Adv. Sci. Res., 6, 271–273, 2011
www.adv-sci-res.net/6/271/2011/
doi:10.5194/asr-6-271-2011

Advances in Science & Research
Open Access Proceedings

10th EMS Annual Meeting and 8th European Conference on Applied Climatology (ECAC) 2010

Issues faced in digitally re-purposing printed archival material

T. M. Dart

The Seeing Ear (Registered Charity 1111371), St. Leonards-on-Sea, UK

Received: 31 December 2010 – Revised: 14 June 2011 – Accepted: 9 September 2011 – Published: 18 October 2011

Abstract. The Seeing Ear On-Line Library (http://www.seeingear.org) is the first general on-line library in the UK and has been providing books in an electronic format to EU users for more than 5 yr. While this provision has been within the context of print-disability, the challenges that the Library has faced and is facing are common to everyone re-purposing printed archival material.

Key issues faced by anyone contemplating such a project are the choices of hardware and software, error rates and reduction, feature analysis and extraction, and archival (meta)format. This paper presents an approach to these questions to facilitate the on-line presentation of popular climatological material. This work was supported by the Big Lottery Fund UK.

1 Introduction

The Seeing Ear is a registered charity based in the south of England. In 2002 the UK Government passed the Copyright Reform Act which allowed people who were unable to read normally printed books to obtain transcribed copies, and The Seeing Ear was set up in 2005 to transcribe and provide those copies. It is a free service, and membership is quite low – perhaps about 400 individual users and about 350 schools, colleges and universities with users predominantly in the UK but also in every member state of the European Union.

The aim is to create a multi-access library. The reason for this is that it quickly became obvious that different people have vastly different and incompatible needs – a person who has weak sight might want larger print, while someone who is deaf and blind will need tactile output such as Braille. Many people might want synthetic speech, while others who are paralysed or have no hands to turn pages may want the whole book with all pictures and diagrams but in such a format that they can navigate around the book by some other means such as blowing into a mouth-operated switch.

The access problem is essentially the same one faced by fully-sighted people trying to access digital information on a mobile device – a smart phone, Blackberry or PDA has a very small amount of screen and it can be very difficult to access, say, large and complex web pages this way. Learned journals are typically printed in two column format which involves scrolling back from the bottom of one column to the top of the next when trying to access the paper on a screen that is too small. Such pages need an entirely different type of presentation – they need re-purposing. This paper therefore concentrates on those aspects of the re-purposing process which are common to the production of specialised texts for the disabled and those for access by normally-sighted persons – it will be seen that these processes are identical until the final stage of document preparation.

It is unfortunately not possible (due to copyright law) to allow access to the On-Line Library materials to non-registered members (registration being restricted to EU print-disabled residents). However, those wishing to see exactly how the Library works (in terms of being able to search for and retrieve items of interest) together with a list of all items held to date are free to try the Guest Library at http://www.seeingear.org.

2 Equipment

The method evolved for bulk scanning older books was to chop the spine off in a guillotine and feed the pages through a document scanner. This is quite a fast method – the scanner is duplex and although it does not achieve anything like the claimed 178 pages a minute it does do about a page a second – both sides – at 600 dpi greyscale. The scanned pages were then read by optical character recognition (OCR) software on a dedicated server.

Correspondence to: T. M. Dart
(tony@seeingear.org)

Published by Copernicus Publications.

This set-up works really well for some books but has a few issues – it destroys the book, it does not do colour well (slow, and the colours are not very faithful), and the quality of reproduction of fine and italic text is just not good enough. This system was therefore upgraded to a dual camera based system.

This scanning system is manually operated and is quite a bit slower than the document scanner – approximately half an hour to scan a book, as opposed to 6 or 7 min. However, not only has the actual image quality increased but also the issues of missing pages (two pages feeding through together) or folded-over pages have been eradicated. Glue and dust from older books also caused scan-lines to appear, causing a large amount of down-time for cleaning and re-scanning. The slower speed of production plus the ability to check each page on-screen has dramatically improved error-reduction times in post-processing.

After scanning, OCR and checking the book is uploaded to the Internet Library together with its' classification data into a special holding area. From here it is automatically picked up and classified by a process that runs every 10 min, and then a user can actually retrieve it.

3 How the Library was built initially

The work of creating the Library began with data capture and analysis – a selection of books was examined and relevant information about publisher, author(s), title, ISBN etc. noted for the purposes of classification, search and retrieval. The data was then normalised or segregated into logical units such as the Title, the Author(s)' surname, the Author(s) forename(s), Publisher etc. Data was segregated to the third degree of normalisation as that appears to be a realistic compromise between data redundancy and database complexity. For example, an address of a publisher would be regarded as one logical unit of data and stored in one table, rather than having individual tables for the towns common to many publishers and using an extra table to tie the Publisher to the various address elements (city, country etc.). The MySQL database tables were then created directly from the logical groupings of the analysed data. No changes have been made to the basic layout in over 5 yr of service.

Once the database had been created and seeded with sample data, the Content Management System for search and retrieval was written in a combination of HTML and PHP, with priority being given to accessibility by assistive technology, and a data production system to automatically put prepared texts in the requisite format, remove any page feed characters etc. and put in the copyright notice was created in Perl.

4 Book production

The key concept for creating a multi-access library is to consider a book as a series of conceptual layers which must be disassociated from each other, then recombined in different ways to suit the final output. There are 4 conceptual layers; the data layer that holds the information, which in a printed book equates to the words and spaces; the presentation layer which determines how those words and spaces are physically arranged on the page – one or more columns, size and style of typeface etc.; the structural layer which is how the book is actually organised – sections, chapters, pages, paragraphs etc., and finally the physical layer which is comprised of the actual book itself – paper, glue and ink.

All these layers are inextricably intertwined in a physical book, which is why we cannot re-purpose it. Firstly the physical layer must be removed by turning the book into a scanned electronic file. The scanned file contains the data, presentation and structural layers bound together – the data can only be accessed by viewing an unmodifiable image. An OCR process will extract the data. However, OCR programs do not generally extract structural information, although they do attempt to extract presentational information. It is essential that the purpose of an element is known rather than what it looks like on the page. For example, a Microsoft Word file that faithfully reproduces the type style and size of the original document's heading, page-number and body-text is not very useful. What is required is a file that explicitly defines what is a heading or body-text in order that a relevant style can be applied to it.

An e-book that is designed to be accessed on a piece of electronic equipment only has three layers. In our model we want to provide a file with just two layers – data and structure – and allow the user to select the items that are required for their particular presentation – for example, a speech file would not have graphical items in it. A small screen might need single columns only – a student text may need to closely follow the physical book layout in order that a student could follow classwork and so on.

In order to achieve this it is necessary to identify each area of interest on each page (manually or automatically) and label it as such to create a rich superset of all features in an intermediate file, then use this file to create the final output format.

5 Problems

In general, simple plain-text books with a single column layout can be transcribed without difficulty and the error-rate can be as low as 1 character in 15 000 (99.993 % accurate). However, such rates are not typical of material that has the features shown in scientific publications such as multiple columns, tables, diagrams, pictures, mathematical equations and unusual symbols.

Multiple columns – while OCR (optical character recognition) of printed material is regarded as a solved problem, the analysis and segmentation of the page into areas of interest prior to such recognition is still very much a field for active research and current commercial OCR software is not able to accurately predict other than very simple layouts. The solution here is to allow the software to recognise the page and then manually adjust the on-screen graphical representation of the column boundaries.

Tables – these can usually be transcribed in their entirety to a format such as Microsoft Word or PDF where in principle the concept of a table has meaning, and commercial software does a good job of doing this automatically. Where spatial relationships do not exist (text and audio are linear media) then the table has to be re-arranged manually. This involves the transcriber taking a decision on the reading order of the table elements and thus destroying table flexibility, which is seen as a non-trivial problem to which there is no current solution the author is aware of.

Graphical elements such as diagrams and pictures are treated the same way currently – as pictures that have only visual meaning and are transcribed as such in a graphical file format (typically embedded JPEG where appropriate). It is possible to manually extract information held in diagrams and graphs, or to vectorise such images using software, but not within the OCR process. There is currently on-going research work to extract data from certain diagrammatic representations such as maps and line drawings.

Mathematical equations are usually misread by commercial OCR software. In fact, embedded numbers in text are often read with far less certainty than alphabetic characters. This is probably due to the fact that most commercial packages are tuned to reading text rather than numbers, and if the package contains a "self-learning" element then such training will be proportionally reinforced as books and scientific papers have a far higher frequency of alphabetic characters than numeric, symbolic or punctuation. Where mathematical equations that span more than one line vertically occur the OCR package is completely lost, and manual resetting is the only recourse.

It should also be noted that an OCR package makes a "best guess" at what the symbol it is trying to interpret actually is, and will frequently return the correct symbol but with a lower degree of certainty for those symbols that occur with lesser-frequency in normal texts – such as the printed symbols for plus, minus, degrees, and so on. It is believed that a solution to this problem would require that OCR engines with mathematical capabilities (recognition of a wider range of symbols, including e.g. Greek characters, different character spacing algorithms, multiple line capabilities etc.) would need to be incorporated within the base OCR package, together with enhanced page segmentation.

6 Conclusions

Social inclusion and mobility both make demands on the accessibility of information which can be solved by a multi-access approach. Multi-access relies on both the data and the structural layer being identified and used together with an access-specific presentational layer. There is currently no universally accepted standard to describe the structural layer (although there are a number of standards, all fairly close, based on XML such as DocBook, epub and DAISY.)

As well as a requirement for a universally accepted standard to describe the structural layer there is also a need for production-level OCR software to recognise more specific areas of interest than just text, pictures, bar-codes and tables. In particular, mathematical equations need to be identified and read from within the OCR package.

Producing multi-access material within a production environment involves a complex pipe-line of processes demanding close integration between the stages (for example, there is a strong need for manual correction but this must be done on the labelled meta-file) and development of these tools is seen as a priority for further work.

Edited by: T. Cegnar
Reviewed by: three anonymous referees

sc | nat The publication of this article is sponsored by the Swiss Academy of Sciences.