AUSTRALIA'S NATIONAL RESEARCH COLLECTION: OVERLAP, UNIQUENESS, AND DISTRIBUTION

Paul Genoni and Janette Wright

This paper reports on the results of an overlap study of Australian research library collections. The study used OCLC's WorldCat Collection Analysis software to mine data recording Australian holdings on the WorldCat database. The data is analysed according to the results obtained for six 'groups' which represent various coalitions of academic, national/state, and special libraries. The data is focussed on the incidence and distribution of overlap and unique items, and the analysis considers the implications for the future management of the nation's research collections.

Paul Genoni, School of Media, Culture and Creative Arts Curtin University of Technology, WA 6845
Email: p.genoni@curtin.edu.au

Janette Wright, CAVAL Ltd, Bundoora, Vic 3083
Email: janette.wright@caval.edu.au

This paper presents the most recent findings of an ongoing study investigating the overlap, uniqueness, and distribution of print collections held in Australian research libraries. The goal of the study is to better understand the distribution of legacy print material amongst Australian research libraries in order to assess the potential for future collaboration in aspects of collection development and collection management. This includes collection development (selection and acquisitions), storage, disposal, and last copy retention. There is a particular focus on storage, and on the potential for creating forms of federated storage that reduce costs for the network of Australian research libraries and produce efficiencies in the discovery and delivery of legacy print items (Genoni, 2007; Genoni, 2008).

An important research method in achieving the goals of the project has been the use of studies of overlap and uniqueness amongst Australia's research collections. Results have been produced (Genoni & Varga, 2009) that supplement those obtained from previous similar studies (Australian Research Libraries Collection Analysis Project, 2004; Missingham & Walls, 2003). Each of these previous

overlap studies have identified both the degree of duplication between collections and the incidence and distribution of uniquely held items for sections of the Australian research library community. While generating useful data, these studies have, however, been restricted in scope and utility by the limited functionality of the bibliographic services and software at their disposal. The primary source of data used has been the National Bibliographic Database (NBD), in the form of Libraries Australia and its predecessors. This service has not been well equipped with either the software or staff support necessary for deep mining of overlap data. As a result, previous Australian overlap studies have been limited to a subset of the university libraries and the National Library, and had limited flexibility in manipulating data.

In recent years, however, an alternative source of data for overlap studies has emerged, coupled with software designed specifically for the task of mining collections-based data. The data source is the OCLC WorldCat, now established as the foremost international union catalogue. As of July 2011, OCLC claim the database consists of over 236.7 million bibliographic records with 1.74 billion holdings provided by over 72,000 libraries (OCLC, 2011).

The potential to use this rich source of data for collaborative collection management has long been recognised, with Lavoie, Dempsey, and Connaway arguing that with the assistance of WorldCat,

> . . . data mining across library collections could open the door to new opportunities for shared collection management. Studies of holdings patterns for institutional clusters at the consortium, regional, or even national level could reveal opportunities to reduce cross-collection redundancies and free up resources to fill gaps in collections (Lavoie, Dempsey & Connaway, 2006: np)

OCLC support WorldCat users with a range of services including 'resource evaluation, comparison and planning' based on the use of the *WorldCat Collection Analysis* software to assess and compare collections (OCLC). Some of the previous uses of WorldCat data have investigated its use in research related to the current study, including for the support of decision making with regard to stored material (Ward & Aagard, 2008); the identification and management of last copies (Connaway, O'Neill & Prabha, 2006); and the profiling of national collections (Dickey, 2011).

METHODOLOGY

One of the challenges faced by overlap studies is that they are only as accurate as the catalogue data they mine (Rochester, 1987). In July 2007, the National Library of Australia entered into an agreement with OCLC that covered all Libraries Australia subscribers. Under this agreement, records in the NBD with attached holdings are uploaded to WorldCat; and WorldCat records with Australian holdings are in turn uploaded to the NBD (National Library of Australia, 2007). The effect is that WorldCat now replicates the holdings of Libraries Australia, and therefore provides data that can be used in research investigating the collections and holdings of Australian research libraries.

It should also be noted that there are reasons why the holdings data on WorldCat may not be a completely accurate reflection of the current records of items and holdings in Libraries Australia. In some cases, local records may be inaccurate in a fashion that prevents matching of the same item resulting in duplicate records; or contributing libraries might have different cataloguing practices that prevent similar items from being identified. It is relevant in this context to appreciate that overlap studies such as that reported here rely on a comparison of the metadata rather than the actual content of an item. There are a number of reasons why metadata may conceal differences between ostensibly similar items, or why items that are for many uses regarded as identical may generate different metadata. The distinction between an 'expression' (the content) and a 'manifestation' (the package) can therefore be lost both in the union catalogue and of course from any research using that catalogue as a data source (Connoway, O'Neill & Prabha, 2006).

There will also be inaccuracies in local catalogues that are subsequently reflected in both Libraries Australia and WorldCat. This may include, for example, libraries not uploading catalogue changes to the NBD when items are deleted from collections, thereby potentially resulting in an overestimation of the level of duplication. There is, however, reason to believe that catalogues are now more accurate than has previously been the case, as many libraries continue the work of retrospective cataloguing or the conversion of catalogue records to digital formats.

A small-scale exploratory study was undertaken in 2010 in order to assess the value of the *WorldCat Collection Analysis* software in assessing the overlap and unique holdings of Australian research libraries (Genoni & Wright, 2010). This exploratory study was limited to four collections (Monash University; Melbourne University; National Library of Australia, and the CARM Centre). On the basis of the results obtained, it was determined that using WorldCat data and the *WorldCat Collection Analysi*s software are currently the best options for conducting a more detailed study of Australian research library holdings than has previously been possible. While this present research undertakes this more detailed examination, and in doing so uses data and software that were not available for previous overlap studies, the methodology should still be regarded as exploratory. At the time of conducting the study, no Australian library or consortium had a subscription to *WorldCat Collection Analysis*, and the research therefore relied upon OCLC to supply the software access and staff support in order to conduct the data collection. This resulted in some limitations on the ability of the researchers to collect and manipulate data.

One of the important features of the *WorldCat Collection Analysis* software used in the current research is that it allows the creation of 'groups' of libraries that can be used for comparative purposes. That is, selected libraries can be grouped together in such a way that their holdings can be evaluated as a single collection. The creation of various groups can therefore allow assessment of, and comparisons between, the collections they collectively represent.

For this purpose, groups were created to represent different categories of research libraries. Four of the groups that were created represent existing coalitions of libraries.

- Group of Eight libraries (G8). Representing the libraries of Australia's eight leading research intensive universities. Consisting of the Universities of Adelaide; Melbourne; New South Wales; Queensland; Sydney, and Western Australia; the Australian National University; and Monash University. This group had records for 5,147,408 items on WorldCat.
- Australian Technology Network (ATN). Representing the libraries of five universities grounded in technology-based education. Consisting of Curtin University; Queensland University of Technology; RMIT University; University of South Australia; University of Technology Sydney. This group had records for 979,505 items on WorldCat.
- Innovative Research Universities Australia (IRU). Representing 'a network of seven comprehensive universities conducting research of national and international standing'. Consisting of Charles Darwin University; Flinders University; Murdoch University; Griffith University; James Cook University; La Trobe University; University of Newcastle. This group had records for 2,147,408 items on WorldCat.
- National and State Libraries Australasia (NSLA). Nine collections representing the national, state and territory libraries of Australia. Consisting of ACT Library and Information Services; National Library of Australia; Northern Territory Library, State Library of New South Wales; State Library of Queensland; State Library of South Australia; State Library of Tasmania; State Library of Victoria; State Library of Western Australia. (The National Library of New Zealand is a member of the NSLA consortium but was *not* included in the group created for this study.) This group had records for 5,201,563 items on WorldCat.

While the libraries represented in each of these groups differ in number and the size and focus of their collections, the fact that they were established groupings with some fundamental basis for collaboration reflected in their collections meant that they were natural 'groups' to use for this study. Together they represent Australia's major research libraries, including the top twenty of the most research active of the nation's 40 universities.

In addition, it was decided to add two other groups representing important Australian research library collections. First, a group to represent the collections of special libraries. This sector has been excluded from previous Australian overlap studies, and a case has recently been made that their collections and interests have been underrepresented in discussions about the management of the nation's research content (Stephens, 2009). Unlike the education-based and NSLA sectors, the special libraries have no existing coalition or consortium that represented an adequate cross section of this sector, and a group was therefore created for this purpose. The principal criterion for inclusion was that a library

should have a minimum of 20,000 holdings recorded in WorldCat as of June 2010, using statistics data from the Libraries Australia website (Libraries Australia). The intention in including larger collections was to select those with a greater research capacity, although it is acknowledged that in the context of special libraries—where small collections can be highly focussed—this assumption may not be true in all cases. This criterion was met by 38 libraries. This number was reduced to 23, in order to make the group a more manageable size and to reduce the risk of domination by particular disciplines. For example, one of the largest collections was that of the National Gallery of Australia (96,609 items), therefore two other art gallery collections that exceeded 20,000 items were removed from the libraries that constituted this group. The libraries included in the Specials group provide a good coverage of subject collections including health, medicine, law, science, education, geosciences, primary industries, art, music, and taxation.

It is acknowledged that in one other regard the principle underlying the inclusion of libraries in the Specials group differs from that which applies in other groups. Whereas the other groups described above are constituted on similarities between institutions and their libraries, those included in Specials have been in part selected due to their difference. This should not, however, have any detrimental impact on the validity of results, as the emphasis is on assessing overlap and uniqueness between groups rather than within groups. The list of libraries included in the Specials group is provided in Appendix A.

Some negotiation was undertaken with OCLC staff in order to ensure that all data recording the holdings of the relevant special libraries were collected, as several of these libraries had items recorded under more than one library 'symbol' on WorldCat. The 23 libraries included in the Specials group had items recorded on WorldCat ranging 20,052 to 114,574, and the group as a whole had 977,749 items recorded.

A final group was created to represent the single collection held by the major Australian repository of print items, the CAVAL Archival and Research Materials (CARM) Centre. The CARM Centre provides a storage facility for member libraries and a document delivery service for the wider research community. At the time the study was conducted, CARM had records for 228,043 items on WorldCat.

The six groups therefore represented the holdings of 53 research library collections. These consisted of nine national, state and territory libraries; 20 university libraries; 23 special libraries, and the CARM Centre. The data was compiled by OCLC staff in June 2011, and represents holdings for printed monographs only. Serial holding were excluded from the research as they raise different issues in terms of their long-term accessibility and management. It is likely, however, that a similar study based on serial holdings may reveal further potential with regard to collaborative management and discard.

RESULTS

Table 1 reports the number of unique items held by the six groups, and the number of items that overlap with other groups. That an item is described as

'Unique' does not mean that it is held by one library only, but that all holdings recorded for an item are within one group only. Similarly, the 'Holdings' does not report the number of all copies held by the libraries within a group, but the number of items recorded as being held within the groups.

Table 1: Items unique to a group

Group	Holdings	*n* Unique	% Unique	*n* Overlap	% Overlap
G8	5147408	2577553	50.1	2569855	49.9
ATN	979505	261912	26.7	717593	73.3
IRU	2154663	581672	27.0	1572991	73.0
NSLA	5201563	3066265	58.9	2135298	41.1
Specials	977749	533470	54.5	444279	45.5
CARM	228043	81306	35.7	146737	64.3
Total	14688931	7102178		7586753	

The six groups have total holdings of 14,688,931. After allowing for the incidence of overlap (see Table 2), there are 10,008,274 separate items, and the number of items that are unique to a group (7,102,178) therefore represents 71% of all items. The number of items that are truly unique would be lower as these figures do not indicate titles that are duplicated (overlapped) within each group.

Table 1 strongly indicates that, in line with expectations, the largest collections are held by the G8 and NSLA. What is perhaps surprising is the high number of items that are unique to one of these groups only, exceeding 50% in both cases. As has been noted in previous overlap studies, larger research focused collections invariably have a higher level of uniqueness and lower incidence of duplication with other collections. This is most starkly illustrated by the results for the three groups of education libraries, G8, ATN and IRU, with the latter two groups both having a level of items unique within their group (26.7% and 27% respectively) that is just over half that of the G8. In terms of the national research collection, however, it is encouraging that the collections of the G8 and NSLA are sufficiently distinctive that they both retain a high level of uniqueness.

The result that runs counter to this is that of the Specials group, which despite having a substantially lower number of total items has a ratio of unique items (54.5%) that is similar to that of the G8 and NSLA. Another reflection of the comparatively high level of items unique to the Specials group is gained by comparing the results of the ATN and Specials groups. These two groups have very similar numbers of total items, but the Specials group has over twice as many items that are unique. These results indicate that in terms of their contribution to the national research collection that Special libraries are making a contribution that is greater than indicated by the number of items and holdings recorded in the NBD.

The results from Table 1 are also indicative of the extent to which the CARM Centre has been used as a last copy repository. Unlike all other libraries in the study, CARM is a group constituted of a single collection. Therefore, any items recorded as being unique to CARM are genuine single copies for the surveyed

libraries. The results indicate that the CARM Centre has 81,306 unique copies amongst the libraries included in the survey, a figure that indicates that in excess of a third of the collection (35.7%) may be either genuinely last copies in Australian collections, or at least copies with a very low level of duplication.

The *WorldCat Collection Analysis* Software also calculates the extent of duplication of individual titles. Table 2 reports the number of the six groups that hold each item listed in WorldCat, and Table 3 presents the same data as a cumulative percentage for each group.

Table 2: Extent of duplication (*n*)

Group	Unique	x 2	x 3	x 4	x 5	x 6
G8	2577553	1371755	772302	318020	102731	5047
ATN	261912	184598	231368	200112	96468	5047
IRU	581672	532005	633276	300359	102304	5047
NSLA	3066265	1095577	631247	301123	102304	5047
Specials	533470	117997	109910	120350	90975	5047
CARM	81306	43432	37293	41892	19073	5047
Holdings	7102178	3345364	2415396	1281856	513855	30282
Items	7102178	1672682	805132	320464	102771	5047

Table 2 indicates that there are 2,906,096 items that are held by more than one group, (calculated by summing columns x 2 – x 6 for 'Items'). For these items there are *at least* 4,680,657 duplicate copies held (calculated by subtracting 'Items' from 'Holdings' for columns x 2 – x 6, and then summing the results). However, the total number of duplicates within the total research library 'system' represented by the groups cannot be calculated as the number of copies duplicated *within* a group is not revealed by the data.

Table 3: Extent of duplication (cumulative %)

Group	Unique	x 2	x 3	x 4	x 5	x 6
G8	50.1	76.7	91.7	97.8	99.9	100
ATN	26.7	45.6	69.2	89.6	99.4	100
IRU	27.0	51.7	81.1	95.0	99.8	100
NSLA	58.9	80.0	92.1	97.9	99.9	100
Specials	54.5	66.6	77.8	90.1	99.5	100
CARM	35.7	54.7	71.1	89.5	97.7	100
All groups	48.3	71.1	87.5	96.2	99.8	100

Table 3 is most interesting for what it reveals about CARM, which is a very different type of 'group' in the context of these results. It is a single (and comparatively) small collection, but one developed with a goal of having a high proportion of unique items. There is some evidence that it has been successful in this in that it has levels of uniqueness similar to, and in a number of cases exceeding, those of groups (ATN and IRU) with substantially larger collections. It is, however, also apparent that CARM is as yet far from being a last copy collection. As reported in

Table 1, there are 146,737 duplicated items held in CARM, and Table 2 (columns x2 – x 6) records that they have an overlap of at least 345,221 copies with other groups. Indeed, as Table 3 indicates, CARM (10.5%) has the highest comparative level of duplication across groups, if this is measured by the number of items in a group that are held by four or more groups (calculated by subtracting column 'x4' from column 'x6'). Together with the results from Table 1, this indicates that the CARM Centre has a schizophrenic nature, consisting of both last copies and heavily overlapped items, the latter likely to consist of curriculum related textbooks with a high incidence of duplication.

The data also reveals the extent of overlap between groups. This data is presented in Tables 4 and 5. Table 4 presents the overlap figures indicating the number of items shared between each pair of groups.

Table 4: Overlap between groups (*n*)

	G8	ATN	IRU	NSLA	Specials	CARM
G8	5147408	625162	1356432	1852705	351993	120286
ATN	625162	979505	455055	414461	136709	27390
IRU	1356432	455055	2154663	995848	243140	83610
NSLA	1852705	414461	995848	5201563	337551	95326
Specials	351993	136709	243140	337551	977749	18609
CARM	120286	27390	83610	95326	18609	228043

Table 5 presents the percentage of overlap between pairs of groups in descending order—for example, in the highest incidence of overlap, 63.8% of the items in the ATN are also held by the G8. The results for CARM have been omitted from Table 5 given the different principle on which that collection is developed.

Table 5 Overlap between groups, highest ten in descending order (CARM omitted)

Overlapped group	Overlapping group	% Overlap
ATN	G8	63.8
IRU	G8	63.0
ATN	IRU	46.5
IRU	NSLA	46.2
ATN	NSLA	42.3
Specials	G8	36.0
G8	NSLA	36.0
NSLA	G8	35.6
Specials	NSLA	34.5
G8	IRU	26.4

It could be anticipated that the highest level of overlap would occur between libraries (or in this case groups of libraries) that are most similar in their

collection development priorities; and that smaller collections will experience a higher incidence of overlap than larger collections. Both of these assumptions are indicated by these results, with the education-based groups, and in particular the ATN and IRU, featuring as the groups with the highest level of overlap. Both of these groups share their highest level of overlap with the third—and largest—of the education groups, the G8. It is notable that despite the considerable difference in the number of items held by the ATN and IRU, the extent of their overlap with the G8 is very similar (63.8% and 63.0% respectively), another result that is strongly suggestive of heavy duplication of 'core' material in university library collections. Further evidence of this extensive core is found in the considerable overlap between the ATN and IRU, with 46.5% of the items held by the ATN also being held by the IRU.

What is perhaps most notable about the overlap results is that they again point to quite different nature of the holdings of the Specials group as evidenced by the comparatively low level of overlap experienced by this group. Although having a total number of items held that is very similar to the ATN (see Table 1), as Table 6 illustrates the ATN experiences a consistently much higher level of overlap with other groups than Specials.

Table 6. ATN and Specials overlapped (%)

	G8	IRU	NSLA
ATN	63.8	46.5	42.3
Specials	36.0	24.9	34.5

Interestingly, the lowest incidence of overlap recorded was between the ATN and Specials. These two groups recorded a two-way overlap of 14%. It can be speculated that this low incidence of overlap is a reflection of the nature of the libraries included in the Specials group, with some of the libraries included representing law, health, and medicine, not areas of teaching and research usually associated with the ATN institutions.

As noted CARM has been excluded from the reporting of the overlap data in Tables 5 and 6. Despite having been created on a repository model in order to permanently retain last copies and thereby encourage de-duplication, the data nonetheless indicate the CARM collection is considerably overlapped by the other groups. Table 4 (figures converted to percentages) indicate this includes considerable overlap with university-based groups, including a high of 120,286 items (52.7%) with the G8, and 83,610 (36.7%) with the IRU. Given that these figures represent the *minimum* number of overlapped holdings (i.e. one for each group), the results indicate that items held in CARM provide considerable potential for de-duplication of other collections.

DATA FOR SUBJECT DIVISIONS

OCLC's WorldCat Collection Analysis software makes it possible to evaluate results for groups according to subject 'divisions' as used by the OCLC Conspectus. OCLC describe the Conspectus as 'a framework to systematically inventory and describe library collections' (OCLC, 2011). There are 32 divisions

within the OCLC Conspectus, which can be mapped to Dewey Decimal, Library of Congress, and National Library of Medicine classification schemes.

For this research, overlap data for the six groups was collected for all 32 divisions. Tables 7-10 present this data for four of the divisions only. These divisions were selected to represent a range of disciplines in the sciences ('Chemistry'); technology ('Engineering and technology'); social sciences ('Law'), and humanities ('Art and architecture'). It should be noted that the 32 OCLC Conspectus divisions include one allocated to items with 'Unknown classification'. There were a number of items with this classification for each of the six groups, ranging from 4.7% of items for the ATN, and up to a very substantial 43.8% of items in Specials.

These 'Unknown classification' items are likely to be explained by the incidence of non-standard cataloguing (very high in some areas of special librarianship) that cannot be mapped to the divisions. It is unclear what the impact of these items will be on the comparative rate and distribution of occurrence of uniqueness and overlap. It is also notable that items included in the 'Unknown classification' have a level of uniqueness above the norm for each of the six groups, with a high of 93.5% for items in Specials. It is again likely that these figures are determined by not only the genuine uniqueness of items but also by local cataloguing practices that result in non-standard description of items and therefore an overstatement of the level of uniqueness (i.e. an inability to identify similar items in union catalogues).

Table 7. Unique holdings and overlap for the division 'Chemistry'

	Holdings	*n* Unique	% Unique	*n* Overlap	% Overlap
G8	83607	30662	26.8	52945	63.2
ATN	19619	4228	21.6	15391	78.4
IRU	45892	8726	19.0	37166	81.0
NSLA	67869	27268	40.2	40601	59.8
Specials	25896	6544	25.3	19352	74.7
CARM	1832	301	16.4	1531	83.6

Table 8. Unique holdings and overlap for the division 'Engineering and technology'

	Holdings	*n* Unique	% Unique	*n* Overlap	% Overlap
G8	229748	86744	37.8	143004	62.2
ATN	91481	25250	27.6	66231	72.4
IRU	99573	20338	20.4	79235	79.6
NSLA	237419	128056	53.9	109363	46.1
Specials	30386	8789	28.9	21597	71.1
CARM	20122	3801	18.9	16321	81.1

Table 9. Unique holdings and overlap for the division 'Law'

	Holdings	*n* Unique	% Unique	*n* Overlap	% Overlap
G8	146530	66289	45.2	80241	54.8
ATN	25333	4635	18.3	20698	71.7
IRU	61296	9189	15.0	52107	85.0
NSLA	109572	48529	44.3	61043	55.7
Specials	37762	7437	19.7	30325	80.3
CARM	2482	588	23.8	1894	76.2

Table 10. Unique holdings and overlap for the division 'Art and architecture'

	Holdings	*n* Unique	% Unique	*n* Overlap	% Overlap
G8	201386	79960	39.7	121426	60.3
ATN	67116	14053	21.0	53063	79.0
IRU	78487	16841	21.5	61646	78.5
NSLA	183224	78223	42.7	105001	57.3
Specials	62004	16299	26.3	45705	73.7
CARM	3727	1383	37.1	2344	62.9

These results indicate the extent of the contribution made by the various groups to the national research collection in these discipline areas. The G8 has the largest number of items recorded for three of the divisions, and NSLA for one; but NSLA has a higher percentage of unique items in three of the divisions. This result undoubtedly reflects the extent of overlap generated by university teaching requirements as recorded in Table 5, which reduces uniqueness for the G8, ATN and IRU. (In all the G8 has the largest number of items recorded for 22 divisions and NSLA for 10).

While larger collections have generally produced a higher percentage of unique items, it is again notable that the Specials group records a comparatively higher incidence of unique items than other groups. This is particularly apparent in the 'Science and technology' division, where despite the very modest number of items held the Specials group has 28.9% (8789 items) that are unique to that group. It should also be kept in mind that as many of the collections included in the 23 libraries in Specials would have negligible holdings relevant to this division, that these unique items are likely to be concentrated in very few libraries and therefore a low likelihood of overlap within the group.

These results also indicate that results for uniqueness and overlap vary considerably within each group depending on the nature of the disciplines represented by divisions. For example, the level of uniqueness reported for the G8 ranges from 26.8% (Chemistry) to 45.2% (Law); while that of the NSLA ranges from 40.2% (Chemistry) to 53.9% (Engineering and Technology). There are a number of

factors likely to influence these results, including the nature of publishing within particular disciplines and the spread of teaching and research across the various universities.

DISCUSSION AND CONCLUSION

The results presented above report only some of the large amount of data mined from the WorldCat data base using the *WorldCat Collection Analysis* software. In particular, the results reporting on the various divisions could only be partially presented. As noted above, there may also be issues with the reliability of some of the data due to issues with the completeness and accuracy of both the local and union catalogues (Libraries Australia and WorldCat) that underpin this research. Nevertheless, the data is sufficiently complete and consistent that sound and defensible conclusions can be drawn.

It is acknowledged that this research is conducted at a time when the whole concept of a 'national research collection' is being challenged by the advent of mass-digitisation and storage programs such as the Google Books Project (Dougherty, 2010) and the HathiTrust Digital Library (Christenson, 2011). These programs require libraries and the broader research sector to question the whole basis on which they conceptualise long-term storage of, and access to, legacy print materials. It is, however, very possible—even likely—that mass-digitisation will prove to be a driver in the reconfiguration of print storage on to a much broader and more collaborative scale. Constance Malpas has recently argued the need for a 'bridge strategy' that will optimise libraries existing investment in print while acknowledging the realities of a digital future. For Malpas, a major component of this strategy is, 'shared print repositories [that] could enable a significant and positive shift in library resources toward a more distinctive and institutionally relevant service portfolio' (Malpas, 2011: 10). If Australian research libraries are to envisage and realise a future that bridges print and digital content and that works to reduce the costs associated with local management of little used print assets, then there is a need to commence planning based on an agreed vision and collaborative action, underpinned by reliable data.

An important conclusion to be drawn from this research is that each of the six groups makes an important contribution to the pool of Australian research content. Although the level of unique items varies between groups (and between disciplines), there is sufficient uniqueness in each group to conclude that all groups are dependent on all other groups in order to optimise access to the national research collection. The contribution made by the Specials group is particularly notable in this regard. Despite the necessarily partial representation of special library collections in this group, it nonetheless contributes a considerably higher percentage of unique items to the national collection than suggested by the comparatively small individual collections represented. These findings are noteworthy at a time when some special library collections are being actively downsized. An examination of the Libraries Australia statistics indicates that of the 38 special libraries with over 20,000 items on the database, eight had experienced a reduction in the number of records between January 2005 and

June 2010 (Libraries Australia, 2010). This very likely understates the number of these libraries that had undergone net withdrawals, as it is clear from the statistics that many of them only began to add substantial numbers of records to the database during this period, thereby masking the effect of any downsizing. This action is being taken at a time when a number of academic libraries are also managing space problems by running 'steady-state' collections that rely heavily on continual withdrawals (Genoni, 2008).

It is also the case that there appears to be a high incidence of uniqueness generally reported by this data. While there is no obvious benchmark for the occurrence of unique items in an extended network of research libraries, the results indicate that a majority of the nation's research collection is held within one group only. It is stressed, however, that this figure under-reports the full extent of duplication as it does not record duplication occurring within groups. It is also the case that there is almost certainly an inflating effect on the level of unique holdings recorded in NSLA, as this group has statutory collecting requirements (for example for children's and young adult's material) that would have a very low level of overlap with other groups used in this study. Nonetheless, it is a result that does not, *prima facie*, indicate any concerning level of duplication in the national research collection.

As noted in the introduction, an aspect of this research is the investigation of the potential for federated storage of low use items. Despite the incidence of items that are unique to groups, the results also indicate the considerable overlap between groups, although the full extent of duplication cannot be assessed on the basis of data that is investigating groups of libraries as opposed to individual collections. Overlap is both necessary and unavoidable, and as with uniqueness there are no benchmarks for desirable levels of overlap. Nonetheless, on the evidence of this research there is encouragement to plan for the collaborative management of legacy print materials, as other countries have with initiatives such as the UK Research Reserve (Boyle & Brown, 2010) and the North Atlantic Storage Trust (Gherman, 2007). The results suggest that the Australian research library collection is sufficiently distributed and inter-dependent that it can be seen as a single network that requires collaborative management in order to optimise its effectiveness. This may be particularly true of decisions relating to the sustainable management of low-use research materials where collaborative repositories provide potential for substantial savings in the cost of long-term storage.

The prospect of a system-wide approach to the management of the national research collection raises the possibility of reducing costs by de-duplication (Courant & Nielsen, 2010). While decisions to delete duplicated items that are available in a secure repository will be made at the local level, it is apparent that there is at least potential for considerable de-duplication of Australian research collections based on the results from this study. This is a matter that requires some further investigation and agreement about what actually constitutes a 'last copy' in the context of Australia's research infrastructure, and indeed whether last (or single) copies will be sufficient to satisfy both demand and the need for secure retention.

It is also apparent that there is scope for considerable further research into the state of Australia's research collections—it is in the nature of overlap studies to leave unanswered questions. Just some of the issues raised by this research that require additional investigation include:

- An extended study of each of the groups used in this study in order to investigate the level of uniqueness and overlap *within* groups, and the implications this has for cross-sectoral collaboration,
- An examination of uniqueness and overlap at the title level in order to gain a better understanding of the characteristics of items (for example, date and place of publication) within these categories, and the implications this might have for storage decisions,
- A focus on the records of special libraries in order to fully explain the incidence of records in the division 'Unknown classification' and the impact this has on the interpretation of other data generated by this research,
- The need for research assessing the overlap, uniqueness, and distribution of Australian content in the national research collection, as opposed to items of non-Australian origin,
- The need for research to be conducted on a regional (i.e., state-by-state) basis in order to aid decision making related to collections at a local level.

On the basis of the research reported above it is apparent that extended use of the WorldCat Collection Analysis software is highly likely to be the most efficient means of undertaking this further research and data collection.

In conjunction with this further research it is necessary to undertake a (re) consideration of the definition of 'last copy' and associated concepts (work, expression, manifestation, and artefact) in the context of Australian and Australasian research collections. This includes the relationship between last print copies and digital surrogates.

ACKNOWLEDGEMENTS

The research was undertaken with support from the Australian Library and Information Association's YBP/Lindsay & Croft Award for Collection Services (2010). The authors would also like to thank OCLC, in particular Meghan Hopkins and Paul Brogger, for their generous assistance with access to the *WorldCat Collection Analysis* software.

APPENDIX A

Parliament of Australia, Parliamentary Library	114,574
National Gallery of Australia	96,609
Australian War Memorial Research Centre	60,804
CSIRO Black Mountain Library	55,936
Department of Justice and Attorney-General (NSW)	42,545
Federal Court of Australia	41,721
Australian Bureau of Statistics	37,908
Department of Conservation and Land Management (WA)	37,479
Australian Taxation Office	35,378
Department of Health (WA)	30,684
National Meteorological Library, Bureau of Meteorology	28,111
Geoscience Australia, N H (Doc) Fisher Library	27,711
Dept of Families, Housing, Community Services and Indigenous Affairs	27,543
Australian Council for Educational Research, Cunningham Library	27,157
Australian Institute of Criminology, JV Barry Library	27,057
Australian Institute of Family Studies Library	25,197
Powerhouse Museum: Research Library	24,848
Australian Museum Research Library	23,968
Supreme Court of Victoria	23,437
Gardiner Library Service: John Hunter Hospital Library	21,195
NSW Police Force Library and Information Service	20,714
Department of Primary Industries (Victoria)	20,598
Australian Music Centre	20,052

REFERENCES

Australian Research Libraries Collection Analysis Project. 2004. *Report*. http://www.library.uwa.edu.au/__data/assets/pdf_file/0004/524794/arlcap_final_report.pdf

Boyle, F. & C. Brown. 2010. The UK Research Reserve (UKRR): machinations, mayhem and magic. *Interlending & Document Supply*, 38(3): 140-146.

Christenson, H. 2011. HathiTrust: a research library at web scale. *Library Resources and Technical Services*, 55(2): 93-102.

Connaway, L., E. T. O'Neill & C. Prabha. 2006. Last copies: what's at risk? *College & Research Libraries*, 67(4): 370-379.

Courant, P. M. & M. Nielsen. 2010. On the cost of keeping a book. In, *The Idea of Order: Transforming Research Collections for the Twenty-first Century* (pp. 81-105). Washington: Council on Libraries and Information Resources.

Dickey, T.J. 2011. Books as expressions of global cultural diversity: data mining for national collection analysis. *Library Resources and Technical Services*, 55(3): 148-162.

Dougherty, W.C. 2010. The Google Books Project: will it make libraries obsolete? *Journal of Academic Librarianship*, 36(1): 86-89.

Genoni, P. 2007. Towards a national print repository for Australia: where from and where to? *Australian Academic & Research Libraries*, 38(2): 84-98.

Genoni, P. 2008. Current and future print storage for Australian academic libraries: results of a survey. *Library Collections, Acquisitions & Technical Services*, 32(1): 31-41.

Genoni, P. & E. Varga. 2009. Assessing the potential for a national print repository: results of an Australian overlap study. *College & Research Libraries*, 70(6): 555-567.

Genoni, P. & J. Wright. 2010. Assessing the collective wealth of Australian research libraries: measuring overlap using WorldCat Collection Analysis. *Australian Library Journal*, 59(4): 197-207.

Gherman, P. 2007. The North Atlantic Storage Trust: maximising space, preserving collections. *Portal: Libraries and the Academy*, 7(3): 273-275.

Lavoie, B., L. Dempsey & L. Connaway. 2006. Making data work harder. *Library Journal*, 131(1). http://www.libraryjournal.com/article/CA6298444.html

Libraries Australia. 2010. *Recent Statistics*. http://www.nla.gov.au/librariesaustralia/about/statistics/

Malpas, C. 2011. *Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment*. Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2011/2011-01.pdf

Missingham, R., & R. Walls. 2003. Australian university libraries: collections overlap study. *Australian Library Journal*, 52(3): 247-260.

National Library of Australia. 2007. *OCLC Agreement*. http://www.nla.gov.au/librariesaustralia/oclc.html

OCLC. 2011. *Introduction to the WorldCat Collection Analysis Service*. http://www.oclc.se/support/documentation/collectionanalysis/using/introduction/introduction.htm

Rochester, M. K. 1987. The ABN database: sampling strategies for collection overlap studies. *Information Technology and Libraries*, 6(3): 190-199.

Stephens, M. 2009. Heritage book collections in Australian libraries: what are they, where are they and why should we care? *Australian Library Journal*, 58(2): 173-189.

Ward, S., & M. C. Aagard. 2008. The dark side of collection management: deselecting serials from a research library's storage facility using WorldCat collection analysis. *Collection Management*, 33(4): 272-287.