

## Project 4 - Replacement for Parent Database

---

### 4.1 Background

The Cochrane Database of Systematic Reviews (CDSR) provides The Cochrane Collaboration with great advantages over regular journals in that all of its published output is stored in a highly structured, consistent, granular format. This makes it possible to perform automatic searches that extract similar data across all reviews or subsets of reviews. There is considerable scope to exploit this unique feature of Cochrane reviews for the benefit of CRGs, The Cochrane Collaboration, current users of our output and potential users.

The Parent Database (PD) was used to assemble CDSR in order to deliver it to the publisher, before the implementation of Archie removed the need for this compilation. The need to compile the PD had the spin off benefit of allowing data extraction for scientific, statistical and monitoring purposes across the whole of CDSR. The publishing efficiencies arising from the introduction of Archie and the removal of the need for the PD for that purpose arise from the ability to store reviews as discrete, whole documents rather than as a collection of individual fields that cut across Cochrane reviews. The downside of this is that it is now more difficult to extract data from all reviews. As an example, it is now time consuming to find and extract all the statistically significant meta-analyses from CDSR since each review document has to be accessed, compared to previously when the relevant fields (comparisons, outcomes, and study results) could be accessed across all reviews.

The move to the new Archie document model and the introduction of the XML format in Cochrane reviews meant that the old PD could no longer be used and a replacement for the PD was developed<sup>1</sup> and launched with Issue 2, 2006 of The Cochrane Library. This 'new' PD was developed like the old PD as a system running in parallel to Archie. For each new Issue, all the reviews from Archie for that issue were imported by running a program that split the documents into individual fields. This new PD contains each issue of CDSR from Issue 2, 2006 to Issue 1, 2008, and has been the data source for numerous data requests<sup>2</sup>.

From issue 2, 2008, reviews in RevMan 5 format started to appear in the CDSR. These reviews (which include Diagnostic and Overviews of reviews) have a quite different structure to RevMan 4 reviews, meaning that they could not be loaded into the PD. Therefore, from Issue 2, 2008, it has not been possible to create a PD and the IMS team do not have the resources for the more laborious task of extracting data from the individual reviews to meet most requests for data for scientific, statistical and monitoring purposes. Therefore, such requests need to be dealt with either using data that can be extracted directly from Archie, or by using the data from issue 1, 2008, which is now substantially out of date and does not reflect the considerable enhancements introduced with RevMan 5.

In hindsight, if we had invested more resources in 2005 when considering how the new PD could be integrated with Archie, it may have been possible to "future proof" this so that it would have lasted longer than two years. However, at that time we decided that it was better to develop a system quickly using the technology we knew, rather than divert resources from the other Archie development work. In addition, the version of the database software we used (and are still using) for Archie did not provide the technology required for integration. The update that introduced this technology was released in 2005, but we did not have sufficient resources to consider the benefits of updating at that time.

### 4.2 Proposal and discussion

A replacement for the PD should be developed, as an integrated module of Archie rather than a separate database. This should consist of a "back-end" where each published version of a Cochrane review is stored in searchable form (as part of the Archie database), and a "front-end" interface for performing searches across the reviews and extracting data.

---

<sup>1</sup> Development took approximately 2 FTE months.

<sup>2</sup> Access to data is restricted to projects approved by the Steering Group Executive. The IMS Team performs the extraction.

The advantage of integrating the new module with Archie, rather than developing another parallel system is that this module will benefit from the existing framework for the storage of data and access to it. It will also reduce the resources that have previously been required for the maintenance of a standalone PD and for designing and running queries on behalf of the people needing data from Cochrane reviews. This change will require guidance on which Archie users should have the rights to perform their own queries, which we can implement when setting up the system.

The technology (XML indexing) required for the development will be available with an update to the database server software as described above, and this project is therefore dependent on the completion of project 1.b (updating the database software).

XML indexing technology can index any type of XML document, independent of its detailed structure. This means that the introduction of the new module will remove the need for a redesign of the PD each time changes to the review structure are introduced, for example in any replacement to RevMan 5. All published review versions stored in Archie (back to 1995) would be indexed, so the new PD would include the data in previous versions of the PD, providing a resource that would be unmatched for investigations of healthcare research.

From the user perspective, some of major features in an integrated module would be:

- The ability to search on specific review versions (e.g. search only in the reviews published in Issue 4, 2008, search in all issues of 2007 for reviews from the Stroke group, or track the impact of updating on reviews with more than five included studies since the first issue of CDSR)
- To allow specification of detailed search criteria relating to the content of the searched document versions (e.g. where one or more outcomes report SMD, or where the Declaration of interest section contains the word 'none').
- To allow extraction (exporting) of specific data from the hits (e.g. the number of included studies, the I-squared value of each meta-analysis, or the Authors' conclusions section)

This project is related to Project 5 (Improve searching functionality) and will benefit from some of the improvements resulting from that project, including:

- The ability to group and sort results (e.g. sort by date edited, or group by number of included studies)
- The ability to report results as counts rather than individual hits (e.g. how many reviews have one included study, how many have two, etc.)
- The ability to combine different Boolean operators (and, or) in a single search

This module will also enable efficiencies in the implementation of other projects including:

- Project A – Review monitoring system (e.g. notify users about changes in specific sections of reviews, such as individual meta-analyses or the Implications for practice)
- Project D - Cross referencing between reviews (e.g. to improve the impact factor and the ease with which users can move between Cochrane reviews)
- Data exchange between Archie and the Cochrane Register of Studies (e.g. extraction and insertion of studies and sharing of information on studies between the new Register and Cochrane reviews)

### 4.3 Summary of recommendations

Replace the PD with an integrated Archie module.

### 4.4 Resource implications

This project is dependent on the completion of project 1.b (updating the database software). We do not expect any additional hardware or software costs. Development time can roughly be divided into three stages: Investigation and design (6 FTE weeks), implementation of working prototype (14 FTE weeks), implementation of first working version

(6 FTE weeks). We estimate that testing and technical documentation will require 6 FTE weeks and end-user documentation will require 2 FTE weeks.

## 4.5 Impact statement

The Cochrane Collaboration as a whole will benefit because it will be easier to find and extract data for research, statistics and monitoring purposes. There is a possibility of generating income for the Collaboration by setting up criteria for charging people for access to the data.

If the PD is not replaced with a new system this will remove the ability to perform detailed searches and data extraction across Cochrane reviews, which has existed since the first release of CDSR. Researches will have to extract data manually from The Cochrane Library (for example by cutting and pasting from the relevant section of each review), and it will not be possible for the IMS team to provide data for monitoring and statistical purposes, including those required by the Steering Group or the Editor in Chief.