Towards a flexible approach to manage varying and altering information representations

Tobias Haubold, Georg Beier, Wolfram Hardt
Zwickau University of Applied Sciences, Informatics, Zwickau, Germany
E-mail: tobias.haubold@fh-zwickau.de

Abstract

Information and content is described using content or document models. Content production always has a particular purpose, target audience and language. Often various forms of representations are created in different data formats and all overlap in content and structure. Fragments of such content have relations across different origins. Such relations refer to a finer granularity than the elementary unit of most management tools. Thus management tools are unable to provide automated support for such relations and consistency must be manually ensured during maintenance.

Using the example of learning and teaching content this paper outlines key challenges of the problem domain, fundamental characteristics of content production and management and necessary concepts of content and document models to address these problems. Related work on content and document models as well as management approaches is discussed and an approach for a repository architecture is presented to target these problems on management level.

1. Motivation

Digital learning and teaching materials are usually the result of target oriented content production processes. They are created using appropriate editing tools (Office-Tools like Word, Powerpoint, Apache OpenOffice, LibreOffice, image processing applications, special tools to produce SCORM-based e-learning content, and more) and managed using the specific data format of the tools.

Every kind of knowledge transfer pursues an intention and follows a didactic approach. Thus teaching in presence (presentation slides), self-study (lecture notes) and interactive online learning (web content) demand all different kinds of content representations. The purpose of materials is significantly characterized by the learner as target audience. Thus materials differ by means of the didactic design, prior knowledge, expert and non-skilled target audience or fundamental or advanced knowledge.

To create such forms of content within a reasonable time copy and paste is usually used to reuse content fragments by duplication. In addition the use of materials in international cooperation demand translation and localization.

On update or revision of content the consistency of partially redundant content fragments spread across different document files need to be ensured manually. The process of locating related content, detecting differences and collecting content for refreshing translation and localization must be done by hand and leads inevitably to time consuming maintenance of content.

This paper focuses on requirements of content and document models as well as content management approaches.

2. Key Challenges in Problem Domain

The process of producing learning and training content starts by defining a main learning path through the knowledge domain. According to this path the material is structured and created using an didactic approach (see [1]). Thus different documents are produced (e.g. presentation slides and scripts) sharing the same structure but contain different representations of the content. During content maintenance the common structure must be ensured manually to keep the materials consistent.

During the production of materials with a similar target audience copy and paste is usually used to reuse content by duplication. Thus content fragments are stored redundantly in different document files. During update or revision all duplicated content fragments must be maintained manually to ensure consistency. The same applies to derived content. Content fragments are duplicated and subsequently changed to better fit other purposes or author styles. During change of the comprised knowledge all derived content must be considered for manual maintenance.

Multilingual materials with a high rate of change need frequent refreshing translations and localizations. When the
latest updates are not yet available, one must deal with incompletely translated and localized materials.

Approaches to address these challenges involve document and content models as well as content management approaches.

3. Fundamental Characteristics of Content Production and Management

3.1. Reuse of Content

By using copy and paste content can be reused by duplication. This results in redundant storage and thus content updates have to be done manually on each copy to ensure consistency.

Content references allow to reuse content without duplication and redundant storage. Instead of the content itself just a reference to it is stored. Whereas copy and paste is a feature of the editing tool, content references are a feature of the document model and must be supported by it.

3.2. Granularity in Content Production

In this context granularity is understood as the extend of content. Document models define concepts to structure content (e.g. headings for hierarchical structures in text documents or slides in presentation software) and capture content (e.g. text in text boxes or paragraphs). Each concept represents a level of granularity. Thus there are levels with finer and coarser granularity.

Especially in document oriented content production processes resulting document files are tailored to particular purposes and have a coarse granularity. Within learning object based content production processes Hamel et al. states that the granularity should be relatively small. In addition Duval et al. states that learning objects with finer granularity are more easily reusable.

3.3. Granularity in Content Management

All management tasks including versioning, categorization, meta data management as well as translation are based on an elementary content management unit. Document Management Systems (DMS) and versioning tools use simple files as management unit. Due to the coarse granularity of document files such tools are not suited to address the key challenges of section 2. Content Management Systems (CMS) use the concepts of their content model as management unit. Thus they provide finer levels of granularity but prescribe particular content models.

4. Solution Approaches in Content Production

To address the key challenges content or document models need the following mechanisms.

4.1. Content References

The document model must provide support for content references. To use them they must be allowed on the desired position and the type of content must be supported.

There are two kinds of content references: references to files and references to file fragments. In the latter case just the referenced fragment is included instead of the whole file. This requires mechanisms to address content fragments defined by the referenced document model.

4.2. Content Aggregations

Mechanisms to aggregate content are based on content references and allow to compose documents out of other documents and document fragments. In addition to content inclusion the content is adapted to the current context if required, i.e. it is included on the desired position in the hierarchical structure of the document and the document wide layout and formatting settings are applied. Layout and formatting settings particularly defined for the included content are not adjusted.

4.3. Content Filtering

Mechanisms for content filtering allow for conditional content inclusion. Some kind of configuration settings specify whether or not particular content fragments are included or not. Usually such mechanisms evaluate meta data attached to content fragments.

4.4. Separation of Content and Presentation

There are two common approaches to separate content and presentation. First, the content contains meta data referencing presentation details. A presentation engine resolves these references and renders the content appropriately. Examples are the HyperText Markup Language (HTML) which references Cascading Style Sheets (CSS) files or Open Document (ODF) which defines format templates referenced by content.

Second, there are transformation based approaches. The content is either described by an XML based document model or an implemented content model. Transformations are described using the Extensible Stylesheet Language Transformations (XSLT) or an appropriate template language. An XSLT or template processor takes the content and transformations as input and produces output documents. Examples are DocBook, DITA or CMSs.
5. Proposed Content Management Approach

To provide automated tooling support for the key challenges mentioned in section 2, content management must detect and manage content references and content aggregations and must support meta data for content filtering on a fine level of granularity. The proposed approach focuses on a repository architecture to manage fine granular content fragments and to capture content references and aggregations in a document and content model independent manner.

5.1. Content Structure Layer

The structural aspects of content are captured by this layer, illustrated in figure 1. Opaque content represents any unstructured content and can be compared to the management unit of other management approaches. Content nodes are used to represent structured content in the form of a tree by having any number of children nodes. Content can be attached directly to content nodes or by referencing other assets. All assets are identifiable by an URI (Uniform Resource Identifier), may have any number of arbitrarily meta data attached and store content binary.

5.2. Semantic Layer

All assets are resources in the semantic sense. This layer is used to classify content and to capture all kinds of content relationships. It resembles a large graph with nodes being content assets or concepts defined by semantic technologies. The edges constitute relationships between content assets, relationships to semantic meta data as well as relationships between semantic concepts.

This layer enables the use of existing meta data standards like Dublin Core [4] or LOM (Learning Object Metadata) [5] as well as their extension or supplement with custom concepts, categories, terminologies or taxonomies defined using SKOS (Simple Knowledge Organization System) [6], RDFS (Resource Description Framework Schema) [7] or OWL (Web Ontology Language) [8]. Having all content assets and semantic meta data accessible in one huge graph allows for simple querying using SPARQL [9] or gremlin [10] as well as easy exploration for similar content.

5.3. Data Storage Layer

The main purpose of this layer is to persist data. All data is stored as binary stream. A separate data storage layer is suited to support deduplication and distribution of data to enable the use as distributed repository.

5.4. Repository Architecture

An adapter mechanism is used to map content into the proposed management model. File based adapters build appropriate content structures based on content or document model concepts as well as content meta data. A plug-in mechanism supports extensible file based mappings. Adapters may be integrated into content production environments like Office-Suites as well.

The adapter concept allows for type as well as instance based mappings. Thus individual content fragments of the same content type may be mapped differently. Furthermore adapters may overcome the limitations of document models by providing extensions to support content references and aggregation mechanisms.

6. Related Work

\LaTeX{} supports content references to files and separates content and presentation. It does not support content aggregations (because explicit outline level for headlines), content filtering and references to file fragments.


DITA [15] supports content references to files and document fragments, content filtering, separation of content and presentation as well as content aggregations directly.

Document models of office suites support content references basically to images, multimedia and OLE objects (Object Linking and Embedding). Content aggregations are not supported. OLE is used to include document fragments of other OLE supported applications in a display oriented manner. If no appropriate application is available, included document fragments are usually replaced by images. Open Document [16] text documents provide the concept of sections for modular editing. It allows to include fragments of other documents by reference and replaces the document wide style settings. Because outline levels are explicitly stated it cannot be considered as content aggregation mechanism.

\(^2\)the element section can be nested and specifies the outline level relatively
Only DITA provides mechanisms for all solution approaches mentioned in section 4. All other document models are not suited by default to target all key challenges of section 2. In addition content management approaches lacking support too.

According [17] there are different kinds of CMSs but there is no clear distinction between them. In accordance with [18] related kinds of CMS are Web CMS, Learning CMS, and Document Management Systems (DMS). Web CMS as well as Learning CMS are limited to web technology based publishing [18]. They might be suited to manage content for interactive online learning but not for materials used for teaching in presence like presentation slides or lecture notes. The proposed approach is suited for any kind of content model and don’t have such limitations.

DMSs are designed to manage document files. They do not consider their internal structure. Thus they are not capable of detecting and managing content references and content aggregations. Alfresco [19] as example provides support to manage translations on document level. Thus it supports the third key challenge with file-based granularity. The proposed approach is able to manage content with relations on any level of granularity. The management software edu-sharing [20] is a distributed repository for teaching content. It is based on Alfresco and therefore bound to the same content management approach [21].

### 7. Summary

Document oriented production of learning and teaching materials is still prevalent in institutions of higher education like universities. With “documents must die” criticized [22] such approaches almost 10 years ago and states that efforts for share and reuse based on simple file oriented document models are bound to fail. He argued that we need more sophisticated content and document models. Though many content and document models evolve to open standards, they still have limitations and need to be supported by a flexible yet scalable content management approach.

Further research includes the mapping of content and document models to the management model including the analysis of consequences, its eligible and effort as well as feasibility to extend document models with aggregation mechanisms.

### References


