Semantic Document Management for Collaborative Learning Object Authoring

Saša Nešić[1], Dragan Gašević[2], Mehdi Jazayeri[1]
Faculty of Informatics, University of Lugano, Switzerland
School of Computing and Information Systems, Athabasca University, Canada
sasa.nesic@lu.unisi.ch, dgasevic@sfu.ca, mehdi.jazayeri@unisi.ch

Abstract

In this paper, we propose the use of semantic documents as learning objects. The core part of our solution is a semantic document model that allows for unique identification of document content units (CUs) and their annotation with different types of metadata. On the top this model, we have developed the Semantic Document Management System (SDMS), which enables efficient collaborative authoring of learning objects within a social network of content authors, by reusing document CUs based on the accumulated metadata.

1. Introduction

Learning object (LO) authoring is a core phase in the LO lifecycle, which may impact the quality of the overall learning process. To create high-quality LOs, a content author has to have a high level of expertise, not only in the subject area, but also in instructional design as well as various learning technologies. As such, authoring is rather an expensive activity. Striving to address this issue, the research in e-learning has focused on increasing the reusability of LOs. The most significant result is the IEEE Learning Object Metadata standards as well as many implementations of LO repositories (e.g., MERLOT) and federated protocols for (e.g., ECL and SQI) for networks of LO repositories (e.g., GLOBE). While these efforts have demonstrated some significant potential to improve today's practices, there are still some issues that are important to be addressed to improve the current state of LO authoring:

- Learning objects are typically created for a specific purpose and to be used in a particular context. This way of authoring of LOs is then inherently used in search engines, so that one can only search for LOs as a whole. Yet, authors usually need a part of a LO, which is related to a certain domain concept and plays a certain pedagogical role (e.g., illustration) [8].
- Reuse of LOs is depended on the format of authoring tools in which they were created. This hampers machines to interpret the meaning of parts of LOs authored in different tools [7].
- Most commonly used authoring tools (e.g., Word) are not integrated with repositories of LOs where the content to be reused can be found. This puts an additional burden on content authors to use several different tools in parallel while creating new LOs, which may reduce their creativity and productivity.
- Repositories of LOs are mainly used as centralized stores of LOs and their metadata. However, today's Web 2.0 technologies demonstrate that a significant content sharing can be achieved by participating in a community and by leveraging social relations with peers (i.e., social networks of content authors).
- Historical changes of LOs and experiences in their use are typically not available in metadata about LOs. The lack of this metadata may limit content authors to understand the contexts of the use of LOs and reasons for changes to LOs. Moreover, this affects collaboration among content authors, who do necessarily not have to work at the same time on revisions of LOs.

To address the above issues, we propose the use of semantic document management, as concept that allows for platform-independent and semantically-rich management of LOs. Representing LOs as semantic documents built on top of the Abstract Compound Content Model (ACCM), we facilitate unique identification of parts of LOs and their different versions (Sect. 2). In that way LO parts can be directly accessed and annotated with different types of metadata, including standardized (e.g. LOM and DC), metadata about the interaction between users and LO parts as well as domain-specific metadata (Sec 3). All this metadata is leveraged in a collaborative authoring process (Sect 4.), which is based on authors' profiles (e.g., social relations) and which is supported by extensions of popular authoring tools (Sect. 5).

2. Semantic Documents as Learning Objects

There are many definitions of a LO, but in a nutshell, most of them characterize a LO as an entity, digital or non-digital, that may be used for learning, education or training. A more specific definition, which we are focused on, defines a LO as a digital, self-contained, reusable entity with a clear learning objective that contains at least three editable

components: ***content***, ***context elements*** and ***instructional activities*** [1]. On the other hand, a semantic document [2] is defined as a combination of electronic documents and ontologies. So far, most of the existing semantic document models have primarily focused on storing the ontologies and relationships between ontological concepts and document parts (i.e. ontological metadata [3]) in the internal document representation. In our approach, we introduce a new semantic document model, which relies on the use of Abstract Compound Content Model (ACCM) [4] as an additional layer in the document content structure representation. On the top of the ACCM we have developed the *ACCM-core ontology,* which formally represents the model, and is used to capture the internal structure of the document. Our semantic document model represents a combination of a regular electronic document (e.g., a Word document) and a semantic layer that contains the document instance of the *ACCM-core ontology*. For each document content unit (CU) there is an instance of the appropriate concept from the ontology, which is identified with the unique resource and version identifiers (i.e., URI and VID). The copies of these identifiers are embedded in the document, thus forming the link between the CUs and their ontological instances. This enables storage of the ontological metadata about the CUs outside the document in which they occur.

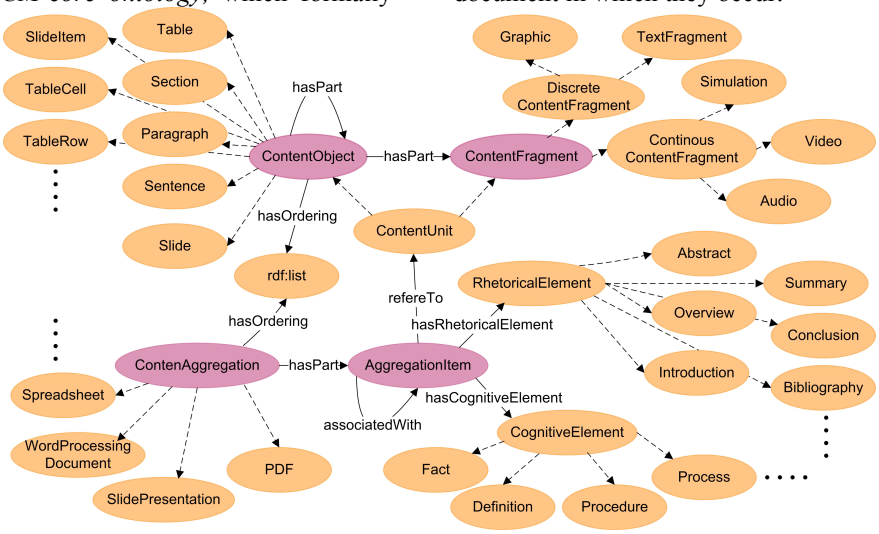

Figure 1. Abstract Compound Content Model (ACCM) – core ontology

Now we explain the *ACCM-core ontology* (Figure 1) and show how our semantic document model can satisfy all three components of a LO as we described in the text above. The *ACCM-core ontology* defines structural elements of document content and possible relationships between them. The ontology contains two groups of concepts: context-independent and context-dependent. *The first group* contains concepts related to the structural elements of a document content, which keep meaning when they are outside a document and can be reused in different contexts (i.e., the ***content*** component of a LO). The most abstract concept in this group is *ContentUnit*, which has two main sub-concepts *ContentFragment*, and *ContentObject*. The *ContentFragment* concept represents CUs in their most basic form (i.e., raw digital resources) that can be further specialized into *DiscreteCF* (e.g., graphics and texts) and *ContinuousCF* (e.g., audio and video). The *ContentObject* concept represents CUs, which collect *ContentFragments* and add navigation (e.g., section and table). *The second group* contains context-dependent concepts, which are related to structural elements that cannot exist without the document. Two core concepts are *AggregationItem* and *ContentAggregation*. An *AggregationItem* holds a reference to an instance of the *ContentUnit* concept via the *refersTo* property and represents the appearance of the CU within a document. The *AggregationItem* enables addition of the ***context elements*** (i.e., the second component of a LO) to a CU when it is considered to be a part of a document. From the prospective of a LO, it enables annotation of CUs with the pedagogical/instructional role(s) that they might have within the context of a document. A CU within a document may have multiple pedagogical roles defined from either a rhetorical (e.g., abstract and introduction,) or a cognitive prospective (e.g., fact and definition).

The *ContentAggregation* defines and organizes the *AggregationItems* and indicates relationships among them. It can also serve as an outline or a table of content deliveries (e.g., a slide presentation or a Word processing document). Besides aggregational relationships, which are expressed with a *hasPart* property, the *ContentAggregation* defines navigational

and associative relationships via the *hasOrdering* and *associateWith* properties. The aggregational and navigational relationships enable sequencing and structuring of the document content in a manner that could enable various learners to achieve the stated learning goals. On the other hand, the associative relationships enable the links among similar aggregation items based on a given criteria, such as the level of difficulty, which allows alternative paths within the *ContentAggregation*. The aggregational, navigational and associative relationships, along with the context-dependent and context-independent semantics, can offer a good way of arranging media and content to help learners and teachers transfer knowledge most effectively (i.e., to satisfy the third component of a LO - ***instructional design***).

The *ACCM-core ontology* is essential but not sufficient to provide all capabilities that we expect from the semantic documents. Besides the *ACCM-core ontology* we have developed two more ontologies, which are the constitutive parts of the semantic layer: *ACCM-changing* and *ACCM-annotating ontology*. The *ACCM-changing* ontology captures possible changes to CUs and creates links between them and ontological instances of CUs. Since the ontology is not essential for this work, we do not explain it in detail. We recommend the interested reader refer to our project Web page [5] for more information. The *ACCM-annotation ontology* is the subject of the next section.

3. Semantic Annotation of Document CUs

One of the main objectives of our semantic document model is to enable software agents to easily discover, access and reuse document CUs of different levels of granularity without affecting the document as a whole. However, prior to the access and reuse, document CUs have to be discovered which demands the annotation of CUs with a substantial set of metadata. Ordinary electronic documents have very limited annotations (e.g., DC or IEEE LOM metadata), and that metadata is related to the document as a whole. We cannot only rely on that metadata to have well annotated CUs. There are three types of metadata that can be used for the annotation and we have defined them in the *ACCM-annotation ontology* [5]:

Derived metadata – metadata derived from the document's metadata and formatting information. Some of this metadata are literally copied from the document's metadata like *dc:creator*, *dcterms:created*, *dc:format* and *dc:language*, referring to the author(s), creation date, media type and language(s) respectively. A value of a dc:title element is generated based on the formatting information. For example, a text fragment with a font style (e.g., *title* or *heading1*) is used as a value for the *dc:title* element of all successive CUs up to the next formatted text fragment. A value of a *dc:description* element is generated out of the values of previously explained elements using the following text pattern: "A content unit of {*dc:format*} media type with a title {*dc:title*} authored by {*dc:creator*}; creation date {*dcterms:created*}"[8].

Interaction metadata – metadata about the interaction between the users (e.g., authors, instructors, learners) and the document CUs. Since our intention is to have a social network of users who can share their documents by interacting (e.g., visiting, modifying, reusing) with the document CUs over Semantic Web protocols [10], this metadata plays an important role in our approach. This metadata is primarily used to determine how discovered CUs correspond to the user's preferences (e.g., if the user prefers recently modified CUs or CUs reused many times, etc.). The *ACCM-annotation ontology* conceptualizes aspects of the interaction by introducing three new concepts: *Modification*, *Reuse*, and *Visit;* and three new properties: *numOfVisits*, *numOfModifications* and *numOfReuses*. All three introduced concepts are characterized by the time of the interaction and person who is involved in it. The *Modification* and *Reuse* concepts also track information about deployed applications. Every time the user interacts with the CU, the interaction metadata is added to the CU.

Ontological metadata – metadata that keeps relationships between ontological concepts from domain ontologies and document CUs. To obtain this metadata, we apply text-mining techniques and try to identify if the CU contains labels of some ontological concepts. For each label found, the CU is annotated with the ontological concept tagged by the label. If the chosen domain ontology has properties, which determine semantically related or equivalent concepts, we annotate the CU with those concepts as well. The *ACCM-annotation ontology* uses the *dc:subject* property to add ontological metadata.

4. Semantic Document Management

Based on the introduced semantic document model, we propose a semantic document management system (SDMS), which provides services for managing a lifecycle of semantic documents, from their authoring and publication to archival. We see the SDMS as a part of a collaborative environment, which supports the sharing and exchanging of document contents and semantics across social and organizational relationships. Now we describe briefly a user profile model of the SDMS, as it has a significant impact on the content authoring process that we describe in the rest of the section.

4.1 User Profile Model

The user profile model is strongly inspired by the notion of collaborative environments and social relationships. The user is a part of a network of people who are interested in similar topics and want to share their documents. To formally represent the user profile model we have developed *User-model ontology* [5]. The ontology combines the vocabulary of the FOAF (Friend of a Friend) and SIOC (Semantically Interlinked On-Line Community) ontologies with the concepts and properties that we have introduced to capture the user's preferences regarding the choice of document CUs for reuse. The examples of these preferences are: the sorted list of preferred authors; the sorted list of preferred document formats; and the information if the user prefers document CUs that are often reused, recently modified or CUs which have many versions. This information helps us determine which document CU is the most suitable for the user from the set of discovered, topic-relevant CUs.

4.2 Learning Content Authoring Scenario

Let us suppose that Claudia is working at university as a teaching instructor and has to prepare a presentation for her next lecture on the topic: "Pattern oriented software architecture". In order to prepare as good presentation as possible, with up to date information, she wants to combine her presentation from last year with presentations of her friends (colleagues) from other universities as well as other related artifacts that she can find in her archive or the archives of her friends. To prepare the presentation, Claudia needs to go through the following steps: to obtain documents from different sources, to read them and determine whether they really contain content relevant for the topic and at the end to copy and paste desired parts in the new presentation. Obviously, such authoring of LOs demands a lot of time and effort.

Let us now suppose that Claudia and her friends have the SDMS installed on their computers and can remotely access and query each repository of semantic documents. Claudia would first sketch an outline of the presentation (e.g., a PowerPoint slide presentation). Than the application, i.e. PowerPoint calls some services of the SDMS and internally maps this outline into a form of the ACCM *ContentAggregation*. For each 'leaf' item of the outline (i.e., each *AggregationItem* in term of the ACCM model), Claudia assigns a domain concept that should be addressed in that part of the presentation. Claudia actually specifies a concept as a set of terms, which are then used by the SDMS to internally search selected domain ontologies for concepts labeled with those terms or some of their synonyms (the synonyms can be obtained through a lexical ontology such as WordNet) and finally assigns discovered concepts to the outline item. In this case, the domain ontologies are about the software architecture domain. Besides the domain concept(s), Claudia might specify the media type of the CU (e.g., *text*, *image*, *audio* and *video*) as well as the CU's pedagogical role (e.g. *definition*, *example*, *exercise* and *algorithm*). The combination of specified concept, media type and pedagogical role is internally transformed into a query in the SPARQL RDF query language [10]. The SDMS search service queries the *User-model ontology* to find out a list of Claudia's friends and then for each specified outline item, executes SPARQL queries against remote SPARQL endpoints using the SPARQL protocol. When the search service retrieves a set of CUs, the SDMS ranking service queries the *User-model ontology* to obtain Claudia's preferences regarding the choice of CUs and ranks the CUs within the retrieved set. The ranking service applies a ranking algorithm based on weighting schemas that we presented in [6]. For each CU from the retrieved set, Claudia can get a preview of its content and all available metadata. By selecting the CUs for each item of the outline, she assembles the whole presentation.

5. Tool support

To prove our conceptual solution, we choose the MS Office 2007 document format (OpenXML) and built the SDMS as a set of six modules: 1) core transformation module; 2) annotation module; 3) indexing module; 4) changing and versioning module; 5) search module; and 6) ranking module. The functionalities of the modules are accessible through the interface of the two main MS Office add-ins: **Transformer add-in** and **Authoring recommender add-in**. In addition to these two, we have also developed the Ontology and User profile manager add-ins to manage ontologies used by the system's modules and the user's personal information.

The **Transformation add-in** transforms office documents (i.e., Word and PowerPoint) into semantic documents by adding a semantic layer that consists of the document instances of the *ACCM-core ontology* and *ACCM-annotation ontology*. In addition, during the transformation process, the indexing module does text indexing for all textual data from the document. The changing and versioning module captures possible changes to the document and its CUs by comparing them to previously transformed document versions.

The **Authoring recommendation add-in** (Figure 2) is a tool that satisfies a set of the SDMS functionalities elaborated in the given content

authoring scenario. The add-in uses functionalities of the search and ranking modules to allow users to simultaneously search their own semantic documents as well as those of their friends for desired CUs. The GUI of the add-in enables users to select domain ontologies whose concepts take part in composing the query along with the other search criteria (Figure 2a). The retrieved CUs are exposed to the author as a ranked list of CUs and the author can get a preview of the CU and its metadata, as well as the evaluation path, by browsing the CU's versions from the versioning tree (Figure 2b). Once the author has selected a CU to reuse, the CU is added in the current cursor position of the active document.

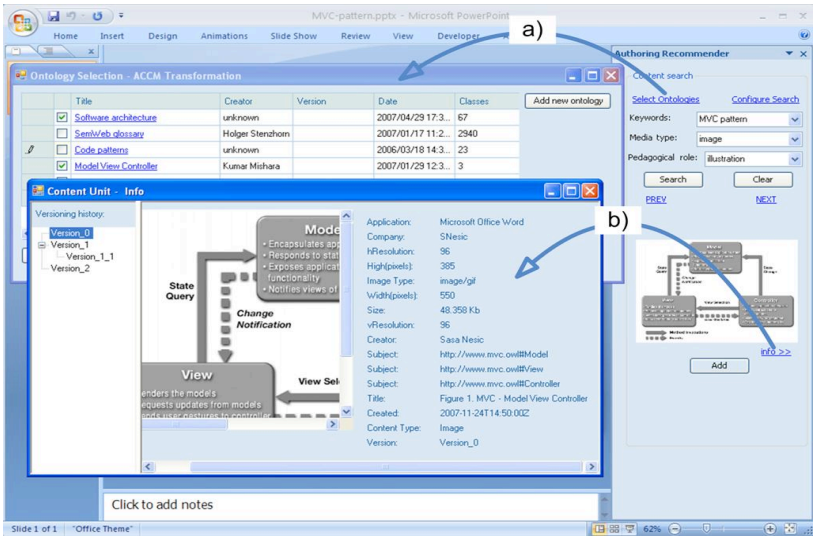

Figure 2. Authoring recommender add-in a) Ontology selection b) Content unit preview

6. Related Work and Conclusions

The ACCM model used in our semantic document approach was partially inspired by the Abstract Learning Object Content Model (ALOCoM) [7] which is the basis for the learning content authoring in the ALOCoM-framework [7]. The framework relies on the centralized repository of extracted document CUs that are managed independently from the documents they originate from. On the contrary, in our approach we have decentralized repositories of whole documents, enriched with semantic layer that enables annotation and direct access and reuse of document CUs. The other approach that can be compared with ours is the approach applied within the CARLOS (Collaborative Authoring of Reusable Learning Objects System) platform [9]. This approach also tries to integrate ontologies into collaborative authoring of LOs, but its main focus are collaborative protocols and mechanisms within the development of the LO.

In summary, we have presented the new semantic document model that enables efficient collaborative authoring of LOs by reusing document CUs of different levels of granularity. We have described the main features of the model and proposed the SDMS with services for the collaborative authoring. The current version of the SDMS has support for the MS Office documents but we plan to add support for other common document as well. We will also work on improving services towards a fully automated authoring process, where social networks of content authors will fully be leveraged.

7. References

[1] Rehak, D. R., et al., "Keeping the learning in LOs," Kogan Page, London, 2003, pp.22-30.

[2] Eriksson, H., "The semantic-document approach to combining documents and ontologies", *Int'l J. of Human-Computer Studies,* 65(7), 2007, pp. 642-639.

[3] Handschuh, S., Volz, R., Staab, S., "Annotation for the Deep Web," *IEEE Intell. Sys.*, 18(5), 2003, pp. 42-48.

[4] Nešić, S., et al., "Ontology-Based Content Model for Scalable Content Reuse," *In Proc. 4th ACM K-CAP Conf.*, 2007, pp. 195-196.

[5] http://www.inf.unisi.ch/phd/nesic/sdms/

[6] Nešić, S., et al., "An Ontology-Based Framework for Authoring Assisted by Recommendation," *In Proc. 7th ICALT Conf.*, 2007, pp. 227-231.

[7] Verbert, K. et al., "Ontology-based Learning Content Repurposing: The ALOCoM Framework," *Int'l J. on E-Learning*, 5(1), 2006, pp. 67-74.

[8] Jovanović, J. et al., "Ontology-based Automatic Annotation of Learning Content," *Int'l J. on Semantic Web and Information Systems*, 2(2), 2006, pp. 91-119.

[9] Dodero, J.M. et al., "Integrating Ontologies into the Collaborative Authoring of Learning Objects," *Int'l J. of Universal Computer Sci.*, 11(9), 2005, pp. 1568-1576.

[10] SPARQL: http://www.w3.org/TR/rdf-sparql-protocol/