Data transformations from CMS to CDP enriched by semantics

Christina Salwitzek[1*] and *Christina* Steuer[1**]

[1]Karlsruhe University of Applied Sciences, Faculty of Information Management and Media, 76133 Karlsruhe, Germany

Abstract. Today's users no longer expect a classic manual, but short, clearly structured pieces of information that fit their application context, use case and role. Instead of conventional documentation, "intelligent information" is required that is modular, format-neutral and can be found via metadata and full-text search. The information is often created in a CMS and provided via CDPs. There are not always compatible interfaces between these systems, especially those of different software manufacturers. Therefore, the information created cannot be processed further. The purpose of this paper is to show that data transformations can provide accessibility for the information from a CMS for different CDPs. On this basis, data transformations were developed, enriched by semantics and implemented within the project. For the enrichment by semantics, metadata were used as well as a further approach based on metadata, called "microDocs". This approach describes the combination and aggregation of different topic-based information that are connected by defined use cases and a logical context. Some CDP manufacturers already support microDocs and it is expected that even more extensions will be implemented in the future. Accordingly, it is highly likely that microDocs will play an important role in the field of information delivery.

1 Introduction

To make information accessible, it must be published in a certain format in a specific system. To provide information that is fast, target group-based and context-based, the information can be published in a Content Delivery Portal (CDP). The data source for a CDP is in most cases a Content Management System (CMS).

For many companies, a CMS is a standard tool for creating documentation. The CMS is tasked with the systematic collection, creation, storage and refinement of structured information and media data in a single, finely granulated stock. [1]

To ensure an efficient delivery of the created information to the user, it is transferred to a CDP. The CDP is an online system for providing this information. The provision is done by modular or aggregated information that are accessed by different target groups using content-related search mechanisms. [2]

1.1 Content retrieval

To ensure that the information is selectively adapted to the user's situation and efficiently accessible, the concept of "Content Retrieval" is used within a CDP. Content Retrieval is the search for information in a specific portal, for example a CDP. Based on Content Retrieval there are two main approaches to search for information within a CDP [2]. These approaches include the direct and the structured search. The main difference of these approaches is the usage of metadata. [2].

1.2 Direct search

The direct search is based on a full-text database which already includes search terms and phrases [3]. It can be optimized, by using ontologies or terminology databases. An example for the direct search is the "Google Search".

1.3 Structured search

The structured search is based on metadata which is handed over by the CMS where the information is initially classified. For the structured search there are two different approaches which differ in their usage of metadata and use cases. The two different approaches are "navigation" and "filtering".

The navigation approach involves the use of navigation trees to follow along the "classical document structure including nested topics" [3]. The basis for the navigation is intrinsic and extrinsic metadata, variant properties and the document structure (see section 3.2.1 "Metadata Concept" for more details about these metadata types).

The filtering approach uses facets to search "corresponding to complex classification taxonomies or more simple metadata sets from CMS" [3]. The basis for

* Christina Salwitzek: sach1023@hs-karlsruhe.de
** Christina Steuer: stch1033@hs-karlsruhe.de

the filtering is intrinsic and extrinsic metadata as well as variant properties or functional metadata.

Within the project the navigation enabled the user to maneuver through modules in a document and localize his position within a document. The data which was used to implement this approach was the document structure, which was built previously in the CMS. The filtering was used to filter a document ("outer facets") and to filter modules within a document ("inner facets"). A more detailed explanation about the faceting possibilities can be found in section 2.3 "Data Transformation Process".

2 Data transformations

Content is often created in a CMS and provided via a CDP. However, there are not always compatible interfaces between these systems, especially between systems from different software manufacturers. These technical barriers prevent a platform-spanning delivery of information. In this case, data transformations enable the transfer of data from a CMS to variant CDPs that differ in the data format. Therefore, data transformations ensure accessibility of the content in several CDPs of different manufacturers.

2.1 Basics of data transformation

In order to understand the basics of data transformation the most important components need to be discussed in more detail (see Fig. 1). These components are:

A: Data input: XML document and media
B: Transformation process: XSLT Processor and XSL Stylesheet
C: Data output: content published in different formats for multiple use cases

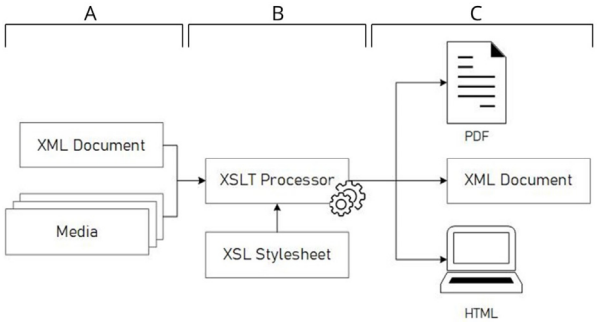

Fig. 1. Components of an XSL Transformation.

The basis of a data transformation is the input data itself. This data can include different content and formats, based on the use case and the source of the data. In order to process the data, an XSLT processor is required. The rules of processing are provided by an XSL stylesheet which is the core of the data transformation. Within the XSL stylesheet all transformation rules are defined for the appropriate output format. For each transformation, a batch file was used to automate this process. The result of the data transformation is the content from the input data in the required output format. This output format is used afterwards to publish the content. [4]

2.2 Data sources

The input for data transformations can be obtained from different data sources. Data sources that can be used for transformations are CMS, ontology editors and other databases (e.g. product information systems) [5]. The most common data source for a CDP is a CMS. Within a CMS the content is typically organized within modules in a certain XML structure, including metadata (see section 1 "Introduction"). Ontology editors as data sources are also getting more popular as ontologies enrich content by relations, which "improves the manual and automated search processes and allows a new and dynamic view on the information" [6]. Within an industrial context, not only CMS or ontology editors should be considered as data sources but also other databases within the company which provide useful information. Depending on the data sources, different output formats are provided. In general, most of the data sources provide the output format XML, which can be further processed via XSLT.

2.3 Data transformation process

The mentioned basics are part of the data transformation process in which certain content objects (also called topics) of a data stock, in this case a CMS, are transformed so that they can be correctly interpreted, indexed and found by the search system of an information portal, here the CDP. The data transformation process of this project can be divided into three essential stages: "preparation", "transformation" and "import".

2.3.1 Preparation

In the first stage, the preparation, the content must be created in the CMS editor in a standardized way and classified with metadata. The modularized content, which is to be published or exported, is compiled in a so-called "book". This book is exported in the descriptive mark-up language XML so it can be transformed with XSLT in the following stage. If the CMS content also contains graphics, a zip file will be created during the export, which includes the XML document and a separate folder with all graphic files (see Fig. 2). In order to be able to filter in the CDP according to different classification properties, an additional book, the metadata book, will be compiled which contains all possible metadata. This book is also exported as an XML document for further processing.

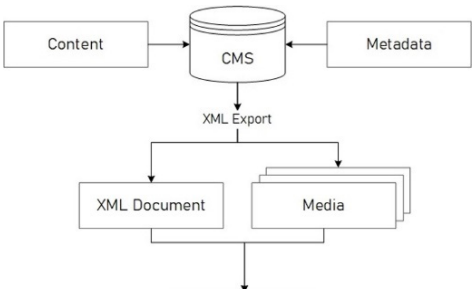

Fig. 2. Schematic illustration of the preparation stage.

2.3.2 Transformation

In the next stage, the actual data transformation, XSLT is used to convert the CMS content into the target format that can be interpreted by the CDP (see Fig. 3). By applying the XSLT stylesheet to the XML document, XSLT uses XPath to define parts of the XML document that should match one or more predefined templates. When a match is found, XSLT transforms the matching part of the XML document into the desired format [4]. As an outcome of the transformation, result documents are created from the source file. The individual transformations developed for the project generate result documents in HTML format for the CDP "CDS" (Schema Company) and "iviews content" (intelligent views Company), whereas the data transformation for the CDP "TopicPilot" (DOCUFY Company) generates documents in the system-specific format "DYXML", which is based on XML. As can be observed, the transformations are similar in their basic structure, but they differ due to system-specific adaptations, which are the result of manufacturer-specific implementations.

The XML document, which was exported from the metadata book in the preparation stage, is transformed into a so-called "facet file" using XSLT. Due to specific structural requirements that need to be fulfilled by the facet file, an additional data transformation was developed. The generated facet file is an XML document that contains all possible metadata and enables filtering in the CDP according to different classification properties, provided that the content has been classified appropriately in the preparation stage. The CDS portal allows a so-called inner faceting (searching for a module within a document) and outer faceting (searching for a document). In this case the facet file is responsible for the outer faceting. The inner faceting refers to filtering within the content packages. According to different classification properties, the system filters by topics that contain the selected property. For the inner faceting to work in the CDP, the corresponding metadata must be written into each topic during the transformation, therefore no separate (facet) file is needed.

The transformed data must be packed as a zip file so it can be imported into the CDP. This zip file must meet the system-specific requirements of the respective CDP manufacturer (e.g. a certain folder structure and designation), otherwise the upload of the zip file will fail. For an automated and therefore faster and error-free generation of the zip file, batch scripts were developed for the data transformations. The commands of the batch scripts not only zip the transformed files in the structure required by the CDP, they also trigger the data transformations. In the CDS portal and the TopicPilot portal, which were used for this project, these zip files are called "content packages". The previously mentioned facet file is not part of the content package, since it is generally valid for all content packages. Accordingly, although this file is transformed by the batch script transformation command, it is not zipped with the other files. This process can be controlled by specifying the batch script commands.

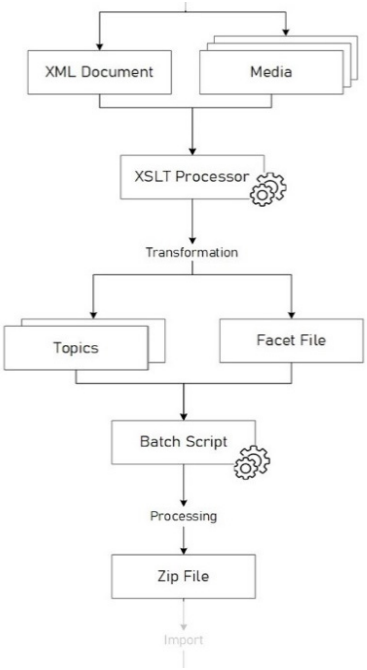

Fig. 3. Schematic illustration of the transformation stage.

2.3.3 Import

The last stage of the data transformation process describes the import of the data into the CDP. The previously mentioned content packages can be uploaded via the CDP import interface, whereas the facet file is imported separately via the facet interface. Once the upload is successful, the content packages and the possible filter options are displayed either in the content collection page or in the overview page (depending on the CDP). The inner and outer facets can now be applied to filter by topics or by content packages. The filter functions also offer the possibility to filter according to one or more properties to further narrow down the result. In addition, a navigation is provided, which is automatically generated by the structure of the content packages.

In summary, an interface was developed during the project between one CMS and the CDPs from several manufacturers. Therefore, the CMS content from Expert Communication Systems was transformed and imported successfully into the CDPs of Schema, intelligent views and DOCUFY.

3 Enrich data transformations

After describing how the first objective of the project was attained, the following sections explain the achievement of the second objective: transferring "intelligent content" from a CMS into CDPs.

3.1 Intelligent content

In general, "intelligent content is designed to be modular, structured, reusable, format free and semantically rich and, as a consequence, discoverable, reconfigurable, and adaptable" [7]. However, it is still difficult to delimit the concept of "intelligent content",

as an unambiguous definition is not easy to present. In order to get a better understanding of the term itself and the multiple concepts of intelligent content in the context of technical communication, it is advisable to take a closer look at the "Intelligence Cascade" [2, 8].

The Intelligence Cascade consists of three levels. The first level is “Native Intelligence” and refers to the semantic structuring and classification of modular and format-neutral content, e.g. to automate processes in the CMS [6]. For a logical and standardized classification of content with metadata, classification methods like the “PI-classification scheme” can be used to “systematically express the validity [...] with regard to product components and information classes” [8]. The classification of content objects makes them machine-readable and supports navigation and filter functions, for example in CDPs. “Augmented Intelligence”, the second level of the Intelligence Cascade, describes additional relations between different content objects, for example by ontologies as already mentioned in section 2.2 “Data Sources”. The third level is “Artificial Intelligence” and refers to the automated extraction of metadata and knowledge by statistical methods [8]. This level was not considered within the project because of the limited time frame.

Regarding the project, content was created modular, format-neutral and structured within the CMS editor. Each content object, as well as the so-called books were classified with structured and semantic metadata. The term “structured” means here that typified metadata have been assigned, which follow a certain classification system, in this case the PI-classification. “Semantic” metadata refers to metadata whose meaning is formally defined. This allows the definition of formal axioms and allows conclusions to be drawn. Accordingly, semantic metadata can be used to make implicit knowledge explicit [9]. However, to ensure consistency and efficiency in creating, customizing and reusing the metadata as well as the content, a standardized metadata concept must first be established, which is described in more detail in the following section.

3.2 Enrich data transformations through a CMS

3.2.1 Metadata concept

As mentioned above, the content from a CMS and correspondingly also the data transformation can be enriched by using metadata. Within the project, the PI-classification scheme was used as a model to assign the content with semantic metadata. Using the PI-classification, a further distinction can be made between “intrinsic” and “extrinsic” metadata [1]. Based on the PI-classification, the metadata types were also extended with functional metadata (e.g. for error messages or code associated with topics) and variant properties (e.g. for technical configuration parameters or installation locations) [1].

Based on these metadata types a taxonomic metadata concept was developed and assigned for all modular information. The metadata concept was also used as a reference for the subsequent data transformation, in which both the facet file with all possible metadata, and the individual content objects or topics with the corresponding metadata were generated (see section 2.3 “Data Transformation Process”).

3.2.2 MicroDocs

The previous paragraphs illustrate that metadata is the core of “intelligent content”. Consequently, it is also possible to enrich content with further approaches based on metadata. One approach highlighted in the project was the approach of “microDocs” [10, 11].

“A microDoc is [...] the (sub)set of topics or other information units that are connected by defined use cases and a logical context and are available as a dynamic connection via a search or delivery system” [10]. Therefore, using the approach of microDocs in information portals like a CDP, further information that could be also relevant for the user, is displayed.

The following figure shows that a single topic usually cannot provide the user with enough information because it is a self-contained content module. On the other hand, the entire manual contains too much information and cognitively overwhelms the user. The microDoc is located in-between and shows not only one topic, but also the additional relevant topics for the specific use case [10, 11]. For instance: an end customer is looking for operating information and additional functions as well as possible operating errors or dangers are shown. So, a microDoc automatically provides him with information that he will eventually need in his situation.

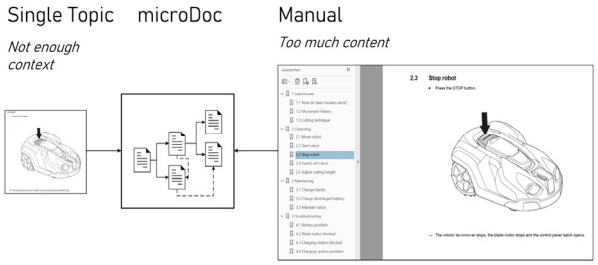

Fig. 4. The approach of microDocs compared to the individual topic and the entire manual. [figure based on 10]

Although the approach of microDocs is universal, there may be differences during implementation and the degree of dynamically generating a microDoc [10].

For example, microDocs can be generated for predefined use cases by a specific and static aggregation of topics by taking content from a CMS and zipping it into application-specific content packages for the CDP.

It is also possible to dynamically aggregate modular information units in CDPs based on semantic metadata, which is assigned in the CMS. The selection of the displayed content objects or links follows logical patterns for specific use cases, for example based on the PI-classification and different user roles.

Contextualization and linking through ontologies for the respective use case make it also possible to dynamically generate microDocs. Accordingly, the previously created ontology must be exported from the ontology editor so that the relations between the topics

can be used for the data transformation. While the first two implementation options refer to the level of Native Intelligence, this implementation type includes the level of Augmented Intelligence (see section 3.1 “Intelligent Content”).

Even though the term "microDocs" refers to a relatively new approach, some CDP manufacturers already have functions that support the approach of logically linking individual topics. In this respect, the focus during the project was on the CDS portal of Schema and it was discovered that different support functions for microDocs are already implemented at this moment.

One feature that supports the approach of microDocs is the "Related Topics" area. In this case, semantic links that intelligently extend the displayed topic are automatically generated by the metadata assignment in the individual topics and by predefined usage scenarios. The user is therefore shown topics that could be interesting for him as well.

4 Summary and outlook

This paper gave an insight regarding the need for data transformations to transfer content from a CMS to different CDPs, and the possibility to extend the content by semantics by using metadata and further approaches based on metadata.

After the basics of data transformations were mentioned, the transformation process of this project was described step by step. By successfully transforming the data from a CMS and importing it into three CDPs of different manufacturers, it was found that although the structure of the transformations is similar, system-specific rules must be met and adjustments need to be made, so that the content is transformed into the appropriate format of the respective CDP.

Content can become “intelligent” by enriching it with semantic metadata and further approaches based on metadata, for example “microDocs” that can support the user in the search for information by providing several topics that have the same logical context and fit to the user’s use case. This logical connection can be created dynamically by assigning semantic metadata to the respective topics, for example by an ontology. Within the project, a specific system, the CDS portal, was focused on in detail and it was discovered that presently there are already integrated support functions for microDocs generated by semantic metadata.

In the future, it is expected that “intelligent information” will become even more important for an efficient and use case specific “Content Retrieval”. As mentioned in this paper, there are already different approaches that are currently available, e.g. microDocs. Some CDPs already have first implementations for microDocs and it is expected that even more extensions will be implemented in the future. Moreover, the dynamic creation of microDocs could be supported by ontologies. A further step would be to get to the third level of the “Intelligence Cascade”: Artificial Intelligence. Machine Learning and Deep Learning could make it possible that metadata is automatically assigned. For future projects also a closer look at ontologies could be interesting for new insights in the enrichment of data.

The authors would like to thank the International Office, the Faculty of Information Management and Media as well as the master’s degree Program Communication and Media Management of Karlsruhe University of Applied Sciences for their support. Thank you to Dominik Kremer (DOCUFY Company), Tim Rausch (Schema Company) and to Lena Padeken (Karlsruhe University of Applied Sciences) for the great collaboration, insightful discussions and for their support. We would also like to thank Prof. Debopriyo Roy (University of Aizu) for the organizational support and Evan Stoddard for proofreading of this publication.

References

1. P. Drewer, W. Ziegler, *Technische Dokumentation: eine Einführung in die übersetzungsgerechte Texterstellung und in das Content-Management,* 2nd Edition, Vogel (2014)
2. W. Ziegler, *The Evolution of Content Management towards Intelligent Delivery Systems for Technical Communication*, Frontier, Official Journal of Japan Technical Communicators Association JTCA, 68-75 (2017)
3. W. Ziegler, *Semantic Information Development for Intelligent Content Delivery*, Frontier, Official Journal of Japan Technical Communicators Association JTCA, 18-29 (2019)
4. W3Schools, *XSL(T) Languages* (Retrieved February 17, 2020 from https://www.w3schools.com/xml/xsl_languages.asp 2020)
5. W. Ziegler, *Drivers of Digital Information Services: Intelligent Information Architectures in Technical Communication*, Proceedings of the ETLTC ACM Chapter International Conference, 48-52 (2019)
6. H. Fischer, L. Krägel, *Modeling of Complex Metadata in Technical Communication by Ontologies*, Proceedings of the ETLTC ACM Chapter International Conference, 11-16 (2019)
7. A. Rockley, C. Cooper, S. Abel, *Intelligent Content: A Primer*, XML Press (2015)
8. W. Ziegler, *Metadaten für intelligenten Content*, Intelligente Information: Schriften zur Technischen Kommunikation, **22**, 51-66 (2017)
9. W. Babik, H. Ohly, K. Weber, *Theorie, Semantik und Organisation von Wissen*, Fortschritte in der Wissensorganisation, **13**, 252-253 (2017)
10. W. Ziegler, *Delivery zwischen Kontext und Content*, technische kommunikation, **6**, 58-61 (2019)
11. W. Ziegler, *Extending intelligent content delivery in technical communication by semantics: micro documents and content services*, Proceedings of the ETLTC ACM Chapter International Conference (2020)