The race to digitize: are we forfeiting quality?

The article describes the errors and deficiencies found in digitized journal back issues. The results are not based on systematic or comprehensive research, but provide a snapshot of the sort of problems librarians and readers can experience when accessing digitized journals.

Errors and deficiencies are classified in the following categories: failed access, inaccurate journal titles, missing elements, insufficient quality of full text images, poor accuracy of OCR and inaccurate metadata.

Observations of the author indicate that digitized back issues of journals vary greatly in their quality. The conclusion contains a general recommendation that all publishers who have entered the 'race for digitization' should carefully review their quality control procedures and make sure that their products are an accurate reflection of their publishing history and not fraught with errors. The author suggests that publishers and providers should develop and adhere to strict quality standards for digitized journals. Only then can libraries really consider removing print journals from their shelves.

ALICE KELLER
Head of Collections Management
Oxford University Library Services

Introduction

The issue of the quality of digitized journals was raised at a meeting in the English Faculty of Oxford University. As a staunch supporter of e-journals I was trying to persuade the academics that duplicate subscriptions should be replaced by electronic access. However, one of the academics voiced his deep concern over the quality of electronic journals, in particular of digitized back issues. The fact that the optical character recognition (OCR) programme of a provider had misread and subsequently misspelled his name swelled his case and heightened his suspicion that new technology was not to be trusted.

I promised to go back and look into this issue, and this article is the sobering result of my investigations. All is not well, and I still need to go back to the academic and admit that he was right in many ways. This example illustrates how the very obvious carelessness of a small number of publishers and providers can wipe out all librarians' efforts to persuade readers to use digital surrogates as soon as they become available.

Furthermore, the results of my research work show that it is not only a small number of 'careless' publishers who deliver unsatisfactory products. It seems that most publishers and providers are not as accurate in their digitization programmes as librarians would wish. It looks as if the race to digitize as much as possible, as quickly as possible, is leaving us with a legacy flawed with errors and deficiencies.

Methodology

The results discussed in this article are not based on systematic or comprehensive research. My work should rather be described as a tentative attempt to classify the errors and deficiencies found in digitized back issues. The results are entirely descriptive, and not quantifiable.

Initially I looked into the errors found by the academic mentioned above. I then broadened my investigation to the whole journal, and other digitized journals with the same provider. Eventually I did spot-checks on as many digitized journals as I could access from my workplace. After gathering many examples from different journals and publishers, I tried classifying these errors and deficiencies into the discrete categories described below.

I tried to limit my research to volumes with (print) publication date earlier than 1990. Very few journals were available in electronic format before

1990. Therefore, full texts of such volumes currently provided in PDF or HTML format will be the product of subsequent scanning, with or without OCR. In some cases I was mistaken, and even volumes earlier than 1990 were clearly digitally 'born'. Such cases were easily recognizable by the nature of the full text file.

As mentioned above, the results are not quantifiable. However, the fact that I could find errors or deficiencies in many of the journals I evaluated indicates that they are plentiful. Indeed, some journals displayed a whole range of faults.

The classification below includes examples from journals in all subject areas and from a large number of sources. The study was carried out during July and August 2005. It is of course possible that some of the errors will be corrected in due course.

Classification of errors and deficiencies

Errors range from missing pages to missing volumes, from typos in article titles to errors in journal titles, from poor image quality to failed access.

Errors and deficiencies are classified in the following categories: failed access, inaccurate journal titles, missing elements, insufficient quality of full text images, poor accuracy of OCR and inaccurate metadata. This classification is tentative and may not be exhaustive.

Failed access

One of the most basic and banal problems is failed access. The error message 'Page cannot be displayed' can cause anything from mild frustration to despair. There is no easy way of knowing whether the provider is experiencing momentary technical difficulties or whether the reader is dealing with long-term unavailability.

In July 2005 the homepage and introductory web pages to *The Canadian Architect and Builder* (1888–1908) appeared without a hitch, but I was unable to browse or search the journal contents[1]. In such situations readers would normally write to the provider and seek clarification. In this case I wrote to the provider, but never received a reply.

Readers of the Internet Library of Early Journals (ILEJ)[2] experience similar problems. Two of the six journals digitized in this project are no longer available. However, in this case readers are informed accordingly: 'Unfortunately, at present some journal titles are not available online. *Annual Register* and *Philosophical Transactions* are currently unavailable. We apologize for the interruption in service.' The date stamp indicates that the journals have been unavailable for twelve months.

In another case Adobe Reader reported: 'There was an error opening this document. This file cannot be found'. This may be a temporary crisis at the publisher's end, it may be caused by my computer set-up, or it may be a more long-term problem. As a reader, I cannot know. All I know is that I do not have desktop access to the journal article I currently need.

Inaccurate journal titles

Inaccurate journal titles present a serious and persistent problem. It is important that the periodical title given in the electronic version is identical to the title of the print journal. Consistency of title information is essential for easy retrieval and correct citation.

This seems an obvious requirement and should not even need mentioning. But mistakes can be found, especially when the publisher or provider fails to track title changes correctly.

A close look at *The Florida Entomologist*[3] (available electronically from Vol. 1, 1917) reveals that this journal was actually called *The Florida Buggist* until Vol. 3, 1920. However, the web site lists all volumes under the later title of *The Florida Entomologist.* The original title features as 'other title'. Interestingly, the PDF full text displays the original title, which will cause confusion amongst both librarians and readers.

Another example of an overlooked title change can be found on the Blackwell Synergy web site. The journal volume cited here by the publisher as *Real Estate Economics,* Vol. 1 (1973) is in fact the digitized version of the *American Real Estate & Urban Economics Association Journal*[4]. The title change to

[1] *The Canadian Architect and Builder* (1888–1908): http://collections.ic.gc.ca/architect/

[2] Internet Library of Early Journals (ILEJ): http://www.bodley.ox.ac.uk/ilej/

[3] *The Florida Entomologist*: http://www.fcla.edu/FlaEnt/

[4] *Real Estate Economics* (Blackwell Synergy): http://www.blackwellpublishing.com/journal.asp?ref=1080-8620

Real Estate Economics occurred only twenty years later in 1995. Interestingly, digitized articles of this journal also appear in EBSCO's full text database Business Source Premier with the correct contemporary journal title. EBSCO goes to great lengths to establish the title history of this journal and lists three title changes. Blackwell's lack of care regarding title changes has also been raised by Steve Shadle who discussed problems relating to the cataloguing of e-journals[5].

ScienceDirect (Elsevier) is very conscious of title changes and lists volumes and issues under the correct title generation. But some title changes seem to defeat the system, and we find Volume 5, Issue 1 of *Deep Sea Research* under the wrong title[6]. Interestingly, the other issues of Volume 5 are listed under the correct title, but making up an incomplete volume, lacking Issue 1.

It needs mentioning that the problem of inaccurate journal titles is not restricted to digitized back issues. In fact, the article by Shadle mentioned above refers to some more recent title changes.

Some publishers have recognized the importance of accurate title information. The web site of the Institute of Physics, for example, gives readers a good overview of the title history. Volumes and issues are listed under the correct title[7].

1977–1998	*Journal of Optics*
1973–1976	*Nouvelle Revue d'Optique*
1970–1972	*Nouvelle Revue d'Optique Appliquée*

This journal merged with

1992–1998	*Pure and Applied Optics: Journal of the European Optical Society Part A*

to form

1999–present	*Journal of Optics A: Pure and Applied Optics*

Missing elements

Digitized journals as seen by the reader are made up of many web pages, PDF full texts and other files. It is not surprising that one or several of these elements can be temporarily or permanently missing. Although we have learnt to live with missing issues or items in our print collection, it is very disconcerting to find such gaps in the digital world.

Missing volumes At the 'highest' level, a whole volume can be missing. Often the reason is not clear. Was the provider unable to find a complete print run? Are there unresolved copyright issues? Or is it just an omission?

On the ScienceDirect web site, Volumes 61 and 62 (both 1994) of *Behavioral and Neural Biology*[8] are missing. No reason is given; they are just not there. Earlier and later volumes are available.

Similarly, the German project DigiZeitschriften fails to display Volumes 1 (1868) and 4 (1871) of *Jahrbücher für Kunstwissenschaft*[9]. Readers only have access to Volumes 2, 3, 5 and 6. Again, no reason is given.

Missing issues At the next level, missing issues can cause a problem. JSTOR lists missing issues as 'Issue Currently Unavailable'. At the moment the digitized back-run of *19th-Century Music*[10] shows three gaps in Volumes 2 and 3 (1978–1980). In this case readers are advised that issues are in production and we can hope that they will become available soon.

No reason is given for the gaps in the *Korean Journal of Parasitology*[11]. Early volumes contain Nos 1 and 3, but no link to No. 2. There are, for example, no references to pages 45–76 of Volume 6 (1968); these would most likely represent the missing No. 2.

Missing articles In order to locate missing articles you have to compare print and online side by side. Alternatively, you can compare entries in bibliographic databases with contents pages of digitized journals.

The latter method made me suspect that some articles in *Real Estate Economics*, Vol. 3 (1975),

[5] Shadle, S., Electronic Journal Forum, Reflection on Wrapping Paper: Random Thoughts on AACR2 and Electronic Serial, *Serials Review*, 2004, 30(1), 51–55.

[6] *Deep Sea Research*: http://www.sciencedirect.com/science/journal/01466291

[7] Example taken from *Journal of Optics*: http://www.iop.org/EJ/journal/JOpt/8

[8] *Behavioral and Neural Biology*: http://www.sciencedirect.com/science/journal/01631047

[9] *Jahrbücher für Kunstwissenschaft*: http://docsrv1.digizeitschriften.de/digitools/loader.php?ID=319180

[10] *19th-Century Music*: http://www.jstor.org/journals/01482076.html

[11] *Korean Journal of Parasitology*: http://www.parasitol.or.kr/kjp/

Issue 2 (actually *American Real Estate & Urban Economics Association Journal*[12], see above) were missing in Blackwell Synergy. According to the table of contents on the Blackwell Synergy web site, this issue only contains one article (pp. 7–29). The full text version in ProQuest lists five further articles, taking readers up to page 96. Finally, the version in EBSCO's full text database Business Source Premier includes a further item 'Presidential Remarks' on pages 99–102.

Looking at the digital archive of *Macromolecules*[13] (American Chemical Society), I noticed that pages 507–508 of Volume 17 (1984) were missing in the table of contents. Consultation of the print volumes showed that these two pages contained the obituary of Walter Hugo Stockmayer, a twentieth-century pioneer of polymer science. This may have been omitted intentionally, but it seems wrong that elements such as obituaries are not included in the digital archive.

Missing pages Publishers and other providers generally focus on digitizing the full text articles. Blank pages, or pages containing masthead, imprint, table of contents, or information on the Editorial Board are normally not included, especially if they are not numbered in sequence.

This means that for some purposes digitized versions cannot be regarded as a cover-to-cover substitute for the print edition. Naturally, one can argue that adding blank pages does not add anything to the journal content. But information on the (historic) Editorial Board can be very revealing; and including the original table of contents could solve questions of missing articles as mentioned above.

ProQuest deals with this issue quite ingeniously by introducing downloadable files with the heading 'Miscellaneous Unindexed Pages'. For example, the section of 'unindexed pages' in *Nonprofit World*[14], Vol. 5 (1987), Issue 6 contains sixteen numbered and unnumbered pages. The section includes title page, adverts, 'Letters to the editor', 'Ask the experts', 'First alert', 'Nonprofit briefs', and 'Welcome to our new members'. Following a similar system, JSTOR has introduced downloadable sections called 'Volume Information', 'Front Matter' or, correspondingly, 'Back Matter'.

I am sure that it would be very much appreciated by librarians and readers if other publishers followed this good example.

Quality of full text images

Poor image quality can cause anything from minor discomfort to serious problems for the reader. If the full text is simply not readable, there is little the reader can do. Similarly, if digitized graphics are not clear, they can become useless.

Obviously, digitizing technology is advancing rapidly and librarians and readers can be confident that the quality of full text images should be improving generally.

This may explain why the image quality in some of the early digitization projects is poor. The Internet Library of Early Journals (ILEJ)[15] can be regarded as one of the earliest projects and suffers to some extent from insufficient image quality. (The project was completed in 1999.) Some of the page images of the early volumes of the *Gentleman's Magazine* (1731–1750), for example, are simply not readable. Digitizing technology has clearly advanced since then.

Another example of a digitized journal which is nearly impossible to read online is the *AMCBT Newsletter* (1964–1974)[16]. Volume 8 defeats all attempts to be intelligible.

Interestingly – or significantly – both journals mentioned above are accessible free of charge. However, this should not serve as an excuse to offer poor image quality to the reader.

Poor image quality can also be found in licensed resources. For example, some articles in *Monthly Labor Review,* Volume 93 (1970) [17] (Business Source Premier) are extremely difficult to read.

Poor accuracy of OCR

Optical character recognition is the recognition of printed or written text characters by a computer.

[12] *Real Estate Economics*, 1975, 3(2): http://www.blackwell-synergy.com/toc/reec/3/2

[13] *Macromolecules*: http://pubs.acs.org/journals/mamobx/index.html

[14] *Nonprofit World*: http://proquest.umi.com/pqdlink?did=670847&sid=7&Fmt=1&clientId=15810&RQT=309&VName=PQD

[15] Internet Library of Early Journals (ILEJ): http://www.bodley.ox.ac.uk/ilej/

[16] *AMCBT (Association of Midwestern College Biology Teachers) Newsletter:* http://acube.org/newsletter.html

[17] *Monthly Labor Review*: http://search.epnet.com/login.aspx?direct=true&db=buh&jid=MLR

This involves analysis of the scanned-in image, and subsequent translation of the character image into character codes, such as ASCII, commonly used in data processing. Digitized journals which have been 'treated' by OCR can be searched, indexed and edited, and therefore offer added value compared to their print counterparts.

OCR involves complex image processing algorithms and rarely achieves 100 percent accuracy. For this reason many publishers and providers choose *not* to display the OCR texts. Instead, they are only used in the background for searching and indexing. Thus JSTOR uses OCR texts for retrieval, but describes them as 'unacceptable for display owing to typographical, word order, formatting, and other elements that are not accurately represented'[18]. The average OCR accuracy rate of JSTOR is 97% (on uncorrected text), but can reach 99.95% with some journals. JSTOR continues: 'The appearance of typographical and other errors could undermine the perception of quality that publishers have worked long and hard to establish and that users of all kinds expect.'

Journals that have been processed with OCR can be displayed as HTML full texts. However, as indicated above, they will most likely include typographical errors if they have not been checked and corrected manually.

Such errors can cause some amusement, but they are, of course, serious errors which undermine the usefulness and trustworthiness of digital surrogacy. A quite humorous example can be found in *Nonprofit World*, Vol. 11 (1993) where the 'Transformational Training Model' has transformed into a 'Transformational Raining Model'[19]. If, as in this example, digitized articles are displayed in PDF *and* HTML format, readers can check the accuracy of the OCR'd text.

Inaccurate metadata

Finally we come to what seems to be the most common area of error: the descriptive metadata.

Correct metadata is crucial for correct identification and citation of articles. All librarians recognize the importance of accurate metadata and take great pride in building flawless catalogues and indexes. It is therefore disappointing to see that some publishers and providers are so careless when it comes to gathering and displaying metadata.

My observations show that the quality of metadata varies greatly. Some publishers and providers supply very accurate metadata, adhering strictly to the standards set by librarians and bibliographic database providers. Others pay very little attention to detail.

Table of contents, title of article It is reassuring to see that the largest provider of digitized journals, JSTOR, offers very high quality metadata[20] (although I did come across an error in JSTOR, which seems to be the unfortunate exception that proves the rule). The French project NUMDAM (Numérisation de documents anciens mathématiques) offers equally flawless metadata[21]. Both get it right, even where foreign languages, diacritics or mathematical formulae are involved.

The quality of metadata in the newly launched German collection DigiZeitschriften, does not quite live up to the standard of its American counterpart JSTOR. Readers will come across several typos, including an article describing a 'Liederhuch'.[22]

Large publishers also tend to understand the importance of accurate metadata, even if you can spot errors here and there. *PROLA* (American Physical Society) includes an article on 'Vacuum Tures'; *Communications in Mathematical Physics* (Springer, Project Euclid) gets the word 'Informationsgössen' wrong; and Elsevier misspells 'coeeficients' in *Dynamics of Atmospheres and Oceans*.

One can, of course, argue that library catalogues also contain typographical errors, and that these examples are no worse than what a reader may find in his or her library's records. This objection seems entirely reasonable. But some

[18] See JSTOR web site, 'Why Images?': http://www.jstor.org/about/images.html

[19] Namaya, T., Training for Transformation in the Nonprofit Sector, *Nonprofit World*, Jan/Feb 1993, 11(1), 26 (ProQuest): http://proquest.umi.com/pqdlink?did=671300&sid=5&Fmt=3&clientId=15810&RQT=309&VName=PQD

[20] JSTOR: http://www.jstor.org/; DigiZeitschriften: http://www.digizeitschriften.de/

[21] NUMDAM: http://www.numdam.org/

[22] *Vierteljahrsschrift für Musikwissenschaft*, Band 7 (1891), Heft 4. http://docsrv1.digizeitschriften.de/digitools/loader.php?ID=336206

publishers and providers clearly do not reach this benchmark.

Surprisingly, even large database providers, who should have a lot of experience with collecting and analysing metadata, make serious errors when displaying title information. Any German speaker looking at the table of contents of *Zeitschrift für systematische Theologie*, Vol. 5 (1928) in PCI Full Text will be shocked to see words such as 'Wahrheitstriterium', 'Gosseserkenntnis', 'Chritus', 'Borsehungsglaubens', or 'Hauptstömungen'[23]. Clearly, readers would expect PCI Full Text, one of the leading full text providers in humanities, to offer more accurate metadata, even if the journals are not in English. Interestingly, other German language journals in PCI Full Text provide excellent metadata; *Zeitschrift für systematische Theologie* may just be an unfortunate slip.

I was equally surprised to find that digitization projects of libraries can be disappointingly poor when recording metadata. Some of the journals produced by the Göttinger Digitalisierungszentrum contain a worrying number of typos[24]. Similarly, projects of universities often lack accuracy. The Polish Biblioteka Wirtualna Nauki, for example, offers digitized back-runs of mathematical journals with comparably inaccurate metadata[25].

Inaccurate volume or issue numbering Numbering, as part of the metadata, can be inaccurate on various levels. Most obvious are wrong volume or issue numbers, or incorrect pagination (see further down). Such errors may seem minor at first, but they can confuse readers and may lead to wrong citations. If readers find contradictory numbering, they will not know which numbers to use in their citation.

As librarians know from experience, even with print publications publishers can make errors when numbering volumes and issues. However, within the context of digitization, we are more concerned about inconsistencies between the numbers appearing on the print issues and in the online version.

The volume listing of *Acta Arithmetica* (Warszawa) (1935–1965)[26] is correct, but the years given to match the volumes are quite obviously wrong, with 1964 appearing against Volumes 9, 10 and 14. Instead, Volumes 11, 12, and 13 are listed with the year 1963. Fortunately, with this journal the scanned images of Volumes 9 to 15 contain a reference to both volume number and year, and can therefore be assigned the correct year.

The digitized archive of *PNAS* (*Proceedings of the National Academy of Sciences of the United States of America*)[27] also contains a wrong date, with two issues supposedly appearing on 15 October 1918.

Inaccurate pagination Some publishers or providers do not include the page numbering in the online version of the table of contents. This is annoying and means that the reader has to consult the PDF full text to get this information. The *Cato Journal* (1981–), for example, does not list page numbering in the table of contents[28]. However, it can be found on the PDF full texts.

In other cases the page numbering is wrong. In the online table of contents of the *Journal of Algebraic Combinatorics* (Kluwer, DigiZeitschriften)[29], Vol. 1 (1992), all the page numbers seem to be shifted by 4. To make things even more annoying, Adobe Reader reports an error, and the PDF full text does not open.

An interesting question can be raised with an article in the *Annales de chimie et de physique (3e série)* (French digitization project Gallica)[30]. An article actually starting in Volume 9 (1843), page 164, is listed as starting on page 165. Further investigation shows that the original table of contents already contained the error. This raises the issue of whether digitization projects should

[23] PCI Full Text: http://pcift.chadwyck.co.uk/

[24] Example: Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen: http://www-gdz.sub.uni-goettingen.de/cgi-bin/digbib.cgi? PPN252457072

[25] Biblioteka Wirtualna Nauki: http://matwbn.icm.edu.pl/. Example: *Acta Arithmetica*.

[26] *Acta Arithmetica*: http://matwbn.icm.edu.pl/spis.php?wyd=6. It appears that volumes 9 to 15 cover the years 1964–1969.

[27] *PNAS*: http://www.pnas.org/. Example: Volume 4 (1918), Issues 10 and 11.

[28] *Cato Journal*: http://www.cato.org/pubs/journal/index.html

[29] *Journal of Algebraic Combinatorics*: http://docsrv1.digizeitschriften.de/digitools/loader.php?ID=325943.

[30] *Annales de chimie et de physique:* http://gallica.bnf.fr/Catalogue/noticesInd/FRBNF34378082.htm.

be used to correct apparent errors of the print texts[31].

Conclusion

The purpose of this article is not to name and shame, although it may seem so at first glance. It is not based on systematic or comprehensive research, and does not offer a consistent comparison of products of different publishers or providers. It is merely a snapshot of the sort of errors I came across when assessing the quality of a variety of digitized journals. The main purpose of the article is to classify the nature of errors and deficiencies found in digitized back issues.

The fact that I easily managed to find so many examples of errors within a short time-span indicates the extent of the problem.

Digitized back issues of journals vary greatly in their quality. The quality of journals included in JSTOR appears to be well above average, which is reassuring for libraries.

The most common errors seem to be typos in the metadata, which indicates that publishers and providers have not implemented strict quality control procedures. These errors may seem minor compared with missing articles or volumes. The problem of poor image quality should disappear as imaging technology advances rapidly.

Coming back to the question asked at the beginning, whether digitized back issues offer reliable surrogates for print journals, the answer seems to be both 'yes' and 'no'. I imagine that in most cases the digitized full text fully meets the reader's requirements. He or she may not worry too much about typos in the table of contents, as long as the PDF full text provides an accurate replica of the print version. Where the requirements are not met, the reader will most likely automatically resort to the print version – providing it is (still) available.

From the librarian's point of view many digitized journals cannot be considered reliable surrogates. Librarians want to be confident that all the information contained in the print volumes is displayed accurately and in its entirety, before accepting a digital surrogate as a true substitute. Even then, many will insist that the long-term availability of digital journals is not yet secure enough to replace the print collection.

As a conclusion I would like to recommend that all publishers who have entered the 'race for digitization' should carefully review their quality control procedures and make sure that their products are an accurate reflection of their publishing history and not fraught with errors. The fact that back issues are available free of charge should not serve as an excuse for poor quality.

I think I am speaking for my colleagues and readers if I recommend that publishers and providers should develop and adhere to strict quality standards for digitized journals. Only then can libraries really consider removing print journals from their shelves.

■ **Alice Keller**
Head of Collection Management
Oxford University Library Services
Bodleian Library
Broad Street
Oxford OX1 3BG, UK
Tel: +44 (0)1865 277 074
Fax: +44 (0)1865 277 187
E-mail: alice.keller@ouls.ox.ac.uk

[31] Considerably more editing is done in the table of contents of the *Bibliothek der schönen Wissenschaften und der freyen Künste* (digitized by the Bielefeld University). Here the titles of book reviews are edited to match the correct title of the original work. Certainly a very useful feature, but how far should we go?