Maine Shared Collection Strategy (MSCS)
Proposal: Collection Analysis Support

January 2, 2013

Document Status:
This document is a revised proposal based on MSCS Project Team feedback to a draft submitted on November 11, 2012, supplemented by a meeting with the Project Team in Orono on December 15th.

Overview
The Maine Shared Collection Strategy (MSCS) project seeks to create a strategy for the shared management of print collections in the state. Participating libraries will collaborate to make decisions about retention, storage, and preservation of print materials, potentially alleviating space concerns while assuring equal or greater access to the information. Among the group’s planned activities is an analysis of the state’s “collective collection”, in which uniquely-held titles, overlap, and institutional strengths are identified.

For monographs, MSCS seeks to answer a number of questions, based on the combined bibliographic, item, and holdings data of eight libraries:

- Colby College
- Bates College
- Bowdoin College
- Portland Public Library
- University of Maine/Orono (URSUS)
- University of Southern Maine (URSUS)
- Bangor Public Library (URSUS)
- Maine State Library (URSUS)

Some of those questions include:

- What monographs should the eight partner libraries designate for long-term retention for the benefit of shared collections in the State of Maine?
- What is an equitable and/or common-sense distribution of retention responsibilities?
- What monographs held by the partners are candidates for incorporating into POD/EOD services (via local or shared catalogs) by virtue of Hathi Trust or Internet Archive programs for public domain material?
- What monograph copies (by library) could optionally be deselected, once retention decisions have been finalized?
It is expected that these larger questions will be informed by data analysis, which will include:

- How many copies of a particular work are owned by partner libraries?
- How many of those are circulating copies?
- How often has the title circulated? What was the last circulation date? Was the circulation local or via Interlibrary Loan?
- How many titles/copies are uniquely held in the group? In Maine? In New England? In WorldCat?
- How do subject strengths compare across the group and the state?
- Which titles are represented in HathiTrust, Internet Archive
- Overlap between general collections and special collections
- Overlap between print and electronic editions
- Others to be determined from the group’s combined data set

SCS proposes to assist MSCS with the creation and use of a monographs data set that will enable the group to answer these questions to the fullest extent that the data supports.

**Current Situation**

MSCS is approximately 18 months into its 3-year grant from IMLS. The group has developed clear ideas about the analysis, and has done initial work on the respective data sets. MSCS would like to be reviewing hard data on its monograph collections by February/March 2013, and could benefit from the additional bandwidth and focus offered by a third party. SCS has done a number of similar group analysis projects, and has developed an approach that we believe can assist in helping MSCS to meet both its analytic and timing objectives. Those techniques are described in the remainder of this document. Because much of SCS’s work involves print deselection, some of the description tends in that direction. But the same tools and approaches serve the other side of the question, quantifying retention, preservation, and shared print decisions.

In July 2012, MSCS prepared its initial data extract and holdings lookups. These have been compiled into a group of delimited files on MSCS’s Dropbox account. The group invited SCS to evaluate these files, to determine whether SCS could adapt them to its own analytical processes, or whether a fresh extract would be needed. SCS notes the absence of several data elements we routinely use:

- MARC Leader data - we won't be able to capture record type or bib level.
- MARC 006 fixed field codes
- Cataloging Source
- GPO Item Number [to help filter out Gov Docs]
- Government Document Classification Number
- Medium Statement. We do have a "format" column but this looks like genre - not sure about the source of the data [used to identify electronic resources]
- LCCN – no data in this column – SCS uses for matching when OCLC # is not present or in question
- Local Catalog Author - we won't be able to do as thorough a validation of the OCLC number without the author from the local catalog record
- Publisher
- Circulation Data (including total charges, YTD charges, last charge date)
- Reserves (to be included ‘just in case’, based on 12/15 meeting in Orono)
Given these gaps, SCS would prefer to start with a fresh data extract. This would assure that we can provide all the outputs described in this proposal, and would make circulation and holdings information current through December rather than July. SCS will provide a specification for the bibliographic and item/circulation data (and related mappings) we would expect from a Millennium system. Based on our experience with other Millennium libraries (echoed by folks in MSCS), we anticipate that existing export tables can be modified to include these additional data elements. The Project Team noted that any analysis of Reserves activity is likely to prove challenging if not impossible, given the inconsistencies in local data practices. The group decided to include reserves data anyway, in case some analysis might be possible.

By design, there are some unique aspects to the MSCS project. Because the group includes a mix of public and academic libraries, some titles will be classified in Library of Congress system, and some in Dewey Decimal. It may be important to view and analyze the full data set through both. This may require assignment of alternate classifications or use of an intermediary mapping tool such as the OCLC Conspectus or a conversion table. On another front, MSCS is the first SCS client to express interest in determining the extent to which the group’s combined holdings are represented in the Internet Archive. Finally, MSCS wants the ability to look at its retention lists both at the edition-specific level and at the FRBR work family level. While SCS always supports this FRBR-off/FRBR-on capability when a title is held by fewer than 20 libraries in the US, to date we have not provided this at higher holdings levels.

In short, there is new ground for SCS in this project, and we look forward to exploring these areas in collaboration with MSCS. These are areas where neither SCS nor MSCS can know everything that may be required up front, and where some experimentation and development may be required to devise the best approach. SCS will be an interested and willing partner, and recognizes that development of some of these capabilities may improve our offering to other libraries. But others may be specific to the MSCS project. In general, therefore, we are proposing that SCS and MSCS share the risks and benefits, and that our two organizations work iteratively, discussing strategies and showing results as we proceed.

In the following sections, we describe the standard SCS approach to group projects, followed by an outline of steps and deliverables specific to the Maine Shared Collections Strategy.

**SCS Processes and Methods: Group Projects**

Shared print projects involve working with data from multiple libraries. The complexity of a project can vary, depending on the participants’ previous history of collaboration, the consistency of their respective data management practices, and other factors. This following table provides a high-level view of the main project components and SCS methods.

In addition to the sample data shown in the table, SCS will include statistics on titles unique to MSCS (i.e., titles not held outside this group). Additional categories may also be defined, based on title protection rules, e.g., for works on or about Maine.

[A more detailed description is included as an Appendix, after the signature page.]
<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
<th>Data/Decisions Needed</th>
<th>Timeframe</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Planning Meeting</strong></td>
<td>Key players discuss data extracts, mappings, anomalies; define peers, title protection rules.</td>
<td>Mappings, peers, Scope of extract</td>
<td>1-2 days</td>
<td>Project outline Preliminary schedule</td>
</tr>
<tr>
<td><strong>Data Preparation</strong></td>
<td>SCS receives data extract, validates, transforms, normalizes, performs holdings lookups</td>
<td>Bib, item, circulation data</td>
<td>Depends on size. Approx. 3 weeks for 2.4 million records</td>
<td>Remediation lists; library &amp; group-wide databases for querying</td>
</tr>
<tr>
<td><strong>Group Collection Summary</strong></td>
<td>Categorical overview of the group data set. Used to gauge opportunities, guide scenario development. Can be further subdivided by subject, location, etc</td>
<td>Review of patterns, benchmarking</td>
<td>Approx. 1 week after group database is available</td>
<td>See example following this table.</td>
</tr>
<tr>
<td><strong>Scenario Development</strong></td>
<td>Project leaders suggest preliminary, retention, preservation &amp; withdrawal criteria, drawing from data elements in the Collection Summary. SCS performs trial iterations, revisions as necessary.</td>
<td>Criteria development, discussion of results, adjustments, repeat process.</td>
<td>Depends on number of iterations &amp; availability of decision makers. 2-3 days per iteration for data work.</td>
<td>Finalized scenario descriptions; criteria for retention, withdrawal &amp; preservation</td>
</tr>
<tr>
<td><strong>Candidate Lists</strong></td>
<td>Based on finalized criteria for retention, withdrawal and preservation. Detailed Excel spreadsheets (sample in Appendix).</td>
<td>Requires final decisions on criteria.</td>
<td>1 week for Retention Candidate lists. 1 day per library for Preservation lists.</td>
<td>Group retention candidate list. Individual preservation candidate lists (for scarcely-held titles).</td>
</tr>
<tr>
<td>Task</td>
<td>Description</td>
<td>Data/Decisions Needed</td>
<td>Timeframe</td>
<td>Output</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Allocation</strong></td>
<td>Responsibility for withdrawal and retention candidates must be divided equitably among participants. This allocation process incorporates many factors: space, distribution of copies, presence in the collection. Allocation can occur based on collection size, proportion of withdrawal candidates, or other factors.</td>
<td>Individual and group goals sometimes come into conflict here. Needed: a group-wide agreement on who benefits and who bears costs and in what proportions.</td>
<td>Unknown. These are difficult decisions, and can require numerous discussions and repeated trials of an allocation algorithm.</td>
<td>Allocation formula that is acceptable to all.</td>
</tr>
<tr>
<td><strong>Discussion/Facilitation</strong></td>
<td>Needed at many points, but especially around scenario development, allocation/retention, and related policy decision (last copies, retention responsibilities, etc)</td>
<td>SCS can help organize &amp; facilitate some conversations by presenting alternatives drawn from the data</td>
<td>Varies, depending on process.</td>
<td>Documented decisions.</td>
</tr>
<tr>
<td><strong>Picklist/Keeplist Production</strong></td>
<td>Once allocation decision has been made, SCS will derive title/item lists for use by individual libraries.</td>
<td>Criteria and allocation scenario finalized.</td>
<td>1-2 weeks.</td>
<td>Two outputs: an annotated group withdrawal list &amp; library-specific retention lists.</td>
</tr>
<tr>
<td><strong>Ongoing Data Management</strong></td>
<td>SCS data sets can be used for 2 years from creation. SCS will maintain (but not update) library data during that period. Data can be used for additional projects.</td>
<td>None</td>
<td>2 years from initial creation.</td>
<td>Data remains available for additional analyses. Pricing for additional work to be negotiated.</td>
</tr>
</tbody>
</table>
### Example of a Group Collection Summary

<table>
<thead>
<tr>
<th>MCLS Library Title Counts</th>
<th>All Libraries</th>
<th>Percent</th>
<th>Wayne State</th>
<th>Western Michigan</th>
<th>Central Michigan</th>
<th>Eastern Michigan</th>
<th>GVSU</th>
<th>SVSU</th>
<th>MTU</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Records - Filtered</td>
<td>3,823,327</td>
<td>100%</td>
<td>1,169,157</td>
<td>971,995</td>
<td>597,046</td>
<td>529,502</td>
<td>226,280</td>
<td>167,707</td>
<td>161,640</td>
</tr>
</tbody>
</table>

#### Circulation Counts

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Charges = 0</td>
<td>1,657,669</td>
<td>43%</td>
<td>538,486</td>
<td>424,284</td>
<td>221,828</td>
<td>259,402</td>
<td>64,436</td>
<td>67,629</td>
<td>81,604</td>
</tr>
<tr>
<td>Total Charges = 1 or less</td>
<td>2,306,924</td>
<td>60%</td>
<td>721,941</td>
<td>588,741</td>
<td>328,917</td>
<td>350,896</td>
<td>101,184</td>
<td>102,650</td>
<td>113,495</td>
</tr>
<tr>
<td>Year-to-Date Charges = 0</td>
<td>3,654,632</td>
<td>96%</td>
<td>1,115,062</td>
<td>933,814</td>
<td>562,286</td>
<td>499,877</td>
<td>226,046</td>
<td>162,657</td>
<td>154,890</td>
</tr>
</tbody>
</table>

#### WorldCat Counts

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 100 holdings in USA</td>
<td>2,366,515</td>
<td>62%</td>
<td>541,951</td>
<td>611,428</td>
<td>430,967</td>
<td>379,881</td>
<td>170,939</td>
<td>123,534</td>
<td>107,815</td>
</tr>
<tr>
<td>&gt; 50 holdings in USA</td>
<td>2,937,455</td>
<td>77%</td>
<td>750,401</td>
<td>769,790</td>
<td>499,997</td>
<td>452,609</td>
<td>194,831</td>
<td>139,938</td>
<td>129,939</td>
</tr>
<tr>
<td>&gt; 20 holdings in USA</td>
<td>3,338,916</td>
<td>87%</td>
<td>928,065</td>
<td>870,487</td>
<td>539,713</td>
<td>493,804</td>
<td>212,220</td>
<td>150,944</td>
<td>143,681</td>
</tr>
<tr>
<td>&gt; 10 holdings in Michigan</td>
<td>576,255</td>
<td>26%</td>
<td>181,336</td>
<td>174,153</td>
<td>157,867</td>
<td>172,426</td>
<td>93,426</td>
<td>71,500</td>
<td>51,543</td>
</tr>
<tr>
<td>&gt; 5 holdings in Michigan</td>
<td>1,515,519</td>
<td>50%</td>
<td>402,238</td>
<td>468,459</td>
<td>361,520</td>
<td>326,391</td>
<td>153,830</td>
<td>111,666</td>
<td>91,415</td>
</tr>
<tr>
<td>&lt; 10 holdings in USA</td>
<td>272,270</td>
<td>7%</td>
<td>143,257</td>
<td>133,839</td>
<td>33,608</td>
<td>16,174</td>
<td>3,760</td>
<td>7,120</td>
<td>10,912</td>
</tr>
<tr>
<td>&lt; 5 holdings in USA</td>
<td>283,515</td>
<td>5%</td>
<td>110,547</td>
<td>97,872</td>
<td>24,946</td>
<td>10,821</td>
<td>3,973</td>
<td>4,612</td>
<td>8,744</td>
</tr>
<tr>
<td>&gt; 5 holdings - MCLS Pilot Group</td>
<td>173,957</td>
<td>5%</td>
<td>26,397</td>
<td>31,527</td>
<td>28,773</td>
<td>27,492</td>
<td>19,263</td>
<td>21,654</td>
<td>18,851</td>
</tr>
<tr>
<td>&gt; 3 holdings - MCLS Pilot Group</td>
<td>989,604</td>
<td>26%</td>
<td>173,311</td>
<td>218,904</td>
<td>189,146</td>
<td>179,301</td>
<td>96,063</td>
<td>72,370</td>
<td>60,509</td>
</tr>
<tr>
<td>&gt; 1 holdings - MCLS Pilot Group</td>
<td>2,456,460</td>
<td>64%</td>
<td>546,918</td>
<td>638,520</td>
<td>443,520</td>
<td>404,101</td>
<td>178,514</td>
<td>127,226</td>
<td>117,661</td>
</tr>
<tr>
<td>Unique Titles in MCLS Pilot Group</td>
<td>1,363,047</td>
<td>36%</td>
<td>622,239</td>
<td>333,475</td>
<td>144,728</td>
<td>125,401</td>
<td>52,744</td>
<td>40,481</td>
<td>43,979</td>
</tr>
</tbody>
</table>

#### Date related counts

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Publication Year before 2005</td>
<td>3,430,618</td>
<td>90%</td>
<td>1,099,755</td>
<td>842,941</td>
<td>530,610</td>
<td>467,682</td>
<td>187,107</td>
<td>151,881</td>
<td>150,662</td>
</tr>
<tr>
<td>Publication Year before 2000</td>
<td>3,061,615</td>
<td>80%</td>
<td>1,030,797</td>
<td>725,169</td>
<td>481,396</td>
<td>402,442</td>
<td>144,239</td>
<td>137,875</td>
<td>139,697</td>
</tr>
<tr>
<td>Publication Year before 1990</td>
<td>2,381,396</td>
<td>62%</td>
<td>871,575</td>
<td>547,126</td>
<td>368,927</td>
<td>280,338</td>
<td>89,136</td>
<td>107,289</td>
<td>117,005</td>
</tr>
<tr>
<td>Added before 2005</td>
<td>3,097,504</td>
<td>85%</td>
<td>1,069,578</td>
<td>805,774</td>
<td>499,860</td>
<td>445,830</td>
<td>N/A</td>
<td>137,080</td>
<td>139,382</td>
</tr>
</tbody>
</table>

#### Electronic Summaries

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hathi Trust Public Domain Match</td>
<td>131,447</td>
<td>3%</td>
<td>51,939</td>
<td>28,109</td>
<td>20,157</td>
<td>17,784</td>
<td>2,667</td>
<td>2,453</td>
<td>8,338</td>
</tr>
<tr>
<td>Hathi Trust In-Copyright Match</td>
<td>1,571,907</td>
<td>41%</td>
<td>470,391</td>
<td>396,657</td>
<td>246,440</td>
<td>226,209</td>
<td>84,628</td>
<td>72,106</td>
<td>75,476</td>
</tr>
</tbody>
</table>
Example: Allocation Scenarios

Scenario 3: Allocable holdings with different circulation thresholds

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Title Set inclusion criteria</th>
<th>Title holding inclusion criteria</th>
<th>Allocable Holdings</th>
</tr>
</thead>
<tbody>
<tr>
<td>3a</td>
<td>No restrictions on aggregate circulation</td>
<td>Holding &lt; 4 circs</td>
<td>534,067</td>
</tr>
<tr>
<td>3b</td>
<td>No restrictions on aggregate circulation</td>
<td>Holding &lt; 5 circs</td>
<td>559,939</td>
</tr>
<tr>
<td>3c</td>
<td>No restrictions on aggregate circulation</td>
<td>Holding &lt; 6 circs</td>
<td>577,318</td>
</tr>
</tbody>
</table>

For all these scenarios:
- Title sets have 3 or more holdings
- Title sets are excluded when publication year > 2004
- Title holdings are excluded when item add date > 2004
- Title holdings are subject to location filters

Scenario 3a: Withdrawal candidates & allocable candidates per 3 allocation schemes (< 4 circs)

<table>
<thead>
<tr>
<th></th>
<th>Withdrawal Candidates (maximum potential withdrawals)</th>
<th>Allocable Candidates (Distributed by % withdrawal candidates)</th>
<th>Allocable Candidates (Distributed by collection size)</th>
<th>Allocable Candidates (Distributed by list agreed to proportions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Western</td>
<td>193,321</td>
<td>111,179</td>
<td>134,416</td>
<td>126,059</td>
</tr>
<tr>
<td>Michigan Tech</td>
<td>60,729</td>
<td>34,925</td>
<td>22,354</td>
<td>46,135</td>
</tr>
<tr>
<td>Central</td>
<td>168,808</td>
<td>97,082</td>
<td>82,556</td>
<td>67,145</td>
</tr>
<tr>
<td>Eastern</td>
<td>212,452</td>
<td>122,182</td>
<td>77,960</td>
<td>67,145</td>
</tr>
<tr>
<td>Wayne</td>
<td>169,844</td>
<td>97,678</td>
<td>161,621</td>
<td>129,723</td>
</tr>
<tr>
<td>GVSU</td>
<td>53,809</td>
<td>30,946</td>
<td>31,976</td>
<td>45,186</td>
</tr>
<tr>
<td>SVSU</td>
<td>60,684</td>
<td>40,075</td>
<td>23,184</td>
<td>52,673</td>
</tr>
<tr>
<td>TOTAL</td>
<td>928,647</td>
<td>534,067</td>
<td>534,067</td>
<td>534,067</td>
</tr>
</tbody>
</table>
Scope/Deliverables
This project will focus on circulating monographs in print form, as well as reference works. Print journals, government documents, audio-visual, microforms, and other non-book materials are out of scope. Special Collections books are included. MSCS and SCS estimate that we will load approximately 2.4 million bibliographic records and 3 million item records. SCS is responsible for analyzing and presenting the data needed to make decisions and formulate policies. SCS will also facilitate discussions related to data analysis, interpretation, and policy options. Decisions and final policies are the responsibility of the libraries and the Maine Shared Collections Strategy group. Specific steps and deliverables include:

1. SCS will load and normalize bibliographic, circulation, and item data for eight library data sets: Colby, Bates, Bowdoin, Portland Public, University of Maine/Orono, University of Southern Maine, Maine State Library, and Bangor Public.

2. SCS will produce data remediation lists for corrections made in the course of normalization. This may include corrected control numbers, identification of records without holdings set in WorldCat, etc. (Since the MSCS partners have recently completed OCLC reclamation projects, however, the number of items identified for remediation is expected to be small.)

3. SCS will compare MSCS data to holdings in OCLC WorldCat, HathiTrust, and to one another. Holdings lookups are typically performed with FRBR groupings off (i.e. edition-specific) unless there are fewer than 20 holdings in the US. We understand that MSCS may want to raise this threshold. SCS can accommodate this, but believes it should be discussed first, as it will have a significant impact on the time required to perform WorldCat holdings lookups.

4. SCS will also investigate comparison of MSCS data to the Internet Archive (IA). SCS foresees some possible issues with this process, such as:

   a. Because the Internet Archive API is not designed for large-scale batch queries, SCS must obtain the full set of Open Library data (of which IA is a subset).

   b. SCS must parse the Open Library records to identify the IA titles. These are large files, e.g., the Open Library Editions file contains 25 million lines. About 6.8 million of these appear to have OCLC numbers. As of 1/2/13, the IA “Texts” division contains 3.7 million items, not all of which are books. It will require some digging to verify the various relationships and the quality of the data. We believe that the actual number of full-text books in IA is between 2.2 and 2.5 million.

   c. SCS must identify items which appear both in IA and in HathiTrust to minimize duplication of counts.

Because the value of this work to the analysis is not yet clear, and because significant development work may be involved, SCS proposes first to dedicate 40 staff hours to this investigation, then report to the Project Team on the efficacy of using IA. We would also propose a process and estimate any additional costs at that point.

5. Holdings counts for all US libraries, all libraries in Maine, all libraries within the group, HathiTrust and (depending on the outcome to #4 above) Internet Archive would be included in
SCS Summaries and reports. SCS can also provide aggregate peer comparison with peer libraries/groups to be determined. SCS can support up to three peer groups, each with up to 20 OCLC holding symbols. However, as we understand it, comparisons to peers outside of Maine may not be of interest to the group. If this is the case, SCS will be able to offer more flexibility in other areas, as management of peer holdings is a very resource-intensive part of our process.

6. SCS will compile all streams of collection, item, circulation, and external holdings data into a consolidated MSCS database, which will be maintained on SCS servers. Reserves data may also be included, provided there is sufficient consistency to support analysis. Depending on the ultimate approach to subject analysis, SCS may need to modify its internal databases to accommodate multiple classifications (DDC and LC) for each item in the MSCS data extract, to support an Internet Archive flag, and to support both FRBR-off and FRBR-on holdings totals.

7. SCS will work with MSCS to develop an approach to subject analysis that works for all eight members of the MSCS group. This may involve supplementing library-supplied call numbers with call numbers from the WorldCat Master Record. But it is likely that a significant number of titles (as much as 40%) will not have a DDC number, even after that has been done. There are several options which might be explored here, but it is difficult to gauge the best approach without some experimentation:

   a. Additional discussion of objectives might be helpful. Are there types of analyses that can only be supported by viewing the data from both an LC and a DDC perspective? Can collection strengths be analyzed using a single scheme rather than augmenting the data to be viewed through both lenses?

   b. Subject analysis could begin by using LC-only. It is likely that LC classification would be available for a much higher percentage of records. This would enable reliable high-level subject sorting, but also use of local call numbers (both LC and DDC) for list production and review. Since this capability already exists, it may be helpful to start here and see how fully it satisfies the need.

   c. OCLC Conspectus: another approach would involve assignment of the 500+ Conspectus Categories to all 2.4 million records. This would provide a broad vocabulary that bridges LC and DDC, but which has acknowledged limitations, especially for the public libraries. (E.g., subjects such as cooking and knitting are subsumed under a category called ‘Domestic Engineering.’)

   d. Assign missing classification numbers in both schemes. Practically, this could only be done at the DDC Thousands and the LC Subclass level. It would be based on a conversion table agreed by MSCS and SCS, and perhaps customized or extended in select areas. (For example, DDC 974.1 and F16-30 represent Maine local history. For the most part, a conversion table would function at the 974 and F level, but it may be important to support more granular analysis. It is unclear how fully this can be supported without experimentation.

It is difficult to predict how much time and effort may be required here. SCS is willing to experiment, discuss options, and proceed to a reasonable degree of effort without additional
charge. However, if significant new development is necessary to support subject analysis, SCS and MSCS will discuss supplementary costs and payments.

8. SCS will work with MSCS to identify areas of strong collecting interest (e.g., any works relating to Maine or the North Atlantic Coast), and to support classification and keyword-based analysis of overlap, etc.

9. The consolidated MSCS database (called a ‘rollup’ by SCS) will collate bibliographic, item, and circulation data, along with holdings and other derived information. The MSCS roll-up will remain available for two years. In consultation with the group, SCS will design and execute queries against this MSCS data set, to produce a Group Collection Summary, similar to the example shown on page 5.

10. SCS will draw from the Summary data to analyze collection overlap by subject, date, circulation level, location and other factors supported by the data. In most cases, it will be possible to drill down to the title or item level, to view and validate results.

11. SCS will present the Summary-level analysis to the Maine Shared Collections [steering group] and libraries’ staffs, either in person or via webinar. SCS will facilitate discussion of deselection and retention scenarios.

12. SCS will iterate queries and retention scenarios as needed, in consultation with the MSCS group and participating libraries. There may be many variations of group-level queries run, adjusted, and repeated. Each scenario will include an estimate of its potential yield for retention, withdrawal candidates and preservation candidates. Based on other criteria, statistics and lists will be produced to identify potential POD/EOD candidates. These iterations are the heart of SCS’s service, enabling libraries to interact with the data until they have evolved criteria that support their collection goals.

13. SCS will recommend specific areas of focus and will estimate the size of various opportunities, based on our experience with other academic libraries and with the data itself. This will occur initially via an in-person visit, which will be supplemented by conference calls, video calls, and additional onsite meetings as needed. SCS will identify issues and strategies for the group to consider, and will help to frame key decisions.

14. SCS will produce a single group-wide list of retention candidates based on criteria specified by the MSCS Libraries. After negotiations with the group regarding allocation priorities, SCS will allocate retention commitments and optional withdrawal candidates, and produce appropriate lists for each library. The allocation database that underlies all of these decisions will be made available to MSCS via login access or, if preferred, a copy will be provided for local hosting by MSCS.

15. SCS will also produce library-specific lists of preservation candidates (titles held scarcely in the US and/or in Maine).

16. SCS will maintain the MSCS data set (as created to prepare the Group Collection Summary) for up to two years. If additional analyses are wanted subsequent to the initial project, the same
data set can be used. Depending on the complexity of subsequent projects, additional charges may apply.

17. In addition to the analyses that SCS will perform on behalf of MSCS, SCS will make the data set available directly to MSCS. Because of its size, this will most likely need to be done by providing MSCS with login access to SCS servers. Use of this database will require facility with SQL and other large-scale database management tools.

**Schedule:**
SCS will accommodate MSCS’s preferences for scheduling to the highest degree possible. The date on which SCS begins its data work depends on several factors, including:

- a) Date SCS receives the data extracts from the library
- b) Finalization of contract and initial payment received
- c) Number of other libraries in production at SCS

At present, SCS could begin work in February 2013. We propose a target date of February 4th for receipt of the MSCS data extracts.

Once SCS has received data extracts from each of the participating libraries, data normalization, holdings lookups, and creation of the MSCS database “roll-up” will require approximately four weeks. (Note that this may be somewhat longer if FRBR-on and FRBR-off results are wanted at all holdings levels.) Creation of an initial Group Collection Summary will require an additional 1-2 weeks. SCS could therefore produce a Group Summary within 5-6 weeks of the data extract. We understand that MSCS would like to be reviewing Summary data by February/March 2013. Assuming SCS receives a data extract by February 4th, we can produce a Group Collection Summary by mid-March.

After presentation and discussion of the Summary, the timing of the process becomes largely driven by the libraries. SCS will produce iterations of the Summary as needed. As decisions are made about allocating withdrawal candidates and retention commitments, SCS will produce lists as needed. The MSCS data set will remain available for querying and analyzing for up to 24 months, after which we recommend it be refreshed and rebuilt to capture additional circulations, newly added titles, etc. The full refresh of the database would again incur set-up and per-record fees.

**Signatures:**

_________________________________________________________________
[Signatory] 
Maine Shared Collection Strategy

_________________________________________________________________
Rick Lugg, President 
Sustainable Collection Services
Appendix: SCS Processes for Group Projects

Shared print projects involve working with data from multiple libraries. The complexity of a project can vary, depending on the participants’ previous history of collaboration, the consistency of their respective data management practices, and other factors. This section provides a detailed description of the main project components and SCS methods, and is intended to supplement the chart on pages 3-4.

1. **Planning meeting:** At the outset of a project, it is essential to bring together key players from each library and SCS to develop a project roadmap and timeline. This can be done in person or via conference call, and typically involves project leaders, systems librarians, representatives from collections, and someone who knows the data well (often a cataloger). The purpose is to discuss the scope of the data extract, field mappings, anomalies, and data keys/legends. It is especially important to identify data variances that result from use of different automated systems, and to determine the degree of common data on circulation and other use. The group will also define peer comparators, title protection rules, and authoritative title lists to be used. SCS will schedule this meeting as soon as an agreement has been finalized. *Output: Project Plan/Outline.*

2. **Data preparation:** For group projects, each library will provide SCS with bibliographic, item, and circulation data for all of its circulating print monographs, plus reference and eBook data if those are included. The Libraries will also provide descriptions of data mappings. SCS will establish an FTP account on its servers for transfer of library data. SCS will then execute its ETL (Extract, Transform, Load) processes. These include:

   - **Unit tests:** Sample loads to check OCLC numbers, confirm data mappings
   - **Transformation:** Split and transform MARC bibliographic files into MARC21XML files to identify encoding errors and to enable use of XSL transformation on smaller batches. Perform XSL transformations on bib, item, and circulation data to create library-specific tab-delimited files.
   - **Normalization:** Some data elements (especially those related to date or time formats, control numbers, and call numbers) tend to vary over time and from library to library. It is necessary to normalize these data elements to assure reliable comparison among and across other libraries. Steps include:

     - Normalize LC Call Number
     - Normalize Local Call Number
     - Normalize SBN to ISBN
     - Normalize LCCN
     - Normalize OCLC Numbers
     - Detect presence of multiple distinct OCLC Numbers
     - Correct year 2000 LCCNs formatted with 4-digit year
     - Extract and normalize publication year from MARC 260|c
     - Retrieve OCLC number from WorldCat when not present in local record
       - Check for valid LCCN
       - Retrieve entire set of records for that LCCN
3. **MSCS:** 'filtered' Group level estimate provide Counts, Summary elements.

- Run title-similarity check (filter down to records that meet title/author risk threshold)
- Look for WorldCat record against which library has set holdings
- If none, accept first record that meets title/author risk threshold

- For MSCS: Extract alternate classifications from the WorldCat Master record.

- Remediation: A major benefit of the SCS normalization routines is the ability to return lists of items with corrected or newly-obtained data to the libraries for remediation of their own data. Examples include corrected or derived OCLC numbers, lists of items with holdings not set in WorldCat, etc. These are typically provided at no additional charge to the library, since they are a by-product of ongoing SCS routines.

- Aggregation: Once transformation and normalization are complete, each library’s tab-delimited files are loaded to a library-specific database on SCS servers. For group projects, the databases of participating libraries are then aggregated into a single, group-wide database.

- Holdings Lookups: Once the aggregate database has been prepared, SCS creates a separate, de-duplicated version, in order to streamline holdings lookups through the WorldCat API. Each OCLC number is matched only once against WorldCat. Initial matching occurs with FRBR-groupings off, assuring edition-specific matches. For titles where fewer than 20 holdings are reported, the matching is repeated with FRBRC-groupings on. In all cases, the OCLC or LCCN matching is supplemented with a title-string similarity check, to minimize the risk of false matches. Holdings information is retrieved and stored for library-defined peers, as well as holdings in the US and in the state. Results are loaded to library-specific databases and the group roll-up. A similar process is used to identify library holdings that appear in the Hathi Trust database, either as public domain or in-copyright titles.

- Rollups: Individual, Group, Aggregate—create views onto data for use by SCS staff to build and execute queries. These may vary from project to project, depending on local needs.

3. **Group Collection Summary:** This is typically the first data reported back to the group. The Group Collection Summary creates a categorical overview of the data: Circulation Counts, WorldCat/Peer Counts, Date-Related Counts, and Hathi Trust overlap. The example below shows only the highest-level view of the data; it can be further subdivided by subject or location. In addition, statistics on ‘filtered’ titles, excluded because they do not fit the scope of the project, can be reported. The Summary is intended primarily as a management tool, to help identify patterns in the data, and to estimate the potential yield of various retention strategies. In some instances, SCS can build preliminary withdrawal and preservation scenarios from this data, by combining various data elements. These trial scenarios can serve to stimulate discussion and ideas among the group, and provide the basis for Scenario Development, the next step in the process.
4. **Scenario Development**: Retention, preservation, and optional withdrawal scenarios are typically constructed from a combination of circulation, date of publication or acquisition, and peer library holdings data. Many combinations are possible, and developing scenarios that balance risk and yield appropriately is perhaps the most important task in any shared print project. These modeling exercises will help inform shared retention agreements, and improve each library’s understanding of respective collection strengths and what must be retained. This is typically an iterative process: a scenario is proposed, the results produced and discussed, and a revised scenario proposed. Sometimes several iterations are necessary. Decisions will be needed from project leaders and participating libraries to shape scenarios that support sufficient redundancy, areas of collecting strength and retention commitment. Possible scenarios will depend on the comparability of circulation data, the degree of collection overlap, and the publication dates and archiving thresholds (number of copies) agreed by the group. A simple example of a group retention scenario follows:
5. **Candidate Lists:** Once criteria are finalized and revised results re-calculated, lists can be produced. In group projects, both retention candidates and withdrawal candidates (if any) are intended to be shared, so a single list of candidates is produced. However, preservation candidates (titles which are scarcely held elsewhere) tend to be held by a single library so individual lists are produced for those.

6. **Allocation:** In group projects, retention responsibilities must be shared equitably, or at least in some manner agreed by the group. Deciding which libraries will retain certain titles depends on a number of variables, including how many copies the group seeks to retain, which library holds which titles, size of collection, space needs, and other factors. New concepts such as title sets, title holdings, maximum potential withdrawals, and allocable candidates are introduced to assist with prioritization and weighting. Working with the participating libraries, SCS will define a sequencing algorithm that assures 1) an agreed number of copies of each title are protected; 2) retention commitments are distributed as agreed by the group; and 3) that the library allocated a title for deselection holds that title (based on the data supplied at the outset of the project).

7. **Picklist and Keeplist Production:** Once it has been established which libraries may withdraw their copies, more targeted lists will be needed: retention lists and withdrawal picklist. Retention lists will identify those titles each library has committed to retain and preserve on behalf of the group. Withdrawal picklists will include only the individual library’s withdrawal candidates, along with the necessary subset of data elements, e.g., location, barcode number, local call number. This will enable the withdrawing library to batch suppress records while items are removed from shelves and record maintenance performed.

8. **Ongoing Data Management:** Although new titles are acquired each year and circulation continues to occur, most libraries find that their SCS data set can serve for up to two years without modification. As part of this agreement, SCS commits to retaining the library’s initial data and corresponding holdings lookups available for two years from the first data load. This means that projects conceived after the initial analysis can be run against that data set, minimizing the additional cost. If a completely refreshed data set is wanted, SCS will be happy to produce a new proposal.