Peeters Online Journals: Unicode Implementation and the Meaning for Online Publishing

Ludovic Janssens, Luc Peeters

Peeters Publishers & Printers
Bondgenotenlaan 153 - B-3000 LEUVEN
poj@peeters-leuven.be

Abstract

Peeters Online Journals offer an online access to the complete contents and texts of scientific journals. A simple, but well structured interface gives access to the journals covered by the system. A MySQL-database and a HTML-based administration module are the sources of this site. Visitors can access the data by ip-recognition or a 24h personal login. The articles are presented in portable data file (PDF) format, a file type which focuses on the presentation of the text. In order to make the full text searchable, Peeters did an optical character recognition of the files of 'tijdschrift voor geneeskunde'. This rose new issues. Implementing a Unicode-based SQL database with XML-data could solve these problems and simplify the workflow severely.

1 Introduction

Already one year Peeters Online Journals offer a online access to the complete contents and texts of scientific journals. These journals appear in an simple, but well structured interface. The uniqueness of Peeters Online Journals is not only in its subjects (it is one of the few databases that cover arts and humanities journals) or its coverage (Peeters Online Journals provides access to French, Dutch and German journal that where never electronically accessible), but also in its structure.

The Peeters Online Journals nowadays covers more than fifty journals that have to be searchable starting from sources varying from bibliographic references to simple keywords. Making this large amount of titles and volumes accessible in a simple way was a challenge that we only could conquer thanks to the creativity and the interaction between the publisher's staff and the Leuven University that offered a testing environment for the website.

The result is a online MySQL-database with a html-based administration module. For visitors the journals are searchable on subject, indexes, title, abstract, author, language, type, year, and language. The results show the full reference to the article and the article abstract. The entry of the data is done manually in the administration module.

Access to the Peeters Online Journals is regulated in an equally simple and clear way. Everybody can search the journals, but only subscribers can access the full text, except for those who make use of the pay-per-view option. As the highest security possible is our goal we provide access through IP-recognition for all institutional subscribers, individuals can obtain an access code linked to their email address that they can use for a 24hr login.

The full text appears in the Adobe PDF-format, which enables the user to view the article in its original layout, but cancels any full text search options. For a single journal, tijdschrift voor Geneeskunde, a project was set up in order to make full text search possible. All files where doubled and next to the PDF-files, html-files where uploaded to a hidden section on the website's server. Full text search is made possible, but shows some inconveniences.

The problem with the tvg-model is the multiplication of data and the greater danger of errors in the database. An even bigger problem for arts and humanities researchers is that there is no way to search on words or signs in foreign or ancient writings and that one always has to rely on the correctness and uniformity of transcriptions and format conversions in order to find specific terms. Certainly since one of the main subject of our publishing house is Oriental Studies.

This problem could be solved with an thoughtful implementation of a Unicode-based SQL database with XML-data. Unicode comes in as the provider of the right character sets and of the language metadata. Ancient and foreign writing can appear en be used in everyone's internet browser without the need of additional software. The only requirement is the use of one single and clear standard.

Implementing Unicode can create even more interesting features. Publishers and printers already use databases. These could be worked out as interactive platforms between authors, printer and publisher. As everyone will work with the same standards in an integrated system, the conversion to different export formats (pdf, html, ...) will not be difficult anymore. In contrast to today's system no useless conversions or extra manual data entry will be necessary. The workflow will be simply linear and more attention can be paid to the correctness of the contents and the functionality of the online system. One would be able to search the XML-data of the database, but get a pdf-file as the full text is requested.

2 The Needs and Answers in Current Online Publishing

In order to give clients what they require, a firm must look at the demands of these clients. In Online Publishing, there is no difference regarding this issue. One has to ask himself: what do scholars want in a academic journals website? As Fletcher (1999, p. 107) puts it: "Online services need to keep this in focus; solve problems by providing answers, not just by providing access to the information."

2.1 Basic Needs in an Online System

First of all an online system must answer to basic needs. In order to do so one should consider the musthaves and don'ts.

2.1.1 Demands

First of all, a scholar does not care by which publisher a journal was made. He looks for the information he needs and wants to find it in an easy, reliable and fast manner. Electronic Journal Services should not be seen as a system to find a particular article, but rather as a cross-referencing powerful research tool, exploiting all advances of digital (not *digitised*) data. The easiness of a system is a major issue: take notice of novice computer users and do not complicate systems unnecessarily. (Fletcher 1999, 112 - 113)

Penfold (1998, 11) points out the demand of thematic online journals services: a researcher, he states, wants to find (almost) all journals regarding his subject.

The conclusion of these studies is that researchers do not mind how online journals are presented. What they want is an easy access to their research field. They want accurate, but simple search modules and websites browsable by subject.

2.1.2 Peeters Online Journal's answers

Peeters Publishers was aware of these issues and tried to provide an answer to them. Easy access is more than just a commercial phrase. Contrary to most major online journal services, Peeters Online Journals focuses on humanities, especially Oriental Studies. Particular to these science is not only the need of vast archives, but also the demand of correct representation and possible retrieval of terms written in foreign or ancient characters.

The classical solution to this demand is the use of Portable Data Files. Articles can be offered in the same way as they where published in their journal. But what about the full text search? The double layering of these files enables to retrieve text data from the file. Starting from this point Peeters Publishers & Printers created a website for 'Tijdschrift voor Geneeskunde'.

For 'Tijdschrift voor Geneeskunde' all Portable Data Files are optically recognised and converted into searchable hypertext mark-up language. These last files are put on a hidden SQL-database on the server, that can be searched, but that is not accessible for the public. The search box on the site searches through this HTML-files and presents the titles containing an entered keyword in their full text. Clients are redirected to the PDF-archive for which they need to log in, so that all full text data remains secured. This process has created a full text search, which answered to the needs of the journal's subscribers.

This full text search solved only a part of the problem. As doctors do not need to write foreign characters or vast mathematical equations, the HTML-format offers them enough features to find the data they need. This is not the case in Oriental Studies, nor in Linguistics or Actuarial Studies. A new solution had to be found.

3 A Future with Unicode?

3.1 What is Unicode?

3.1.1 Introduction

Unicode's websites define the standard as following:

The Unicode Standard is a character coding system designed to support the worldwide interchange, processing, and display of the written texts of the diverse languages and technical disciplines of the modern world. In addition it supports classical and historical texts of many written languages. (Unicode inc., 22-03-2005a)

In their Standard Unicode (Unicode inc., 22-03-2005b) the Standard sets out three directives for further development of the system. The Unicode Standard is meant to be:

- *Universal*. Unicode must contain all characters likely to be used in general text interchange.
- *Efficient*. Unicode should be used without additional software. It should synchronize all systems for sorting, displaying, searching and editing a text.
- *Unambiguous*. A Unicode code point always contains the same character

The definition clearly sets borders. The primary goal is to create a system for data interchange. Contrary to most other projects Unicode is a *standard* describing the place of a certain character in a layout table. It does not tell how a character should be presented. Therefore Unicode can be used in a multiplatform way, as font developers can use it as reference for their fonts.

3.1.2 How is Unicode organised? (Unicode inc., 22-03-2005c)

First of all, one has to look at the needs of text interchange. Different demands rise studying the application of Unicode. The primary goals include, as mentioned above, sorting, displaying, searching and editing a text. All four options require certain features.

The most basic feature is the display of certain character sets. This simply requires a table mentioning which character can be found on which cell of the layout table. Standardisation of lay out tables of fonts can solve the problem of displaying characters easily.

Next to displaying a character set, one should also be able to search, sort and even edit a text. This requires a more complicated standard. Unicode contains metadata dividing text elements into different characters, noticing the demands of a particular language, such as the addition of an diacritic sign, or display, such as hyphenation or contextual changes. Unicode contains character sets defined by tags, which can be used in a mark-up language enabling changes in lay out within a single text. Examples of these adaptations are bidirectional ordering and the insertion of text in foreign writings.

The combination of this particularities enables for example words containing "é" if the search string was entered only mentioning an "e". The edition of a text is also simplified as transcription is included into Unicode's metadata. Characters from the one set are connected to those of another. This once again enhancing search possibilities.

3.1.3 Unicode and Peeters Online Journals

Implementing Unicode from the beginning of a publishing process, would increase possibilities for online journals tremendously. First of all incoming data could be organised into a standardised XML-based SQL-data structure. Authors could directly connect to this database and make preprint corrections and changes, more or less in the same way as the papers of this congress where entered.

Starting from this corrected XML-files, publisher and printer can export the data to various export types: paper volumes, online Portable Data Files or even printable draft versions of an article. This would simplify the workflow opposing to the current situation (Figure 1).

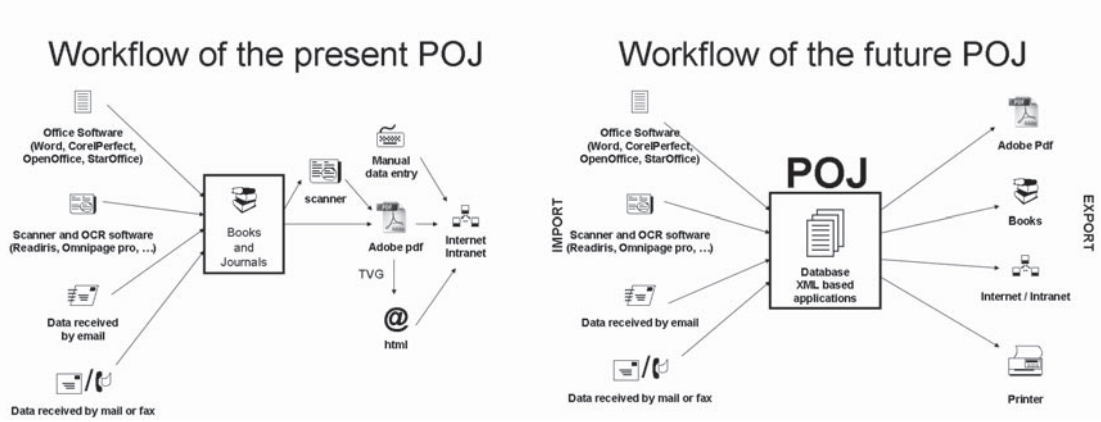

Figure 1: Current and future workflow

Materials on the website will be presented as they appeared on paper, displaying all characters without any additional software requirements.

Next to the presented material, also search options can be extended thanks to Unicode. First of all, one will be able to search for any type of character, even more, one can search for the foreign character without even being able to type this character. Language search and definition will be much more easy as characters or mark-up language will contain this data.

3.2 Unicode and OpenType

OpenType, a font format based on TrueType, but adding support for PostScript data, can be used for the implementation of Unicode. The OpenType font system allows to use Unicode encoding. Whereas Unicode defines characters, OpenType defines glyphs and thus the display of a font. The main goals of this font type are the following:

- broader multi-platform support
- better support for international character sets
- better protection for font data
- smaller file sizes to make font distribution more efficient
- broader support for advanced typographic control

The OpenType fonts use layout tables containing glyph substitution, glyph positioning, justification, and baseline positioning, which enables text processors to improve text layout on paper and online. Last, the fonts contain information on their source and are protected to keep their integrity. (An., 22-03-2005a)

OpenType enables to implement Unicode in a simple way, as the OpenType fonts are used by all Microsoft and Adobe Word Processing Software. Nevertheless questions remain on the multi-platform compatibility of the system. If this is a preferable option for use in Peeters Online Journals remains an open question.

4 Conclusion

For *Peeters Online Journals* implementing Unicode would mean a kind of rebirth. Particular demands of scholars would be answered in a simple way, without giving up the idea of easy access. Unfortunately implementing this systems demands a synchronisation of not only the publisher's or printer's software, but also the author's computers and attitude towards this rather technical issue. Nevertheless part of this implementation is already fact. Within the coming years this standard will increasingly become important. Peeters Online Journals will not fail to follow this evolution.

References

An., from http://www.microsoft.com/OpenType/OTSpec/otover.htm, 22-03-2005a.

Fletcher, L.A. (1999). Developing an integrated approach to electronic publishing: tailoring your content for the web. *Learned Publishing*, 12(2), 107-117.

Penfold, D. (1998). Requirements for an Online electronic journal service, *Learned Publishing*, 11(1), 9-16.

Unicode inc., from http://www.unicode.org/standard/standard.html , 22-03-2005a.

Unicode inc., from http://www.unicode.org/versions/Unicode4.0.1/ch01.pdf, 22-03-2005b.

Unicode inc., from http://www.unicode.org/versions/Unicode4.0.1/ch02.pdf, 22-03-2005c.