

Update Software's System for Publishing CENTRAL Records

Drafted by: Mark Starr, 5 October 2000
Latest revision: Hazim Timimi, 9 September 2002

Archived
1 Jan 2006

Introduction

MEDLINE, EMBASE and specialised register records undergo two levels of processing at Update Software ("Update"). This happens before their publication each quarter in the Cochrane Collaboration's CENTRAL database of trials in The Cochrane Library.

1. Level 1 processing

MEDLINE, EMBASE and specialised register records are imported into the relevant 'source' table within Update's CENTRAL database. There are 4 distinct source tables:

- Table 1 holds only MEDLINE records;
- Table 2 holds only EMBASE records;
- Table 3 holds specialised register records and the hand search results register;
- Table 4 is referred to as the 'Controller'.

Table 4 is used by Update in Level 2 processing, described in Section 2 below.

1.1 MEDLINE records

Update runs a search strategy each quarter against OVID's MEDLINE database on CD-ROM to identify 'new' (i.e., newly identifiable within MEDLINE since the last OVID search) relevant MEDLINE records. The results of the following search strategy, limited by date, are saved to a tagged-text file and then imported into the MEDLINE table (table 1):

1. randomized controlled trial.pt.
2. controlled clinical trial.pt.
3. #1 or #2
4. limit to human

This MEDLINE search strategy is also re-run against the entire MEDLINE database (from 1966 onwards) whenever OVID release the latest annual "Global Reload" for MEDLINE on CD-ROM. All records from the "Global Reload" are loaded into table 1, replacing the previous set of MEDLINE records in CENTRAL. This yearly reload is necessary to capture changes in thesaurus terms for MEDLINE records. (To reflect changes in health-care practice and to conform with terminology in current use, the MEDLINE thesaurus changes each year, with new terms added and old terms removed.)

1.2 EMBASE records

A search strategy covering the years 1974-1999 was run by the UK Cochrane Centre in 1999 against EMBASE to identify reports of trials. The search strategy consisted of the free-text search terms random*, crossover*, cross-over*, factorial* and placebo* and the results were de-duplicated against records already tagged as randomized-controlled-trial or controlled-clinical-trial in the publication type field in MEDLINE, since these are already included in CENTRAL. The results of this search were saved to a text file and then imported into the EMBASE table (table 2). The retrieved set of records is published on CENTRAL for each quarterly release. Additional searching of EMBASE began in December 2000, and this stage of the project will include searching using additional free text and Emtree thesaurus terms. A standardised tagged-text file containing the latest batch of EMBASE records retrieved by the new search is sent quarterly to Update by the UK Cochrane Centre. These records are imported into the EMBASE table (table 2).

1.3 Specialised registers and hand search results

Update Software receives standardised tagged-text export files of Cochrane entities' specialised registers each quarter from the US Cochrane Centre ("USCC"). The format of these files was agreed upon by Update and USCC and is detailed in Appendix A. These register files are saved in a separate register 'holding area' outside of the 4 source tables. If an updated register has been sent, it automatically overwrites the previously submitted register file in the holding area. If an updated register has not been sent, the entity's earlier register submission is kept in the holding area. The register table (table 3) is emptied each quarter and then all the files found in the register holding area are imported into the register table (table 3).

This system works because the USCC uses standard names for the registers based on the "SR-code" for each register. For example, the code for the Cancer Network is SR-CANCER, so the filename used for their register is SR-CANCER.txt. An up-to-date list of all SR-codes can be found on the Data Providers Pages of the Update Software website (<http://www.update-software.com/>).

The USCC maintains and updates a single cumulative database of all hand search results received from Cochrane entities since September 1998. A tagged-text file (SR-HANDSRCH.txt) is exported quarterly from the database and is imported by Update into the register table (table 3).

2. Level 2 processing: the 'controller' table

The 'controller' table (table 4) is used to decide which of the other 3 source tables (i.e., MEDLINE, EMBASE or specialised registers) should be used to create each CENTRAL record published in The Cochrane Library.

Publication criteria

All source records from tables 1, 2 and must have a title and a valid year of publication (i.e., greater than 1000 and less than or equal to the current year). All records in table 3 must have an SR code. Records that don't satisfy these criteria are ignored by the 'controller' and are not published.

Building the 'controller'

The 'controller' table is re-built each quarter using the three 'source' tables, in the following order of precedence:

1. MEDLINE records from table 1
2. EMBASE records from table 2
3. Hand Search register records from table 3
4. Specialised register records from table 3

The text published in CENTRAL comes from the source highest in the order of precedence (the "primary source") as determined by the 'controller' table. If the 'controller' identifies other records lower down in the order of precedence ("secondary sources") that match a "primary source" then extra information is published on the CENTRAL record, as follows:

- EMBASE records from table 2: only the EMBASE number is added to the published record
- SR-HANDSRCH records from table 3: only the SR-HANDSRCH code is added to the published record
- Other SR- records from table 3: only the appropriate SR-code is added to the published record

Note that MEDLINE or EMBASE accession numbers submitted via the handsearch and specialised register files are not published in CENTRAL.

The 'new' tag is assigned to records that have not previously been published in CENTRAL (i.e., records not in the previous version of the 'Controller' table). In the future, there are plans to assign a 'CCTR' tag to quality-assessed records in CENTRAL, as described in the CCTR Coding and Error Removal section of the CENTRAL Management Plan. (Note: The CCTR tags are not currently assigned. Prior to the Issue 3, 2001 release of *The Cochrane Library*, CCTR tags were assigned to records uploaded from either MEDLINE or EMBASE.)

Matching "secondary" sources to a "primary" source

The correct matching of "secondary" to "primary" sources is crucial to Level 2 processing and the process has been refined over time. The latest changes were made in August 2002 to remove around 17,700 unnecessary duplicates that had appeared on CENTRAL and to prevent their reappearance in future.

The matching algorithm looks for identical information in the year and title field of each publishable record in the source tables. The first four characters in the year field are used. Titles are first searched for the presence of extra expressions that are often appended by various database providers. If such an expression is found it is deliberately ignored by the algorithm. The list of expressions ignored by the algorithm is given in Appendix B. Titles are then converted to upper case, all punctuation characters are removed and then the first 236 characters of what remains are used for matching.

How can users of CENTRAL identify the "primary source" of a record?

The "primary source" used for each record on CENTRAL can be deduced as follows:

- Any record published with a MEDLINE accession number will have come from table 1.
- Any record published without a MEDLINE accession number but with an EMBASE number will have come from table 2.
- Any record published without either a MEDLINE or an EMBASE number and which shows the SR-HANDSRCH code will have come from the latest SR-HANDSRCH.txt file.
- Any record published without either a MEDLINE or an EMBASE number and without the SR-HANDSRCH code will have come from the latest specialised register files.

3. Feedback reports sent to USCC

Update sends USCC automated feedback reports about the specialised register and handsearch files submitted for each quarter. The reports list the following information for each file (ie each SR code) as identified by Update (note: these figures may conflict with USCC's figures):

- Total number of records that can be published on *The Cochrane Library*.
- Number of records with a missing or invalid year.
- Number of records with a missing source (eg, journal name).
- Number of records with the SR code and title.

Feedback reports directly concern the citations Update Software will use to create CENTRAL. USCC communicates information on the feedback reports to the Review Groups, to notify them of problems with their submitted registers and to assist them with making changes for the next submission.

Appendix A. Data format for the USCC's submission of specialised registers to Update Software.

Specialised Registers and hand search results to be published in The Cochrane Library CENTRAL database must be provided by USCC to Update Software in the "tagged text" format specified below.

Each file must be a ASCII text file, not rich text format or word document.

A record delimiter of %%%% should separate each record in the file. The record delimiter must be on a separate line.

A tag should only appear once per record. Please do not embed tags within fields.

The following table lists the acceptable tags and fields that Update Software can process from USCC's submissions.

| Tag | Description | Format | Comments |
|-----|-----------------------------|------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| AU: | Authors | All authors in form last name initial. Separate names with comma. | |
| TI: | English title | Full English title. If possible, try not to have " or [as the first character of the title. | Record must have a title because these are used in the "index" display of The Cochrane Library |
| TO: | Original language title | Format as for English title | |
| SO: | Name of journal/book/source | | |
| YR: | Year of publication | 4 character year | Years only please. Records must have a year that is greater than 1000 and less than or equal to the current year. Values such as 19 th May 1999 will cause the record to be rejected. |
| VL: | Volume | Can also contain volume supplement e.g. 50 Suppl 5 | |
| NO: | Issue | Can also contain issue supplement e.g. 1 Suppl Can also be used for volume part, e.g. Pt 69 | |
| PG: | Pages | Start and end pages, e.g. 50-55 | |
| EN: | Edition | Edition of book, e.g. 2 nd or software platform e.g. 3.1 for Windows | |
| ED: | Editors | Format as for authors | |
| PB: | Publisher name | | |
| CY: | City of publication | | |
| MD: | Medium | Only relevant if reference type is Computer software e.g. floppy disk, CD-ROM | |
| AN: | Medline ID | Numeric | |

Update Software's System for Publishing CENTRAL records

| Tag | Description | Format | Comments |
|------------|--------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| ID: | Other id numbers | If more than one separate with semi-colon. | |
| AB: | Abstract | Free text | Embedded tabs will be removed on publication. |
| KY: | Keyword terms | Each term should be separated with a semi-colon. | |
| DE: | Study design | | Not the same as MEDLINE publication type. |
| PT: | MEDLINE publication type | Free text | |
| AD: | Contact address | Complete address on one line. Use commas to separate address lines. No carriage returns | |
| RT: | Reference type | Please standardise to the following values: Journal article Book Section of book Conference proceedings Correspondence Computer program Unpublished data Cochrane review Other Scientific or technical report Patent Newspaper article Audiovisual Legal material Map Unpublished material, in press Publications on the Internet | Refer to the RevMan 4.0 User Guide Appendix G for more detailed information. |
| SN: | Study name | Free text | |
| PD: | Proceedings date | | |
| PL: | Proceedings location | | |
| UF: | URL of full text article | Only one URL should be included. No other text should be in the field. e.g. http://www.update-software.com | If the first 4 characters of the supplied field are not "http" then will not be able to create hyperlink |
| US: | URL of summary or abstract | As for UF | As for UF |
| CO: | For comments for a particular reference that one often gets when downloading from MEDLINE. | | |
| CC: | Collaboration code. | The SR code used to indicate the Cochrane entity that is linked to the record | Must be upper case and included in the list of agreed codes. e.g. SR-PREG |
| AK | | | Copyright statements, etc |

If a field is empty, you can either provide the tag with no information following (record 1 below) or leave the tag out altogether (record 2 below), both options will be processed successfully.

Example file format:

AU: Wones RG
TI: Failure of low-cost audits with feedback to reduce laboratory test utilization
TO:
SO: Medical Care
YR: 1996
VL: 25
NO: 1
PG: 10-11
EN:
ED:
PB:
CY:
MD:
AN: 87114199
ID:
AB:
SN: Wones 1996
DE: RCT
RT: Journal article
CC: SR-UPDATE
%%%%
AU: Diener HC
SO: Drug-induced headache
YR: 1988
PB: Springer-Verlag
PG: 25-30
CY: New York
RT: Book
CC: SR-UPDATE
%%%%
AU: Weinstein L
TI: Pathologic properties of invading microorganisms
SO: Pathologic physiology: mechanisms of disease
YR: 1974
PB: Saunders
PG: 457-72
ED: Sodeman WA Jr, Sodeman WA
CY: Philadelphia
RT: Section of book
CC: SR-UPDATE
%%%%
AU: Demicheli, V. and Jefferson, T.
YR: 1997
TI: An exploratory review of the economics of recombinant vaccines against hepatitis C
SO: Organisation for economic co-operation and development. The economic aspect of biotechnologies related to human health, Part 1: Biotechnology and medical innovation
RT: Scientific or technical report
CC: SR-UPDATE

Archived
1 Jan 2006

Appendix B. Expressions ignored by the 'controller' matching and de-duplication algorithm.

The algorithm searches for the presence of each of the expressions in the list *in the order shown below*. The first time one of these expressions is found in a title the algorithm stops searching the list and removes the expression from the title for matching purposes only (ie not from the published record). The search is case sensitive.

[abstract]
[see comments.] (This is the OVID 2002 CD-format)
[see comments] (This is the OVID pre-2002 CD-format)
[German]
[French]
[Italian]
[Japanese]
[letter]
[Spanish]
[Russian]
[letter; comment.] (This is the OVID 2002 CD-format)
[letter; comment] (This is the OVID pre-2002 CD-format)
[Polish]
[Danish]
[Chinese]
[EASL abstract]
[Review]
[AASLD abstract]
[Norwegian]
[Swedish]
[Portuguese]
[Dutch]
[Czech]
[proceedings]
[Hungarian]
[translation]
[IASL abstract]
[news]
[Serbo-Croatian (Roman)]
[Finnish]
[editorial; comment]
[editorial]
[Hebrew]
[comment]
[Bulgarian]
[In Process Citation]
[UEGW abstract]
[Slovak]
[DISSERTATION]
[extract]
[Romanian]
[Ukrainian]
[Prior annotation incorrect]
[Serbo-Croatian (Cyrillic)]
[see comments] (two spaces separating the words)
[meeting report]
[Poster]

Archived
1 Jan 2006