Nature - A Prototype Digital Archive

Ross MacIntyre[1], Simon Tanner[2].

[1]*Manchester Computing, University of Manchester, UK, email:r.macintyre@mcc.ac.uk*
[2] *Higher Education Digitisation Service (HEDS), University of Hertfordshire, UK, email:s.g.tanner@herts.ac.uk*

Abstract

In response to the proposal to digitise the journal *Nature* (1869->1992), published by Macmillan, a pilot project was commissioned to discover the technical issues and ascertain costs.

The initial conversion and digitisation elements are being provided by the Higher Education Digitisation Service, University of Hertfordshire, (HEDS) whilst Manchester Computing, University of Manchester, is providing all the data management and access elements.

This report details the processes so far undertaken, the results ascertained from these pilot processes and the techniques used. The pilot, though not complete, is now at the stage where certain conclusions can be drawn from the progress made so far. This is also the point at which further progress requires certain decisions regarding format and techniques to be focused and validated.

---

1 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 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 one of the most widely cited interdisciplinary science journals in the world today.

The establishment of a Higher Education archive would provide the following:

- A valuable research tool
- An aid to teaching - Macmillan advise that they receive many requests for access to papers from *Nature* for teaching purposes and making these available electronically would make such access more cost effective
- Provide students and researchers with a vast backstore of source material
- A unique view of scientific history and would therefore also be an aid to historians and sociologists of science.

2 Pilot Objectives

The *Nature* digitisation pilot project has the following objectives as expressed in the initial proposal:

- 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 and techniques required to meet the above objectives.

3 Method

The basic scan specification used was:

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

Each page to be scanned was assessed for its content and the specification above applied. For all the 400 DPI samples an “image and text” PDF plus a text file was created and delivered to Manchester Computing. The 600 DPI samples were processed to these formats to discover the process but not shipped to Manchester Computing. The process was measured, refined and measured again to gain production metrics.

Following discussions HEDS has taken several pages through the processes to create downsampled “image and text” PDF’s.

HEDS has also visited the Macmillans Publisher’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 pages types identified in the table above.

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 review. The TIFF images are also to be retained for archiving and used again if a different technology for viewing/searching is selected at a later date. Other approaches using page image rendition, e.g. JSTOR [1], 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.

4 Processes and Techniques

A number of processes and techniques were used to produce the samples, 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 is 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 in the 1970's and early 1980's (not fully assayed yet) where the paper of the publication is thin and there is significant show through.

As *Nature* has been presented in bound volumes for scanning, HEDS has utilised its Zeutschel Omniscan 3000 Bookscanner to gain the best possible output from the journal with minimum degradation to the originals. HEDS are constrained by the technology available in the Bookscanner market. HEDS 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. Currently, for the 600 DPI or colour scans, HEDS has had to flatbed the *Nature* samples provided, with the additional time premium and attendant degradation on the originals.

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, but not so easy to achieve with flatbed processes.

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

As there is some skew in many of the image files created, the need for post-processing the image increased. This includes deskewing the image, removing some dirt or speckling from the image, and output of the image file in the correct format for further processes. The despeckling or dirt removal has 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 with the text is not at a true 90 degree angle 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 are 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 deskew 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 independently test with confidence what the exact resolution is within the PDF file, only the means by which it was created. The Capture 2.x engine is slow at converting files into PDF format, sometimes as bad as 10 minutes machine time per page image converted. However, when compared with the quicker production times possible using Capture 1.x engine, the Capture 2.x engine is far more reliable, and the output requires lower levels of QA. Another issue with Capture 2.x is that the 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 which are 600 DPI greyscale or colour. Whilst Capture 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 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 4.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. As the image file sizes average between 20Mb and 30Mb this places a big overhead on the machine processing the image files and also slows the processing of colour and greyscale 600 DPI images relative to the other samples completed to date.

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 that the text appears blurred. This may possibly be due to interpolation. To achieve downsampled PDF files would require the Adobe Capture process to be repeated in its entirety for all PDF's, doubling the costs of this portion of the process.

4.4 ASCII Text Production

The ASCII text from each page is required to assist Manchester Computing create search resources to find individual pages from *Nature* and validate the results. This file is being created at the same point at which the "image and text" PDF is being 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 are a number of preparation functions that need to be completed before the originals can be converted. These include 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. Obviously, 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 are a number of points in the process where there are basic checks made to ensure the quality of the output. These are to ensure that every page has been scanned and that all the pages are in the correct order. There are further checks on the content of every TIFF file output to ensure that the content is representative of the original and that the correct file name has been assigned to the TIFF image. Similar checks to the file name and content are 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 due to the total volume of files to be controlled and checked against every page scanned.

5 Assay of *Nature* Collection at Macmillan

HEDS have done a survey of one issue of *Nature* per year of the publication from 1869 to 1992 to ascertain the proportions of greyscale and colour format pages in the whole collection. From such a survey it has been possible to gain an estimate of the total number of such pages in the collection and thus estimate 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 for the majority of the publication run, making them unsuitable for scanning due to the obscuring of the text.)

- 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 has been 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. Then by adding up the results, a figure of 298,600 pages is found which cross checks 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 website

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. It is intended to replace this with IP address checking should the prototype be more widely available. This has been loaded with the selected issues from *Nature*, provided by HEDS.

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.

6.3 Data Loading Approach

The application currently contains two versions of digitised *Nature*:

1) page-by-page viewing

2) article-by-article viewing

Both offer 'traditional' hierarchical browsing, i.e. Year->Issue->Table of Contents-> Header and/or Article.

The data as it arrives is loaded into the 'page-by-page' viewing interface, remembering that the unit of digitisation is one file per page, so nothing else is known about the contents at this stage.

As metadata is defined and received, and the PDFs are combined (see 6.5 below), the data is then accessible by the second viewing interface. This gets content on-line as soon as possible, following digitisation, removing the existence of metadata from the critical path. It is recognised

that the content at that stage is cumbersome to navigate, but is accessible via searching. The creation of the metadata is, however, vital for sensible browsing.

6. 4 Metadata

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

1st pass - automatic creation of page-level data

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

2nd pass - manual creation of (minimal) article-level metadata.

Article metadata files have been created by Manchester Computing for an initial, small set of issues. However, an external 'keying agency', Saztec Europe Ltd., has been appointed to create and validate further header records. Their prime objective is to identify the editorial contents of each issue at an acceptable level. This could be viewed as little more than the re-keying of the table of contents for each issue, though it should be noted that the data present in the Table of Contents is not actually sufficient. The data is again inserted into the PDF files, for consistency and to improve the presentation of search results.

3rd pass - on-going improvement of archive based on experience and feedback received.

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 have defined a minimal set of metadata for the archive and created an SGML DTD to reflect.

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 would need to be manually created and passed to Manchester Computing. Item 8, internal scheme, would be the appropriate one 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 not applicable or can be generated or derived by Manchester Computing.

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 are combined, via Acrobat Exchange, 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.

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

The combination has been done manually so far, but a command line batch process will be developed.

The full classification of *Nature*'s contents has not been undertaken. Classifications noted include: Letters to the Editor, News & Views, Book Reviews, On Our Bookshelf. Macmillan will help better classify over time.

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 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 applied to the software on the server, but all have worked successfully.

Software Scientific [7] have code that given a set of text documents, will generate a 'topic tree'. This can be used in conjunction with Verity to assist in searching. Effectively the search for 'like' terms is biased based on the meanings of 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.

6.7 Themed Access - 'Nature Trails'

The Electronic Publishing Research Group at Nottingham University[8] and the Multimedia Research Group at University of Southampton[9], have agreed to a formal collaboration in support of the 'themed' access required. Southampton's DLS software (available commercially from Multicosm[10]) will be used to establish links within the archive, as defined at the outset and also to support the subsequent definition of linkbases in support of teaching. The dynamic generation and insertion of links into the PDF files has already been demonstrated at Southampton using *Nature* test files. The technology was the subject of a recent paper presented at EP98[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 originals should be stripped and tested to gain production metrics.

Stripping enables the optimisation of the scanning phase, eases 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. An acquisition cost may be incurred, but this can be offset against the savings in the cost of processing.

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 drives the resolution requirements for *Nature* is 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.

[Aside: HEDS is using a Bookscanner with optical technology specifically for 400DPI or iterations of 400DPI. Using the HEDS equipment to gain 600DPI images would incur interpolation in the characters which would actually be less faithful to the original than a lower

resolution. Unless HEDS can strip the bound volumes of *Nature,* the best standard of image will be obtained by using a resolution of 400DPI, to optimise the Bookscanners optical characteristics.]

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. Therefore, the decision is whether the additional cost is warranted for the security of the next 10 years.

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. Therefore, whatever the resolution chosen for the black and white text pages of *Nature*, 400 DPI should be used for greyscale and colour type pages.

7.4 Further experimentation to reduce PDF file sizes is required.

The size of the PDF files can be onerous, e.g. 150Kb->3Mb per page, so avenues to reduce should 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, have ruled this out as an option. Nevertheless, perhaps certain PDF files could be OCR corrected. Candidates would be all articles included in a '*Nature* Trail'.

Experiments with downsampling 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.

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.oclc.org/metadata/dublin_core (last updated 2/11/97).

[5] TASI - http://www.tasi.ac.uk

[6] Verity - http://www.verity.com

[7] Software Scientific - http://ourworld.compuserve.com/homepages/swsci

[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] S.Probets, D.F.Brailsford, L.Carr and W.Hall, *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)

Acknowledgements

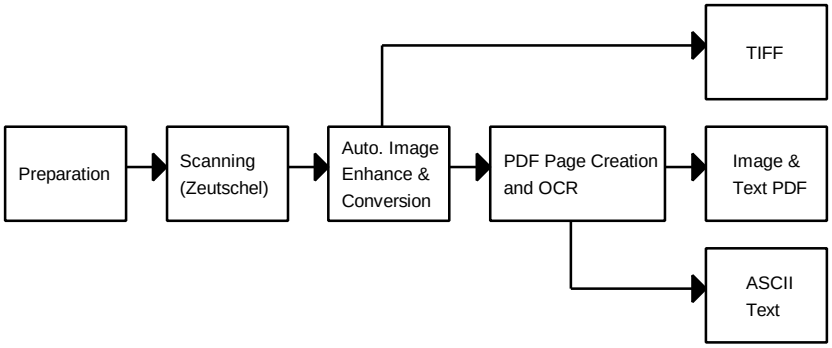
Prof.D.Pullinger, Macmillan Publishers Ltd, provided inspiration and guidance, and continues to do so.

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

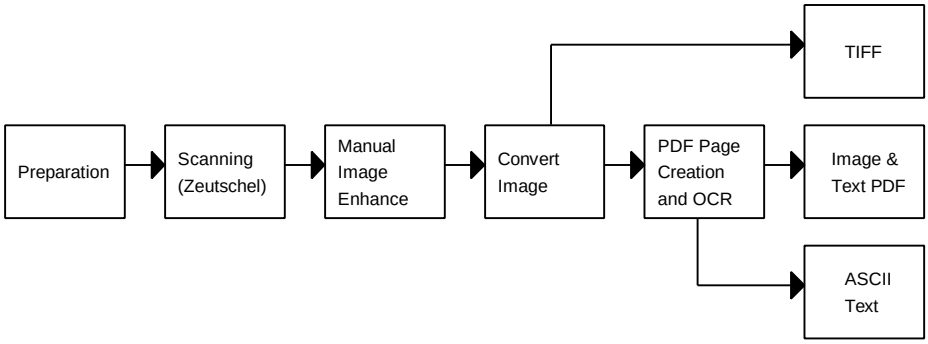
Appendix A Production Process Maps

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.

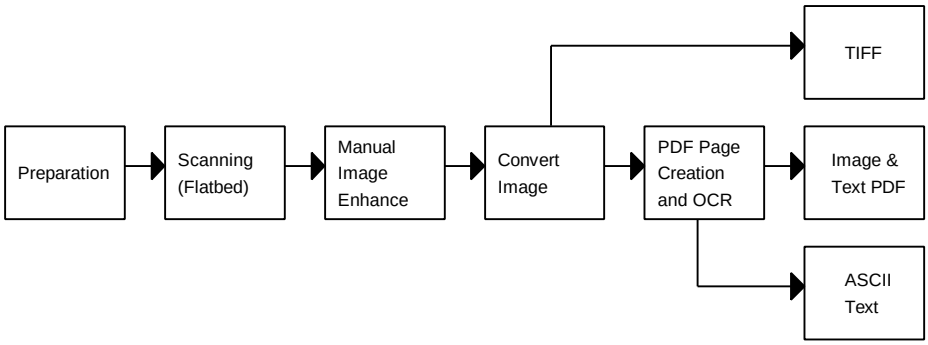
400 DPI Black and White Pages



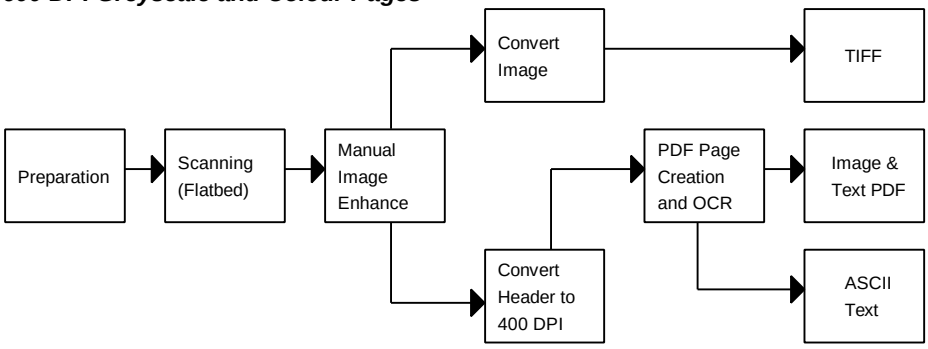
400 DPI Greyscale Pages



400 DPI Colour and 600 DPI Black & White Pages



600 DPI Greyscale and Colour Pages



Appendix B

Nature - Estimated Number of Page Format per Year