Newspaper digitalization - Hierarchical storage levels and long-term preservation

Kathrin Huber
Sächsische Landesbibliothek - Staats- und Universitätsbibiothek Dresden (SLUB)
Team Digital Library
Kathrin.Huber@slub-dresden.de
+49 351 4677-242

Gerald Hübsch
Sächsische Landesbibliothek - Staats- und Universitätsbibiothek Dresden (SLUB)
Head of Digital Infrastructure and Long-term Preservation Division
Gerald.Huebsch@slub-dresden.de
+49 351 4677-237

Abstract

Efficient digitization and long-term digital archiving of printed newspapers are building blocks that contribute towards the creation and maintenance of a collective memory for news media. Due to the large volume of digitized files per newspaper, there is a demand for their structured digitization, tool support, and integration with digital long-term archives. However, the state-of-the-art in digitization of newspapers with the open source tools like Kitodo.Production lacks support for structuring digitization processes. Such structures are necessary to make large scale projects with thousands of newspaper edition manageable and to enable a proper integration with long-term digital archives and presentation systems. Within this work, we analyze the existing problems and pitfalls of digitizing newspapers with the Kitodo.Production and propose a concept that supports flexibly structured digitization processes.

1 Motivation

Efforts and complexity of newspaper digitization are significantly higher than those for the digitization of single books. A single book, once scanned, can be indexed and exported for presentation and long-term archiving with manageable effort in one single process, even in case it

consists of a large number of pages. The process includes a small number of steps: creation of the process, importing the book's metadata from the library catalogue, importing scans images, metadata enrichment to include and link the content structure of the book to the scan images, automated OCR, and finally the export of the result towards presentation and the long-term archive. This workflow is well supported by existing tools like Kitodo.Production.

In contrast to books, newspapers and other periodicals are characterized by a large number of editions with a rather small number of pages. The indexing process for a single edition is more or less identical to that of a book. Because of the large amount of editions, the number of indexing processes necessary to fully digitize a newspaper is significantly higher. As an example, consider a newspaper that appeared for one decade with two editions per day, each of which has six pages. To fully digitalize such a newspaper, the digitization project consists of 7300 individual processes, one per edition, with a total of almost 44000 pages.

Existing indexing tools still lack support to efficiently manage such bulky work loads. Today, it is only possible to bulk-create the 7300 processes necessary for the digitalization of the whole newspaper, each carrying an own full copy of bibliographic metadata. Because these processes are not linked, they can't be centrally managed and edited.

Physical copies of newspaper editions are often bound in multiple volumes, e.g. one volume per quarter year. To make scan operations as efficient as possible, every volume should be scanned once as one unit. In such cases, the indexing tool must provide a function to import all editions that are contained in one volume at once. It must also provide means to make the imported images exclusively available to the processes that were created for the editions belonging to that quarter.

We claim that the possibility to flexibly define a hierarchy of structural elements ("structure trees") for digitization projects in indexing tools will resolve these issues. Such a hierarchy links and structures the large number of processes required for a newspaper digitization project. Linking and structuring enables bulk actions that target multiple newspaper editions at once. It avoids redundant metadata by moving and attaching common parts to the appropriate hierarchy level.

Structure trees enable users to reflect the logical and physical structure of newspaper to be digitized, thus matching the structure of the physical world and the structure of the digitization project as closely as it is deemed to be necessary. Fig.1 shows three examples of structure trees, all of them using calendar units like year, month, quarter, day, and daytime in different useful combinations.

The remainder of this paper is strustured as follows. In chapter 2, we introduce Kitodo.Production as the a state-of-the-art open source software tool for indexing and analyze its shortcomings regarding support for newspaper digitalization. In chapter 3, we present the concept of flexibly structured digitization processes and structure trees. Chapter 4 discusses how this concept can contribute to enhance the integration of indexing tools with digital long-term archives. Chapter 5 concludes this

paper.

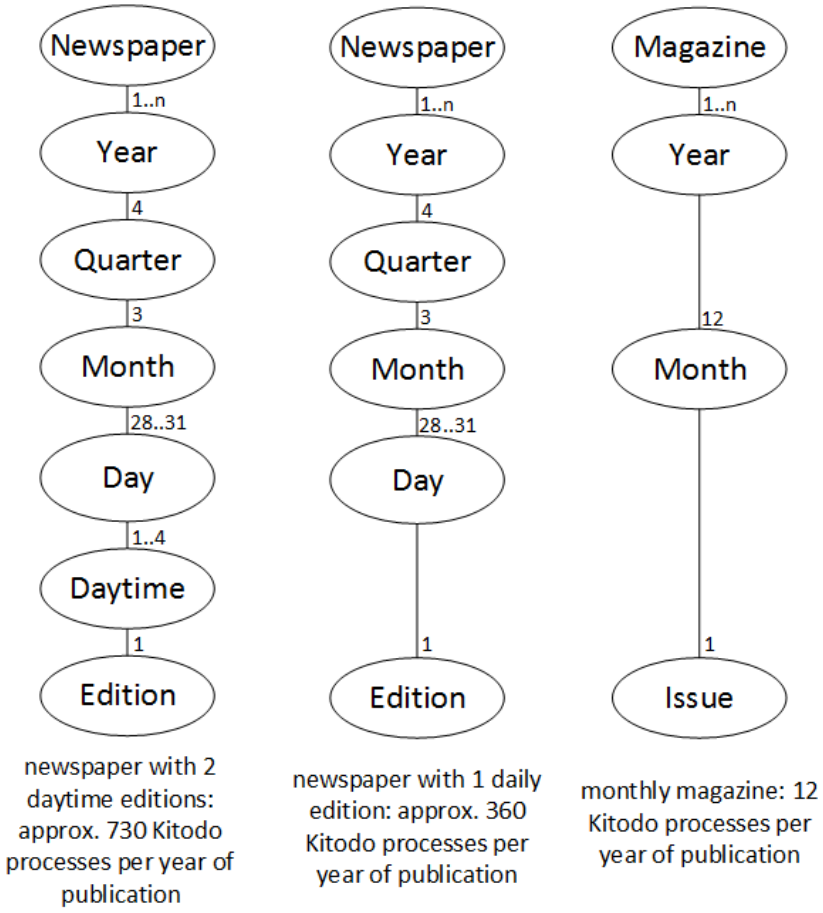

Fig. 1: Examples of structure trees for indexing newspapers and periodicals

2 State-of-the-Art

Kitodo.Production is an open source software to support and manage workflows in digitization projects for cultural heritage institutions. It accompanies the digitization process from creating a docket for the physical object, to exporting the results of the finished process to the long-term archive and presentation systems.

In Kitodo.Production, digitization projects are structured into separate processes. Every process represents, for example, a book, a manuscript or a newspaper edition. In our introductory example, 730 Kitodo processes are required to digitize all newspaper editions of one year.

Kitodo.Production already supports newspaper digitization with a specific calendar tool. It was introduced to simplify the creation of processes for periodicals. When importing data from the bibliographic database, Kitodo.Production recognizes the media type newspaper and opens a calendar, where the user can set the frequency of appearance and select all days on which the

newspaper appeared. In the next step, Kitodo.Production generates one individual process for every chosen date. For each process, the metadata is copied from the bibliographic database and some additional metadata added by the user.

The processes are automatically structured into “Newspaper - year - month - day - edition”. This structure is hardwired in Kitodo.Production and cannot be adjusted. Regarding our introductory example, the actual physical books, each containing the editions of a quarter of a year, are not represented. Consequently, managing scan operations per book is not possible with Kitodo.Production.

In the automatic process creation mechanism of Kitodo.Production, the hierarchical structures in the bibliographic database are lost. Because of that, already existing structures and edition-specific metadata are not available to the automatically generated processes. The only way to maintain this information is their manual generation. In manual generation however, only edition-specific metadata from the bibliographic database can be imported from the bibliographic database. Root-level metadata must be added by hand. It is obvious that this approaches is totally infeasible for large-scale projects with tens or hundreds of processes.

Another serious problem is related to the occurrence of errors. Before initiating the automatic process creation, the user typically amends additional information to the metadata from the bibliographic database. In case the user makes any mistakes, or forgets to add information, the only way to fix this error is to open every single process and correct the faulty information. That is because all processes store an individual copy of the metadata.

Kitodo.Production offers an export function to integrate with other systems, like presentation and long-term archives. Only single processes can be exported as bundles of Metadata Encoding & Transmission Standard (METS) encoded metadata files, image files, and OCR data. For newspapers, the export is therefore strictly limited to the edition-level. This limits presentation systems to the same level of granularity. For example, putting all editions of one year in one presentation unit for printing is impossible. The same limitation applies for export towards the long-term archive. It is, for example, impossible to create submission information packages with a scope identical to that of the physical volumes.

3 Flexibly Structured Digitization Processes

The concept of structure trees introduces hierarchies that structure digitization processes in order to address the shortcomings identified in chapter 2. Flexible hierarchies facilitate the direct, automatic and lossless import of structures from bibliographic databases, including the takeover of metadata available in the catalogue and its assignment to the appropriate hierarchy level.

Structures that exist bibliographic databases serve as a template for the initial structure tree of a

newspaper digitization project. The template is instantiated by the indexing tool during the automatic process creation. As a side benefit, the created structure trees profit from data quality assurance in the bibliographic database.

In the structure tree, metadata is generally stored at the hierarchy level to which it belongs. Instead of redundant storage, all nodes of the structure tree reference and inherit the metadata belonging to their ancestors. As a result, editing metadata becomes easy and robust, as any metadata edit needs only to be performed in one place. All lower levels in the hierarchy inherit it.

Because newspapers and digitization projects alike have individual structures, the automatically created structure tree must be customizable by adding hierarchy levels (cf. 2). As an example, the a "Quarter" level may be required in the digitization project to include the physical structure (four band per year) of the material to be processed.

The hierarchy levels in the structure tree can be exploited to trigger group actions. Members of the group to which the action applies those nodes that belong to the subtree rooted at the node on which the action is triggered. Essential group actions are docket generation, scan image import, search, and export towards presentation systems and long-term archives.

To continue with the example in fig. 2, one docket per quarterly volume must exist to initiate and track the scanning process of all volumes. The docket generation action for the whole newspaper project is triggered on the "Newspaper" level to generate one docket for each node that references physical material to be scanned, i.e. for all "Quarter" volumes. As soon as scanned images are available, they can be added to the project by triggering an image import action on their corresponding "Quarter" nodes. In case there should be quality problems with some images, dockets for correction requests can be easily generated by triggering the docket generation only for the concerned editions.

The search action is required to retrieve all processes that correspond to some search criterion, e.g. those that have a specific year of publication. As this information is contained in the structure tree, the search can be performed more efficiently and performant than a search that has to scan the metadata sets of all editions.

The hierarchical levels are also suitable for exporting processes to presentation. There are usually different use cases with different requirements towards the presentation of a newspaper. While an online presentation may only show one edition at a time, it is desirable to export all editions of one year in one PDF file for reproduction by means of printing. Because all necessary information is stored in the structure tree, the scope of the export action can be freely determined by selecting the appropriate hierarchy level, giving a good indication of the usefulness of the structure tree approach in this regard.

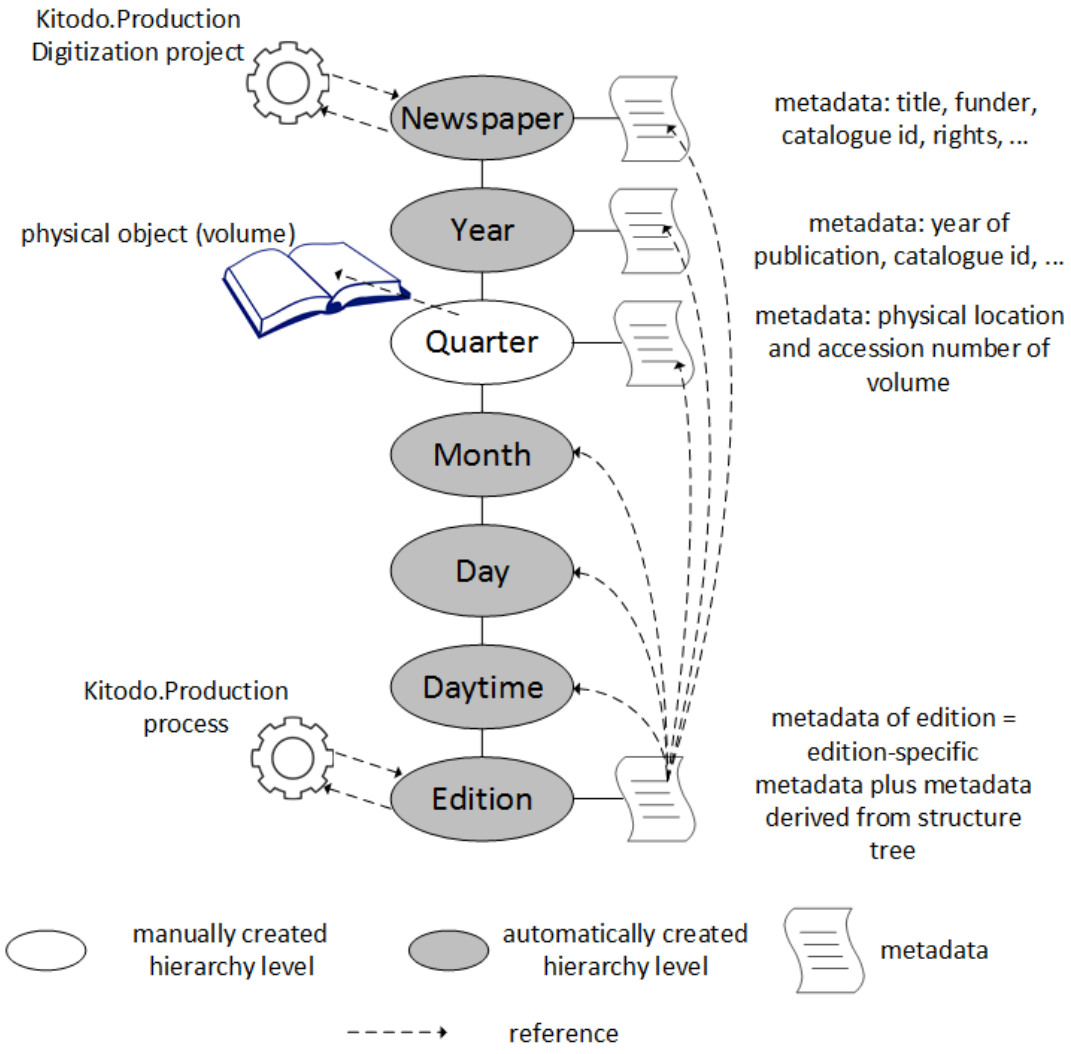

Fig. 2: References and metadata assignments in structure trees

4 Long-term Preservation

Technically, the digital long-term preservation of newspapers and periodicals is not much different from that of other printed matter. General concepts such as metadata validation, integrity checks, and format validation as part of the submission process, content- and bitstream-preservation, preservation planning and preservation action apply unchanged.

For a proper integration between indexing tool and the long-term archive, it is essential that purposeful Intellectual Entities and complete submission information packages can be formed when processes are exported from the indexing tool. Typically, Intellectual Entities are oriented at the boundaries of bibliographic units. As bibliographic units for newspapers and periodicals can be modeled in structure tree, the approach is feasible in this regard. All necessary information, i.e.

administrative metadata, descriptive metadata, structural metadata, and images are available from the structure tree to build complete submission information packages.

Things become more interesting when large digitization projects and mass digitization come into play. In such scenarios, it has shown to be an essential ability of the long-term archive to process multiple submissions (updates) of one and the same Intellectual Entity. Typical scenarios that necessitating updates are corrections of errors in metadata, additions of missing images, and replacements of low-quality scans. Modern long-term archives provide update functions to process such updates as full or differential updates. Full updates contain a revised full copy of the Intellectual Entity, independent of the extend of the actual change. Differential updates do only contain changed files. Because of their smaller size and better traceability, they are generally prefered over full updates.

Full updates can be created at any time without additional information from structure trees. Differential updates are more complex, as they require knowledge about the history of exports, metadata and images. To implement differential updates, the indexing tool must record the timestamp of the last export that was successfully ingested by the long-term archive. It must compare this export timestamp to the update timestamps of all nodes in the structure tree, i.e. all metadata items and all images. In case the update timestamp is younger than the export timestamp, the node is packaged into the differential update, otherwise it is ignored. Because the success status of the ingest of the previous export is relevant for this decision, a bidirectional data flow between the indexing tool and the long-term archive is required.

Without the non-redundant storage of metadata facilitated by structure trees, consistent metadata updates and timestamp-based change tracking would be practically impossible to implement. Both are preconditions for the correct export of differential updates.

5 Conclusion

The DFG project “Development and improvement of Kitodo.Production” encompasses the refinement of the Kitodo.Production code base and introduces new core functionalities, including flexible hierarchies for digitization projects.

In this paper, we have shown how flexible hierarchies contribute to the fulfilment of a multitude of requirements imposed on next-generation indexing tools. Their main areas of application are process creation, redundancy-free metadata management, scan process management and group actions.

Modeling these hierarchies based on existing structures from bibliographic database has shown to be an appropriate approach with regard to the automation of process creation and tool support for the management of large-scale digitization projects.

The complete set of actions that an indexing tool must support on such hierarchies has been identified and described based on practical requirements. The export of results to presentation systems and digital long-term archive is crucial for the integration of the indexing tool into the ecosystem of a modern digital library. For both export targets, structure trees show clear benefits when compared to existing solutions. The

In summary, it can be stated that the proposed concept is a sound basis for the design and implementation of indexing tools that support the digitization of newspapers and periodicals.