

Nature - A Prototype Digital Archive

Ross MacIntyre¹, Simon Tanner².

¹Manchester Computing, University of Manchester, UK, email:r.macintyre@mcc.ac.uk

²Higher Education Digitisation Service (HEDS), University of Hertfordshire, UK,
email:s.g.tanner@herts.ac.uk

Abstract

In response to a proposal to digitise the entire backrun of the journal *Nature* (1869-1992), a pilot project was commissioned to discover the technical issues and to ascertain costs.

The conversion and digitisation were performed by the Higher Education Digitisation Service (HEDS) whilst Manchester Computing undertook the data management and application development.

The paper describes the processes undertaken, the results ascertained from these pilot processes and the techniques used. The pilot, though not complete at the time of reporting, could draw some conclusions from the progress made. They include recommendations to strip bound volumes for digitisation and that 400dpi is sufficient for greyscale and colour. There is, however, no satisfactory solution for the large file sizes produced in PDF.

Keywords : digitisation, digital archive, Nature.

1 Introduction: The Journal *Nature*

Nature was first published in 1869 and has been a weekly science journal to the current day. The journal's objectives were stated in the first issue by founding editor Norman Lockyer, who then remained at the helm for fifty years: to bring the accomplishments of science to the "general public". *Nature* became the international journal of first choice for the presentation of original results and discoveries during the inter-war years, coinciding with the change of editor to Richard Gregory. *Nature* remains the most widely cited interdisciplinary weekly science journal in the world.

A digitised archive of *Nature* made available to higher education institutions would provide the following:

- A valuable research tool
- An aid to teaching – the publishers of *Nature* advise that they receive many requests for access to papers for teaching purposes and making these available electronically would make such access more cost effective
- A store of source material for students and teachers
- A unique view of scientific history as an aid to historians and sociologists of science.

2 Pilot Objectives

In order to assess technical requirements, the cost and timescale in a digital archive of nature, a pilot project was established in 1997. The pilot project had three UK partners, the Higher Education Digitisation Service (HEDS), Manchester Computing, University of Manchester, and

Macmillan Publishers Ltd ('Macmillan'), the publishers of *Nature*. It had the following objectives:

- To test the digitisation of historical scientific material with textual, graphical, photographic and formulaic content. This content being expressed in many fonts, styles and standards in changing paper formats.
- To test whether the source material can be digitised, accessed and archived to standards sufficient to support teaching and research with in Higher Education in the UK.
- To compare various types of search technology on the source material.
- To test possible delivery mechanisms.
- To ascertain the costs, time and techniques required to meet the above objectives.

3 Method

The first general decision was to provide a page image to the user with background searching and indexing. Why "image and text" PDF? This format gives many of the benefits of PDF conversion with the main difference from full PDF being that it is viewed on the screen as a bitmap image of the original. All the features of text search and retrieval are available, but hidden from immediate view. In searching the hit words can be displayed. Other approaches using page image rendition, e.g. JSTOR [1] and Internet Library of Early Journals [2], store the OCRd text in separate, searchable files, but there is no indication on the image itself of hit terms.

HEDS visited Macmillan's offices to assay the complete collection of *Nature* to ascertain an idea of the numbers of pages in total for the project and the proportions of the various types of pages. These were divided into four groups: text-only pages, pages containing bi-tonal graphical images (e.g. graphs, sketches or reproductions of etchings), pages with photographic or any half tone or other greyscale graphical content, and pages with colour content.

The basic scan specification used was the following.

Page Format	Tone / Colour	Resolutions
Text only.	1 bit, Black and White	400 & 600 DPI
Text with line drawing or bi-tonal graphic.	1 bit, Black and White	400 & 600 DPI
Pages with photographic or any half tone or other greyscale graphical content.	8 bit, 256 Greyscales	400 & 600 DPI
Pages with colour content	24 bit, RGB Colour	400 & 600 DPI
	8 bit, 256 Greyscales	400 & 600 DPI

Each page to be scanned was assessed for its content and the specification above applied. For all the samples an "image and text" PDF plus a text file was created and delivered to Manchester Computing. The process was measured, refined and measured again to gain production metrics.

The TIFF images are to be retained for archiving and for using again if a different technology for viewing/searching is selected at a later date.

4 Processes and Techniques

A number of processes and techniques were used to produce the samples and this section details these and the reasons for following certain routes. (See Appendix A for Production Process Maps.)

4.1 Scanning

Nature is complex in the density of its layout and the font sizes used over the lifetime of the publication. In more recent years the content has become more graphical, with heavy use of photographs and colour images. This was especially challenging in the genetic subject areas where the gradations of tone used in graphic representations are critical to the scientific accuracy of the material. There is also a period of years from the mid 1970's to the mid 1980's where the paper of the publication is very thin. This affects the preparation and handling of the pages and there is significant show through of text from the reverse side of pages.

As *Nature* had been presented in bound volumes for scanning, HEDS utilised its Zeutschel Omniscan 5000 Bookscanner to gain the best possible output from the journal with minimum degradation to the originals. HEDS were constrained by the technology available in the Bookscanner market, although they were the first in Europe to receive the Zeutschel greyscale scanning capability and fully expect to be among the first to gain 600 DPI capability in the Bookscanner. There does not appear to be any Bookscanner capable of 600 dpi optical resolution in 24 bit colour or 8 bit greyscale with TIFF as the native file format.

For colour and greyscale scans at 600 dpi, HEDS had two options, one that proved more viable than the other. The less attractive option was to flatbed scan the *Nature* samples in the bound volumes with all the attendant degradation of the originals and an additional production time premium. The better option proved to be to strip the *Nature* bound volumes into single pages, scan those pages and then reconstitute the bound volume. HEDS tested this process using volumes kindly sourced from the library of The Royal Society in London. It proved cheaper than bookscanning despite the strip and rebind costs, enabled higher resolution image outputs and lower overall degradation to originals than flatbed scanning of bound volumes. The challenges faced in this method was the conservation standard stripping and rebinding of the original *Nature* volumes carried out by the HEDS supplier, Riley, Dunn and Wilson Limited. For the period from the mid 1970's onwards the paper used for *Nature* proved too thin to rebind effectively to the Royal Society's requirements: if the entire digitisation were to go ahead, then volumes covering these years would have to be sourced from elsewhere.

Nature is also tightly bound and the pages are slightly larger than A4 with rather small margins. This results in non standard scan image file dimensions that have to be taken into account in all other software processes to ensure that the whole page content is retained in the PDF file. It also means that the content of the originals may appear to run very close to the gutter of the bound volume with curvature in the paper. This must be resolved in the scan process to get the best image from the page as possible. This is possible with the handling capabilities available from Bookscanner technology with the use of a book cradle or with pages from stripped bound volumes.

The time and cost of scanning are distinctly affected by the standard to which the image is being reproduced. There is a marked increase in file size and attendant scan time for increases in resolution. These file sizes are increased again by any increase in tonality from black and white to greyscale and to colour. This is demonstrated in the table below:

Average TIFF file sizes for various resolutions and tones / colours

Tone / Colour	400 DPI Resolution	600 DPI Resolution
Black and White (bi-tonal)	1,300 Kb	3,400 Kb
Greyscale (256 tones)	11,000 Kb	20,000 Kb
Colour (24 bit)	15,000 Kb	30,000 Kb

The file size detrimentally affects the speed of writing and retrieval of the file to and from disk for any processing that is required from the scan stage onwards.

4.2 Post Processing

Post-processing is necessary for a number of reasons: deskewing the image, removing some dirt or speckling from the image, and output of the image file in the correct format for further processes. Many of the image files contained some skewing. The despeckling or dirt removal had to be managed with care to ensure that the process does not affect the content of the text or graphical elements of the original content.

The deskewing process when automated is very quick and effective but the software can have some occasional problems dealing with certain types of page formats. *Nature* pages have a horizontal line across the top of each page with the text arranged in columns below. Where the alignment of the horizontal line is not at a true 90 degree angle with the text then the post processing is quite likely to align the page with the horizontal line and thereby introduce a skew to the page image. Also where there are vertical lines or dirt in vertical lines in the image then this could also introduce skew into the image file.

HEDS were also constrained by the technology available for post processing greyscale and colour image files. These types of files cannot be automatically processed to remove skew etc. by the top products in the market, ScanFix and PixEdit. Both of these software tools are developing greyscale capabilities, but there is no timescale for availability as yet. Any deskewing of greyscale or colour images has to be done manually at the scan stage and this adds an additional time element to the process. The manual skew available within the scan engines is quite basic and not as effective as can be achieved with tools such as ScanFix. This means that some colour and greyscale image files will retain some small level of skew in a production process.

4.3 “Image and Text” PDF Production

The conversion of the TIFF files into “image and text” PDF format is being done using the Adobe Capture 2.x conversion tool. The Adobe PDFWriter module is set at 600 DPI with no compression or downsampling selected to achieve the maximum resource representation into the PDF file. It is not possible to test independently what the exact resolution is within a PDF file, only the means by which it was created. The Capture 2.x engine was slow at converting files into PDF format, sometimes as bad as 10 minutes machine time per page image converted. However the Capture 2.x engine is far more reliable, and the output requires lower levels of QA, when compared with the quicker production times possible using Capture 1.x. The Capture 2.x engine has a hardware dongle that effectively adds an additional cost per page converted.

The Adobe Capture process also adds complexity to the production process for image files that are 600 DPI greyscale or colour. Whilst Capture 2.x can process 600 DPI bi-tonal files, it is not capable of converting greyscale or colour files sourced at above 400 DPI. Therefore, if HEDS creates 600 DPI originals they have to be converted to 400 DPI prior to conversion into PDF format. This does not necessarily involve downsampling the original file, but merely changing the TIFF header information through a batch save process using a product such as Paint Shop Pro 5.x. The processing of such large files through Capture requires a very large amount of

memory to be available, approximately 5 times the image file size being converted. There is a big overhead on the machine processing the image files as the image file sizes average between 20Mb and 30Mb. This particularly slows the processing of colour and greyscale 600 DPI images relative to the other samples.

Average PDF file sizes for various conversion processes:

Tone / Colour	Average PDF at 600DPI conversion.	Downsampled size - minimum PDF
Black and White (bi-tonal)	340 Kb	222 Kb
Greyscale (256 tones)	2,225 Kb	249 Kb
Colour (24 bit)	4,200 Kb	851 Kb

The downsampling process in Adobe Capture means setting lower resolutions and maximum compression for the output PDF file. These changes mean that information content is being lost in the conversion. The result is that picture elements still look reasonable, but text might appear blurred. This may possibly be due to interpolation. To achieve downsampled PDF files as well as full resolution PDF files would require the Adobe Capture process to be repeated in its entirety for all PDF's, almost doubling the costs of this portion of the process.

4.4 ASCII Text Production

The ASCII text from each page was required to create search indexes to find individual pages. This file was created at the same point at which the "image and text" PDF was written, using the OCR'd content from the PDF file. The results show a higher level of accuracy in the OCR than expected, so that the indexes will be a richer search tool.

4.5 Preparation and Quality Checks

There were a number of preparation functions that needed to be completed before the originals can be converted. These included marking all advertising pages to ensure they are not scanned, checking for colour prints and marking them for different processing, setting up a production log and data entry of document pages and data structures. There is also the set up time for each of the production processes to ensure the machinery is at the optimum setting for the originals to be converted or data processed.

There were a number of points in the process where there were basic checks made to ensure the quality of the output. These were to ensure that every page had been scanned and that all the pages were in the correct order. There were further checks on the content of every TIFF file output to ensure that the content was representative of the original and that the correct file name had been assigned to the TIFF image. Similar checks to the file name and content were carried out for the PDF and text files. The fact that a single page creates three separate single output files adds to the cost of quality assurance because of the total volume of files to be controlled and checked against every page scanned.

5 Assay of Nature Collection at Macmillan

HEDS did a survey of one issue of *Nature* per year of the publication from 1869 to 1992 in order to find the proportions of black and white, greyscale and colour format pages in the whole collection. From such a survey it was possible to estimate of the total number of such pages in the collection and thus, with higher accuracy, the total costs of digitising the whole collection. The results of the survey are presented in graphical form in Appendix B.

(It is interesting to note that the complete set held in the Editor's office had been written on, in ink, over most text passages thereby obscuring of the text for the majority of the publication

run, making them unsuitable for scanning. This is also one of the main reasons that HEDS sourced originals from The Royal Society.)

- There are an estimated **298,950 pages** in the total production run of *Nature* from 1869 - 1992. This figure was found by taking the number of pages per issue per year sampled and multiplying that by 52 for each year and then adding up the results. This figure was cross checked by dividing the publication into blocks according to the chronological changes in design and layout of *Nature* and then averaging the number of pages per issue per year across each design change. By adding up the results, a figure of 298,600 pages was found which cross checked favourably with the above figure.
- That the proportions of black and white, greyscale and colour pages are:

Tone / Colour	Percentage of total	Total no. of pages
Black and White (bi-tonal)	87.6%	262,028
Greyscale (256 tones)	11.8%	35,204
Colour (24 bit)	0.6%	1,196

6 Application Development

6.1 Outline Application Specification

The following documents the prototype's specification drafted by Prof. David Pullinger from Macmillan. The sections that follow subsequently describe the application development undertaken and some planned future directions.

Home Page :

- Paragraph explaining the contents and pilot project.
- Invitation to fill in feedback form
- Three routes to articles in the archive
- Link to *Nature's* web site

Navigation by introductory pieces and indexes:

- Introductory piece explaining what has been in *Nature* and its value and whom it might interest
- Links to focus area
- Each focus area has introductory paragraph explaining the interest of this section and a linked index of articles

Navigation by tables of contents:

- When each article is scanned, the whole issue is done at the same time.
- Agreement on header information will lead to the automatic construction of table of contents.

Navigation by search:

- Search by text
- Search by bibliographic citation

6.2 Technical Development

A prototype application has been created, available on the WWW, with access restricted by password and UK academic network address. This could be replaced by the UK academic standard access control mechanism, ATHENS[13], should the prototype be made more widely available. This has been loaded with the selected issues from *Nature*, provided by HEDS, supplemented with additional descriptive data.

The application infrastructure was developed as part of another research project and was not specifically designed for this purpose. Some tailoring has been necessary, in particular, the use of objectbase management system software added a level of complexity to the archive which is unnecessary. The data is highly ordered, its hierarchical structure being reflected explicitly in the directory structures defined. The files are assigned unique, self-identifying, names. In addition to the PDF files, the header data is held in consistently named files in SGML which is dynamically converted to HTML for display using (OmniMark[3]) scripts. Latterly, the SGML source files have been converted to XML and experimentation will continue.

6.3 Data Loading Approach

The unit of digitisation is one file per page and upon receipt at Manchester Computing, effectively nothing is known about the contents, other than page number and other 'standing data' such as journal name, publisher, etc.

The creation of the metadata is vital for sensible browsing and its creation is a separate process, involving an external organisation, described in the next section. As metadata is defined and received, the PDFs are combined (see 6.5 below) and the data is then accessible.

6.4 Metadata

To give a context for discussing metadata, we note here that articles might have no author, or publishing house author that was noted only by initials. It might have no title, or a title that stayed the same in each issue (e.g. Report from The Royal Society) or be part of a longer piece which was split up and spread over several issues. Moreover the type of article appearing in the journal changed according to editorial policy and, sometimes, whim.

The creation of metadata was performed as a 3-pass operation:

6.4.1 Automatic Creation of Page-level Metadata

The aim is to serve the digitised content at the earliest opportunity. When individual pages were received, a header record was created. Initially this contained just 'standing data', e.g. journal name, publisher, ISSN, plus the page number. The header data was loaded and inserted into the PDF files for consistency.

6.4.2 Manual Creation of Minimal Article-level Metadata.

Article metadata files were created by Manchester Computing for an initial, small set of issues. However, an external 'keying agency', Saztec Europe Ltd., was appointed to create and validate further header records. Their prime objective was to identify the editorial contents of each issue at an acceptable level. Specification detail is given in section 6.5, below. It would be incorrect to view this as little more than the re-keying of the table of contents for each issue, since the data presented in the published Table of Contents is not actually sufficient. This is true for authors, whose names are often omitted; for collections small pieces, such as "Our Astronomical Column", which are not subdivided; and most significantly for file handling, the end page is not listed.

Reference volumes were withdrawn from the John Rylands University Library at Manchester and transported to Saztec, who then worked from the bound volumes. The header files were returned as simple tagged files on CD.

The data received was then expanded to incorporate the standing and derived data, creating SGML header files. Again, the data was inserted into the PDF files, for consistency and to improve the presentation of search results.

6.4.3 On-going Improvement.

The Archive will present numerous possibilities for subsequent cataloguing. The amount and type of work undertaken could be the subject of separately funded initiatives outside the scope of the pilot. The infrastructure within the application should support such embellishment.

6.5 Minimal Metadata Defined

Manchester Computing defined a minimal set of metadata for the archive and created an SGML DTD to reflect this. Initially this focused solely on Dublin Core, though RDF has been used for experimental purposes.

Dublin Core [4] Tagged Items:

- 1) * <TITLE scheme="Internal"> Article Title </TITLE>
- 2) * <CREATOR scheme="Internal">
 - <FNMS> AuthorForename(s) </FNMS>
 - <SNM> AuthorSurname </SNM>
 - <SFX> PostNomial </SFX>
 - <AFF> Affiliation </AFF>
</CREATOR>
- 3) <SUBJECT> n/a for *Nature* </SUBJECT>
- 4) * <DESCRIPTION scheme="Internal"> Description </DESCRIPTION>
- 5) <PUBLISHER scheme="Internal"> Macmillan </PUBLISHER>
- 6) <CONTRIBUTOR> n/a for *Nature* </CONTRIBUTOR>
- 7) * <DATE scheme="ISO 8601"> Cover Date YYYY-MM-DD </DATE>
- 8) <TYPE scheme="DCObjects"> "Article" </TYPE>
 - * <TYPE scheme="Internal"> TypeOfContent </TYPE>
- 9) <FORMAT scheme="IMT"> "application/pdf" </FORMAT>
- 10) <IDENTIFIER scheme="SICI"> SICI </IDENTIFIER>
 - <IDENTIFIER scheme="Internal"> PhysicalFileIdentifier </IDENTIFIER>
- 11) * <SOURCE scheme="Internal">
 - <JTL> JournalTitle </JTL>
 - <VID> Volume </VID>
 - <IID> Issue </IID>
 - <PPF> StartPage </PPF>
 - <PPL> EndPage </PPL>
</SOURCE>
- 12) <LANGUAGE scheme="ISO 639"> "EN" </LANGUAGE>
- 13) <RELATION scheme="ISSN" relation="IsPartOf"> ISSN </RELATION>
- 14) <COVERAGE> n/a for *Nature* </COVERAGE>
- 15) <RIGHTS scheme="Freetext"> Copyright String</RIGHTS>

Note that additional labels have been introduced within the SOURCE and CREATOR elements for clarity, though they would not necessarily appear explicitly in future instances of the data, e.g. as HTML 'meta' elements.

* Items 1, 2, 4, 7, 8 and 11 are created by Saztec and passed to Manchester Computing in the simple tagged file referred to in 6.4.2. Item 8 (TypeOfContent) is drawn from the types of contribution *Nature* publishes: Articles, Letters, Review Articles, Progress Articles, Scientific Correspondence, News & Views, and Supplementary Information, etc. The remaining items are either derived by MC or are not applicable.

Example: corresponding HTML v4.0, simple, unstructured metadata elements:

- 1) <meta name = "DC.Title" content = "The Comet">
- 2) <meta name = "DC.Creator" content = "Hind,J.R., FRS, Greenwich Observatory">
- 3) n/a
- 4) <meta name = "DC.Description" content = "The Comet by Prof.J.Brocklehurst, University College, Dublin. ">
- 5) <meta name = "DC.Publisher" content = "Macmillan Publishers Ltd, Crinan St, London">
- 6) n/a

- 7) <meta name = "DC.Date" scheme = "ISO 8601" content = "1874-06-25">
- 8) <meta name = "DC.Type" scheme = "DCObjects" content = "Article">
<meta name = "DC.Type" scheme = "Internal" content = "Book Review">
- 9) <meta name = "DC.Format" scheme = "MIME" content = "application/pdf">
- 10) <meta name = "DC.Identifier" scheme = "SICI" content = "0028-0836(18740625)10:243">
- 11) <meta name = "DC.Source" content = "Nature 10-243 pp149-150">
- 12) <meta name = "DC.Language" scheme = "ISO 639" content = "EN">
- 13) <meta name = "DC.Relation.IsPartOf" scheme = "ISSN" content = "0028-0836">
- 14) n/a
- 15) <meta name = "DC.Rights" content = "Macmillan Publishers Ltd. 1874">

The PDF files were combined, via a command line batch process, using Acrobat Developers Association library utilities, based on the metadata created during the '2nd pass'. There is 1 file per distinct start page. So, if article 1 runs from p2-p3, article 2 is only on p3 and article 3 runs from p3-p4, two files are created 2 files, 1 containing p2 & p3 PDFs and the other p3 and p4 PDFs. It is acknowledged that this can lead to the creation of duplicate pages, but it establishes a one-to-one relationship between header and article.

The reason for combining the physical files include: likelihood of retrieval of subsequent page(s); availability of byte-serving, so pages are downloaded only as required or in the background; logical consistency with metadata; averting problems in printing and obtaining 'next page'.

The metadata being proposed has been discussed with TASI, the Technical Advisory Service for Imaging[5]. They are a recently established, JISC-funded body who help support image-based development projects involving the UK Higher Education community.

6.6 Searching

Both a bibliographic and a free text search capability have been implemented, using Verity's Search97 Information Server software[6]. It is refreshing to report that, in the main, the software worked as expected, including 'hit-term highlighting'. This means that even though an image is being displayed on screen, the term(s) searched for are highlighted, due to the presence of the text within the PDF file. The facility requires later versions of web browsers and the Acrobat Reader v3 plug-in. There have been some browser version specific patches required to the software on the server and Microsoft Internet Explorer has been problematic.

Early experimentation was done in generating a 'topic tree' from the text using Software Scientific[7] code. This can be used in conjunction with Verity to assist in searching. The search for 'like' terms is biased towards the words in context, i.e. terminology actually used in the documents. This appears to help with problems associated with terminology changing over time. It could possibly be used in time-slices, as appropriate. It is planned to explore the extraction of keywords from the uncorrected OCRd text.

6.7 Themed Access - 'Nature Trails'

Nature commissioned some leading historians of science to identify sections of *Nature* that both remained as hot research topics and that had a long history. These were considered as useful 'themes' to help introduce users to the riches to be found in historical archives. The question was how to provide links between the pages comprising such themes.

The Electronic Publishing Research Group at Nottingham University[8] and the Multimedia Research Group at University of Southampton[9], agreed to a formal collaboration in support of the 'themed' access required. Southampton's DLS software (available commercially from Multicosm[10]) was used to establish links within the archive. The method used was to dynamically insert next and previous bookmarks into the PDF when served. This does not affect the appearance of the image and does not permanently affect the PDF file itself.

It is planned to exploit the software further to support the subsequent definition of linkbases in support of teaching. The technology was the subject of a previous paper [11]. Coincidentally, Software Scientific also offer code to assist in the development of linkbases and identify 'themes' across collections of documents. It is intended to explore use of their software in conjunction with DLS.

7 Observations

7.1 *The advantages of stripping the original bound volumes for scanning.*

Stripping enables the optimisation of the scanning phase, eases some aspects of preparation, handling and reduces overall costs substantially. Stripping also allows for a wider choice of scan equipment and supplier of scan services to be considered. There is an acquisition cost and the cost of the conservation stripping and rebinding, but this is offset against the savings in the cost of overall processing.

The main downside of having stripped pages of *Nature* is that the later issues have substantial advertising content that must not be scanned. The correct and accurate inventory of all pages to ensure that they are rebound in the exact order received in the bound volume is a detailed and time-consuming task. This is because advertising pages are not numbered and the inventory creates an additional production cost not expected at the outset of the trial.

7.2 *400 versus 600 DPI*

600 DPI has been established as a standard for full archive scanning of black and white text in the USA on projects such as JSTOR and at Cornell University. The reason for Cornell and JSTOR recommending 600 DPI is that in bi-tonal scanning there is a risk of losing some data at lower resolutions.

The main issue that drove the resolution requirements for *Nature* was related to the information content of the resource. The resolution of 400 DPI will represent all the textual information in the journal. 600DPI would add a level of detail that would not add to the readable content but would add a small amount of character edge smoothing. This effect in the TIFF files is negligible from the perspective of the end user of the material. Please also note that the end user will only ever view the PDF's which will not be at a measurable 600 DPI whatever the input source file used.

Therefore, the only remaining reason for 600DPI must be to futureproof the TIFF files for potential uses not defined at this time. It is doubtful that increasing resolution alone, when not needed for the immediate application, will ever futureproof image files. TIFF file standards and the 400 or 600 DPI standards have a maximum shelf life of about 7-10 years and it is likely after this period that two outcomes will have occurred. First, the images will be deemed of too low resolution whatever choice is made for 400 or 600 DPI at this time, creating a potential rescan requirement. Or, the other possibility is that technology will develop such that for OCR or other post scanning processes, resolution becomes a non-issue due to the fuzzy nature of the technology. Thus, the choice of 600 DPI for archive is a belt and braces approach to archiving but with a shelf life of 10 years maximum. The decision is whether the additional cost is warranted for the security of the next 10 years.

HEDS considers it essential that any decision on resolution specifications be based not just on technical issues, but on the structured evaluation and assessment by subject specialists and other interested parties with a deep knowledge of the journal *Nature*. This recommendation fits with the keynote paper given by Anne Kenney at the Joint RLG and NPO Preservation Conference: Guidelines for Digital Imaging [12]. She argued against the development of guidelines that are too rigid or too narrowly conceived; instead, she suggested that experts should be working toward agreements on providing guidance rather than guidelines. Anne Kenney also encouraged the group to undertake that which is doable, and to acknowledge the gap between the ideal and the possible.

7.3 600 DPI would be unsuitable and unnecessary for the colour and greyscale pages.

There are no fixed standards for colour or greyscale images, but discussions by HEDS with Anne Kenney, Associate Director of the Department of Preservation and Conservation at Cornell University has derived a recommendation that any higher resolution than 400 DPI for colour or greyscale would not add any further content to the scanned images. This is because the increased bit depth produces more information and covers the potential for lost information of scanning at lower resolutions. Therefore, whatever the resolution chosen for the black and white text pages of *Nature*, 400 DPI should be used for greyscale and colour format pages.

7.4 Further experimentation to reduce PDF file sizes is required.

The size of the PDF files can be onerous, e.g. from 150Kb to 3Mb per page, so avenues to reduce these continue to be explored. Though the archive could be digitised to produce 'normal' PDF, the cost associated with OCR correction and the almost absolute certainty of error ruled this out as an option. Nevertheless, perhaps certain 'cherry-picked' PDF files could be OCR corrected, such as those all included in a specific '*Nature* Trail', or papers by Darwin, Watson/Crick etc.

[Experimentally, an all-text page from 1874 was Captured as PDF Normal using New Century Schoolbook(NSB). Recognition was surprisingly good given the small point size (equivalent of 8 point). The character spacings and other font metrics of NSB gave a pretty harmonious match. The page ended up as 13Kbytes as opposed to about 300 Kbytes for the Image+Text PDF.]

Experiments with down-sampling so far have been disappointing, giving a blurred appearance to the text and this would be unsuitable as the main means of access to the information in *Nature*, but further work is recommended.

7.5 Use of OmniMark.

The SGML header files are currently converted to HTML for display using OmniMark scripts. This has proved to be a technically satisfactory method, though the academic price of OmniMark has risen approximately ten-fold since first acquired in 1996, making upgrade unlikely. The reduced functionality offered by the free OmniMark Lite (OMLE) software may be adequate and is being used, though the conversion code has also been written in C for comparison. This does provide advantages in terms of speed, but is much less 'elegant'.

8 Concluding remarks

The pilot project has achieved its first aims of exploring the technical procedures for the digitisation of historical scientific content from 1869 to 1992 with a mixture of types of content and presentation. In particular the source material can be digitised and archived to the standards sufficient to support teaching and research in UK Higher Education institutions. Moreover, we have been able to determine the costs of creating a full back-run. We have not, so far, been able to test the accessibility of this material and this would be the next stage prior to major work. In particular, the standard solutions for reducing the size of PDF files do not work easily with this type of historical scientific material and more work is required on this front.

References

- [1] JSTOR - <http://www.jstor.org>
- [2] ILEJ - <http://www.bodley.ox.ac.uk/ilej/>
- [3] OmniMark - <http://www.omnimark.com>
- [4] The tag labels are from "Dublin Core Element Set: Reference Description" <http://purl.org/DC> (last update 1997-10-2)
- [5] TASI - <http://www.tasi.ac.uk>
- [6] Verity - <http://www.verity.com>
- [7] Software Scientific - <http://www.scientific.co.uk>
- [8] The Electronic Publishing Research Group at Nottingham University - <http://www.ep.cs.nott.ac.uk>
- [9] The Multimedia Research Group at University of Southampton - <http://www.mmrg.ecs.soton.ac.uk>
- [10] Multicosm - <http://www.multicosm.com>
- [11] Proberts, S., Brailsford, D.F., Carr, L., Hall, W.: *Dynamic Link Inclusion in Online PDF Journals*, EP98, International Conference on Electronic Publishing, Document Manipulation and Typography, April 1998, Saint-Malo, France. (<http://www.ep.cs.nott.ac.uk/~sgp/ep98.pdf>)
- [12] Joint RLG and NPO Preservation Conference: Guidelines for Digital Imaging, Warwick, UK, September 1998. Reported by Elkington, N., Beagrie, N.: *RLG DigiNews 2:5*, October 15, 1998. (<http://www.thames.rlg.org/preserv/diginews/diginews2-5.html>)
- [13] ATHENS - <http://www.athens.ac.uk>

Acknowledgements

Prof.D.Pullinger, Macmillan Publishers Ltd, provided inspiration and guidance, and continues to do so.

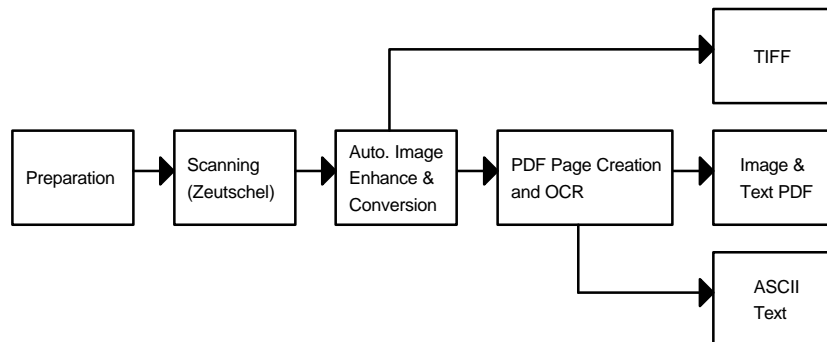
The Royal Society kindly provided access to their complete collection of Nature in bound volumes to enable the pilot to progress further. Thanks are due to the Library Manager, Sandra Cumming, for her assistance and flexibility.

This pilot project is funded by the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils.

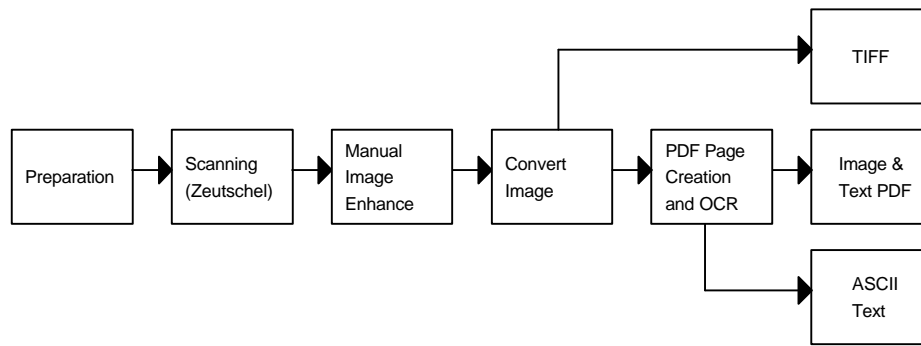
Appendix A Production Process Maps

The following are graphic descriptions of the production process for each of the types of sample completed. They show functional and output file paths. Please assume quality checks throughout the process and before the final delivery of the output files, not shown here to keep the graphics simple.

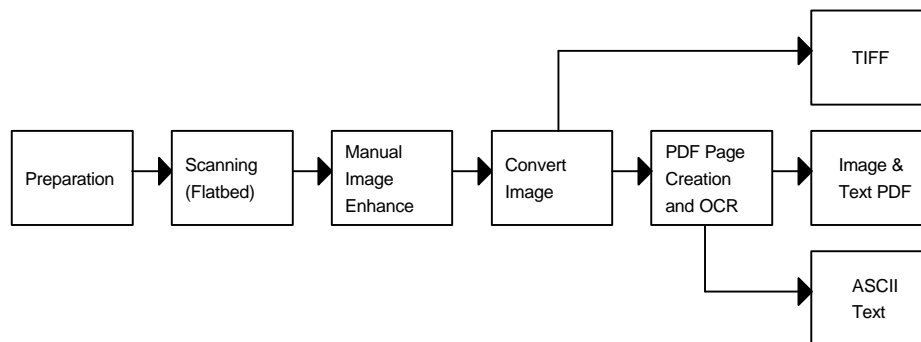
400 DPI Black and White Pages



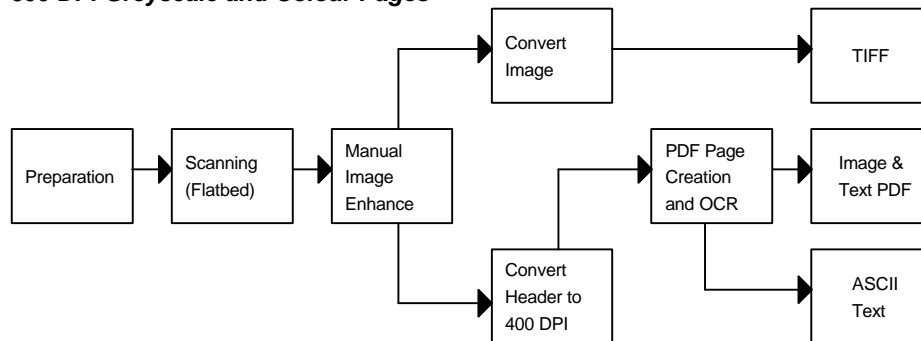
400 DPI Greyscale Pages



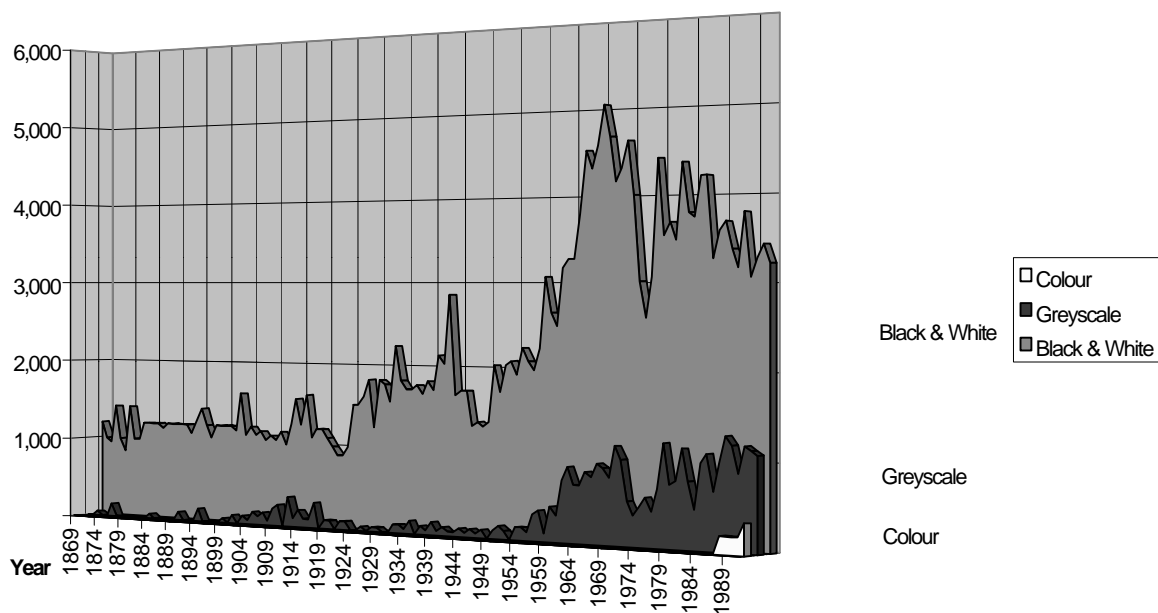
400 DPI Colour and 600 DPI Black & White Pages



600 DPI Greyscale and Colour Pages



Nature - Estimated Number of Page Format per Year



Appendix B