Library and Information Research
Volume 31 Number 98 2007

The CLA HE Trial Scanning Licence – how we're using it.

Jenny Delasalle

Abstract

The Copyright Licensing Agency (CLA) Higher Education (HE) Trial Scanning Licence has been implemented to varying degrees across UK HE institutions. The UK HE community has sought to share expertise in the technical, practical and compliance issues associated with the licence since its introduction. The University of Warwick Research and Innovation Unit (RIU) conducted a survey of scanning practice which revealed different approaches and levels of scanning in practice. This, taken with the CLA's own data about how the licence has been used, presents a picture of how UK HE Institutions are providing electronic extracts to their students. Issues such as who is using the licence, how much they are scanning, what the nature of that scanned content is and how they are creating scanned extracts, reveal much that can help those considering whether and how to use the CLA HE Trial Scanning Licence.

Introduction

The CLA Higher Education (HE) Photocopying and Trial Scanning Licence (CLA, 2006) was introduced in 2005, effective from 1 August. This licence built on the previous photocopying licence and was negotiated by Universities UK and GuildHE, with the CLA acting as an agent for the Publishers Licensing Society (representing publishers), the Authors Lending and Collecting Society (representing the authors) and the Design and Artists Copyright Society Limited (representing visual artists), and has bilateral agreements with more than 20 foreign reproduction rights organisations (e.g. Copyright Clearance Center in the United States). The licence is available for UK HE institutions to subscribe to for a fee based on the number of students at the institution.

Although the licence covers both photocopying and scanning practices, the study described here is concerned primarily with the scanning activities covered by the licence. In short, the part of the licence relating to scanning allows subscribing UK HE institutions to scan extracts of books and journal articles from print, and make them available to students on a particular course, without having to approach publishers directly in each and every instance. The extracts must be no more than five percent or one chapter or article, and there are various other restrictions to the coverage of the licence, not least that the material must have

Author

Jenny Delasalle is a Service Innovation Officer at the University of Warwick Library Research and Innovation Unit and has a background in subject support and project work.

Email: J.Delasalle@warwick.ac.uk

Received 27 July 2007

Accepted 02 August 2007

been published in the UK. While the repertoire available for photocopying consists of material published in the UK, USA and more than 20 other mandating territories, scanning is limited to books, journals and magazines published in the UK. Other key conditions are: the publisher or the work itself must not be excluded from the licence (Penguin is the only UK publisher to have opted out from scanning altogether); the print original must be owned by the institution (although this can be as a fee paid copy from the British Library); and that the scanned extracts must only be available to University members, displayed with cover-sheet copyright statements.

The full licence covers the making of multiple photocopies and / or the preparation and distribution of digital copies scanned from extracts of printed books, journals and magazines. It is a requirement of the licence that HE institutions must report what has been scanned for each course, in order that the licence fee money is distributed to the rights holders in due proportion. Rights holders' interests are protected through an audit process, which is carried out by the CLA. As a new licence in a still relatively new digital arena, there have been many lessons for the UK HE community in making the most of the licence. The CLA user guidelines and documentation have also been developed further since the introduction of the licence, and they explain the requirements in full detail (CLA, 2007).

The two years in which the licence has been effective so far have been an opportunity for us all to explore the procedures required by the licence, the technologies available to us and the needs of our staff and students. Some HE institutions were early adopters with experience of scanning before the licence came into effect, whilst others took the opportunity that the licence afforded to explore the possibilities of providing scanned extracts to their students.

In January 2007 a survey was carried out by the University of Warwick Library's Research and Innovation Unit (RIU) into scanning practice at UK institutions. The purpose of the survey was to learn about how others had interpreted the licence and overcome the procedural hurdles involved, in order to share the lessons learnt through the LIS-Copyseek Jiscmail list (LIS-Copyseek, 2007) and at the Scanning Practice event that was hosted by the University of Warwick on 20th March 2007.

Thirty five survey responses were received, which revealed a mixed picture up to that date. This article details the results of this survey, together with the information collated by the CLA themselves into how the licence has been used, to present a picture of scanning practice at UK HE institutions.

About the survey

The RIU survey was posted onto the University of Warwick website, and advertised through LIS-Copyseek (Delasalle, 2007) and LIS-Link (Chiner Arias, 2007), as well as to known contacts at particular institutions.

The only compulsory question was the one which asked whether the respondents' institution had already subscribed to the licence or not. The survey was also carried out in such a way that results were anonymous, in order to encourage people to be open about their practice.

What we wanted to find out

Some of the key issues we were investigating were to do with interpretation of the licence itself, whilst others related to practical and technological issues about scanning generally. The first section of the survey was therefore relevant to scanning in a general sense and the second, larger section was about using the scanning licence specifically.

One issue that we wanted to investigate was how many extracts could be digitised in return for how much investment in terms of staff time, equipment costs and/or outsourcing fees. Accordingly, we asked about how many extracts had been produced over a year, what equipment had been used and we asked about staffing levels involved in scanning work. There is a dilemma for any scanning operation about the balance of quality against file size, so we wanted to explore how importantly other Universities rated Optical Character Recognition.

In terms of administering the licence, we were interested in the level of control in the selection of what was to be scanned, and in what way the thorough checking requirements of the licence were carried out. We also wanted to explore if there was any possibility for the UK HE community to share the burden of scanning.

Who has a CLA Licence?

The survey began by asking whether or not the respondents used the CLA HE Trial Scanning Licence.

Of our 35 respondents to the survey, 32 used the licence, 14 of whom had used it since the beginning in August 2005. CLA data show that as of 31 Jan 2007, 146 institutions had taken up the CLA HE Trial Scanning Licence.

One important feature of the scanning licence is that it only covers UK published content[1]. There are also two excluded works lists, and the list of US publishers associated with the photocopying aspect of the licence defines those publishers as categorically not UK publishers, therefore their content is not covered by the scanning aspect of the licence. This means that there is a considerable body of material that is not eligible for scanning under the CLA HE Trial Scanning Licence. We asked survey respondents to indicate whether or not they carried out scanning outwith the licence, either through Heron[2] or by contacting the publishers directly. Nine people were using Heron to get copyright clearance and eight were contacting publishers directly, with only two of those respondents doing both.

The scale of handling requests beyond the scope of the CLA Trial Scanning Licence varied from “none yet” to “200-300”. Those handling larger numbers of requests external to the licence were also generally handling larger numbers of requests under the licence. The library that reported “200-300” explained that this was because they had “a lot of requests for US publishers’ extracts” and reported a similar number of extracts being scanned under the licence.

[1] The CLA User Guidelines (CLA, 2007) clarify how UK content is identified.

[2] Heron is a copyright clearance service (Heron, 2007a)

How much is being scanned?

The survey asked "How many items do you scan under the licence, per academic year?" The answers revealed a disparity between the respondents. The highest number reported was "around 900" and the lowest was "none so far". Some respondents were trialling or piloting the introduction of the licence and so had deliberately kept numbers low. Two answers were about what they were aiming to achieve: one was aiming for 100, the other for 2000 a year until their photocopy collection had been converted. Five respondents reported a rising trend in numbers being scanned, either by stating this or by providing figures for the previous year as well.

Some of the scanning taking place at these HE institutions took place in house, whilst some was outsourced. With outsourcing, one respondent pointed out, the cost is covered from a finite sum so you can predict what you expect to achieve. Whereas if you use existing staff resources you can potentially achieve more as they might have capacity at different times of year that can be used to increase your scanning volume. Although we did ask about which staff were involved in scanning that was carried out in-house, and how much time they spent, the survey revealed too complex a picture to extrapolate any average number of extracts likely to be provided by a certain number of full time equivalent staff.

The largest number of extracts reportedly achieved up to the survey date within the academic year with the smallest number of staff reported was 903 extracts, with:

Scanning done by computer help desk assistants c.1.2 fte (when not dealing with computing problems) - grade 1/2 - varies depending on time of year currently around 20%. Administration done by …Librarian - grade 6 - around 40%.

It may or may not be significant that the scanner used in this instance is the Plustek Opticbook 3600. (Scanning equipment is discussed in the section below, entitled "Scanners used"). The next most efficient appearing results were from those who outsourced all or part of their scanning.

The CLA data up to January 2007 show that 273,063 pages had been scanned in total, to that point, which indicates that the licence is indeed being used on a large scale at some institutions.

A dedicated unit for scanning?

Nine library services declared that they run their own digitisation unit. Five of these also reported that they outsource their scanning, which is more than half of the total number who reported outsourcing their scanning (also a total of nine). Only one respondent had a digitisation unit external to the library, and the remaining 17 did not claim to have a central digitisation unit.

In fact, when people described their units or the staff who were involved in digitisation, the most common model was of staff from other areas of the library taking on digitisation work, whether or not the respondent to the survey classed them as a digitisation unit. Typically, short loan, inter library loan, photocopying or audio visual services staff are also involved in scanning work, with some

involvement of a professional librarian either from within those teams, a specialist role, or from a subject team background.

Scanners used

One significant factor in how quickly extracts can be scanned is of course, the equipment that is used. The way the equipment works and its cost is a significant factor for any digitisation unit to consider.

Those who didn't have a central digitisation unit answered that various scanners were used across campus. Others used multi-function-printers that handled printing, photocopying and scanning all in one.

Some reportedly used sheet feed scanners, presumably scanning from photocopies, whilst others used ordinary flatbed scanners. The Plustek Opticbook 3600 was a popular flatbed model, used by 5 respondents. The Opticbook was recommended as being affordable as well as adapted to scanning books because you can put books on right up to the edge of the glass.

One respondent used expensive, specialist large book scanners, the Zeutschel models (Zeutschel, 2007), that capture the open pages image and correct for page curvature.

To OCR or not to OCR?

Optical Character Recognition (OCR) is a way of capturing text as a text file, rather than as an image. Scanning is much like taking a digital photograph of a page, and indeed in some scanning operations a digital camera is the equipment actually used. Once you have an image of the page you can either leave it as just that, an image, or you can run some software to go through an OCR process, in order to make the file something that you can copy and paste from, add text to and delete from, search for words in and generally handle in the same way as you would any other text file.

The dilemma with Optical Character Recognition (OCR) is that it takes time and effort to correct the machine-read text: it's clever but not completely accurate. Under the Special Educational Needs and Disability Act, 2001, HE institutions are obliged to make their provision accessible in order not to place disabled students at a substantial disadvantage in comparison with students who are not disabled. Also, text files are often smaller in size than scanned image files. There is a balance to be found with the time it takes to produce OCR text files instead of image files, though, which might disadvantage all students as far fewer scanned articles would be produced in total if all were OCR-ed, so we wanted to find out where other institutions drew the line.

Only six institutions replied that they did use OCR software on their scans, whilst 17 did not use OCR. We did not ask for reasons for this decision, but one respondent did elaborate:

Sometimes we have in the past as I was worried about file sizes… However this is very time consuming and when we chatted to other Libraries [they] were just producing optimised searchable image scans. Now we offer OCRed scans to units where there are visually impaired users or other users who require to read the

text at a high magnification on screen - the OCRed text is less blurry. So we can do it, but are sticking with the much quicker image scans.

Optimised searchable image scans are enabled in Acrobat 8 Professional (Baker, 2007) and keep the formatted page exactly as the image, with searchable text in a hidden layer – which means that the scanned extract handles slightly more like a text file.

Software reportedly used by those who do OCR includes Adobe Capture, Adobe Professional and Abby fine reader. Some respondents also recommended using PdfCompressor compression software to reduce the size of the files and make them faster to download.

Designated persons

The CLA HE Trial Scanning Licence requires HE institutions to nominate "designated persons" internally, responsible for authorising the distribution of a digital copy. These need not be named individuals, but rather posts or categories of staff at the institution. The survey asked who the designated persons were at the respondents' institutions.

The largest group was nine who answered that library staff (at all levels) are designated persons. Six others answered that only senior library staff (three of whom included senior library assistants in that category) were designated persons. Five respondents named a person with a particular role as the designated person, whilst four answered that all academic grade staff were designated persons. Therefore a greater proportion (20 out of the 24 who answered this question) of the institutions using the licence were trying to retain a high level of control over the approval of scanning under the licence, and it is apparent that the majority were involving the library in this control.

Who accesses extracts?

The terms and conditions of the CLA HE Trial Scanning Licence restrict printing and saving of anything that is scanned to members of the course for which an extract was scanned. However, any member of the institution may view a scanned extract, which is particularly useful for students who are choosing a course to study. There is no easy technical way to allow viewing to all whilst restricting printing and saving to course members. The two main ways to comply with the licence are: a) to restrict access to only those registered on the course for which an extract was scanned, or b) to allow access for all, but warn those reading the extract that only course members can print and save it. We wanted to find out how other institutions were handling this particular dilemma.

Figure 1 below shows that, of the 25 responses to our question on restricting access, more people used the copyright notice to warn readers of the restrictions (20) than used the technical restrictions (14), although four of these used both methods. Two others ticked the option that they inform their users through other means. There appeared to be no consensus on the way to interpret this feature of the licence.

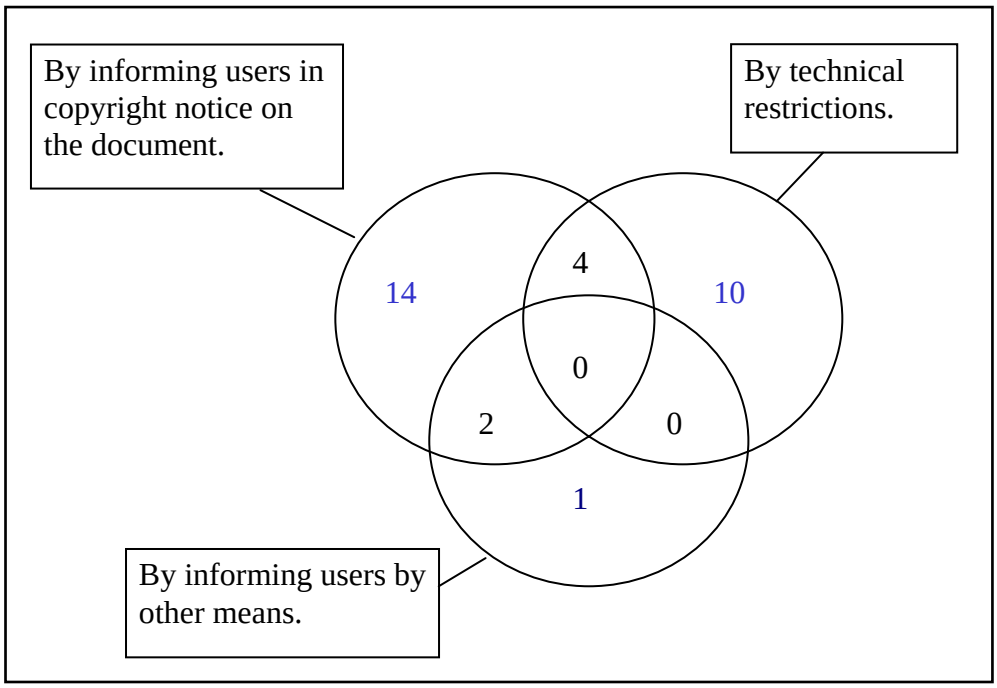

Figure 1: Do you limit the ability to print and save a document by technical restrictions or by informing your users?

At the University of Warwick we currently restrict access to students on the course *and* use the copyright notice cover sheet. We decided not to leave access open to the whole University membership because we do not have particularly good ways of measuring the number of accesses of our web pages in order to demonstrate, in the event of an audit by the CLA, that only the same number of people registered on the course or fewer are accessing our extracts.

We used the survey to ask about the statistics that other institutions gathered on the use of their scanned extracts. The answers demonstrated that most (17) did not record statistics, although some of those intended to eventually as their implementation was at an early stage at the time of the survey. Amongst the ten who did record statistics, reported uses of the access statistics included calculating the cost per student, weeding the collection, reporting to course coordinators either automatically or upon request, and as part of general reported library statistics.

CLA data give us an insight into the number of students benefiting from scanned extracts, across the whole country: in the period 1st August 2006 – 31st January 2007, extracts were provided under the licence for 2,452 courses, on which there were 189,159 registered students, as of January 2007.

Record keeping

Compiling the report that is required by the CLA every 6 months is a significant part of implementing the licence, and internal record keeping procedures are an additional burden. The CLA report cannot accommodate any internal procedures, nor can it track items not covered under the CLA Trial Scanning Licence. The CLA report records what has been scanned and made available in the previous 6 months, so it is of little use as a way of tracking back-up copies of items

previously scanned but not currently made available. We used the survey to ask how other institutions handled the reporting requirements of the scanning licence in conjunction with their own administrative requirements for the licence.

Six respondents used Heron's specialist PackTracker software. (Heron, 2007b) This is a significant investment, however, in terms of both money and expertise in learning to use the package and all of these respondents were amongst those who reported having a central digitisation unit. Three other respondents reported that they used "Heron" without specifying that they use PackTracker.

Three respondents reported that they use MS Excel, two that they use MS Access, and one mentioned MS Office generally. Other tracking and record keeping methods reported include paper print-outs in folders, suppressed Talis catalogue records that create a CSV file to import into the CLA spreadsheet, an existing ILL system, just the CLA reports and extracts themselves on the VLE, and one "we haven't worked it out yet"!

What is scanned?

The survey asked: "What proportion of the total number of items you have scanned are from books, and how many are from journals?"

We asked this question really because we had expected that the generally higher level of electronic provision of journals supplied by publishers would incline people towards scanning more book extracts, but wondered if this was in fact the case.

What we found overall backed up this impression, although for four institutions the balance is the other way, with more journal articles being scanned than book extracts. The mean average percentages, in spite of the four large balances in the opposite direction, were:

Books: 67%

Journals: 33%

This would seem to support our original expectation.

CLA data are harder to interpret on this matter: for the period, 1st August 2006 – 31st January 2007, material had been scanned from 1018 ISSNs and 7898 ISBNs, but this does not indicate for either type of standard number whether more than one extract has been scanned from an individual number, and this is perhaps more likely for serials than for books due to the larger amount of content associated with one ISSN. However, CLA data do indicate the age of the printed material that is being scanned under the licence for the same period: where the date of publication is known, for serials (or journals), 65% of the extracts are from the pre 2000 era, and for books pre-2000 extracts make up 56% of the total. This is an interesting trend to note, particularly when taken with the statistic that 66% of the extracts reported in that 06/07 period were from out of print materials.

CLA data are also very enlightening as to what subjects the scanned extracts relate to. Figures 2 and 3, below give a very detailed picture, which might help those who are piloting scanning or beginning to market a new service to decide where to target their efforts. It is particularly interesting to note the difference between books and serials in the field of Medicine, which is 19% of the total for

serials, but only 7% of the total for books. Social sciences show a marked change the other way, from 11% of the total for serials to 21% of the total for books.

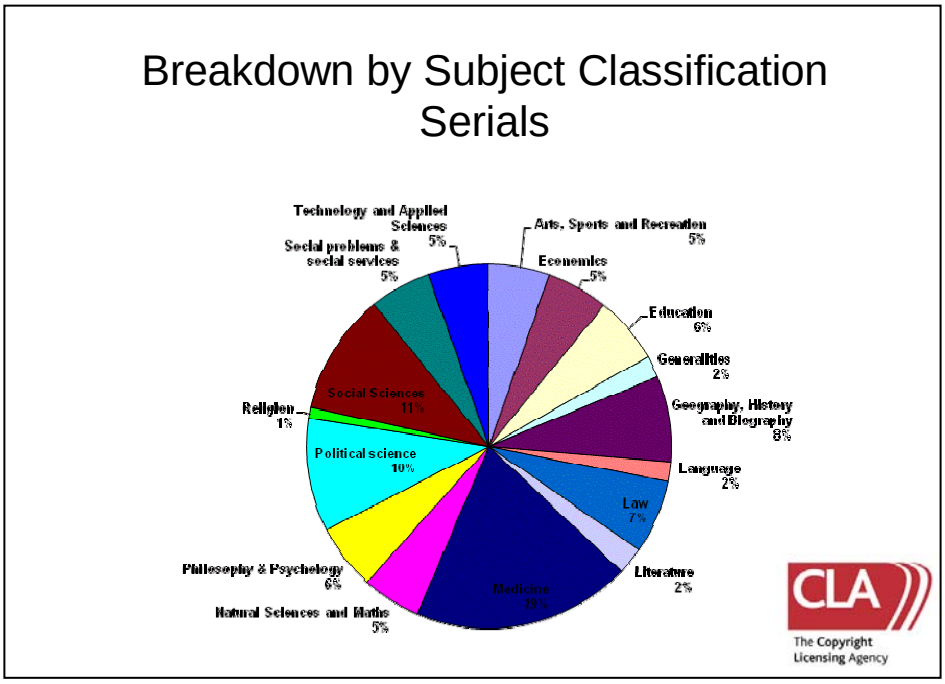

Figure 2. Breakdown of scanned extracts by Subject Classification, for Serials. (Source: CLA, used with permission)

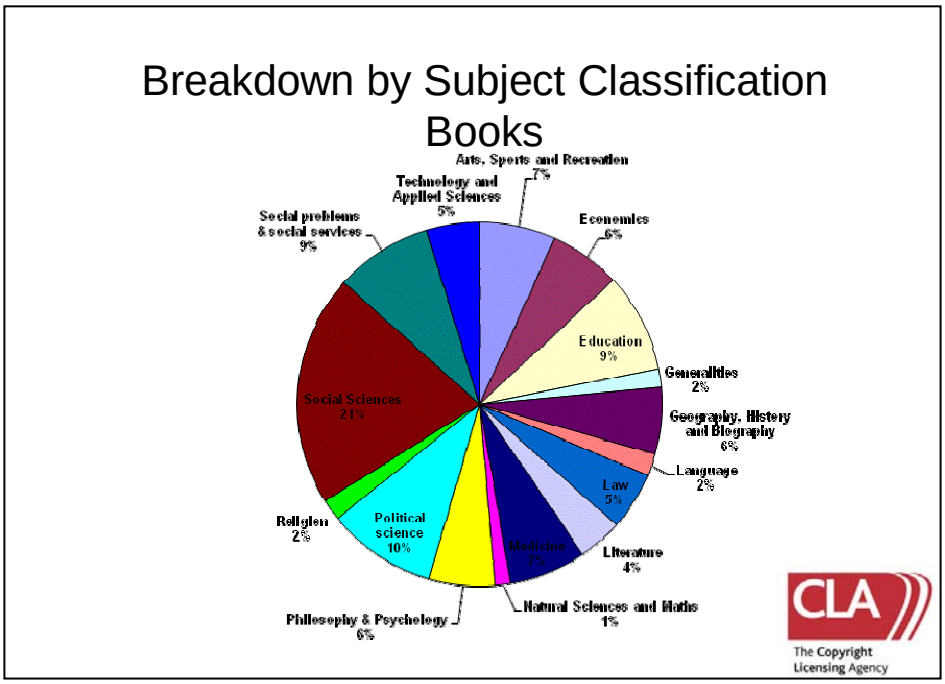

Figure 3. Breakdown of scanned extracts by Subject Classification, for Books. (Source: CLA, used with permission)

How do you prioritise what to scan?

We asked "How do you identify which course(s) to scan for?" and the responses that we got were largely about selection criteria, but many respondents also included marketing activities in their answers. Most used contact with academics through Subject Librarians and Library web pages as marketing tools, whilst one respondent also reported contacting Heads of Department.

The largest category of answers was 15 responses where the scanning was based on lecturers' requests, with some including using additional selection methods in addition to this. Five answers described targeting courses and departments with high book use and/or large student numbers. Three respondents mentioned using their short loan collection as the basis of their scanning work, and other factors taken into consideration were distance learning, Subject Librarians' priorities, the advice of a learning and teaching unit, only courses on the VLE, using reading lists and working with those academics who were willing to take on some of the administration themselves.

Only four institutions reported a limit set on the number of items they would scan, and mostly they were concerned with cost, although one limited to 20 articles per module.

Checking compliance

Another point that we wanted to explore was that the scanning licence actually allows institutions to scan extracts from publications that are already available electronically, as long as you report your reason for doing so. So, how do you know that an item is already available electronically when you are asked to scan from it? We asked how licence users checked for existing digital versions, and the list below includes all the places that were mentioned, although no one institution checked in all these places!

- Archive of documents already scanned
- Own catalogue
- COPAC
- SUNCAT
- Publisher's web-site
- Heronweb
- NetLibrary & Dawson e-books (main e-book suppliers)
- "Internet search" – Google & Google Scholar
- "Metalib search" / "List of sources"
- Amazon

Two respondents reported that they knew that their Subject Librarians had previously checked whether the library could purchase an electronic copy.

We also asked how the institutions checked for the most recent edition, and the list above was duplicated. One institution took the lecturer's request literally and

scanned that edition, and another only checked the edition if it was apparent that there might be a new edition.

Sharing extracts

The scanning licence has provision for HE institutions to share their scanned extracts with each other, so the survey asked whether respondents would be willing to share extracts, and if so what their ideas were as to how they might be shared. Eleven respondents were willing to share, whilst sixteen were not, and there were numerous ideas about how it could be done, through Heron or the CLA or by direct request to institutions. There was no clear way forward on this, especially as one institution pointed out that they clean and OCR their scans, and would expect the same quality from other institutions, which is a considerable hurdle as only six of the respondents to this survey used OCR, as discussed earlier.

Conclusion

The survey revealed that UK HE institutions were growing their scanning operations, using the licence as a pilot project, or beginning to grow numbers of scanned extracts. Generally, those with larger collections had outsourced at least part of their scanning. Most survey respondents were using staff from other areas of the library service provision to carry out scanning, rather than setting up dedicated units, and only one had invested in expensive scanning equipment.

The results of the survey were presented at the Scanning Practice event held at Warwick in March, and the Powerpoint slides were published online and advertised to the Jiscmail LIS-Copyseek list. They were a discussion point amongst those at the event who had not yet set up scanning operations using the licence, and indeed they helped to inform decision making at the University of Warwick library.

The Scanning Practice event also included discussion of the British Library's proposal to provide scanned extracts to libraries that they can use under the CLA scanning licence. Such a service might help libraries to incorporate outsourcing into their scanning model as most already have a relationship with the British Library. By outsourcing, libraries might be able to provide more scanned extracts to their students, as the survey revealed that it was generally the outsourcing libraries that had more scanned extracts. CLA data for 1 Aug – 31 Jan 2007 also indicate that 20% of serials scanning is from British Library fee paid copies which is a not insignificant amount, although only 2% of extracts scanned from books in that period came from the British Library.

The survey questions reflected concerns prevalent amongst those beginning to use the licence at the end of 2006 and beginning of 2007, and there are other concerns that could be covered in any future surveys. The matter of how the licence can be improved is already being considered by the CLA and Universities UK (UUK) as they enter negotiations for a new version of the licence to be launched in 2008. Matters such as extending the licence to include copying of born digital content, arranging for the licence to cover content from other countries, allowing cataloguing of scanned extracts and streamlining reporting requirements are all concerns that have been expressed amongst the UK HE community. The time is

now right to consider and communicate these matters to the rights holders through the CLA and UUK.

Apart from new outsourcing opportunities and changing features of the licence itself, CLA HE Trial Scanning Licence users would benefit from a definitive list of e-books. Something like the comprehensive books in print databases that we currently use, but for e-books, would be a very useful source. There is a reporting requirement in the licence to check if something is already available electronically and report why we are scanning it anyway. With so many potential places to look, and ever increasing amounts of e-books being published, there might be occasions when we might miss the fact that a book is available to buy electronically.

The Scanning Practice event, and an earlier CLA sponsored event held in February were both occasions when practitioners were able to share their experiences of implementing the photocopying and trial scanning licence, including discussing their experiences of a CLA audit. At both of these events, the CLA also shared information about how the licence was used from their own data, as reported to them by HE institutions: in person at their own event, and with kind permission at our March event. It is very interesting for us to be able to see what other Universities are doing as well as exploring how they are doing it through surveys like the one written up here. For those institutions just starting to implement the licence, we are able to gauge what might be possible for us to achieve, and what kinds of materials we might expect to be scanning, which helps us with our planning. Through sharing information we are able to support each other to make the most of this potentially very useful licence. A further study on how the scanned extracts are being used by students would no doubt be a valuable follow on from this investigation.

References

Baker, D. L. (2007) *The skinny on scanning and touchups* [online], Adobe Systems Incorporated. URL: http://www.acrobatusers.com/tutorials/2007/skinny_on_scanning_touchup/ [3 August 2007].

CLA (2006) *Photocopying and Trial Scanning Licence* [online], London, The Copyright Licensing Agency Ltd. URL: http://www.cla.co.uk/support/he/HE_TrialPhotocopyingandScanningLicence.pdf [26 July 2007].

CLA (2007) *Higher Education Support Material* [online], London, The Copyright Licensing Agency Ltd. URL: http://www.cla.co.uk/support/he/index.html [3 August 2007].

Chiner Arias, A. 5 Jan 2007. Scanning Practice Survey & Event, *LIS-Link* [online]. URL: http://www.jiscmail.ac.uk [26 July 2007].

Delasalle, J. 15 Feb 2007. Re: trial scanning licence event, 20th March, *LIS-Copyseek* [online]. URL: http://www.jiscmail.ac.uk [26 July 2007].

Heron (2007a) *HERON: Home* [online], Oxford, Ingenta. URL: http://www.heron.ingenta.com/ [26 July 2007].

Heron (2007b) *HERON: PackTracker* [online], Oxford, Ingenta. URL: http://www.heron.ingenta.com/about/about_packtracker.html [3 August 2007].

Special Educational Needs and Disability Act, 2001, London, HMSO. URL: http://www.opsi.gov.uk/acts/acts2001/20010010.htm [3 August 2007].

Zeutschel (2007) *The future of the past* [online], Tübingen-Hirschau: Zeutschel GmbH. URL: http://www.zeutschel.com/produkte/os5000-s.html [3 August 2007].

Acknowledgements

Many thanks to the Copyright Licensing Agency for supplying their data on how the licence is being used.