performance. This adds a bottleneck to our system as the amount of nested data objects become large. For example, to get a list of contributors of a deck, we need (1) to recursively go through the whole deck tree and get all its subdecks and slides, (2) to get contributors of each slide and (3) to aggregate the contributors. Using existing NoSQL databases with native support for multilevel JSON could remove the complexity of ORM layer and increase the performance.

- Proving a high write time and high availability. Due to the wiki-based nature of SlideWiki, number of updates on data is very high. Existing MySQL database with support for transactions in ACID model comes with the cost of lower write time and lower availability. On the other hand, current NoSQL databases by default prefer high insert rate over transaction safety which would have a better fit in our SlideWiki requirements.
- Providing a Linked Data interface. One of the novel features in SlideWiki is providing highly structured e-learning content which could be queried and integrated with other educational material. To achieve this goal, we are providing an RDF-based version of content on SlideWiki supported by a Linked Data interface which enables accessing and querying data in a machine-readable format. In the existing version, we used an RDB2RDF tool to convert the underlying relational model to a graph-based data model. Exploiting a NoSQL database can facilitate this task to a great extent because of the underlying graph-based data model provided.

In order to solve the challenges, we propose to build the next generation of SlideWiki based on recently emergent MERN³⁰ technological stack in couple with microservice architecture.

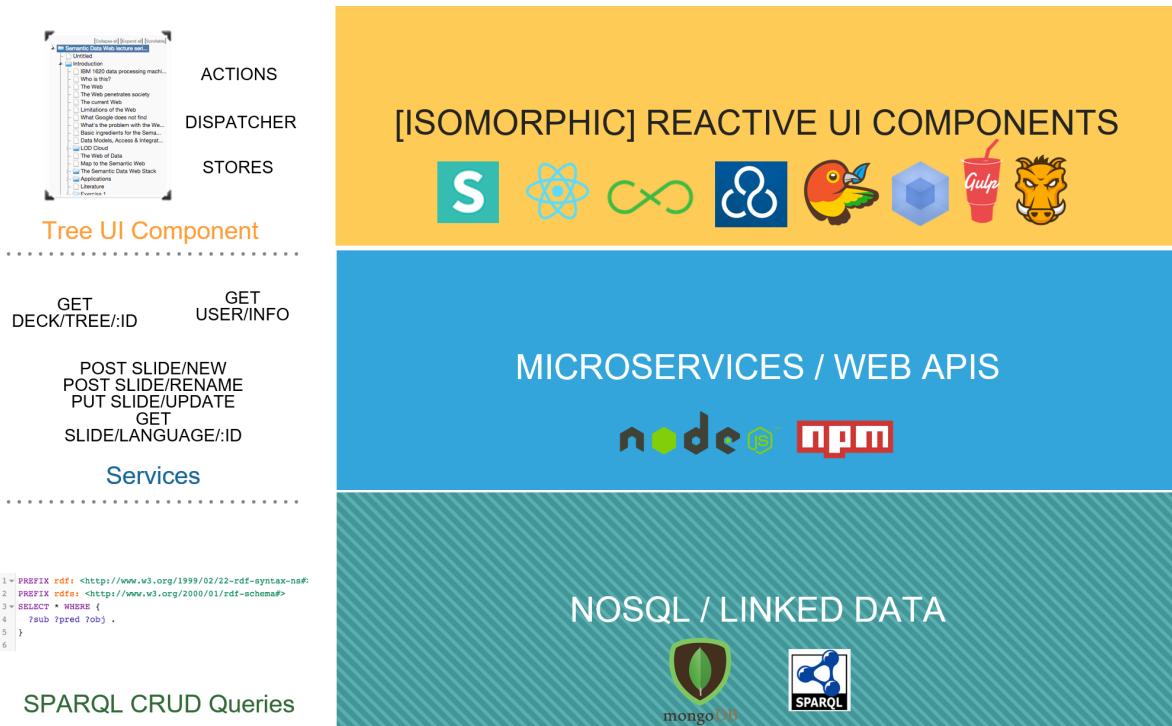


Figure 5.19: New Architecture of SlideWiki platform with an example usage (left side)

As shown in Figure 5.19, the new architecture relies heavily on microservices which aim to distribute the complexity of our data layer among self-contained replicable Web services supporting horizontal scalability. The main benefit of the microservice architecture is that it dramatically improves agility and velocity. This is because when our system is correctly decomposed into a set of microservices and their dependent data schemas, we can develop and deploy each microservice independently and in parallel with the other services. In order to enhance the interoperability, the proposed 3-tier architecture uses JSON as the communication language between user interface, services and data layer.

Apart from improved performance, the proposed architecture facilitates the maintaining of the SlideWiki's code due to the component-based attitude of ReactJS.

³⁰ MongoDB-ExpressJS-ReactJS-NodeJS

CHAPTER 6

Discussion and Conclusion

A major obstacle of increasing the efficiency, effectiveness and quality of education in Europe is the lack of widely available, accessible, multilingual, timely, engaging and high-quality educational material (i.e. OpenCourseWare). The creation and maintenance of comprehensive OCW is tedious, time-consuming and expensive, with the effect that often courseware employed by teachers, instructors and professors is incomplete, outdated, dull and inaccessible to those with disabilities. Universities create much of the world's intellectual capital and are eager to share this knowledge beyond the walls of the academy and to grant access to education for everyone. Unfortunately, cash-strapped institutions and schools have found it difficult to scale the significant organizational, technical, and cost barriers to distribution of rich OpenCourseWare plus also support interoperability and fully satisfy accessibility demands.

In the current thesis we aimed to contribute to the problem by providing a complex solution for the collaborative OCW development. In order to do so, we have firstly defined the major technological and conceptual gaps. Then we have developed missing concepts and evaluated them with an innovative SlideWiki platform.

The results of systematic literature study showed us the lack of conceptual and technological solutions for features supporting the collaborative OCW authoring. Namely, the issues of content localization, remixing and repurposing, as well as user engagement and coordination were not discussed enough in the previously existing research. In the scope of the thesis we have provided our own solutions to these issues:

- In order to engage and coordinate collaborators we have developed a CrowdLearn concept, that applies social networking techniques to the structured content development.
- To facilitate the content reuse and re-purpose we have developed the WikiApp data model, that presents the content as a sequence of content pieces, each of which can be operated and reused independently.
- In order to enable a fully-featured collaboration on multilingual educational content we have developed the CoSMEC concept, which allows synchronization and co-evolution of the content between its versions in different languages.

We have implemented and evaluated the developed concepts within a mid-size proof-of-concept application SlideWiki. The application deals with two types of structured OERs: slide sets and self-assessment item sets attached to the slides. Both content types can be operated collaboratively, with enhanced possibilities for cross-lingual reuse and repurpose. The results of the evaluation showed us that the developed concepts do work together with a synergistic effect, even though the implementation lacked technical maturity.

Although many of the techniques employed by SlideWiki (e.g. crowdsourcing, semi-automatic translation, semantic structuring, cloud-computing) are well-known, their combination and tailored application for OpenCourseWare authoring is novel and innovative:

- The **crowdsourced creation of OpenCourseWare** as performed with SlideWiki applies concepts such as fine-grained branching and merging to learning content (remixing and repurposing), which are already established in other fields such as software development (Github) or content management (Wikis).
- The **semantic structuring** of the content in SlideWiki allows novel reuse and remixing scenarios and also facilitates the accessible, inclusive and ubiquitous representation of content stored in the system.
- The **semi-automatic multilingual authoring** enables the sharing of learning content translation and maintenance effort between many community members. Since all content is versioned and semantically structured, it is trivial to (automatically) translate content and to keep track of changes in various multilingual versions of the same content object.
- The **intertwining of the technological advances and accompanying community building**, fostering and cross-fertilization functions helps establishing a socio-technological ecosystem, which makes the SlideWiki service an innovative, world-wide used educational content facility.

The innovative SlideWiki functionality is beneficial for different types of users, including:

- **Educators, Lecturers and Teachers** can significantly increase the user base by making the content accessible to a worldwide audience, obtain translations of their high-quality content in many different languages. They can engage students in contributing to and discussing the courseware and easily create (self-)assessment tests for students. Peer-educators can be involved in improving and maintaining the quality and attractiveness of the learning content. Ultimately sharing and contributing to qualitative content will increase the educator's reputation in the community.
- **Students and Learners** obtain access to an unprecedented repository of freely available, multilingual and high-quality OpenCourseWare. They can view rich-learning content right in a browser, discuss particular content (e.g. a slide or question) with other students and instructors and contribute additional content, improvements and feedback. They are empowered to assess their learning progress using the questionnaires attached to presentations.
- **Schools and Universities** can make courseware easily accessible (since each presentation and slide has its own URL), leverage the wisdom of educator crowd, which can collaborate efficiently in creating rich educational content, make the e-learning content produced in the organization really re-usable and re-mixable as well as increase the reputation of the school by sharing quality content.
- For **Companies** presentations are a crucial element of the corporate knowledge exchange. SlideWiki can help to make the knowledge available in a company easily accessible to everyone (in the company), engage employees in sharing and reusing knowledge available in presentations. In addition to the public SlideWiki portal companies can install and run a SlideWiki instance inside their corporate Intranet with access control.

Due to these innovations SlideWiki won the OpenCourseWare Consortium's Excellence Award in 2014 and is already used by hundreds of educators and thousands of learners. Several hundred comprehensive course materials are available on SlideWiki in dozens of languages. The maturity and innovative nature of the concepts behind SlideWiki allowed us to further develop and extend the platform in scope of EU funded project. In order to turn SlideWiki into a large-scale European platform the consortium of 17 European educational and research organizations will collaborate on all aspects of the development, management and evaluation. The consortium along with its European and international partner organizations is committed to support and sustain SlideWiki. As well, it aims to raise the community awareness for an easy-to-use and barrier-free possibility for the creation of Open Educational content that can be shared, mixed and re-edited across platforms and device borders.

The expected project impacts on European education after successful objectives completion include (but not limited to):

Impact on secondary Education and Higher-Education. The intended system components and services will invite learners, educators, and other professionals to evaluate its outcomes and use the community created content in their own context. All three levels of tertiary education are affected, namely, undergraduate, postgraduate, and further. This is of paramount importance as it allows an individual to follow SlideWiki and its advancements throughout his/her entire professional life, that is from college to work and staff development life periods. Moreover, in order to derive the best practices for educational content reuse and sharing, the SlideWiki needs a critical mass of educational content as well as content types. To this extent, that fact that all levels of education are targeted in SlideWiki is crucial for the establishment of the aforementioned critical mass.

Impact on professional and vocational training. Existing online learning resources often used by life-long learners lack engagement mechanisms. For instance, Massive Open Online Courses show really big dropout rates (90% following some studies). This constitutes one of the main risks for online educational initiatives given the amount of resources (platform, bandwidth, teaching assistants time, learners time etc.) that are wasted. Consequently, one of the main contributions of a successful online educational initiative like SlideWiki is to reduce dropout rates by engaging learners. This will be done by better adapting content and the learning experience to user needs, something possible and easier as content is really open and contributions encouraged. Moreover, the social components included will also engage users through group learning, gamification and social interactions.

Impact on Open Education Community. We expect SlideWiki and its surrounding services to become a catalyst in the establishment of an European Open Education community. The value produced by the large amount of high-quality multilingual courseware being freely available through the service, the large user community and the positive overall image this no-for-profit service accumulates, will attract a large number of stakeholders to join this initiative and its aim of establishing a European Open Education community. We expect universities, companies, institutions of higher education, school authorities, grass-roots initiatives, companies in the education field and many others to join the activities and events started and spun-off from this project.

Further expected economic impacts Wikipedia and other projects have been perfecting and evolving the community-source approach to pooling resources, knowledge sharing, leveraging and aggregating wisdom with user support to create attractive solutions that used by thousands of different users with manifold experience and backgrounds everyday. We envision to pool individual, community, and university resource contribution and intellectual capital investments and leverage them to create sharable and accessible Open Educational solutions that are of, by, and for schools, teachers, higher education as well as companies and learners around the world. As a community driven project, SlideWiki will employ the development principles of open source software. We are aiming at solving challenges together with our collaborating networks to make SlideWiki a reliable service companion for existing enterprises as well as for beginners that want to build upon this investment adapting the content and formats to their needs.

Further expected societal impacts Universities want to go where the learners are to share their rich scientific and intellectual knowledge beyond the walls of the academy and to expand the boundaries of the classroom. Many people have just experienced education as courses they learn from. Often OERs are just thought of as free courses, but the opportunity to involve users within the creation process (without losing track of changes and authorship) gives us a chance to make learning more engaging and active. OERs need to be useful in many different channels (mobile, social web, blog, wikis, internal company networks, etc.) to rise the societal impact by providing direct benefit to the context they are being used. Further aspects include the demand to insert cultural specific references to make a concept easier to understand for a different group (school kids, foreign culture, etc.). SlideWiki as a platform depends on partnering organizations and their established network outreach. Through the projects participating

6 Discussion and Conclusion

organizations and involved parties we expect an easy and welcome uptake of the described idea and technology which increases the societal impact.

APPENDIX A

Research Paper: Measuring the quality of relational to RDF mappings

A.1 Abstract

Mapping between relational data and linked data (RDB2RDF) is an essential prerequisite for evolving the World Wide Web into the Web of Data. We propose a methodology to evaluate the quality of such mappings against a set of objective metrics. Our methodology, whose key principles have been approved by a survey among RDB2RDF experts, can be applied to evaluate both automatically and manually performed mappings regardless of their representation. The main contributions of the paper are: (1) assessing the quality requirements for mappings between relational databases and RDF, and (2) proposing methods for measuring how well a given mapping meets the requirements. We showcase the usage of the individual metrics with illustrative examples from a real-life application. Additionally, we provide guidelines to improve a given mapping with regard to the specific metrics.

A.2 Introduction

Translating the data stored in relational databases (RDB) to the linked data format is an essential prerequisite for evolving the current Web of documents into a Web of Data. In order to be effectively and efficiently reusable, linked data should meet certain data quality requirements [161]. Assessing the quality of a linked dataset created from RDB may be a laborious, repetitive task if the dataset is frequently recreated from its RDB source, e.g. after any update to the RDB. We therefore claim that it is possible to positively influence the quality of a linked dataset created from RDB by improving the quality of the RDB2RDF *mapping* that produces the linked data¹. To the best of our knowledge there has not been prior research towards collecting and describing quality requirements for RDB2RDF mappings. Thus, this paper aims to fill the gap by providing a system of metrics to measure the quality of RDB2RDF mappings.

A large number of tools for mapping relational data to RDF exist already²; they implement different mapping approaches, often allowing to customize the mapping. To standardize the description of mappings, the W3C RDB2RDF Working Group has released two recommendations in 2012 (more

¹ Here, we use the term “mapping” to refer to the function that maps relational database columns to RDF properties, but not to materialisations of such mappings in concrete languages or representations, nor to tools that would execute such a function.

² One of the listings is available at <https://github.com/timrdf/csv2rdf4lod-automation/wiki/Alternative-Tabular-to-RDF-converters>.

details in Section A.4.1): Direct Mapping [8], which produces RDF graph representations directly from a relational database (data and schema) and the Relational Database to RDF Mapping Language (R2RML) [37] for expressing customized mappings.

In this paper, we aim at developing a system of quality metrics that can be applied both to direct and customized mappings that works with different representations of mappings (R2RML or proprietary formats, visual diagrams, tables, etc.). Any of these representations are suitable for being evaluated with the proposed system as long as one can derive a list of database columns (including unmapped ones) and their corresponding RDF properties (for mapped ones).

To determine the scope of requirements that can be posed to RDB2RDF mappings, we studied the state of the art in related research fields such as ontology matching, linked data quality assessment and RDB2XML mappings. However, as Section A.3 shows, these metrics collected from the scientific community do not solve the given problem entirely. In order to fulfil the list, we followed the Direct and R2RML Mapping recommendations. Some of the RDB2RDF features standardized in the documents are connected with the output data quality dimension, while others influence the quality of the mapping itself. For example, the ability to define R2RML views allows to incorporate domain semantics into the mapping – an opportunity that we consider to increase the quality of mapping. We propose not only a system of metrics but also define means of measuring them, along with guidelines to improve the rating of a dataset w.r.t. each metric.

The paper is structured as follows: in Section A.3 we discuss the existing approaches, in Section A.4 we summarize the quality requirements for mappings and the metrics for measuring the quality. Section A.5 presents the results of the survey conducted in order to collect the community feedback. In Section A.6, we conclude and propose future research directions.

A.3 Related work

RDB2RDF mapping approaches. Although previous research has not explicitly focused on collecting requirements for RDB2RDF mappings, many such requirements can be found in best practice and mapping approach descriptions. For instance, in [70], the importance of reusing existing vocabularies to enhance the interoperability of the output dataset is explained. We extend this statement to a set of requirements combined in the *interoperability* dimension, taking into account not only quantitative, but also qualitative metrics.

We notice that existing literature on the topic uses incoherent terminology. Often, the term *mapping* is used instead of mapping *language*, e.g. in [10, 50]. Discussing the requirements for mapping *language*, both articles mention the following requirements for RDB2RDF *mapping* as well:

- presence of both ETL and on-demand implementation
- incorporating domain semantics that is often implicit or not captured at all in the RDB schema
- indication of time metadata about the dataset during the mapping creation process to control the dataset currency (we subsume this metric under “data quality”)
- intensive reuse of the existing ontologies

Requirements for mappings between relational data and XML. The topic of mapping relational data to *XML* is related to RDB2RDF mapping due to the similarity between the XML and RDF data models. The earliest articles on the quality of RDB2XML mappings (e.g. [133, 135]) only take into account the performance of the mapping algorithms. This is because RDB2XML mappings are mostly produced automatically and do not allow customization. Only few authors (e.g. [129]) propose metrics for evaluating the quality of the output data. However, the proposed metrics evaluate only the syntactic correctness of the output XML and therefore cannot be applied to documents in RDF (except for *presence of syntax errors*).

Other approaches (e.g. [136]) prove the need for customizable techniques of mapping due to the wide gap between these two formats. As their major evaluation criterion, they use the *efficiency of query*

processing. We adopt this approach and base our *simplicity* metric on it. Liu et al. [97] describe an approach to design a high-quality XML Schema from ER diagrams. The authors define a list of requirements for the design; the most relevant one for our purposes is *information preservation*. However, the metrics for *measuring* such requirements are not discussed. We study the information preservation requirement in detail and provide objective metrics in the *Faithfulness of Output* dimension.

Requirements for ontology matching. Mapping between a database schema and an RDF vocabulary can also be viewed as a special case of ontology matching. Several studies propose requirements for measuring the quality of an ontology matching. According to [52], all proposed measures can be distinguished between compliance measures, measures concerned with system usability and performance measures focusing on runtime or memory usage. As our study focuses on evaluating the quality of mapping itself, we do not take the system usability or performance dimensions into account.

In the evaluation of ontology matching, compliance measures are based on the two classical information retrieval measures of *precision* and *recall* [153]. We adapt these metrics to the evaluation of RDB2RDF mappings and include them in the *faithfulness of output* dimension as *coverage* and *accuracy of data presentation* metrics.

Measuring linked data quality. The quality of mapping correlates with the quality of output linked data produced by the mapping. Poor design decisions made at the stage of defining a mapping, such as using deprecated classes, redundant attributes or badly formed URIs, decreases the quality of the output data. Therefore, metrics for the quality of the output data can be viewed as metrics for quality of the mapping.

The field of assessing the quality of linked data is still in its infancy, but several articles have addressed it already. Our recent survey [161] collects and compares the existing papers, the most complete and detailed ones being Bizer’s PhD thesis [17], Flemming’s master’s thesis [54] and empirical evaluations by Hogan et al. [72, 73]. The survey provides formal definitions of data quality dimensions and metrics as well as evaluates existing tools for (semi-)automatic measurement of linked data quality.

The current paper assumes that the quality of a linked dataset is influenced by the mapping that produces it and thus categorizes the metrics from the survey from the perspective of the RDB2RDF mapping. We select those metrics that are related to the mapping process and adapt them to the RDB2RDF domain.

A.4 Quality Requirements

This section gives a detailed overview of the proposed quality requirements for RDB2RDF mappings, ways of measuring them (metrics) and guidelines for improving mappings w.r.t. these metrics. Our proposed system incorporates four quality dimensions with 14 objective metrics overall. For assessing the overall quality of a mapping, one would, in practice, assign *weights* to the metrics. Their choice depends on the goal of the mapping process. For example, when the goal of the mapping is to accurately represent the relational data, the metrics in the “faithfulness of the output” dimension should be assigned the highest weight.

A.4.1 Quality of the mapping implementation and representation

The requirement of mapping quality implementation and representation combines the requirements for resultant data accessibility and standard compliance of the mapping representation.

Data accessibility. Data accessibility describes how the result of the mapping is accessed. This metric is also known in the literature as “access paradigm”, “mapping implementation” or “data exposition” [139]. There are two possibilities: (i) Extract Transform Load (ETL) and (ii) on-demand mapping. According to [50], ETL means physically storing triples produced from relational data in an RDF store.

Requirement	Description	Measure
Quality of the mapping implementation and representation		
Data accessibility	Describes how the mapping result can be accessed.	ETL/on-demand/both
Standard compliance	Characterizes if the mapping representation is standard compliant.	boolean
Faithfulness of the output		
Coverage	Characterizes the mapping completeness	percentage of DB columns mapped
Accuracy of data representation	Characterizes the mapping correctness	percentage of correctly mapped DB columns
Incorporation of domain semantics	Shows level of domain semantics incorporation	percentage of properties that link to the results of SQL queries
Quality of the output		
Simplicity	Shows the simplicity of SPARQL queries returning the frequently demanded values	percentage of complex SQL queries results integrated into the mapping
Data quality	Characterizes the quality of output data	aggregation of linked data quality metrics (→ Table A.2)
Data integration	Characterizes the interlinking degree of the output data	percentage of external instances integrated into the resultant dataset
Interoperability		
Reuse of existing ontologies	Shows the amount of reused vocabulary elements	percentage of reused properties and classes
Quality of reused vocabulary elements	Characterizes the quality of chosen for reuse properties and classes	accumulated quality and popularity of reused vocabulary elements
Accuracy of reused properties	Characterizes the accuracy of properties reuse	percentage of accurately reused properties
Accuracy of reused classes	Characterizes the accuracy of classes reuse	percentage of accurately reused classes
Quality of declared classes/properties	Shows the quality of ontology documentation	accumulated quality of declared classes/properties

Table A.1: Summary table of proposed metrics system

The disadvantage of ETL is that that, whenever the RDB is updated, you have to re-run the entire ETL process, even if just one RDB record has changed, carrying out an often redundant synchronization process. However, in the ETL case nothing more than the RDF store is needed to answer a query.

On-demand mapping is realized by translating a SPARQL query into one or more SQL queries at query-time, evaluating these against (a set of) unmodified relational database(s) and constructing a SPARQL result set from the result sets of such SQL queries. In contrast to the ETL implementation, on-demand mapping requires more resources for processing each query. However, the on-demand implementation does not face the synchronization issue and does not replicate the data. In the light of these advantages and disadvantages, we claim that the best solution is to implement a mapping with both data access approaches. Thus, the index of implementation takes a value equal to 1, if both implementations are present and 0.5 otherwise.

Standard compliance. An RDB2RDF mapping can either be represented as a set of customized rules or through a generic set of rules (as defined by the W3C Direct Mapping standard [8, 16]). Often, the output of a Direct Mapping may not be useful, that is, it may not adequately take the structural/semantic complexity of the database schema into account. Thus, while the applicability of a Direct Mapping satisfies the requirements for the “standard representation” metric, not using the customized rules will lead to a loss of points w.r.t. other metrics.

Languages for representing customized mappings have been surveyed in [139]. Until recently, no standard representation language for RDB2RDF mappings existed, however, as of 2012, R2RML [37] has been released as a W3C recommendation. As it is the only mapping language that has been standardized to date, we do not consider other (proprietary) formats reasonable.

Thus, we propose that if there is no material representation of a mapping available, it should be assumed that a Direct Mapping is carried out. We define the “standard representation” metric to be one if the mapping is represented in *R2RML* or if there is no representation (and therefore a Direct Mapping is applicable), and zero otherwise.

A.4.2 Faithfulness of the output

In terms of RDB2RDF mapping we define the faithfulness of output as an abstract measure of similarity between the source data and resultant dataset.

Coverage. This metric indicates the ratio of the database columns mapped to the RDF. In general, a high coverage increases the faithfulness of the output data. However, reaching the highest level of coverage may conflict with privacy restrictions. Therefore, great care should be taken when mapping any kind of personal (sensitive) data that might be linked with other data sources or ontologies. It may even be illegal to create or to process such linked information in countries with data protection laws without having a valid legal reason for such processing [40].

Moreover, application-specific data such as statistics of usage or service tables often are not included into the mapping due to its application-related nature. Thus, in real-world applications, the coverage metric never reaches one.

Accuracy of data representation. When mapping relational data to RDF, the accuracy of data representation is tightly connected to data formatting issues. For example, differences in the representation of numeric values between the database and RDF can cause inaccuracies in the output data. Inaccurate dealing with not-Latin characters leads to loss of data meaning. We propose the following algorithm to evaluate the accuracy index of the mapping:

1. for each mapped database column containing numeric data, compute the average value of the numbers in this column;
2. find an average value for the corresponding property in the RDF output, and compare it to the average computed in step 1;

3. add a point for each coinciding average (considering the statistical error)
4. for each mapped database column containing literal data, analyze the readability of the corresponding properties; add a point for each completely readable property;
5. calculate a sum of the points obtained in steps 3 and 4 and divide the sum on the total number of columns mapped.

Incorporating domain semantics. The incorporation of domain semantics is one of the crucial aspects that indicate the quality difference between direct and customized mappings. Measuring this requirement helps to estimate the direct benefit of mapping customization. Generally, a direct mapping does not incorporate domain semantics; in this case the metric takes a value of zero. The extent of domain semantics incorporation can be measured by counting number of *class properties* which take values from any database table *other than* the table that corresponds to a class. The metric is then calculated as the ratio between this number and the count of properties summed up over all classes in the mapping.

To improve the mapping w.r.t. this metric, implicit relations between the data in the database should be explicitly modelled in the mapping. This process also increases the simplicity of the mapping, as it requires integration of the SQL query results into the mapping, thereby simplifying (future) SPARQL queries for obtaining these values.

A.4.3 Quality of the output

Quality of the output combines requirements of the output data quality in aspects of simplicity of usage, level of interlinking and objective data quality metrics.

Simplicity. One often has the task of mapping highly complex data models to RDF. In that case, Direct Mapping produces an RDF output that may be difficult to use, especially when considering the limitations of the SPARQL query language. Thus, useful information from the source database may lose its value in the RDF dataset that results from mapping. Metrics for the quality of the output should therefore take the simplicity of the RDF dataset into account. By simplicity of RDF data, we mean the simplicity of *operating* on it. In other words, the simplicity of data determines the simplicity (or length in terms of abstract syntax) of the SPARQL queries returning frequently demanded values. To increase the simplicity, the mapping should aim to produce a dataset, that can be queried for frequently demanded values by relatively simple SPARQL constructions. To do this, the *frequently demanded values returning by the complex SQL queries* should be integrated into the mapping. Such preparations not only increase the simplicity of the output, but also *incorporate domain semantics* not presented in the relational database explicitly (cf. Section A.4.2).

To calculate the index of simplicity, first a list of complex SQL queries should be assembled. We propose to consider a query “complex” in the following cases:

- a query joins at least two tables or views,
- a query supposes recursive computations over one table or view.

After the list is assembled, frequently demanded values, returned by the queries should be selected. This selection is subjective and should be carried out by a group of developers and administrators of the project. The simplicity metric is then calculated as the percentage of *frequently demanded values returned by complicated SQL queries* that have been integrated into the mapping.

Data quality. The quality of a mapping can partly be assessed by evaluating the quality of the output RDF data. Section A.3 points to approaches for assessing data quality. However, when evaluating the quality of a mapping, not all aspects of output data quality need to be taken into account. This is due to the fact that not all aspects of linked data quality are affected by the mapping stage.

Requirements for linked data quality and their metrics	
1. Influenced by mapping	
<ul style="list-style-type: none"> • Completeness: influenced by coverage • Validity-of-documents: influenced by quality and accuracy of reused and newly determined classes • Interlinking: influenced by data integration • Availability: influenced by dereferenceability of reused and newly defined classes • Consistency: influenced by reuse 	
2. Depend on mapping	
<ul style="list-style-type: none"> • Provenance: indication of metadata about a dataset • Verifiability: verifying publisher information, authenticity of the dataset, usage of digital signatures • Licensing: machine-readable/human-readable license, permissions to use the dataset, attribution, Copyleft or ShareAlike • Validity-of-documents: no syntax errors • Consistency: entities as members of disjoint classes, usage of homogeneous data-types, misplaced classes or properties, misuse of <code>owl:datatypeProperty</code> or <code>owl:objectProperty</code>, bogus <code>owl:Inverse-FunctionalProperty</code> values, ontology hijacking • Conciseness: redundant properties/instances, not-unique values for functional properties, not-unique annotations • Performance: no usage of slash-URIs, no use of prolix RDF features • Understandability: human-readable labelling of classes, properties and entities by providing <code>rdfs:label</code>, human readable metadata • Interpretability: misinterpretation of missing values, atypical use of collections, containers and reification Currency: currency of data source • Volatility: no timestamp associated with the source 	
	3. Depend on relational data
	<ul style="list-style-type: none"> • Completeness: values for a property are not missing • Amount-of-data: see [161] • Provenance: trustworthiness of RDF statements, trust of an entity, trust between two entities, trust from users, assigning trust values to data/-sources/rules, trust value for data • Believability: see [161] • Accuracy: see [161] • Consistency: no stating of inconsistent property values for entities, literals incompatible with datatype range • Interpretability: interpretability of data • Versatility: provision of the data in various languages
	4. Depend on publishing
	<ul style="list-style-type: none"> • Availability: accessibility of the server, accessibility of the SPARQL end-point, accessibility of the RDF dumps, no structured data available, no dereferenced back-links • Performance: see [161] • Security: see [161] • Response-time: see [161] • Conciseness: keeping URIs short • Understandability: indication of one or more exemplary URIs, indication of a regular expression that matches the URIs of a dataset, indication of an exemplary SPARQL query, indication of the vocabularies used in the dataset, provision of message boards and mailing lists • Versatility: provision of the data in different serialization formats, application of content negotiation • Currency: see [161] • Timeliness: see [161]

Table A.2: Data quality metrics divided into four groups according to the stage of mapping creation, on which they can be influenced.

We confine ourselves to the objective requirements discussed in [161] and divide them into four groups (cf. [Table A.2](#)):

- Requirements that are influenced by mapping quality implicitly. Increasing the quality of the mapping in turn increases the quality of data.
- Requirements that are influenced by mapping explicitly. The proposed data quality index discussed below aggregates these metrics.
- Requirements that depend on the quality of data stored in the database. We consider these metrics to be out of the scope of this paper.
- Requirements that depend on the quality of linked data publishing. We do consider these metrics to be out of the scope of this paper.

The metrics and methods of measuring them are discussed in detail in [161]. To evaluate the quality of mapping with regard to output data quality we propose to aggregate only the metrics explicitly depending on the mapping. There are two possible strategies for calculating our overall data quality metric based on the individual metrics: assigning the same weight to all individual metrics, or assigning different weights to different dimensions or even to individual metrics.

Data integration. The key advantage of Linked Data is its ability to effectively integrate data from disparate, heterogeneous data sources. In most cases, the output dataset resulting from a mapping can be linked to existing datasets via explicitly modelled relationships between entities. As mentioned in [128], the use of domain ontologies along with user defined inference rules for reconciling heterogeneity between multiple RDB sources, is an effective integration approach for creating an RDF. A simple metric for data integration can be defined as the number of external datasets linked to the one evaluated. However, this metric cannot be mapped to a value in the interval $[0, 1]$ in an obvious way, as needed for computing the overall mapping quality from the values of the individual metrics. Thus, we propose an alternative measure to evaluate the data integration level: index of data integration. It can be calculated as the ratio between the number of external instances integrated in the resulting RDF dataset and the total number of instances.

A.4.4 Interoperability

The interoperability requirement tackles the aspect of a mapping to produce interoperable data, which can be easily linked to and integrated with other ontologies and datasets.

Reuse of existing ontologies. The reuse of existing ontology elements (i.e. classes and properties), increases the interoperability of the output dataset. Additionally, this reuse prevents one from having to introduce new elements. This requirement can be measured using two metrics: the ratio of reused properties and the ratio of reused classes. However, not only the quantity, but also the *quality* of reused properties and classes is important. This aspect is covered by the next metric.

Quality of reused vocabulary elements. Measuring the quality of reused properties or classes is not trivial. As a large number of vocabularies has been published on the Web, many vocabulary elements are repeatedly declared instead of being reused. We claim that the best choice of a class or property to reuse depends on two parameters: (i) the quality of the class or property itself and (ii) the frequency of its usage in LOD in comparison with semantically similar alternatives. Thus, to measure the quality of each individual property or class we propose the following workflow:

1. Measure the quality of the chosen vocabulary element (class/property)
2. Find alternatives to the chosen element and measure their quality
3. Compare the quality metrics of the chosen element and the best alternatives

We propose to compute the quality index (i_{qual}) of the chosen vocabulary element as the product of the following two indexes: the index of documentation quality (i_{doc}) and the index of popularity (i_{pop}). The proposed index of documentation quality is computed differently for classes and properties. Both for classes and properties, the index of documentation quality takes into account dereferenceability as well as documentation by `rdfs:label` and `rdfs:comment` (preferably in multiple languages). If the class or property is explicitly deprecated, i.e. its `owl:deprecated` annotation property equals `true`, we define its quality index as zero. Additionally, for classes, their relation to other classes should be indicated (`rdfs:subClassOf` or `owl:equivalentClass`).

For properties, `rdfs:range` and `rdfs:domain` should be defined. Based on the presence of the properties definitions mentioned above, a decision about quality of the documentation for a vocabulary element (possibly automated) should be made. The index of popularity aims to measure the frequency of usage of the class or property on the Web in comparison with semantically similar alternatives. This index can be taken from services such as Linked Open Vocabularies³.

As an example, we evaluate the property `agrelon:hasChief`⁴ that links an *organization* to its *chief*. The property is dereferenceable, it has labels available in four languages; however, no `rdfs:range`, `rdfs:domain` or `rdfs:comment` is specified. Based on this, we define its i_{doc} as 0.5. For measuring the frequency of usage we type the term ‘chief’ into the search field of the Linked Open Vocabularies service. Looking through the results, we can conclude that on the Web of Data the term ‘chief’ is most commonly used to model a military commander and the chosen property `agrelon:hasChief` has a low popularity metric value of 0.271. Thus, the quality index for this evaluated property is $i_{pop} * i_{doc} = 0.271 \cdot 0.5 = 0.1355$.

The next task is to find the most commonly used term for the evaluated property. The list of possible substitutes will include properties matching ‘chief’, ‘leader’, ‘head’, or ‘boss’. For each item on the list the property with the highest i_{pop} and satisfying the following requirements, should be found:

- `rdfs:comment`, `rdfs:domain` and `rdfs:range` statements should correspond to the domain model
- depending on the application settings, special requirements such as presence of an inverse property in the same vocabulary may need to be satisfied.

In our case, the metrics for the best candidates are represented in Section A.3. Due to space limitations, we only take into account the first two synonyms from that list, as the two other ones do not have matches that satisfy the requirements given above. According to the investigation, the `agrelon:hasChief` property is not the best possible alternative for linking an organization to its chief. The `swpo:hasLeader` property should be used instead, as its quality index is $i_{pop} * i_{doc} = 0.682 \cdot 0.8 = 0.5456$.

	‘chief’	‘leader’
best candidate	<code>agrelon:hasChief</code>	<code>swpo:hasLeader</code>
i_{pop}	0.271	0.682
i_{doc}	0.5	0.8
i_{qual}	0.1355	0.5456

Table A.3: Metrics for semantically identical properties to link an organization and its chief

The last step is to compare the quality of the chosen property and the best possible alternatives. The resultant score for the chosen property is calculated as the ratio between its quality index and the quality index of the best possible alternative: $i_{qual_{chosen}} / i_{qual_{best}}$ where $i_{qual_{best}}$ is the quality index of the best alternative and $i_{qual_{chosen}}$ is the quality index of the evaluated property. In our case, choosing the `agrelon:hasChief` property adds $0.1355 / 0.5456 = 0.248$ points to the interoperability metric. Thus, the overall value of the *quality of reused vocabulary elements* metric is defined as the ratio between the sum of the quality indexes of all reused vocabulary elements and the number of reused vocabulary elements.

³ <http://lov.okfn.org>

⁴ full URI: <http://d-nb.info/standards/elementset/agrelon.owl#hasChief>

Accuracy of reused properties. We determine two requirements for accurate reuse of existing properties: (i) respecting the property definition and (ii) unambiguous meaning of the property. Respecting the property definition means that the domain model must be consistent with the definition of the property, namely its `rdfs:range`, `rdfs:domain` and `rdfs:comment` relations. Unambiguous meaning is important when similar relations are presented in the domain model. For example, consider the domain model of our SlideWiki OpenCourseWare authoring platform⁵ and the following relations between decks (collections of slides):

- Deck1 `sw:isTranslationOf` Deck2
- Deck2 `sw:isRevisionOf` Deck3
- Deck3 `sw:isVersionOf` Deck4

Here, Deck2 results from *translating* Deck1 into a different language, Deck3 is a result of *editing* Deck2, and Deck4 results from *moving* Deck3 into another e-learning platform (for example, Moodle⁶). If we decide to reuse the property `dcterms:isVersionOf` instead of one of these three properties from SlideWiki's domain-specific vocabulary, the meaning of the property will be ambiguous. A better solution in this case is to declare all three new properties to be sub-properties of `dcterms:isVersionOf`.

The index of accuracy of reused properties can be calculated as the number of accurately reused properties (i.e. the properties, that satisfy both requirements) in relation to the total number of reused properties in the dataset.

Accuracy of reused classes. An important aspect of reusing classes is the semantic *compatibility*. By this we mean the compatibility of the meaning of a reused class and a domain object declared to be an instance of the class. The index of accuracy of reused classes is defined as the ratio between semantically compatible reused classes and the total number of reused classes. We illustrate the issue with an example below.

Let us assume that we need to link a deck of slides to its CSS style. The property `oa:styledBy` from the *Open Annotation Data Model ontology*⁷ could be chosen to model the link. However, its domain is `oa:Annotation`, which is not compatible with the `sw:Deck` class in terms of its intended semantic. If such a statement occurs within the mapping, the reused class `oa:Annotation` should be considered as incompatible for reuse.

To measure the requirement we propose to calculate the index of the accuracy of reused classes. It can be calculated as the number of reused classes semantically compatible with domain objects in relation to the total number of reused classes.

Quality of declared classes/properties. In order to obtain high quality, the classes/properties whose *definition* is *introduced* by the mapping should meet the requirements of documentation quality analogically to reused vocabulary elements (see Section A.4.4). Thus, the proposed index of quality of declared vocabulary elements is computed separately for class and properties and accumulates the documentation quality score in relation to the total number of properties/classes declared.

A.5 Evaluation

In order to evaluate how well the proposed methodology agrees with real-life common practices, we conducted a survey using a questionnaire⁸. The survey was formed of statements, one per requirement from Section A.4. Each requirement was phrased in the style of a practical guideline, in that complying with this guideline when designing a mapping would help the mapping to obtain the highest possible

⁵ <http://www.slidewiki.org>

⁶ <http://moodle.org>

⁷ <http://www.w3.org/ns/oa>

⁸ <http://slidewiki.org/application/questionnaire.php>

score for the corresponding metric. For example, the requirement “accuracy of reused properties” (Section A.4.4), evaluated by combination of two metrics, was divided into two statements: “The properties chosen for reuse should be available, dereferenceable and have unambiguous meaning in the application domain.” and “When choosing the properties for reuse, *rdfs:domain*, *rdfs:range* and *rdfs:comment* must be taken into account.” For each statement the experts had to express to what extent they agreed with it. Thus, the agreement of a majority of experts with a statement would prove the relevancy of both the requirement and the proposed way of its measuring. We did not include requirements for *output data quality*, *coverage* and *accuracy of data* in the survey, as their metrics had been proposed by prior research. In addition to the opinions about the statements we collected open feedback.

Requirement	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	Mode
Data accessibility	4	2	5	5	4	5	5	4	4	2	4	2	5	5, 4
Representation	2	2	4	5	1	5	4	4	3	4	3	2	5	4
Incorporation of domain semantics	4	3	3	4	3	3	3	4	2	5	5	3	5	3
Simplicity	3	3	3	4	3	3	4	4	5	2	1	4	4	4, 3
Data integration	4	3	4	3	5	2	5	4	3	4	2	4	3	4
Reuse of existing ontologies	5	3	4	3	5	4	5	4	5	2	2	4	2	5
Quality of reused properties (frequency)	4	4	4	3	4	3	4	4	3	2	2	4	2	4
Quality of reused properties (dereferenceability)	5	1	3	4	5	3	5	4	5	4	4	4	4	4
Accuracy of reused properties	4	4	3	4	5	4	5	4	4	4	3	4	5	4
Accuracy of reused classes	5	4	4	3	4	4	4	3	4	4	3	4	5	4
Quality of declared classes/properties	4	3	3	4	2	4	5	4	5	2	2	4	4	4

Table A.4: Results of survey on degree of agreement with the key concepts of the proposed methodology. Numbers represent the degree of agreement from *absolutely agree* (5) to *absolutely disagree* (1). The colors indicate the self-estimated level of expertise in the RDB2RDF domain from *expert* (darkest) to *experienced* (lightest).

The survey was announced to the RDB2RDF community via mailing lists and personal e-mails. We received 13 individual responses. All participants assessed their level of RDB2RDF expertise as either *experienced* or *expert* (the two highest out of four possible levels). They could decide either to stay incognito or to fill in their name, affiliation and institution. Section A.4 shows the results.

Due to the high level of participants’ experience in the domain, we chose not to leverage the individual opinions by calculating their means or medians. Instead, we base our analysis on the *mode value*. We did not calculate separate mode values for responses from participants with different experience levels (“experienced” vs. “expert”), as we do not consider the difference between these two groups sufficiently significant to influence the responses (however, we indicate the level of experience in the result table).

As Section A.4 shows, the experts approved most of the requirements. The ones that the experts were doubtful about were the *incorporation of domain semantics* and *simplicity* requirements. This outcome can, however, be explained by the circumstance that these two requirements are difficult to explain in a short statement. On the other hand, there were experts who accepted both these requirements, even with a value of *absolutely agree*. Thus, we consider the low overall agreement with these two requirements to result from incomplete understanding of the statements, as the requirements and their metrics are not trivial. Finally, within the open feedback no participant suggested further objective requirements (besides ones aggregated in the data quality requirement), which provides evidence of the completeness of our system, at least in view of today’s state-of-the-art.

A.6 Conclusion

In this paper, we proposed a methodology to evaluate the quality of mappings between relational and linked data. In particular, we proposed a set of 14 requirements for mapping quality and described ways of measuring them. We believe that the most of the measures can be done (semi-)automatically and will attempt to prove that in our future work. As we show in the paper, the proposed system can not only be used to evaluate the mappings, but also as a guideline to increase the quality of the mapping. Additionally, we evaluated the relevance of our proposed set of requirements by conducting a survey. A total of 13 experienced individuals, mainly from the RDB2RDF community, participated in our survey. The analysis of their responses allows us to claim that the community accepts our requirements and their metrics.

Acknowledgements.

We would like to thank the anonymous experts as well as Richard Cyganiak (INSIGHT@NUI Galway), Ravi Kumar (TCS), Mr. Spenser Kao (Bureau of Meteorology, Australia), Dr. Nikolaos Konstantinou (National Technical University of Athens), Freddy Priyatna (Ontology Engineering Group), Juan Squeda (University of Texas at Austin), Tim Lebo (Tetherless World/RPI), Oscar Corcho (UPM), Dr. Pedro Szekely and Jose Luis Ambite (University of Southern California). We are grateful to Mr. Nicolas Van Peteghem for providing us with one of the test databases.

APPENDIX B

Research Paper: Balanced scoring method for multiple-mark questions

B.1 Abstract

Advantages and disadvantages of a learning assessment based on multiple-choice questions (MCQs) are a long and widely discussed issue in the scientific community. However, in practice this type of questions is very popular due to the possibility of automatic evaluation and scoring. Consequently, an important research question is to exploit the strengths and mitigate the weaknesses of MCQs. In this work we discuss one particularly important issue of MCQs, namely methods for scoring results in the case, when the MCQ has several correct alternatives (multiple-mark questions, MMQs). We propose a general approach and mathematical model to score MMQs, that aims at recognizing guessing while at the same time resulting in a balanced score. In our approach conventional MCQs are viewed as a particular case of multiple-mark questions, thus, the formulas can be applied to tests mixing MCQs and MMQs. The rational of our approach is that scoring should be based on the guessing level of the question. Our approach can be added as an option, or even as a replacement for manual penalization. We show that our scoring method outperforms existing methods and demonstrate that with synthetic and real experiments.

B.2 Introduction

Advantages and disadvantages of a learning assessment based on multiple-choice questions (MCQs) are a long and widely discussed issue in the scientific community. However, in practice this type of questions is very popular due to the possibility of automatic evaluation and scoring [53]. Consequently, an important research question is to exploit the strengths and mitigate the weaknesses of MCQs. Some systems (e.g. Moodle¹) allow teachers to create MCQs with multiple correct options. This type of questions we will call multiple-mark questions (MMQs), to distinguish them from the conventional MCQs, where there is always only one correct option. Multiple-mark questions were already recommended by Cronbach [34]. Other research [120, 115, 74] considers MMQs to be more reliable, when compare them with conventional MCQs. However, even though the advantages of MMQs are meanwhile widely accepted, up to our knowledge there are no balanced methods for scoring multiple-mark questions available to date.

One possible approach to score the MMQs is to use dichotomous scoring system. The dichotomous scoring awards the constant amount of points, when the question is answered correctly and zero points

¹ <https://moodle.org/>

in a case of *any* mistake. However, the partial scoring is preferable to the dichotomous, especially in case of MMQs. [120, 82, 14, 15]

The second possible approach is to use the methods, developed for scoring the multiple true-false questions (MTFs). However, despite the possibility to convert the MMQs into MTFs, the studies [34, 47] show the differences between two formats. Moreover, the researches mentioned above named the following disadvantages of MTF questions compared to MMQs:

- The multiple true-false format "clouds" from the learners the possibility of marking several options as true.
- The level of reliability in multiple true-false questions is not equal for true and false answers.
- The multiple true-false format requires more resources to store the answers.

In the paper we show that the differences prevent the applying methods developed for the MTFs to the MMQs scoring.

Another possible approach is to use the penalties, similarly to the paper-based assessment where the teacher can analyze the student answers and decide how much points she deserves. The method was proposed by Serlin [134]. For example, in Moodle a teacher has to determine what penalty applies for choosing each distractor. However, this work is an additional, unpopular burden for teachers, since not required in paper-based tests. Instead of asking the teacher, some systems calculate the penalties automatically. However, computer-based assessment opens additional possibilities to guess, for example choosing all options. Often the scoring algorithms do not take into account such ways of guessing.

Consequently, we are facing the challenge to find a scoring method, that is able to recognize and properly penalize guessing. Previously proposed algorithms suffer from imbalance and skewness as we show in section B.4.

The task to find the scoring method can be divided into two steps:

1. Find a method to determine points for the correctly marked options.
2. Find a method to determine the penalty for the incorrectly marked options.

For the first part a reasonable approach was proposed by Ripkey [120]. Thus our research aims to provide a method for the second part (determining penalties). We propose a general approach and a mathematical model, that takes into account the most common ways of guessing and behaves balanced at the same time.

Our concept is based on the assumption, that scoring can be based on the guessing level of the question. Each question is associated with a difficulty to guess a (partially) correct answer. To accommodate the difficulty level of guessing in the scoring method, we propose to determine the penalty only when a student marks more options, than the actual number of correct ones. We argue that our approach can be added as an option, or even as a replacement of manual designation of penalties. We claim that our algorithm behaves better, than existing ones and prove that with both synthetic and real experiments. In our approach conventional MCQs are viewed as a particular case of multiple-mark questions, thus, the formulas can be applied to the tests mixed of MCQs and MMQs. As the scoring of conventional MCQs is a trivial task, we do not consider such type of questions in our experiments.

The paper is structured as follows: First, we present the terminology we use. Then we discuss existing algorithms for scoring MMQs, which we have found in the research literature and real applications. After that we describe our approach on conceptual and mathematical levels. Finally we show and discuss the results of synthetic and real-life experiments.

B.3 Terminology

In the following we define the key concepts building the basis for our MMQ scoring method:

Dichotomous scoring – the concept of scoring the results, that allows users to get either the full amount of points or zero in a case of any mistake;

Partial scoring – the concept of scoring the results in a way that allows users to obtain some points for a question, which they answered only partially correct;

Difficulty – a difficulty weight of the question in the questionnaire in the interval $(0, 1]$. The difficulty can be determined automatically and dynamically based on prior scoring. In our implementation, for example, difficulty is dynamically updated after one student provided an answer, according with the formula:

$$d' = \frac{\text{incorr}}{\text{all}}$$

In a case of dichotomous scoring, the values of *incorr* and *all* mean, respectively, the accumulated number of *incorrect* and *all* responses on the question by any user. In a case of partial scoring, the definition of *incorr* changes as follows:

$$\text{incorr} = \sum_i 1 - d_i$$

where i is a counter from 1 to the number of attempts for the question and D_i is the difficulty, that the question had at the moment, when the i^{th} -attempt was made. After the difficulty is determined, it is scaled to the interval $(1, d_{max}]$, where d_{max} is the maximal difficulty, that a question can have.

$$d = f(d') = (d' * (d_{max} - 1) + 1$$

The scaling is performed for better usability. For example, d_{max} can be set to 10 to obtain a difficulty level between 1 and 10.

Guessing level – the theoretical probability to guess the correct answer from the list of options. In partial scoring, we determine the guessing level as the probability to obtain more than zero points.

Basic question points – an absolute value of points for the correctly checked options or the percentage of correctly checked options within all correct options. Basic points = $f(d)$.

Penalty – the value, that should be deducted from the basic points due to the logic of the applied algorithm. In our approach we propose, that penalty should be only deducted, when user checks more options, than the number of correct ones.

Total question points – the amount of points for the question, gained by the user after the deduction of penalty. Total question score = $f(p, s)$.

B.4 Related work

There are several existing platforms, that use multiple-mark type of questions as well as several approaches to score them. We collected such approaches to describe, discuss and compare them. Existing approaches for scoring the multiple-mark questions implement four base concepts. In the section we describe the basic ideas, advantages and disadvantages of these concepts.

B.4.1 Dichotomous scoring

This method is often used in paper-based questionnaires, where the good quality of questionnaires allows teacher to be more strict when score the results. In the case choosing a wrong option indicates, that a student hopes to guess the correct response as she does not know the material behind the question well. In e-based learning the quality of questionnaires is not perfect, especially in the systems with collaborative authoring. That is why the dichotomous scoring can punish the learners for the teachers mistakes too much. As the aim of questionnaires is not only to score the results, but to catch the gaps of knowledge, the scoring of partially correct responses shows the actual knowledge of the student better. Also, dichotomous scoring does not show the accurate progress of the student. However, when dealing with multiple-mark questions dichotomous scoring almost excludes the possibility of guessing, that is why we use it as a standard of reference when evaluating our approach with real users.

B.4.2 Morgan algorithm

One of the historically first methods for scoring the MMQs was described in the 1979 by Morgan [105]. In the accordance to the method, the scores are determined by the following algorithm:

1. for each option chosen which the setter also considers correct, the student scores +1.
2. for each option chosen which the setter considers to be incorrect, the student scores -1.
3. for each option not chosen no score, positive or negative, is recorded regardless of whether the setter considers the response to be correct or incorrect.

The algorithm can be improved by changing the constant 1 to dynamically determined amount of points:

1. for each option chosen which the setter also considers correct, the student scores $+(p_{max}/n)$, where n is a number of correct options
2. for each option chosen which the setter considers to be incorrect, the student scores $-(p_{max}/k)$, where k is a number of distractors.

We use this improved algorithm for our experiments. However, the experiments show a large dependence between number of options (correct and incorrect) and amount of penalty, that indicates the skewness of the method (see subsection B.6.1).

B.4.3 MTF scoring

Multiple-mark questions can be scored with the approaches developed for multiple true-false items. The base approach to score the MTF items is to determine, how close is the student response to the correct one. Tsai [150] evaluated six different implementations of the approach. Later his findings were confirmed by Itten [80]. Although both researches found partial crediting to be superior to dichotomous scoring in a case of MTFs, they do not consider any of the algorithms to be preferable. This fact allows us to use the most base of them for our experiments.

All the MTF scoring algorithms imply that any item has n options and a fully correct response is awarded with full amount of points p_{max} . If the user did not mark a correct option or marked a distractor, she is deducted with the penalty $s = p_{max}/n$ points. Thus a student receives points for not-choosing a distractor as well as for choosing a correct option. This point does not fit perfect to multiple-mark questions because of the differences between two types [115, 34, 57]. Our experiments (see subsection B.6.1) confirm the studies and show the skewness of the concept when deal with MMQs. The main problem of the MTF scoring method, when applied to MMQs, is that a user obtains points, even if she did not chose any options. Although the problem can be solved by creating an additional rule, the experiments show the further problems of the algorithm, when used for MMQ items.

B.4.4 Ripkey algorithm

Ripkey [120] suggested a simple partial crediting algorithm, that we named by the author. In the approach a fraction of one point depending on the total number of correct options is awarded for each correct option identified. The approach assumes no point deduction for wrong choices, but items with more options chosen than allowed are awarded zero points.

The Ripkey's research showed promising results in a real-life evaluation. However, later researches (e.g. Bauer [14]) notice the limitations of the Ripkey's study. The main issue in the Ripkey algorithm is the not well-balanced penalty. Our experiments show that in many cases the algorithm penalizes so severely, that learners could consider it to be the dichotomous scoring. We aim to improve the Ripkey's algorithm by adding the mathematical approach for evaluating the size of penalty.

B.5 Balanced scoring method for MMQs

B.5.1 Concepts

As shown above, existing approaches do not solve the problem of scoring MMQs perfectly. Our concept is based on the assumption, that scoring can be based on the guessing level of the question. Thus, when a student marks all possible options, she increases the guessing level up to 1. In this case the student should obtain either the full amount of points (if all the options are considered to be correct by the teacher), or zero, if the question has at least one distractor. However, if a student did not mark any option, the score should be always zero, as we assume that all the questions have at least one correct option. Thus, the task is to find the correctness percentage of the response and decrease it with a penalty, if the guessing level was artificially increased by marking too many options.

Questions have the native level of guessing, and we propose to deduct the penalty only if after the student's response the guessing level increases. In other words, we determine the penalty only when a student marks more options, than the number of correct ones.

B.5.2 Mathematical model

In this section we present the mathematical model as well as an algorithm, that can be used for its implementation.

Assumptions and restrictions

We propose to use our approach only in systems, that comply with the following requirements for assessment items:

- all the item's options have the same weight;
- there is at least one correct option;
- there are no options excluding all other (e.g. "all above are correct")

Scoring the basic points

To score the basic points we use the approach, described by Ripkey. Below we present it mathematically in accordance with the following designations:

- $d \in \mathbb{R}, d \in (1..d_{max}]$ – difficulty of the current question, for our experiments we set $d_{max} = 5$
- $C \subseteq A$ – set of the *correct* options c_i for the current question, where A – set of the options a_j for the current question,
- $c_{max} = |C|, c_{max} \in \mathbb{N}$ – number of *correct* options for the current question
- C_{ch} – set of the *correctly checked* options
- $c_{ch} = |C_{ch}|, c_{ch} \in \mathbb{N}, c_{ch} \in [0, c_{max}]$ – number of *correctly checked* options for the current question
- $p_{max} = f(d) = d * K_{points}$ – maximal possible points for the current question, in our system we set $K_{points} = 1$
- p_c – points for the correctly checked option c . As we assume all the correct options have the equal weight,

$$\forall c \in C_{ch} | p_c = \frac{p_{max}}{c_{max}}$$

- $p \in \mathbb{R} \wedge p \in [0, p_{max}]$ – the basic points for the current question,

$$p = \sum_{c \in C_{ch}} p_c \Rightarrow$$

$$p = \sum_{c \in C_{ch}} \frac{p_{max}}{c_{max}} = \frac{p_{max}}{c_{max}} * c_{ch} = p_c * c_{ch}$$

Scoring of the penalty

Below we present our approach for scoring the penalty. We use the following designations:

- $a_{max} \in \mathbb{N}, a_{max} = |A|$ – number of options $a \in A$
- $Ch \subseteq A$ – set of *checked* options
- $ch = |Ch|, ch \in \mathbb{N}, ch \in [0, a_{max}]$ – number of checked options for the current question
- $b \in \mathbb{R}, b \in [0, 1]$ – basic level of guessing for the current question,

$$b = \frac{c_{max}}{a_{max}}$$

- $n \in \mathbb{R}, n \in [b, 1]$ – measure, that shows the possibility, that user tries to guess the correct response by choosing too much options; we do not evaluate it in the cases, when $n \leq b$,

$$n = \frac{ch}{a_{max}}$$

- s – penalty for the guessing,

$$s = n - b \Rightarrow s \in [0, 1 - b]$$

- $s_k \in [0, p_{max}]$ – the penalty, mapped to the maximal possible points.

A mapping function is calculated as follows:

$$f : s_k \rightarrow s$$

Given, $s_k \in [0, p_{max}]$ and $s \in [0, 1 - b]$, then

$$f : s_k \rightarrow s = f : [0, 1 - b] \rightarrow [0, p_{max}] \Rightarrow$$

$$s_k = f(s) = s * \frac{p_{max}}{1 - b} = (n - b) * \frac{p_{max}}{1 - b}$$

Scoring the total question score

The absolute score for the question is trivially determined as

$$T = f(p, s_k) = p - s_k$$

The percentage representation of the total score is determined as follows:

$$T\% = \frac{p - s_k}{p_{max}} * 100\%$$

B.6 Evaluation

B.6.1 Synthetic experiments

In the subsection we describe our experiments with synthetic data and compare the behavior of different methods. For shorter presentation, we use the following reductions:

- Dich. – dichotomous scoring;
- Balanced – the proposed balanced scoring method

We consider all the questions to have the difficulty $d = 1$, then the maximal possible points $p_{max} = 1$ as well.

Question 1 of 7				
An item with 5 options, 2 of which are correct				
Dich.	MTF	Morgan	Ripkey	Balanced
0	0.4	0	0	0

Table B.1: Comparison of the proposed approach with other existing approaches

Example 2 (Case: 5 options, 2 correct, 5 marked) *In the case the student chose all the options and should obtain zero points. However, we see that MTF method does not recognize this type of guessing and considers the questions to be answered partially correct, awarding the points for two correct options, that were marked.*

Question 2 of 7				
An item with 5 options, 2 of which are correct				
Dich.	MTF	Morgan	Ripkey	Balanced
0	0.6	0	0	0

Table B.2: Comparison of the proposed approach with other existing approaches

Example 3 (Case: 5 options, 2 correct, 0 marked) *The situation is opposite to the previous: in the case the student chose none of the options. As we assume that question must have at least one correct option, in case of not choosing any options a student also should obtain zero points. However, we see that MTF method awards the points for three distractors, that were not marked. Although the situation is absurd, we faced it within real learning platforms, for example within several on-line courses of the Stanford University².*

Two examples below are trivial and the problem could be solved by adding the rules. However, the MTF scoring also suffers from skewness, when applied to MMQs, as it is shown below.

Question 3 of 7					
An item with 6 options, 2 of which are correct					
<input type="checkbox"/> Incorrect					
<input type="checkbox"/> Incorrect					
<input type="checkbox"/> Incorrect					
<input type="checkbox"/> Incorrect					
<input checked="" type="checkbox"/> Correct					
<input type="checkbox"/> Correct					
Dich.	MTF	Morgan	Ripkey	Balanced	
0	0.83	0.5	0.5	0.5	

Table B.3: Comparison of the proposed approach with other existing approaches

Example 4 (Case: 6 options, 2 correct, 1 correct marked) *This case proves, that the MTF method has a dependency from a number of correct and incorrect options. Thus, in a case of 6 options two of which are correct, a student is awarded 0.833 points for choosing only one correct option. In a case of 5 options two of which are correct, she would be awarded 0.80 points for the same. Moreover, if she choose only one incorrect option in a case of 6 alternatives, she obtains 0.5 points; in a case of 5 options she will be awarded 0.4 for the same.*

Thus, our experiments prove, that multiple-mark questions can not be scored properly with the algorithms, developed for multiple true-false items. Moreover, a teacher should be careful when creating multiple true-false questions and create them in such a manner, that not-choosing a distractor deserves awarding. However, the MTF scoring is the only existing approach of partial scoring that can be used in a case, when a question does not have any correct options.

Example 5 (Case: 4 options, 2 correct, 1 correct and 1 incorrect marked) *This case illustrates the issues of using the Morgan algorithm. The Morgan algorithm deducts penalties for choosing the incorrect option, as well as the proposed approach. There are two main issues:*

- Does the response deserve penalty?
- If deserves, how big the penalty should be?

In that case we are facing the situation, that penalty has the same size, as the basic points, and the student is awarded zero. We consider the penalty to be needlessly high, especially because the penalty

² <http://online.stanford.edu/courses>

Question 4 of 7				
An item with 4 options, 2 of which are correct				
Dich.	MTF	Morgan	Ripkey	Balanced
0	0.5	0	0.5	0.5

Table B.4: Comparison of the proposed approach with other existing approaches

depends on the number of incorrect options. Thus, if the question has 3 incorrect options, choosing one of them would be fined on 0.33, and in case of 2 incorrect options, the penalty is 0.5. After recognizing behavior of the algorithm, students will mark only the options, they are sure in, because choosing an incorrect one may cost them a full amount of points, they collected with correct options.

The next two examples show mainly the differences between the proposed approach and Ripkey algorithm. Namely, we show the situations, when Ripkey algorithm awards zero points, while we consider that it should award more.

Question 5 of 7				
An item with 4 options, 2 of which are correct				
Dich.	MTF	Morgan	Ripkey	Balanced
0	0.75	0.5	0	0.5

Table B.5: Comparison of the proposed approach with other existing approaches

Example 6 (Case: 4 options, 2 correct, 2 correct and 1 incorrect marked) In this case the student chose more options, than the number of correct ones, and according to the Ripkey, the answer should be awarded zero. Our claim is, that until the student have not chosen all the options, she could have some points. However, choosing three of four options could mean a try of guessing. Although in this case the student gets the full amount of basic points, she is fined on a half of them.

Example 7 (Case: 5 options, 2 correct, 2 correct and 1 incorrect marked) The example shows the disadvantage of the Ripkey algorithm more clear. It is not clear for the student, why she was awarded zero points, as she did not try to guess and answered partially correct.

Example 8 (Case: 5 options, 3 correct, 2 correct and 1 incorrect marked) In that case balanced scoring and Ripkey algorithms behave the same, as none of them deducts a penalty.

Question 6 of 7				
An item with 5 options, 2 of which are correct				
Dich.	MTF	Morgan	Ripkey	Balanced
0	0.8	0.67	0	0.67

Table B.6: Comparison of the proposed approach with other existing approaches

Question 7 of 7				
An item with 5 options, 3 of which are correct				
Dich.	MTF	Morgan	Ripkey	Balanced
0	0.6	0.17	0.67	0.67

Table B.7: Comparison of the proposed approach with other existing approaches

B.7 Conclusions

In the paper we evaluate the existing approaches for scoring the multiple-mark questions and propose a new one. The proposed approach has a list of restrictions, however it has advantages when compare with the discussed approaches. One of the main advantages is its clearness for the students, that was proven by the user evaluation. Also, our approach is based on the mathematical model, it does not suffer from the skewness, as it has the same formula for all cases. At the same time, the proposed approach recognizes the attempts to guess the correct answer, for example choosing all the possible options. When compare with the existing approaches, the advantages of the proposed algorithm could be summarized as follows:

- The approach allows to score both multiple-mark and conventional multiple-choice questions.
- The approach is based on the partial scoring concept.
- The algorithm can be easily implemented, it is pure mathematical.
- The score does not highly depend on the amount of correct and incorrect options.
- The value of the penalty is in balance with the possibility, that the student is trying to guess.
- Due to the balance, the results are clear for the students.

However, we suppose our algorithm to be optional together with other discussed approaches. This is due to the fact, that teachers create questions in their own manner and should be able to choose an appropriate method to score the results. Also, the different situations require different levels of severity, and the proposed approach might be too lenient.

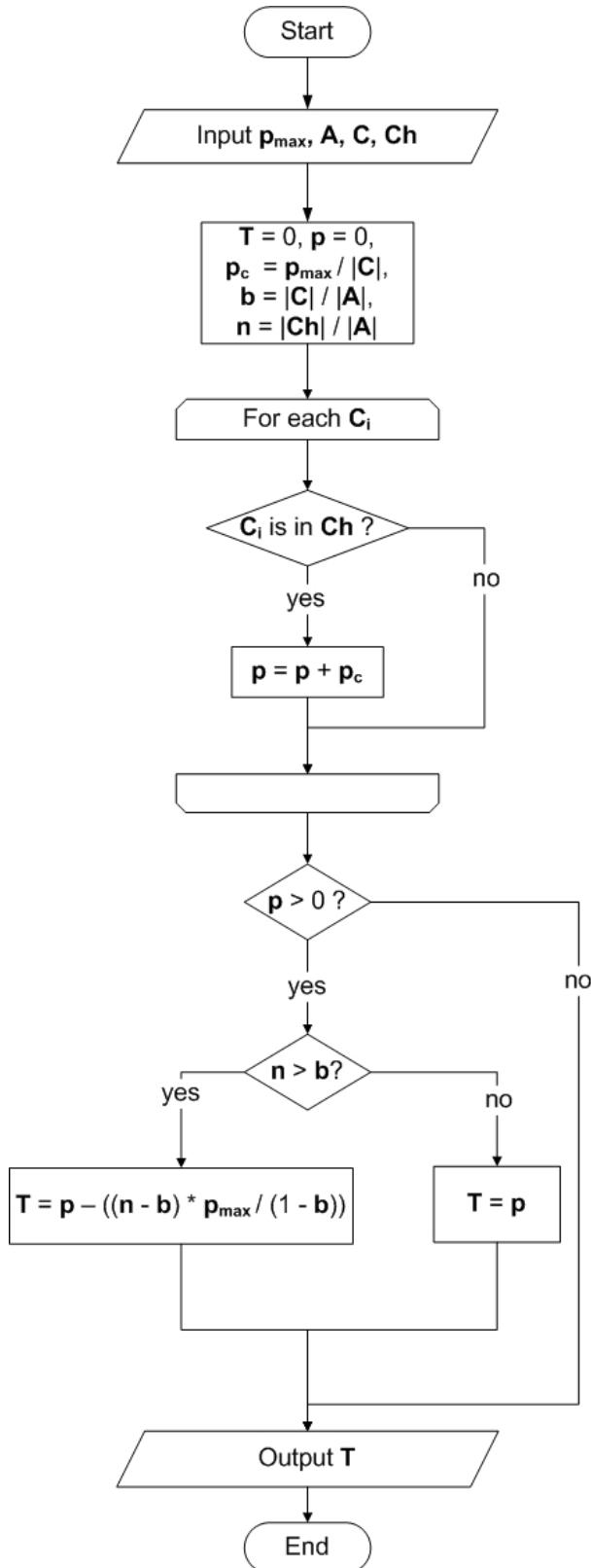


Figure B.1: Flow chart of the Balanced scoring algorithm.

APPENDIX C

RDB2RDF Mapping Documentation

C.1 R2RML mapping representation

```
prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix sw: <http://slidewiki.org/rdf/sw#>.
@prefix wa: <http://slidewiki.org/rdf/wa#>.
@prefix sa: <http://slidewiki.org/rdf(sa#)>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix cnt: <http://www.w3.org/2011/content#>.
@prefix oa: <http://www.w3.org/ns/oa#>.

<#DeckContainerIncludes>
    rr:logicalTable [ rr:tableName "deck_revision" ];
    rr:subjectMap [
        rr:template "http://slidewiki.org/rdf/deckContainer/{deck_id}";
        rr:class sw:DeckContainer;
    ];
    rr:predicateObjectMap [
        rr:predicate wa:includes;
        rr:objectMap [
            rr:template "http://slidewiki.org/rdf/deck/{id}";
        ];
    ];
</#DeckContainer>

<#DeckContainer>
    rr:logicalTable [ rr:tableName "deck" ];
    rr:subjectMap [
        rr:template "http://slidewiki.org/rdf/deckContainer/{id}";
        rr:class sw:DeckContainer;
    ];
    rr:predicateObjectMap [
        rr:predicate dcterms:created;
        rr:objectMap [ rr:column "timestamp" ];
    ];

```

```

rr : predicateObjectMap [
  rr : predicate dcterms : creator ;
  rr : objectMap [
    rr : parentTriplesMap <#User>;
    rr : joinCondition [
      rr : child "user_id";
      rr : parent "id";
    ];
    ];
  ];
  rr : predicateObjectMap [
    rr : predicate wa : translationOf ;
    rr : objectMap [
      rr : template "http:// slidewiki.org/rdf/deckContainer/{translated}";
    ];
  ].

<#SlideContainerIncludes>
  rr : logicalTable [ rr : tableName "slide_revision" ];
  rr : subjectMap [
    rr : template "http:// slidewiki.org/rdf/slidesContainer/{slide}";
    rr : class sw : SlideContainer;
  ];
  rr : predicateObjectMap [
    rr : predicate wa : includes ;
    rr : objectMap [
      rr : template "http:// slidewiki.org/rdf/slides/{id}";
    ];
  ].

<#SlideContainer>
  rr : logicalTable [ rr : tableName "slide" ];
  rr : subjectMap [
    rr : template "http:// slidewiki.org/rdf/slidesContainer/{id}";
    rr : class sw : SlideContainer;
  ];
  rr : predicateObjectMap [
    rr : predicate dcterms : creator ;
    rr : objectMap [
      rr : parentTriplesMap <#User>;
      rr : joinCondition [
        rr : child "user_id";
        rr : parent "id";
      ];
      ];
    ];
  ];
  rr : predicateObjectMap [
    rr : predicate wa : translationOf ;
    rr : objectMap [
      rr : template "http:// slidewiki.org/rdf/slidesContainer/{translated}";
    ];
  ];
  rr : predicateObjectMap [

```

```

rr : predicate dcterms:source ;
rr : objectMap [ rr : column "description" ] ;
] .
```



```

<#TagDeck>
    rr : logicalTable [ rr : sqlQuery """
        SELECT *
        FROM tag WHERE tag.item_type='deck' """
    ];
    rr : subjectMap [
        rr : template "http://slidewiki.org/rdf/deck/{item_id}";
        rr : class sw:Deck;
    ];
    rr : predicateObjectMap [
        rr : predicate dcterms:subject;
        rr : objectMap [ rr : column "tag" ];
    ].
```



```

<#DeckPartOf>
    rr : logicalTable [ rr : sqlQuery """
        SELECT *
        FROM deck_content WHERE deck_content.item_type='deck' """
    ];
    rr : subjectMap [
        rr : template "http://slidewiki.org/rdf/deck/{item_id}";
        rr : class sw:Deck;
    ];
    rr : predicateObjectMap [
        rr : predicate dcterms:isPartOf;
        rr : objectMap [
            rr : template "http://slidewiki.org/rdf/deck/{deck_revision_id}"
        ];
    ].
```



```

<#SlidePartOf>
    rr : logicalTable [ rr : sqlQuery """
        SELECT *
        FROM deck_content WHERE deck_content.item_type='slide' """
    ];
    rr : subjectMap [
        rr : template "http://slidewiki.org/rdf/slides/{item_id}";
        rr : class sw:Slide;
    ];
    rr : predicateObjectMap [
        rr : predicate dcterms:isPartOf;
        rr : objectMap [
            rr : template "http://slidewiki.org/rdf/deck/{deck_revision_id}";
        ];
    ].
```

```

<#DeckHasPartDecks>
    rr:logicalTable [ rr:sqlQuery """
        SELECT *
        FROM deck_content WHERE deck_content.item_type='deck' """
    ];
    rr:subjectMap [
        rr:template "http://slidewiki.org/rdf/deck/{deck_revision_id}";
        rr:class sw:Deck;
    ];
    rr:predicateObjectMap [
        rr:predicate dcterms:hasPart;
        rr:objectMap [
            rr:template "http://slidewiki.org/rdf/deck/{item_id}";
        ];
    ].

<#DeckHasPartSlides>
    rr:logicalTable [ rr:sqlQuery """
        SELECT *
        FROM deck_content WHERE deck_content.item_type='slide' """
    ];
    rr:subjectMap [
        rr:template "http://slidewiki.org/rdf/deck/{deck_revision_id}";
        rr:class sw:Deck;
    ];
    rr:predicateObjectMap [
        rr:predicate dcterms:hasPart;
        rr:objectMap [
            rr:template "http://slidewiki.org/rdf/slides/{item_id}";
        ];
    ].

<#DeckContributor>
    rr:logicalTable [ rr:sqlQuery """
        SELECT deck_revision.id, deck_revision.user_id as deck_user, user_group.user_id
        FROM deck_revision
        JOIN user_group ON deck_revision.id = user_group.deck_revision_id
        WHERE user_group.user_id != deck_revision.user_id """
    ];
    rr:subjectMap [
        rr:template "http://slidewiki.org/rdf/deck/{id}";
        rr:class sw:Deck;
    ];
    rr:predicateObjectMap [
        rr:predicate dcterms:contributor;
        rr:objectMap [
            rr:parentTriplesMap <#User>;
            rr:joinCondition [
                rr:child "user_user";
                rr:parent "id";
            ];
        ];
    ];

```

].

```
<#Deck>
rr:logicalTable [ rr:sqlQuery """
    SELECT deck.language , deck_revision.* 
    FROM deck
    JOIN deck_revision ON deck_revision.deck_id = deck.id """
];
rr:subjectMap [
    rr:template "http://slidewiki.org/rdf/deck/{id}";
    rr:class sw:Deck;
];
rr:predicateObjectMap [
    rr:predicate dcterms:created;
    rr:objectMap [ rr:column "timestamp" ];
];
rr:predicateObjectMap [
    rr:predicate dcterms:language;
    rr:objectMap [ rr:column "language" ];
];
rr:predicateObjectMap [
    rr:predicate dcterms:title;
    rr:objectMap [ rr:column "title" ];
];
rr:predicateObjectMap [
    rr:predicate sw:popularity;
    rr:objectMap [ rr:column "popularity" ];
];
rr:predicateObjectMap [
    rr:predicate dcterms:description;
    rr:objectMap [ rr:column "abstract" ];
];
rr:predicateObjectMap [
    rr:predicate wa:instanceOf;
    rr:objectMap [
        rr:parentTriplesMap <#DeckContainer>;
        rr:joinCondition [
            rr:child "deck_id";
            rr:parent "id";
        ];
    ];
];
rr:predicateObjectMap [
    rr:predicate dcterms:creator;
    rr:objectMap [
        rr:parentTriplesMap <#User>;
        rr:joinCondition [
            rr:child "user_id";
            rr:parent "id";
        ];
    ];
];
];
```

```

rr : predicateObjectMap [
  rr : predicate dcterms:contributor ;
  rr : objectMap [
    rr : parentTriplesMap <#User>;
    rr : joinCondition [
      rr : child "user_id";
      rr : parent "id";
    ];
    ];
];
rr : predicateObjectMap [
  rr : predicate wa:translationOf;
  rr : objectMap [
    rr : template "http://slidewiki.org/rdf/deck/{translated_from_re";
  ];
];
rr : predicateObjectMap [
  rr : predicate wa:revisionOf;
  rr : objectMap [
    rr : template "http://slidewiki.org/rdf/deck/{based_on}";
  ];
];
rr : predicateObjectMap [
  rr : predicate oa:styledBy;
  rr : objectMap [
    rr : template "http://slidewiki.org/rdf/style/{default_theme}";
  ];
];
].
```

<#Slide>

```

rr : logicalTable [ rr : sqlQuery """
  SELECT slide.language , slide_revision.*;
  FROM slide
  JOIN slide_revision ON slide_revision.slide = slide.id """
];
rr : subjectMap [
  rr : template "http://slidewiki.org/rdf/slides/{id}";
  rr : class sw:Slide
];
rr : predicateObjectMap [
  rr : predicate dcterms:created;
  rr : objectMap [ rr : column "timestamp" ];
];
rr : predicateObjectMap [
  rr : predicate dcterms:language;
  rr : objectMap [ rr : column "language" ];
];
rr : predicateObjectMap [
  rr : predicate cnt:chars;
  rr : objectMap [ rr : column "content" ];
];
rr : predicateObjectMap [
  rr : predicate sw:popularity;
]
```

```

        rr :objectMap [ rr :column "popularity" ];
];
rr :predicateObjectMap [
    rr :predicate dcterms:description;
    rr :objectMap [ rr :column "comment" ];
];
rr :predicateObjectMap [
    rr :predicate sw:slideNote;
    rr :objectMap [ rr :column "note" ];
];
rr :predicateObjectMap [
    rr :predicate wa:instanceOf;
    rr :objectMap [
        rr :parentTriplesMap <#SlideContainer>;
        rr :joinCondition [
            rr :child "slide";
            rr :parent "id";
        ];
        ];
];
rr :predicateObjectMap [
    rr :predicate dcterms:creator;
    rr :objectMap [
        rr :parentTriplesMap <#User>;
        rr :joinCondition [
            rr :child "user_id";
            rr :parent "id";
        ];
        ];
];
rr :predicateObjectMap [
    rr :predicate sw:translator;
    rr :objectMap [
        rr :parentTriplesMap <#User>;
        rr :joinCondition [
            rr :child "translator_id";
            rr :parent "id";
        ];
        ];
];
rr :predicateObjectMap [
    rr :predicate wa:translationOf;
    rr :objectMap [
        rr :template "http://slidewiki.org/rdf/slide/{translated_from_rev}";
    ];
];
rr :predicateObjectMap [
    rr :predicate wa:revisionOf;
    rr :objectMap [
        rr :template "http://slidewiki.org/rdf/slide/{based_on}";
    ];
].

```

```

<#QuestionsAssigned>
    rr : logicalTable [ rr : sqlQuery """
        SELECT *
        FROM questions WHERE questions.id NOT IN
        (SELECT ‘based_on’ FROM ‘questions‘ WHERE based_on IS NOT NULL)"""
    ];
    rr : subjectMap [
        rr : template "http://slidewiki.org/rdf/question/{id}";
        rr : class sw:Question;
    ];
    rr : predicateObjectMap [
        rr : predicate wa:assignedTo;
        rr : objectMap [
            rr : template "http://slidewiki.org/rdf/slidesContainer/{item_id}"
        ];
    ];
<#CommentDecks>
    rr : logicalTable [ rr : sqlQuery """
        SELECT * FROM comment
        WHERE item_type='deck'
    """];
    rr : subjectMap [
        rr : template "http://slidewiki.org/rdf/comment/{id}";
        rr : class sw:Comment;
    ];
    rr : predicateObjectMap [
        rr : predicate wa:assignedTo;
        rr : objectMap [
            rr : template "http://slidewiki.org/rdf/deck/{item_id}";
        ];
    ];
    rr : predicateObjectMap [
        rr : predicate dcterms:creator;
        rr : objectMap [
            rr : parentTriplesMap <#User>;
            rr : joinCondition [
                rr : child "user_id";
                rr : parent "id";
            ];
            ];
    ];
    rr : predicateObjectMap [
        rr : predicate cnt:chars;
        rr : objectMap [ rr : column "text" ];
    ];
    rr : predicateObjectMap [
        rr : predicate dcterms:title;
        rr : objectMap [ rr : column "title" ];
    ];
    rr : predicateObjectMap [
        rr : predicate dcterms:created;
    ];

```

```

        rr :objectMap [ rr :column "timestamp" ];
    ].

<#CommentSlides>
    rr :logicalTable [ rr :sqlQuery """
SELECT * FROM comment
WHERE item_type='slide'
"""];
    rr :subjectMap [
        rr :template "http://slidewiki.org/rdf/comment/{id}";
        rr :class sw:Comment;
    ];
    rr :predicateObjectMap [
        rr :predicate wa:assignedTo;
        rr :objectMap [
            rr :template "http://slidewiki.org/rdf/slides/{item_id}";
        ];
    ];
    rr :predicateObjectMap [
        rr :predicate dcterms:creator;
        rr :objectMap [
            rr :parentTriplesMap <#User>;
            rr :joinCondition [
                rr :child "user_id";
                rr :parent "id";
            ];
            ];
    ];
    rr :predicateObjectMap [
        rr :predicate cnt:chars;
        rr :objectMap [ rr :column "text" ];
    ];
    rr :predicateObjectMap [
        rr :predicate dcterms:title;
        rr :objectMap [ rr :column "title" ];
    ];
    rr :predicateObjectMap [
        rr :predicate dcterms:created;
        rr :objectMap [ rr :column "timestamp" ];
    ].

<#StyleHasRevision>
    rr :logicalTable [ rr :sqlQuery """
SELECT * FROM style
WHERE based_on IS NOT NULL
"""];
    rr :subjectMap [
        rr :template "http://slidewiki.org/rdf/style/{based_on}";
        rr :class sw:Style
    ];
    rr :predicateObjectMap [
        rr :predicate wa:hasRevision;
    ];

```

```

        rr:objectMap [
            rr:template "http://slidewiki.org/rdf/style/{id}";
        ];
    ];
    rr:predicateObjectMap [
        rr:predicate dcterms:creator;
        rr:objectMap [
            rr:parentTriplesMap <#User>;
            rr:joinCondition [
                rr:child "user_id";
                rr:parent "id";
            ];
            ];
        ];
    ].

<#Style>
    rr:logicalTable [ rr:tableName "style" ];
    rr:subjectMap [
        rr:template "http://slidewiki.org/rdf/style/{id}";
        rr:class sw:Style;
    ];
    rr:predicateObjectMap [
        rr:predicate dcterms:created;
        rr:objectMap [ rr:column "timestamp" ];
    ];
    rr:predicateObjectMap [
        rr:predicate cnt:chars;
        rr:objectMap [ rr:column "css" ];
    ];
    rr:predicateObjectMap [
        rr:predicate dcterms:description;
        rr:objectMap [ rr:column "comment" ];
    ];
    rr:predicateObjectMap [
        rr:predicate dcterms:title;
        rr:objectMap [ rr:column "name" ];
    ];
    rr:predicateObjectMap [
        rr:predicate dcterms:creator;
        rr:objectMap [
            rr:parentTriplesMap <#User>;
            rr:joinCondition [
                rr:child "user_id";
                rr:parent "id";
            ];
            ];
        ];
    ];
    rr:predicateObjectMap [
        rr:predicate wa:revisionOf;
        rr:objectMap [
            rr:template "http://slidewiki.org/rdf/style/{based_on}";
        ];
    ];

```

].

```
<#Question>
    rr:logicalTable [ rr:tableName "questions" ];
    rr:subjectMap [
        rr:template "http://slidewiki.org/rdf/question/{id}";
        rr:class sw:Queston;
            rr:class sa:Queston;
    ];
    rr:predicateObjectMap [
        rr:predicate dcterms:created;
        rr:objectMap [ rr:column "timestamp" ];
    ];
    rr:predicateObjectMap [
        rr:predicate sw:defaultDifficulty;
        rr:objectMap [ rr:column "difficulty" ];
    ];
    rr:predicateObjectMap [
        rr:predicate cnt:chars;
        rr:objectMap [ rr:column "question" ];
    ];
    rr:predicateObjectMap [
        rr:predicate sa:difficultyLevel;
        rr:objectMap [ rr:column "diff_count" ];
    ];
    rr:predicateObjectMap [
        rr:predicate wa:revisionOf;
        rr:objectMap [
            rr:template "http://slidewiki.org/rdf/question/{based_on}";
        ];
    ];
    rr:predicateObjectMap [
        rr:predicate dcterms:creator;
        rr:objectMap [
            rr:parentTriplesMap <#User>;
            rr:joinCondition [
                rr:child "user_id";
                rr:parent "id";
            ];
        ];
    ];
].
```

```
<#UserTestHasPart>
    rr:logicalTable [ rr:tableName "test_content" ];
    rr:subjectMap [
        rr:template "http://slidewiki.org/rdf/userTest/{test_id}";
        rr:class sw:UserTest;
```

```

];
rr : predicateObjectMap [
    rr : predicate dcterms : hasPart ;
    rr : objectMap [
        rr : parentTriplesMap <#Question>;
        rr : joinCondition [
            rr : child "question_id";
            rr : parent "id";
        ];
    ];
];
.

<#QuestionPartOftest>
    rr : logicalTable [ rr : tableName "test_content" ];
    rr : subjectMap [
        rr : template "http://slidewiki.org/rdf/question/{question_id}";
        rr : class sw : Question;
    ];
    rr : predicateObjectMap [
        rr : predicate dcterms : isPartOf;
        rr : objectMap [
            rr : parentTriplesMap <#UserTest>;
            rr : joinCondition [
                rr : child "test_id";
                rr : parent "id";
            ];
        ];
    ];
];
.

<#Choice>
    rr : logicalTable [ rr : tableName "answers" ];
    rr : subjectMap [
        rr : template "http://slidewiki.org/rdf/choice/{id}";
        rr : class sa : Choice;
    ];
    rr : predicateObjectMap [
        rr : predicate cnt : chars;
        rr : objectMap [ rr : column "answer" ];
    ];
    rr : predicateObjectMap [
        rr : predicate sa : explanation;
        rr : objectMap [ rr : column "explanation" ];
    ];
];
.

<#Distractors>
    rr : logicalTable [ rr : sqlQuery """
        SELECT *
        FROM answers
        WHERE is_right='no'
    """;
    rr : subjectMap [
        rr : template "http://slidewiki.org/rdf/question/{question_id}";
        rr : class sw : Question;
    ];
];
.
```

```

];
rr : predicateObjectMap [
  rr : predicate sa : distractor ;
  rr : objectMap [
    rr : parentTriplesMap <#Choice>;
    rr : joinCondition [
      rr : child "id";
      rr : parent "id";
    ];
  ];
].
<#CorrectChoices>
  rr : logicalTable [ rr : sqlQuery """
    SELECT *
    FROM answers
    WHERE is_right='yes' """
  ];
  rr : subjectMap [
    rr : template "http://slidewiki.org/rdf/question/{question_id}";
    rr : class sw : Question;
  ];
  rr : predicateObjectMap [
    rr : predicate sa : correctChoice ;
    rr : objectMap [
      rr : parentTriplesMap <#Choice>;
      rr : joinCondition [
        rr : child "id";
        rr : parent "id";
      ];
    ];
  ];
].
<#UserTest>
  rr : logicalTable [ rr : tableName "user_tests" ];
  rr : subjectMap [
    rr : template "http://slidewiki.org/rdf/userTest/{id}";
    rr : class sw : UserTest;
  ];
  rr : predicateObjectMap [
    rr : predicate dcterms : created ;
    rr : objectMap [ rr : column "timestamp" ];
  ];
  rr : predicateObjectMap [
    rr : predicate dcterms : title ;
    rr : objectMap [ rr : column "title" ];
  ];
  rr : predicateObjectMap [
    rr : predicate dcterms : creator ;
    rr : objectMap [
      rr : parentTriplesMap <#User>;
      rr : joinCondition [
        rr : child "user_id";
      ];
    ];
  ];

```

```
        rr:parent "id";
    ];
];
].
<#User>
rr:logicalTable [ rr:tableName "users" ];
rr:subjectMap [
    rr:template "http://slidewiki.org/rdf/user/{id}";
    rr:class sw:User;
];
rr:predicateObjectMap [
    rr:predicate foaf:mbox;
    rr:objectMap [ rr:column "email" ];
];
rr:predicateObjectMap [
    rr:predicate foaf:nick;
    rr:objectMap [ rr:column "username" ];
];
rr:predicateObjectMap [
    rr:predicate dcterms:date;
    rr:objectMap [ rr:column "registered" ];
];
rr:predicateObjectMap [
    rr:predicate foaf:firstName;
    rr:objectMap [ rr:column "first_name" ];
];
rr:predicateObjectMap [
    rr:predicate foaf:familyName;
    rr:objectMap [ rr:column "last_name" ];
];
rr:predicateObjectMap [
    rr:predicate dcterms:language;
    rr:objectMap [ rr:column "locale" ];
];
rr:predicateObjectMap [
    rr:predicate foaf:gender;
    rr:objectMap [ rr:column "gender" ];
];
rr:predicateObjectMap [
    rr:predicate sw:speaks;
    rr:objectMap [ rr:column "languages" ];
];
rr:predicateObjectMap [
    rr:predicate foaf:based_near;
    rr:objectMap [ rr:column "location" ];
];
rr:predicateObjectMap [
    rr:predicate foaf:topic_interest;
    rr:objectMap [ rr:column "interests" ];
];
rr:predicateObjectMap [
    rr:predicate foaf:depiction;
```

```
    rr:objectMap [ rr:column "picture" ];  
];  
    rr:predicateObjectMap [  
    rr:predicate sw:homeLocation;  
    rr:objectMap [ rr:column "hometown" ];  
];  
    rr:predicateObjectMap [  
    rr:predicate foaf:birthday;  
    rr:objectMap [ rr:column "birthday" ];  
];  
    rr:predicateObjectMap [  
    rr:predicate dcterms:description;  
    rr:objectMap [ rr:column "description" ];  
];  
    rr:predicateObjectMap [  
    rr:predicate foaf:page;  
    rr:objectMap [ rr:column "infodeck" ];  
].
```

C.2 Tabular mapping representation

C.3 SlideWiki knowledge graph

Table Column	Class Property	Table Column	Class Property
Answers	sa:Choice	Deck_revision	sw:Deck
question_id	-	footer_text	-
answer	cnt:chars	license	-
is_correct	-	is_featured	-
explanation	sa:explanation	visibility	-
Comment	sw:Comment	translation_status	-
user_id	dcterms:creator	translated_from_revision	wa:translationOf
item_type	-	deck_content.item_id	dcterms:hasPart dcterms:ispartOf
item_id	sw:assignedToSlide sw:assignedToDeck	deck_content.position	-
title	dcterms:title	brand.image	-
text	cnt:chars	brand.url	-
timestamp	dcterms:created	brand.description	-
Deck	sw:DeckContainer	deck.language	dcterms:language
timestamp	dcterms:created	short_urls.short	-
user_id	dcterms:creator	short_urls.url	-
description	dcterms:source	short_urls.stamped	-
translated_from	wa:translationOf	short_urls.user_id	-
translated_from_revision	-	tag.tag	dcterms:subject
Deck_revision	sw:Deck	user_group.user_id	dcterms:contributor
deck_id	wa:instanceOf wa:includes	user_group.category	-
title	dcterms:title	user_group.timestamp	-
timestamp	dcterms:created	Impress_transition	-
user_id	dcterms:creator	user_id	-
based_on	wa:revisionOf wa:hasRevision	slide_position	-
popularity	sw:popularity	css	-
default_theme	oa:styledBy	Media	-
default_transition	-	user_id	-
comment	-	type	-
abstract	dcterms:description		

Figure C.1: Matching the DB tables and columns to WikiApp ontologie classes and properties

Table Column	Class Property	Table Column	Class Property
Media	-	Slide	sw:SlideContainer
uri	-	description	dcterms:source
timestamp	-	language	dcterms:language
title	-	translated_from	wa:translationOf
original_width	-	translated_from_r	-
original_height	-	revision	-
media_relations.slide_rev_id	-	Slide_revision	sw:Slide
media_relations.deck_rev_id	-	slide	wa:instanceOf wa:includes
Message	-	timestamp	dcterms:created
sender_id	-	content	cnt:chars dcterms:title
receiver_id	-	user_id	dcterms:creator
title	-	based_on	wa:revisionOf wa:hasRevision
content	-	popularity	sw:popularity
msg_type	-	comment	-
date	-	note	sw:slideNote
Questions		license	dcterms:license
item_id	wa:assignedTo	translation_status	-
question	cnt:chars	translator_id	sw:translator
difficulty	sw:defaultDifficulty	translated_from_r	wa:translationOf
guessing	-	Style	sw:Style
user_id	dcterms:creator	user_id	dcterms:creator
based_on	wa:revisionOf wa:hasRevision	name	dcterms:title
timestamp	dcterms:created	scss_varfunc	-
mark	-	scss	-
diff_count	-	css	cnt:chars
incorrect	-	timestamp	dcterms:created
all_answers	-	comment	dcterms:description
Slide	sw:SlideContainer	based_on	wa:revisionOf wa:hasRevision
user_id	dcterms:creator		

Figure C.2: Matching the DB tables and columns to WikiApp ontologie classes and properties (cont.)

Table Column	Class Property	Table Column	Class Property
Testing	-	Users	sw:User
attempt_id	-	password	-
user_id	-	registered	dcterms:date
item_id	-	default_theme	-
title	-	fb_id	-
timestamp	-	first_name	foaf:firstName
wiki_app	-	last_name	foaf:familyName
mtf	-	gender	foaf:gender
dich	-	locale	dcterms:language
morgan	-	hometown	sw:homeLocation
ripkey	-	location	foaf:based_near
max_points	-	languages	sw:speaks
type	-	picture	foaf:depiction
mode	-	interests	foaf:interests
limit	-	description	dcterms:description
test_results.attempt_id	-	birthday	foaf:birthday
test_results.quest_id	-	infodeck	foaf:page
test_results.checked	-	preferences.preference	-
Transition	-	preferences.value	-
user_id	-	subscription.item_id	-
name	-	subscription.item_type	-
scss_varfunc	-	subscription.timesamp	-
scss	-	User_tests	sw:UserTest
css	-	user_id	dcterms:creator
timestamp	-	title	dcterms:title
comment	-	timestamp	dcterms:created
based_on	-	test_content.question_id	dcterms:hasPart dcterms:isPartOf
Users	sw:User		
email	foaf:mbox		
username	foaf:nick		

Figure C.3: Matching the DB tables and columns to WikiApp ontologie classes and properties (cont.)

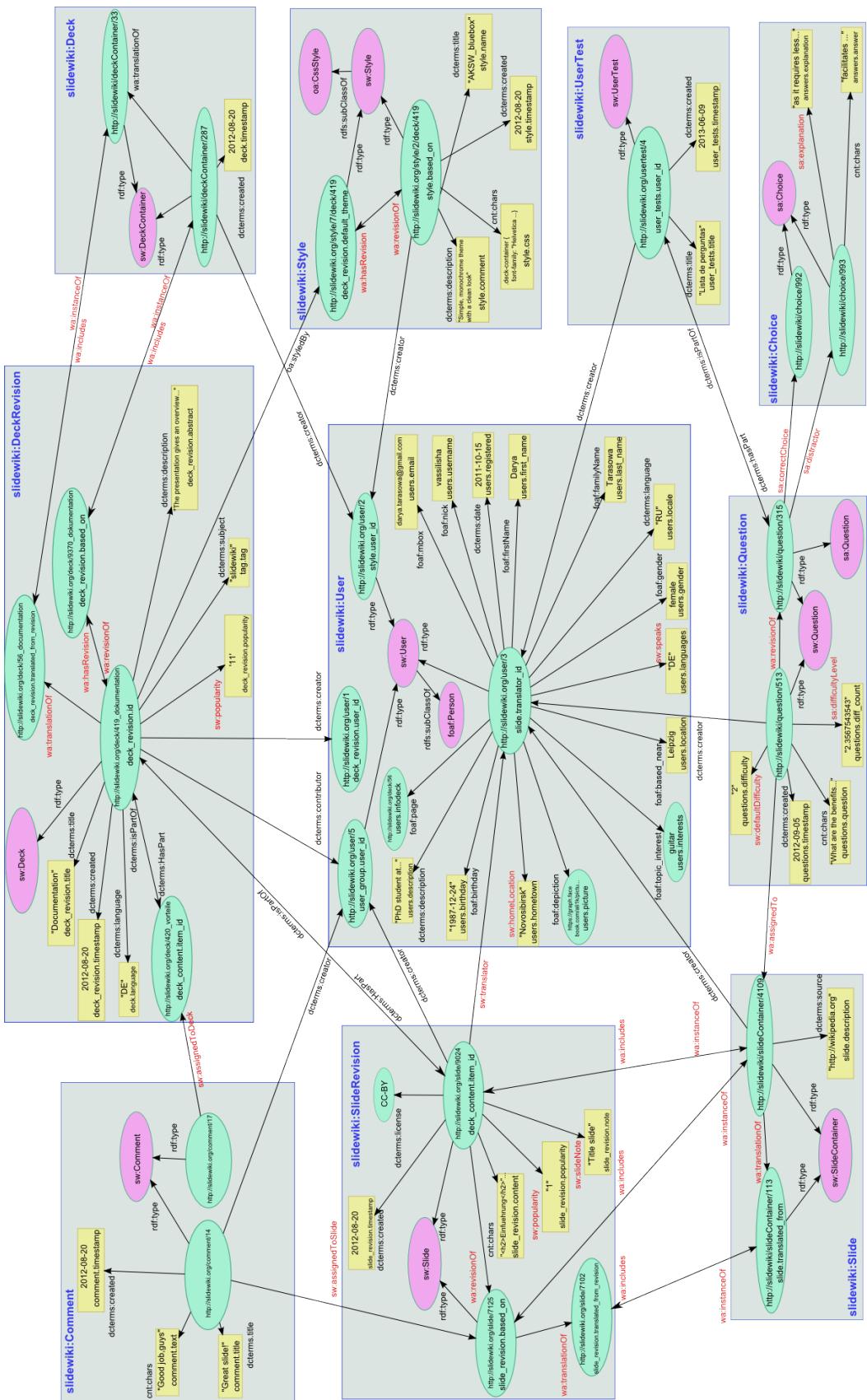


Figure C.4: SlideWiki Knowledge graph

Bibliography

- [1] SCORM 2004 4th Edition Specification. Technical report, ADL, 2011.
- [2] SCORM Users guide for programmers. Technical report, ADL, 2011.
- [3] *OpenCourseWare Observatory – Does the Quality of OpenCourseWare Live up to its Promise?* Zenodo, February 2015.
- [4] Palakorn Achananuparp, Katherine W McCain, and Robert B Allen. Supporting student collaboration for image indexing. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pages 24–34. Springer, 2007.
- [5] Grace Agnew, Chad M Mills, and Carolyn A Maher. Vmcanalytic: Developing a collaborative video analysis tool for education faculty and practicing educators. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.
- [6] Per Andersen. *What is Web 2.0?: ideas, technologies and implications for education*, volume 1. JISC Bristol, UK, 2007.
- [7] Anupriya Ankolekar, Markus Krötzsch, Thanh Tran, and Denny Vrandecic. The two cultures: Mashing up web 2.0 and the semantic web. In *Proceedings of the 16th international conference on World Wide Web*, pages 825–834. ACM, 2007.
- [8] Marcelo Arenas, , Alexandre Bertails, Eric Prud’hommeaux, and Juan Sequeda. A direct mapping of relational data to rdf. Working draft, W3C, 2012.
- [9] Sören Auer, Ali Khalili, and Darya Tarasowa. Crowd-sourced open courseware authoring with slidewiki.org. *International Journal of Emerging Technologies in Learning (iJET)*, 8(1), 2013.
- [10] Sören Auer, Lee Feigenbaum, Daniel Miranker, Angela Fogarolli, and Juan Sequeda. Use cases and requirements for mapping relational databases to rdf, 2010.
- [11] Andreas Auinger, Martin Ebner, Dietmar Nedbal, and Andreas Holzinger. Mixing content and endless collaboration–mashups: Towards future personal learning environments. In *Universal Access in Human-Computer Interaction. Applications and Services*, pages 14–23. Springer, 2009.
- [12] Georgia Bafoutsou and Gregoris Mentzas. Review and functional classification of collaborative systems. *International journal of information management*, 22(4):281–305, 2002.
- [13] Elena García Barriocanal, Miguel-Ángel Sicilia, Salvador Sánchez Alonso, and Miltiadis D. Lytras. Semantic annotation of video fragments as learning objects. *Interactive Learning Environments*, 19(1):25–44, 2011.
- [14] Daniel Bauer, Matthias Holzer, Veronika Kopp, and Martin R Fischer. Pick-N multiple choice-exams: a comparison of scoring algorithms. *Advances in health sciences education : theory and practice*, 16(2):211–21, May 2011.

- [15] A. Ben-Simon, D. V. Budescu, and B. Nevo. A Comparative Study of Measures of Partial Knowledge in Multiple-Choice Tests. *Applied Psychological Measurement*, 21(1):65–88, March 1997.
- [16] Alexandre Bertails and Eric Gordon Prud'hommeaux. Interpreting relational databases in the rdf domain. In *K-CAP*, pages 129–136, 2011.
- [17] Christian Bizer. *Quality Driven Information Filtering: In the Context of Web Based Information Systems*. PhD thesis, Free University of Berlin, 2007.
- [18] Christian Bizer and Andreas Schultz. The berlin sparql benchmark. *Int. J. Semantic Web Inf. Syst.*, 5(2):1–24, 2009.
- [19] Ina Blau. E-collaboration within, between, and without institutions: Towards better functioning of. *Interdisciplinary Applications of Electronic Collaboration Approaches and Technologies*, page 188, 2013.
- [20] Hansen C Bow, Jonathan R Dattilo, Andrea M Jonas, and Christoph U Lehmann. A crowdsourcing model for creating preclinical medical education study tools. *Academic Medicine*, 88(6):766–770, 2013.
- [21] Danah Boyd. Why youth (heart) social network sites: The role of networked publics in teenage social life. *MacArthur foundation series on digital learning–Youth, identity, and digital media volume*, pages 119–142, 2007.
- [22] danah m. boyd and Nicole B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.
- [23] Rovy F Branon and Christopher Essex. Synchronous and asynchronous communication tools in distance education. *TechTrends*, 45(1):36–36, 2001.
- [24] Christopher Brooks, Rupi Panesar, and Jim Greer. *Awareness and collaboration in the ihelp courses content management system*. Springer, 2006.
- [25] Amy Bruckman. Social support for creativity and learning online. In *Digital Games and Intelligent Toys Based Education, 2008 Second IEEE International Conference on*, pages 3–7. IEEE, 2008.
- [26] Peter Brusilovsky. Adaptive hypermedia. *User modeling and user-adapted interaction*, 11(1-2):87–110, 2001.
- [27] Peter Brusilovsky. Developing adaptive educational hypermedia systems: From design models to authoring tools. In *Authoring Tools for Advanced Technology Learning Environments*, pages 377–409. Springer, 2003.
- [28] Gregoire Burel, Amparo E. Cano1, and Vitaveska Lanfranchi. Ozone browser: Augmenting the web with semantic overlays. volume 449 of *CEUR Workshop Proceedings ISSN 1613-0073*, June 2009.
- [29] Ramón Zatarain Cabada, María Lucía Barrón Estrada, and Carlos Alberto Reyes García. Educa: A web 2.0 authoring tool for developing adaptive and intelligent tutoring systems using a kohonen network. *Expert Systems with Applications*, 38(8):9522–9529, 2011.
- [30] Jing Cai, Yueliang Zhou, and Gaofeng Ruan. Learning activities design of a blended learning course based on public internet. In *Education Technology and Computer (ICETC), 2010 2nd International Conference on*, volume 3, pages V3–297. IEEE, 2010.

- [31] Dan Chiribuca, Daniel Hunyadi, and Emil M Popa. The educational semantic web. In *Proceedings of the 8th conference on Applied informatics and communications (AIC 08)*, Nikos E. Mastorakis, Metin Demiralp, Valeri Mladenov, and Zoran Bojkovic (Eds.). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, pages 314–319, 2008.
- [32] Ivan Claros and Ruth Cobos. Social media learning: An approach for composition of multimedia interactive object in a collaborative learning environment. In *Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on*, pages 570–575. IEEE, 2013.
- [33] Learning Technology Standards Committee et al. Ieee standard for learning object metadata. *IEEE Standard*, 1484(1):2007–04, 2002.
- [34] L. J. Cronbach. An experimental comparison of the multiple true-false and multiple multiple-choice tests. *Journal of Educational Psychology*, 32:533–543, 1941.
- [35] Sarah Currier. Integrating information resources and online learning in the uk. In *Computers in Education, 2002. Proceedings. International Conference on*, pages 818–822. IEEE, 2002.
- [36] Halina M Dahlan, Ismail Mohamed Ali, and Ab Razak Che Hussin. Suitability of collaborative learning activities in web 2.0 environment. In *User Science and Engineering (i-USER), 2010 International Conference on*, pages 65–70. IEEE, 2010.
- [37] Souripriya Das, Seema Sundara, and Richard Cyganiak. R2rml: Rdb to rdf mapping language. Recommendation, W3C, 2012.
- [38] Ross Dawson and Steve Byngahall. *Getting results from crowds*. Advanced Human Technologies San Francisco, 2012.
- [39] Carmen De Maio, Giuseppe Fenza, V Loia, S Senatore, M Gaeta, and F Orciuoli. Rss-generated contents through personalizing e-learning agents. In *Intelligent Systems Design and Applications, 2009. ISDA '09. Ninth International Conference on*, pages 49–54. IEEE, 2009.
- [40] Mike Dean, Guus Schreiber, Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. Owl web ontology language reference. Recommendation, W3C, 2004.
- [41] Vladan Devedzic. *Semantic Web and Education (Integrated Series in Information Systems)*. Springer, Secaucus, NJ, USA, 2006.
- [42] Darina Dicheva, Lora Aroyo, and AI Cristea. Collaborative courseware authoring support. *CATE*, 2002.
- [43] Stefan Dietze, Honq Qing Yu, Daniela Giordano, Eleni Kaldoudi, Nikolas Dovrolis, and Davide Taibi. Linked education: Interlinking educational resources and the web of data. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pages 366–371, New York, NY, USA, 2012. ACM.
- [44] D DiNucci. Fragmented future, 1999.
- [45] Darcy DiNucci. Design & new media: Fragmented future-web development faces a process of mitosis, mutation, and natural selection. *PRINT-NEW YORK-*, 53:32–35, 1999.
- [46] Anhai Doan, Raghu Ramakrishnan, and Alon Y Halevy. Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4):86–96, 2011.

- [47] PL Dressel and J Schmid. Some modifications of the multiple-choice item. *Educational and Psychological Measurement*, 13(4):574–595, 1953.
- [48] Tore Dyba, Torgeir Dingsoyr, and Geir K. Hanssen. Applying systematic reviews to diverse study types: An experience report. In *Proceedings of the First International Symposium on Empirical Software Engineering and Measurement*, ESEM ’07, pages 225–234, Washington, DC, USA, 2007. IEEE Computer Society.
- [49] Elharith Elrufaie and David A Turner. A wiki paradigm for use in it courses. In *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*, volume 2, pages 770–771. IEEE, 2005.
- [50] Orri Erling. Requirements for relational to rdf mapping, 2008.
- [51] Enrique Estellés-Arolas and Fernando González-Ladrón-de Guevara. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200, 2012.
- [52] Jérôme Euzenat and P Shvaiko. *Ontology alignment*. Springer, 2007.
- [53] DW Farthing, DM Jones, and D McPhee. Permutational multiple-choice questions: an objective and efficient alternative to essay-type examination questions. *ACM SIGCSE Bulletin*, 30(3):81–85, 1998.
- [54] Annika Flemming. Quality characteristics of linked data publishing datasources. Master’s thesis, Humboldt-Universität of Berlin, 2010.
- [55] D Foord. The stolen principle tick list, a6 training and consultancy, 2007.
- [56] International Institute for Educational Planning. Open educational resources: Open content for higher education. Technical report, 2005.
- [57] David A. Frisbie. The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice*, 11(4):21–26, 1992.
- [58] Francisco José García and Joaquín García. Educational hypermedia resources facilitator. *Computers & Education*, 44(3):301–325, 2005.
- [59] R Stuart Geiger. The lives of bots. *Lovink and Tkacz*, 2011.
- [60] Fawaz Ghali and Alexandra I Cristea. Social reference model for adaptive web learning. In *Advances in Web Based Learning–ICWL 2009*, pages 162–171. Springer, 2009.
- [61] Ian Glover, Zhijie Xu, and Glenn Hardaker. Online annotation–research and practices. *Computers & Education*, 49(4):1308–1320, 2007.
- [62] Sahin Gokcearslan and Seher Ozcan. Place of wikis in learning and teaching process. *Procedia-Social and Behavioral Sciences*, 28:481–485, 2011.
- [63] Christine Greenhow, Beth Robelia, and Joan E Hughes. Learning, teaching, and scholarship in a digital age web 2.0 and classroom research: What path should we take now? *Educational Researcher*, 38(4):246–259, 2009.
- [64] Anja Haake, Stephan Lukosch, and Till Schümmer. Wiki-templates: adding structure support to wikis on demand. In *Proceedings of the 2005 International Symposium on Wikis*, pages 41–51. ACM, 2005.

-
- [65] Khaled Halimi, Hassina Seridi, and Catherine Faron-Zucker. Solearn: A social learning network. In *Computational Aspects of Social Networks (CASoN), 2011 International Conference on*, pages 130–135. IEEE, 2011.
 - [66] Richard Hall. Towards a fusion of formal and informal learning environments: the impact of the read/write web. 2009.
 - [67] Awang Hamid, Dayang Tiawa, Md Islam, Norma Alias, Abdul Hafiz Omar, et al. An efficient authoring activities infrastructure design through grid portal technology. *The 7th WSEAS International Conference on Engineering Education*, pages 146–151, 2010.
 - [68] Farid Hasanov. *Measuring and improving the quality of user interaction with learning management systems*. PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, 2015.
 - [69] Norman Heino, Sebastian Dietzold, Michael Martin, and Sören Auer. Developing semantic web applications with the ontowiki framework. In *Networked Knowledge - Networked Media*, volume 221 of *Studies in Computational Intelligence*, pages 61–77. Springer, Berlin / Heidelberg, 2009.
 - [70] Guillaume Hillairet, Frédéric Bertrand, and Jean-Yves Lafaye. Mde for publishing data on the semantic web. In *TWOMDE*, pages 32–46.
 - [71] Matthias Hirth, Tobias Hoffeld, and Phuoc Tran-Gia. Human cloud as emerging internet application-anatomy of the microworkers crowdsourcing platform. *University of Wurzburg, Institute of Computer Science, Am Hubland, D-97074 Wurzburg, Germany, Research Report*, (478):45–46, 2011.
 - [72] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. Weaving the pedantic web. In *LDOW*, 2010.
 - [73] Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. An empirical survey of linked data conformance. *Web Semantics*, 14:14–44, 2012.
 - [74] C. Hohensinn and K. D. Kubinger. Applying Item Response Theory Methods to Examine the Impact of Different Response Formats. *Educational and Psychological Measurement*, 71(4):732–746, January 2011.
 - [75] Chuleui Hong and Yeongjoon Kim. The multimedia authoring in collaborative e-learning system. In *Information Science and Service Science and Data Mining (ISSDM), 2012 6th International Conference on New Trends in*, pages 158–161. IEEE, 2012.
 - [76] Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.
 - [77] Jeff Howe. The rise of crowdsourcing. *Wired Magazine*, 14(6), 06 2006.
 - [78] Teresa Hübscher-Younger and N Hari Narayanan. Constructive and collaborative learning of algorithms. *ACM SIGCSE Bulletin*, 35(1):6–10, 2003.
 - [79] Neil Ireson and Gregoire Burel. Knowledge sharing in e-collaboration. In *Electronic Government*, pages 351–362. Springer, 2010.
 - [80] S. Itten and R Krebs. *Messqualitaet der verschiedenen MC-Itemtypen in den beiden Vorpruefungen des Medizinstudiums an der Universitaet Bern 1997/2*. Bern: IAWF, 1997.
 - [81] Lv Jiake, Zhou Wei, and Wang Xuan. Design and evaluation of a wiki-based collaborative learning environment for colleges computers compulsory education. In *Computer Science and Education (ICCSE), 2010 5th International Conference on*, pages 695–699. IEEE, 2010.

- [82] H Jiao, J Liu, and K Haynie. Comparison Between Dichotomous and Polytomous Scoring of Innovative Items in a Large-Scale Computerized Adaptive Test. *Educational and Psychological Measurement*, 72(3):493–509, 2012.
- [83] Cynthia Jimes. Travel well open educational resources: a presentation of ongoing research. *iSummit 08*, 2008.
- [84] Christopher M Johnson. A survey of current research on online communities of practice. *The Internet and Higher Education*, 4(1):45 – 60, 2001.
- [85] Gerald C Kane and Robert G Fichman. The shoemaker’s children: using wikis for information systems teaching, research, and publication. *MIS quarterly*, pages 1–17, 2009.
- [86] Ali Khalili. A semantics-based user interface model for content annotation, authoring and exploration. 2014.
- [87] Ali Khalili, Sören Auer, Darya Tarasowa, and Ivan Ermilov. Slidewiki: Elicitation and sharing of corporate knowledge using presentations. In *Proceedings of the EKAW 2012*, Lecture Notes in Computer Science (LNCS) 7603. Springer, 2012.
- [88] B. Kitchenham. Procedures for performing systematic reviews. Technical report, Keele University and NICTA, 2004.
- [89] Jeni Klugman. Human development report 2010–20th anniversary edition. the real wealth of nations: Pathways to human development. 2010.
- [90] Markus Krötzsch, Denny Vrandečić, Max Völkel, Heiko Haller, and Rudi Studer. Semantic Wikipedia. *Journal of Web Semantics*, 5(4):251–261, 2007.
- [91] Yen-Hung Kuo, Qing Tan Kinshuk, Yueh-Min Huang, Tzu-Chien Liu, and Maiga Chang. Collaborative creation of authentic examples with location for u-learning. In *International Conference e-Learning 2008*, pages 16 – 20. IADIS, 2008.
- [92] Baozhen Lee and Shilun Ge. Personalisation and sociability of open knowledge management based on social tagging. *Online Information Review*, 34(4):618–625, 2010.
- [93] Teemu Leinonen, Jukka Purma, Hans Poldoja, and Tarmo Toikkanen. Information architecture and design solutions scaffolding authoring of open educational resources. *Learning Technologies, IEEE Transactions on*, 3(2):116–128, 2010.
- [94] Bo Leuf and Ward Cunningham. *The Wiki way: quick collaboration on the Web*. Addison-Wesley, London, 2001.
- [95] James Lewis and Jeff Sauro. The Factor Structure of the System Usability Scale. In *Human Centered Design*, volume 5619 of *LNCS*, chapter 12, pages 94–103. Springer, 2009.
- [96] Yanyan Li, Mingkai Dong, and Ronghuai Huang. Developing a collaborative e-learning environment based upon semantic wiki: from design models to application scenarios. In *Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on*, pages 222–226. IEEE, 2010.
- [97] Chengfei Liu and Jianxin Li. Designing quality xml schemas from e-r diagrams. In *Advances in Web-Age Information Management*, number 4016 in LNCS, pages 508–519. Springer, 2006.
- [98] Vanessa Lopez, Victoria Uren, Marta Sabou, and Enrico Motta. Is question answering fit for the semantic web? a survey. *Semantic Web ? Interoperability, Usability, Applicability*, 2(2):125–155, September 2011.

- [99] IEEE LTSC. Learning object definition. Technical report, LTSC Learning Technology Standards Committee of the Institute of Electrical and Electronics Engineers (IEEE),, 2002.
- [100] Airong Luo, Dick Ng'ambi, and Ted Hanss. Towards building a productive, scalable and sustainable collaboration model for open educational resources. In *Proceedings of the 16th ACM international conference on Supporting group work*, pages 273–282. ACM, 2010.
- [101] Saenz Marcilla et al. Framework for the development and reuse of educational resources in the area of information systems management for different degrees at a technical university. In *Proceedings of the 2012 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–9, 2012.
- [102] Florian Matthes, Christian Neubert, and Alexander Steinhoff. Hybrid wikis: Empowering users to collaboratively structure information. In *ICSOFT (1)*, pages 250–259. SciTePress, 2011.
- [103] Carnegie Melon. SCORM Best Practices Guide for Content Developers. Technical report, Learning Systems Architecture Lab, 2004.
- [104] Wolfgang Müller, Isabel Rojas, Andreas Eberhart, Peter Haase, and Michael Schmidt. A-r-e: The author-review-execute environment. *Procedia Computer Science*, 4:627 – 636, 2011. ICCS 2011.
- [105] MRJ Morgan. MCQ: An interactive computer program for multiple-choice self-testing. *Biochemical Education*, 7(3):67–69, 1979.
- [106] Róbert Móro, Ivan Srba, Maroš Unčík, Mária Bieliková, and Marián Šimko. Towards collaborative metadata enrichment for adaptive web-based learning. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 03*, pages 106–109. IEEE Computer Society, 2011.
- [107] Joseph Moxley. Datagogies, writing spaces, and the age of peer production. *Computers and Composition*, 25(2):182–202, 2008.
- [108] San Murugesan. Understanding web 2.0. *IT professional*, 9(4):34–41, 2007.
- [109] Sasa Nesic, Mehdi Jazayeri, Monica Landoni, and Dragan Gasevic. Using semantic documents and social networking in authoring of course material: An empirical study. In *Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on*, pages 666–670. IEEE, 2010.
- [110] Dade Nurjanah, Hugh C Davis, and Thanassis Tiropanis. Collaborative authoring of adaptive learning resources: Opportunities and challenges. In *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2011 7th International Conference on*, pages 538–542. IEEE, 2011.
- [111] Ade G Ola, Abiodun O Bada, Emmanuel Omojokun, and A Adekoya. Actualizing learning and teaching best practices in online education with open architecture and standards. In *Proc. of International Conference on Information Security and Privacy*, pages 1790–511, 2009.
- [112] Luciano TE Pansanato and Renata PM Fortes. System description: An orienteering strategy to browse semantically-enhanced educational wiki pages. In *The Semantic Web: Research and Applications*, pages 809–818. Springer, 2007.
- [113] Nieves Pedreira, José Ramón Méndez Salgueiro, and Manuel Martínez Carballo. E-learning in new technologies. In *Encyclopedia of Artificial Intelligence*, pages 532–535. IGI Global, 2009.
- [114] Isabella Peters. *Folksonomies: Indexing and retrieval in Web 2.0*, volume 1. Walter de Gruyter, 2009.

- [115] M. Pomplun and M. H. Omar. Multiple-Mark Items: An Alternative Objective Item Format? *Educational and Psychological Measurement*, 57(6):949–962, December 1997.
- [116] Darrell Porcello and Sherry Hsi. Crowdsourcing and curating online education resources. *Science*, 341(6143):240–241, 2013.
- [117] Ilona R Posner and Ronald M Baecker. How people write together [groupware]. In *System Sciences, 1992. Proceedings of the Twenty-Fifth Hawaii International Conference on*, volume 4, pages 127–138. IEEE, 1992.
- [118] Peter J Rich and Michael Hannafin. Video annotation tools technologies to scaffold, structure, and transform teacher reflection. *Journal of Teacher Education*, 60(1):52–67, 2009.
- [119] Matthew Richardson and Pedro Domingos. Building large knowledge bases by mass collaboration. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 129–137. ACM, 2003.
- [120] DR Ripkey and SM Case. A" new" item format for assessing aspects of clinical competence. *Academic Medicine*, 71(10):34–36, 1996.
- [121] Adexe Rivera, Santiago Candela, and Carmelo R García. Wikidis: a case of collaborative content management system for educative community. *Education*, 2010.
- [122] G Rodriguez and S Cueva. Social network and oer. In *OCWC Global*, 2009.
- [123] Luisa M Romero-Moreno, F Javier Ortega, and José A Troyano. Obtaining adaptation of virtual courses by using a collaborative tool and learning design. In *Proceedings of the 2007 Euro American conference on Telematics and information systems*, page 42. ACM, 2007.
- [124] Jeremy Roschelle and Stephanie D Teasley. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*, pages 69–97. Springer, 1995.
- [125] Dewi Rosmala et al. Study of social networking usage in higher education environment. *Procedia-Social and Behavioral Sciences*, 67:156–166, 2012.
- [126] Przemysław Różewski. Model of community-build system for knowledge development. In *Computational Collective Intelligence. Technologies and Applications*, pages 50–59. Springer, 2011.
- [127] Melike Sah, Wendy Hall, Nicholas M Gibbins, and David C De Roure. Semport ? a personalized semantic portal. In *18th ACM Conference on Hypertext and Hypermedia*, pages 31–32. Sheridan printing, 2007.
- [128] Satya S. Sahoo, Wolfgang Halb, Sebastian Hellmann, Kingsley Idehen, Ted Thibodeau Jr, Sören Auer, Juan Sequeda, and Ahmed Ezzat. A survey of current approaches for mapping of relational databases to rdf. Rdb2rdf incubator group report, W3C, 2009.
- [129] Arnaud Sahuguet. Everything you ever wanted to know about dtlds, but were afraid to ask. In *The World Wide Web and Databases*, pages 171–183. Springer, 2001.
- [130] Gail Salaway and Judy Caruso. The ecar study of undergraduate students and nformation technology. Technical report, 2008.
- [131] Sebastian Schaffert, François Bry, Joachim Baumeister, and Malte Kiesel. Semantic wikis. *IEEE Software*, 25(4):8–11, 2008.

- [132] Sebastian Schaffert, Julia Eder, Szaby Grünwald, Thomas Kurz, Mihai Radulescu, Rolf Sint, and Stephanie Stroka. Kiwi - a platform for semantic social software. In *SemWiki*, volume 464 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.
- [133] Albrecht Schmidt, Florian Waas, Martin Kersten, Michael J. Carey, Ioana Manolescu, and Ralph Busse. Xmark: a benchmark for xml data management. In *VLDB*, pages 974–985, 2002.
- [134] RC Serlin and HF Kaiser. A method for increasing the reliability of a short multiple-choice test. *Educational and Psychological Measurement*, 38(2):337–340, 1978.
- [135] Jayavel Shanmugasundaram, Eugene Shekita, Rimon Barr, Michael Carey, Bruce Lindsay, Hamid Pirahesh, and Berthold Reinwald. Efficiently publishing relational data as xml documents. *The VLDB Journal*, 10(2–3):133–154, 2001.
- [136] Jayavel Shanmugasundaram, Kristin Tufte, Chun Zhang, Gang He, David J. DeWitt, and Jeffrey F. Naughton. Relational databases for querying xml documents: Limitations and opportunities. In *VLDB*, pages 302–314. Morgan Kaufmann, 1999.
- [137] Clive Shepherd. Objects of interest. *Brighton East Sussex: Fastrak Consulting Limited*. Retrieved June, 14:2005, 2000.
- [138] Karl A Smith. Cooperative learning: Effective teamwork for engineering classrooms. In *2013 IEEE Frontiers in Education Conference (FIE)*, volume 1, pages 2b5–13. IEEE, 1995.
- [139] Dimitrios-Emmanuel Spanos, Periklis Stavrou, and Nikolas Mitrou. Bringing relational databases into the semantic web: A survey. *Semantic Web*, 3(2):169–209, 2012.
- [140] Johannes Strobel, David H Jonassen, and Ioan Gelu Ionas. The evolution of a collaborative authoring system for non-linear hypertext: A design-based research study. *Computers & Education*, 51(1):67–85, 2008.
- [141] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.
- [142] Darya Tarasowa and Sören Auer. Balanced scoring method for multiple-mark questions. In *5th International Conference on Computer Supported Education (CSEDU 2013)*, 2013.
- [143] Darya Tarasowa and Sören Auer. Balanced scoring method for multiple-mark questions. In *CSEDU*, pages 411–416, 2013.
- [144] Darya Tarasowa and Sören Auer. Collaborative authoring of opencourseware: The best practices and complex solution. In *Open Data for Education*, pages 103–131. Springer International Publishing, 2016.
- [145] Darya Tarasowa, Sören Auer, Ali Khalili, and Jörg Unbehauen. Crowd-sourcing (semantically) structured multilingual educational content (cosmec). *Open Praxis*, 6(2):159–170, 2014.
- [146] Darya Tarasowa, Ali Khalili, and Sören Auer. Crowdlearn: Collaborative engineering of (semi-)structured learning objects. In *Proceedings of the KESW 2012*, 2012.
- [147] Darya Tarasowa, Ali Khalili, and Sören Auer. Crowdlearn: Crowd-sourcing the creation of highly-structured e-learning content. *iJEP*, 5(4):47–54, 2015.
- [148] Darya Tarasowa, Ali Khalili, Sören Auer, and Jörg Unbehauen. Crowdlearn: Crowd-sourcing the creation of highly-structured e-learning content. In *5th International Conference on Computer Supported Education (CSEDU 2013)*, pages 33–42, 2013.
- [149] Darya Tarasowa, Christoph Lange, and Sören Auer. Measuring the quality of relational-to-rdf mappings. In *Proceedings of the KESW 2015*, pages 210–224, 2015.

- [150] FJ Tsai and HK Suen. A brief report on a comparison of six scoring methods for multiple true-false items. *Educational and psychological measurement*, 53(2):399–404, 1993.
- [151] Daniel Tunkelang. Faceted search. *Synthesis lectures on information concepts, retrieval, and services*, 1(1):1–80, 2009.
- [152] Jörg Unbehauen, Claus Stadler, and Sören Auer. Accessing relational data on the web with sparqlmap. In *Semantic Technology*, pages 65–80. Springer, 2012.
- [153] C.J. Van Rijsbergen. *Information Retrieval*. 1979.
- [154] Graham Vickery and Sacha Wunsch-Vincent. *Participative Web And User-Created Content: Web 2.0 Wikis and Social Networking*. Organization for Economic Cooperation and Development (OECD), Paris, France, 2007.
- [155] Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004.
- [156] Chien-min Wang and David Turner. Extending the wiki paradigm for use in the classroom. In *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on*, volume 1, pages 255–259. IEEE, 2004.
- [157] Xiangyu Wang, Peter E.D. Love, Robert Kline, Mi Jeong Kim, and Peter Rex Davis. Integration of e – learning 2.0 with web 2.0. *ITcon - Special Issue eLearning 2.0: Web 2.0-based social learning in built environment*, 17:387–396, 2012.
- [158] Zhao-hong Wang. The study of the educational resources sharing model based on grids. In *Database Technology and Applications (DBTA), 2010 2nd International Workshop on*, pages 1–4. IEEE, 2010.
- [159] Etienne Wenger. *Communities of practice: Learning, meaning, and identity*. Cambridge university press, 1998.
- [160] D Woods. The myth of crowdsourcing: Crowds don't innovate—individuals do, 2009.
- [161] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment methodologies for linked open data. *Semantic Web Journal*, major revision, 2013.
- [162] Hong Zhang and Hongjun Su. A collaborative system for software engineering education. In *Computer Software and Applications Conference, 2007. COMPSAC 2007. 31st Annual International*, volume 2, pages 313–318. IEEE, 2007.

List of Figures

1.1	Collaborative authoring platforms in various domains	6
2.1	Social tools [122].	17
3.1	Steps followed to scope the search results.	23
3.2	Distribution of selected papers between a priori identified concepts	27
3.3	The mind map for Collaborative authoring of reusable OCW	29
4.1	CrowdLearn concept.	43
4.2	Conceptual view of the WikiApp data model.	45
4.3	Overview of the CoSMEC concept	48
4.4	CoSMEC mechanism of the multilingual content co-evolution: 1 - initial translation, 2 - partial translation to synchronize the content of translation with the source, 3 - merging the revisions to avoid the repeatedly translated content	49
5.1	SlideWiki conceptual scheme.	54
5.2	SlideWiki architecture overview.	55
5.3	Decision flow during the creation of new slide and deck revisions.	56
5.4	Four screenshots of SlideWiki features. <i>1 - Tree structure of the presentation, inline WYSIWYG editor; 2 - Editing of a question and manual assigning to a test using lists; 3 - Question in learning mode with correct answers displayed; 4 - Module-based scoring of an assessment test.</i>	58
5.5	Example object diagram for Slidewiki. <i>1 - connection between questionnaires and deck revisions to enable module-based scoring; 2 - connection between questions and slides to allow users see the related material when answering the questionnaires</i>	60
5.6	An overview of SlideWiki MySQL tables	62
5.7	WikiApp data model implemented as an ontology	64
5.8	SlideWiki deck imported in Moodle as a SCORM package	69
5.9	Interface of translations management: 1 - language of selected object; 2 - a drop-down list with links to languages available fro the selected object; 3 - button for translation; 4 - dynamically updated list of the languages supported by Google Translate service	69
5.10	Two scenarios of content synchronization between translation and source: 1 - automatic; 2 - manual synchronization	70
5.11	Interface for comparing and merging deck revisions	71
5.12	Percentage distribution between original and translated versions vs. total number of content objects for slides and decks. Blue line shows an example time moment, discussed in the paragraph above.	72
5.13	Distribution of languages for content objects and new visitors	72
5.14	The statistics of the evaluation	75
5.15	Results of SlideWiki evaluation survey: mean μ and standard deviation σ	77
5.16	The deck tree navigation component: current state (left) and proposed improvements (right)	81

5.17	The search component: current state	81
5.18	The search component: proposed improvements	82
5.19	New Architecture of SlideWiki platform with an example usage (left side)	83
B.1	Flow chart of the Balanced scoring algorithm.	112
C.1	Matching the DB tables and columns to WikiApp ontologie classes and properties	128
C.2	Matching the DB tables and columns to WikiApp ontologie classes and properties (cont.)	129
C.3	Matching the DB tables and columns to WikiApp ontologie classes and properties (cont.)	130
C.4	SlideWiki Knowledge graph	131

List of Tables

3.1	Surveys and essays studying technologies in application to collaborative OCW authoring	26
3.2	Surveys and essays addressing the issues of collaborative OCW authoring	26
3.3	OCW collaborative authoring papers distribution	28
4.1	Chosen approaches for solving OCW collaborative authoring challenges	42
5.1	Description of SlideWiki Database Tables	61
5.2	Summary table of the proposed metrics system	65
5.3	Mapping quality evaluation.	66
5.4	Incorporation of frequently used complicated queries results into the mapping to simplify usage the linked data	68
5.5	The multiple choice questions and the average of the answers (1 = strongly disagree and 5 = strongly agree), categorized by the cell colors: dark-gray referring to “Responsiveness of the website”, (2) light-gray referring to “Design and usability” and (3) white referring to “Functionality and awareness of the website features and usage”	79
5.6	The most liked aspects of SlideWiki according to the final user evaluation	79
5.7	Demographical data of users participating in the single-user observation study.	80
A.1	Summary table of proposed metrics system	92
A.2	Data quality metrics divided into four groups according to the stage of mapping creation, on which they can be influenced.	95
A.3	Metrics for semantically identical properties to link an organization and its chief	97
A.4	Results of survey on degree of agreement with the key concepts of the proposed methodology. Numbers represent the degree of agreement from <i>absolutely agree</i> (5) to <i>absolutely disagree</i> (1). The colors indicate the self-estimated level of expertise in the RDB2RDF domain from <i>expert</i> (darkest) to <i>experienced</i> (lightest).	99
B.1	Comparison of the proposed approach with other existing approaches	107
B.2	Comparison of the proposed approach with other existing approaches	107
B.3	Comparison of the proposed approach with other existing approaches	108
B.4	Comparison of the proposed approach with other existing approaches	109
B.5	Comparison of the proposed approach with other existing approaches	109
B.6	Comparison of the proposed approach with other existing approaches	110
B.7	Comparison of the proposed approach with other existing approaches	110