DAISY PRODUCER: AN INTEGRATED PRODUCTION MANAGEMENT SYSTEM FOR ACCESSIBLE MEDIA

Christian Egli

Swiss Library for the Blind and Visually Impaired Zurich

Grubenstrasse 12

CH-8045 Zurich

SWITZERLAND

ABSTRACT

Large scale production of accessible media above and beyond DAISY Talking Books requires management of the workflow from the initial scan to the output of the media production. DAISY Producer was created to help manage this process. It tracks the transformation of hard copy or electronic content to DTBook XML at any stage of the workflow and interfaces to existing order processing systems. Making use of DAISY Pipeline and Liblouis, DAISY Producer fully automates the generation of on-demand, user-specific DAISY Talking Books, Large Print and Braille. This paper introduces DAISY Producer and shows how creators of accessible media can benefit from this open source tool.

1 Introduction

The typical production of an accessible media involves a number of processes such as acquisition, markup and output generation. In any medium to large organization a number of people will be involved in this process, maybe in different locations and with different roles. They need to collaborate and share intermediate artifacts of the process. With all these factors taken into consideration, the management of this workflow becomes increasingly complex when scaling to a large production.

Parts of this process, such as the output generation, have very good tool support in the form of the DAISY Pipeline (DAISY Consortium 2009). Others, such as integrated workflow management and collaboration, currently are lacking.

DAISY Producer sets out to fill this gap. It helps to manage the production process by managing the workflow, providing a shared workspace for process artifacts and a versioned storage for documents and finally by integrating with back-end systems such as the DAISY Pipeline and Liblouis (Liblouis 2009) for the generation of output media, such as DAISY Talking Books, Large Print or Braille.

In this paper Section 3 provides some background to the problem that large scale producers of accessible media face. Section 4 outlines the idea for a solution, while Section 5 presents the implementation of *DAISY Producer*, explains what use cases it covers and how it helps in the production process. Finally, Section 7 shows the roadmap for *DAISY Producer* and how you can get involved.

2 The problem

The Swiss Library for the Blind and Visually Impaired produces accessible media for the blind in multiple formats, namely DAISY Talking Books, Braille and Large Print. As our production usually starts with a printed hard copy, the production workflow involves acquisition in the form of scanning and OCR, followed by markup, archiving and versioning and finally the generation of the output media, as shown in Figure 1. A separate workflow management tool keeps track of the progress of each production.
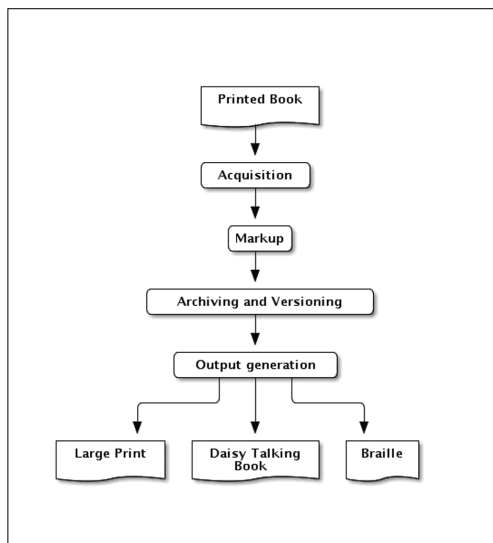
Figure 1: Production process for accessible media

Let's see how these steps would look in an *ideal* world:

Acquisition: The first step in the production involves scanning the hard copy and doing optical character recognition (OCR) on it to extract the text.

Markup: Once the text has been extracted from a hard copy it is enriched with context information, such as author and publisher, meta data of the book and with structural information such as sections, headlines, tables, etc. to make it accessible. Ideally the result of this process is a DTBook XML [Fehler! Verweisquelle konnte nicht gefunden werden.].

Archiving and Versioning: The result of the markup process forms the basis for the output generation and obviously goes through a quality control process. Thus the DTBook XML needs to be versioned to handle proof reading feedback and the finalized XML is archived for future reference.

Output Generation: Once the document has been approved by quality control it is ready to be sent to the output generation process, such as the DAISY Pipeline, which handles the generation of DAISY Talking Books and Large Print or Liblouis, which takes care of generating Braille.

Workflow Management: An integrated system keeps track of the state of each book in production and accordingly assigns the next pending task to the right person.

Unfortunately, the current work process at Swiss Library for the Blind and Visually Impaired doesn't look exactly like this. There are two main problems.

No unified production: For historical reasons, the production process is not centered around a common (XML) format that shares markup, archiving and versioning, but rather is separated by output format. For Braille, for example, the markup process is specific to that format and cannot be used for other formats. Similarly, the archiving is done differently for all the output formats. While this process works and has so for quite a few years, it does not scale to large production numbers, causes duplicated effort and misses out on a lot of potential for synergies between the workflows.

No integrated workflow management: The other main problem is that there is no integrated and automated workflow management included in the process, so it is difficult to know at what point in the workflow a document currently stands. There is an external order processing system which keeps track of all the books in production, yet since it is not integrated in the workflow, the data can easily get out of sync.

Thus, if we want to scale our production to large numbers of books, we need to unify and automate this process as much as possible.

3 The idea

In order to unify and automate the production process we need to address three issues:

Unify production: The production of accessible media needs to be single source publishing; i. e. centered around one common document format that contains all the information for the automatic generation of all the output formats. Not surprisingly, we chose DTBook XML as the central document format. It is a widely deployed standard and has a lot of tool support in the form of the DAISY Pipeline. While it does not yet have support for all possible markup (that would be needed, for example, for Braille markup), it is clear that support for this will be added in upcoming revisions (Zednext 2009) of the standard.

Automatic output generation: The actual generation of the output is to be fully automated based on the common document format and is to happen on-demand. By choosing DTBook XML we can profit from the DAISY Pipeline and Liblouis where a lot of the automated output generation has already been implemented. Where needed, we enhance the Pipeline and contribute the code back, such as in the case of Large Print.

Web-based management tool: An integrated management tool keeps track of the state of all books in production, helps to assign jobs like markup or quality assurance to the respective experts and guides users through the workflow. We implement a web-based tool that handles document workflow, versioning and archiving and ties in the DAISY Pipeline and Liblouis as back ends to generate all the output formats.

4 DAISY Producer

There are a number of projects that in part address the problems mentioned in Section 3. They are discussed in Section 6. Unfortunately, none of them cover all our requirements as outlined above.

For that reason the Swiss Library for the Blind and Visually Impaired is implementing an open source software solution named *DAISY Producer* to remedy the problems mentioned above. We base our work on existing work, such as the DAISY Pipeline and Liblouis and profit from the tools that the DAISY Consortium and others are providing.

4.1 Limited scope

To unify the different production lines that we currently have, we decided to center the workflow around a common XML document format, namely DTBook XML. We are fully aware that the current version of the standard is not adequate to capture the entire markup needed for Braille for example. But rather than wait for the standard to evolve or to add our own proprietary extensions we took a different approach. As a large portion of our production deals with simple, fiction books, we decided to focus the project initially on fiction books not containing any math, tables or images, knowing that future versions of the DTBook XML standard will be able to handle more if not all of the markup needed to produce books of any complexity.

4.2 Unified production

In order to achieve single source publishing for all our accessible media, we center our production around DTBook XML as outlined above. This allows us to benefit from synergies by sharing the acquisition and markup process and using common procedures for versioning and archiving.

4.3 Automatic output generation

If we limit ourselves to fiction books and use DTBook XML as the common document format, we can use the DAISY Pipeline and Liblouis almost as is. Thus the work we need to do in order to get to a fully automated output generation is manageable. *DAISY Producer* currently supports user-specific output to Large Print using the DTBook to LATEX Pipeline script and to Braille using Liblouis. The output is user-specific in the sense that the user is, for example, able to choose the font, font size and page style in the case of Large Print. The output is then generated according to the users' specifications on-the-fly. A similar process occurs for Braille. Other output formats such as XHTML, DAISY Talking Books or EPUB (IDPF 2009) will be added later (see Section 7.1).

4.4 Web-based management tool

All the user interaction in *DAISY Producer* is handled through a web interface. In the future some functionality of the web interface might be replaced with web services as we integrate the production with the library system and the order processing system.

DAISY Producer is implemented in Python (Python Software Foundation 2009) using the Django (Django Software Foundation 2009) web framework. The state information for all the productions is persisted to a relational database. It should run on all the major platforms such as Windows, Linux and Mac and support all major databases (MySQL, PostgreSQL, Oracle).

4.5 Use cases

DAISY Producer essentially satisfies three use cases:

Management: Initiate a new production and oversee its progress.

Production: Carry out the steps needed to make the book accessible, i. e. acquisition, markup and quality control. This process is supported by enabling collaboration through a shared space for temporary artifacts such as output from OCR and a versioned repository for the DTBook XML.

Consumption: Consume a finished document, i.e. make it available for download by generating it on-the-fly based on user-supplied specifications.

These use cases are reflected in the user interface of DAISY Producer with the three tabs "Management", "TODO" and "Browse" (see Figure 2).

4.5.1 MANAGEMENT

The "Manage" tab gives a management overview of the progress of all documents in production i. e. an overview of all documents that have been or are currently produced, their state and who is assigned to work on it (see Figure 2). It also allows the user to initiate a new production by entering all the meta data.

Figure 2: Management overview of all documents in production

4.5.2 PRODUCTION

At the time of writing, the acquisition and markup process in DAISY Producer offers basic support for a client-side process where intermediate files can be attached to a document that is in production and DT-Book XML files can be uploaded and versioned. The user can, for example, do scanning and OCR with tools on the client work station and use *DAISY Producer* as a shared work space for collaboration by uploading the resulting RTF. The next user sees in her TODO list that she has been assigned to markup a document (see Figure 3). She can then download the result of the OCR and use it to markup the document using either a full-blown XML editor or simply Microsoft Word or OpenOffice.org for example. The latter two tools have support for Save-As-DAISY-XML (DAISY Consortium 2009a; Spiewak / Archambault, 2008) which can be used to generate and upload DTBook XML. The XML is then validated and versioned by DAISY Producer (see Figure 4). In future we would like to move to a more server-side approach as discussed in Section 7.1.

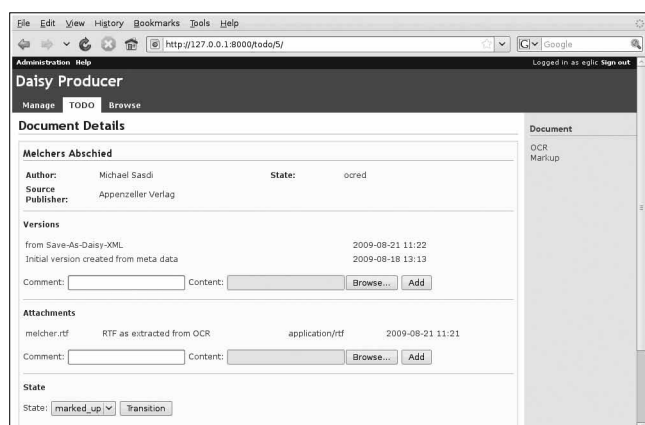Figure 3: Overview of all pending documents

Figure 4: Detailed view of a pending document

4.5.3 CONSUMPTION

DAISY Producer wraps the back end tools for the output generation in a web-based front end that allows the user to browse all available books (see Figure 5) and download a specific book in the form that the user chooses. See, for example, Figure 6 where a document can be downloaded as user-specific Large Print (PDF) or Braille. It would be fairly straightforward to also wrap other parts of the pipeline to be able to consume a DAISY Talking Book or an XHTML rendering of the document in the future.

Related work

There are a number of interesting projects which address similar problem domains. For example, there is the DAISY PipeOnline project (PipeOnline 2009) which aims to develop a distributed application to create and run DAISY Pipeline jobs. While it also offers a web-based interface to the DAISY Pipeline, it does not address the production workflow and management issues. In that sense it is probably more of a complement than a competition to *DAISY Producer*.

Another very interesting project is DAISY4Dspace (Ribera / Golkhosrav 2008) which integrates the DAISY Pipeline in the DSpace (Duraspace 2009) digital repository system. However, this approach has a few drawbacks in our view: the workflow of DSpace does not fit that well into our production and the integration as it is done currently precludes the ability to generate on-the-fly user-specific versions of a document (as they are pre-generated at the time of submission).

Figure 5: Overview of all finished documents

BrailleNet (BrailleNet 2009) has been doing very advanced work for years. They have the Helene Server (Braillenet 2009a) which not only handles the automatic production of Braille and other accessible formats, but also manages the secure distribution of it. Unfortunately, probably since it predates the DAISY Pipeline, it is not based on it, so it doesn't share any synergies with the efforts from DAISY Consortium and has no clear upgrade path to upcoming standards of the DTBook XML.
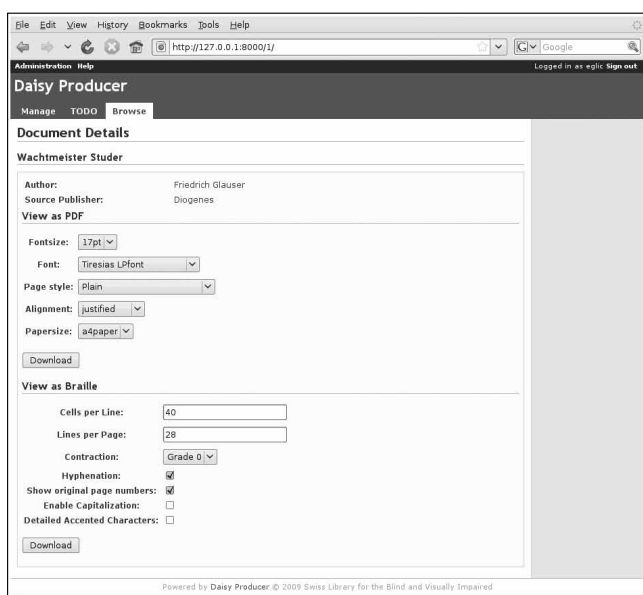

Figure 6: Detailed view of a finished document

5 Conclusions and further work

DAISY Producer helps to manage your production of accessible media and allows it to scale to higher volumes. It currently offers basic support for the entire process from acquisition all the way to output generation. It is, however, still a work-in-progress and is planned to go into production at Swiss Library for the Blind and Visually Impaired at the beginning of next year.

DAISY Producer is freely available from http://www.daisyproducer.org and released under an Open Source license. Any form of feedback, be it bug reports, customizations or even enhancements is highly welcome.

5.1 Further work

As mentioned before, *DAISY Producer* is not yet finished. While it has basic functionality in place, there are a number of areas that are still being worked on namely acquisition, archiving, support for more output formats and for complex books.

For acquisition *DAISY Producer* currently supports a client-side approach where all the work for acquisition and markup basically happens on the client machine with the server providing a workspace for document sharing and versioning. Future versions will have support for a more server-side scenario where OCR and maybe even markup happen on the server. This will offer several benefits, for example, that no client-side software has to be installed. Markup could be done with in-browser XML editing or with Wiki editing as other projects (Bernier 2009) are suggesting. The wiki approach offers a simple and compelling in-browser solution. It is questionable, however, if the wiki syntax can really handle all of DTBook XML.

In terms of archiving, we are considering moving to a JSR-170 (Java Community Process 2009) compliant back end instead of the simple file system based back end that we currently use.

The currently supported output formats include Large Print and Braille. It is not very hard to also add the other formats such as XHTML, TTS DAISY Talking Books or EPUB that the DAISY Pipeline already supports.

In a later stage we will want to move beyond the limited subset of DTBook XML that we are currently supporting (see Section 5.1). However, this will mostly involve work in the supporting libraries such as the DAISY Pipeline and Liblouis. We also want to integrate with our order processing and library system. This will be done in such a way that other institutions can adapt it to their back end systems.

References

BrailleNet (2009). BrailleNet Accueil. Retrieved from: http://www.Braillenet.org/

Braillenet (2009a). Serveur Hélène. Retrieved from:http://www.serveur-helene.org

DAISY Consortium (2009). Project DAISY Pipeline. Retrieved from: http://www.daisy.org/projects/pipeline/

Django Software Foundation (2009). Django – The Web framework for perfectionists with deadlines. Retrieved from: http://www.djangoproject.com/

Duraspace (2009). DSpace software. Retrieved from: http://www.dspace.org/

IDPF (2009). Open Publication Structure (OPS) 2.0. Retrieved from: http://www.idpf.org/2007/ops/index.htm

Java Community Process (2009). Content Repository for Java technology API. Retrieved from: http://jcp.org/en/jsr/detail?id=170

Liblouis (2009). A Braille translation and back-translation library. Retrieved from: http://code.google.com/p/liblouis/

PipeOnline (2009). A web application for creating and executing Pipeline jobs over the wire. http://daisy-trac.cvsdude.com/pipeline/wiki/PipeOnline.

Python Software Foundation (2009). Python Programming Language. Retrieved from: http://www.python.org/

DAISY Consortium (2009a). "Save as DAISY" Add-in for Microsoft Office Word. Retrieved from: http://www.daisy.org/projects/save-as-daisy-microsoft/.

DAISY Consortium (2009b). DAISY/NISO Standard. ANSI/NISO Z39.86 Specifications for the Digital Talking Book. Retrieved from: http://www.daisy.org/z3986/.

Zednext (2009). ZedNext Home Page. Retrieved from: http://www.digitaltalkingbook.com/zw/Main_Page.

Alex Bernier (2009). The potential of Wikis in the DAISY context. In: Proceedings of the DAISY International Technical Conference 2009. Retrieved from: http://www.daisy2009.de/files/presentations/itc/bernier__the_potentials_of_wikis_in_the_daisy_context/daisy2009.html#%281%29

Burger, Dominique/du Bourguet, Guillaume/Guillon, Benoît (2003). Hélène: A collaborative server to create and securely deliver documents for the blind. Technical report, Association BrailleNet, Université Pierre et Marie Curie, INSERM U483, 2003, Retrieved from: http://www.snv.jussieu.fr/inova/publi/2003/helene-server.pdf

Ribera, Mireia / Golkhosrav, Mehrad (2008). An automated workflow to publish accessible scientific papers: integrating DAISY Pipeline within DSpace. Retrieved from: http://www.euain.org/files/

Spiewak, Vincent / Archambault, Dominique (2008). odt2dtbook-OpenDocument To DAISY DTBook. Retrieved from: http://odt2dtbook.sourceforge.net/