HHDB III : a review

Anders Thulin, Malmö

Introduction

In April 2005, Harold van der Heijden published the latest major edition of his chess endgame study database (see *http://home.studieaccess.nl/heijd336/home.html* for details).

The two previous editions, published by ChessBase, are widely regarded as must-haves for anyone with an interest in endgame studies. The latest edition is called HHDB III and contains information of almost 68000 endgame studies: the publisher estimates this is approximately 80% of all studies published up to the publication date.

The purpose of this review is to take a closer look at what is included in the database, how well it fulfills its purpose, and also if there are any obvious improvements that could be made.

The CD

HHDB III is delivered on a single CD. The copy I bought contained 8 files:

readme.pdf — a short introduction to the CD, with information about the special codes used for information in the files, and also the other files found on the CD.

hhdbIII.pgn — the main database, the contents of which is reviewed below.

codes.txt — documents the source and event codes used in the main database to get around shortcomings in the software used to produce the database. As an example 'b013' is used to refer to Lommer's book *1357 Endgame Studies*, and 'jr01' is used to refer to the Roycroft Jubilee Tourney. If you must know where a study was published, or what award it has won, this is the file to use.

hhdbIIIa.pgn — a combination of the main database and the *codes.txt* file: codes in *hhdbIII.pgn* have been replaced by full text from *codes.txt*. Unfortunately, in a few cases, this was not possible, and so some information has been truncated. For general usage this version is more convenient to use than the standard

database, but it should not be used where completeness of information is required.

In addition to these files, some material of historical interest is also included:

hhdbI.pgn — PGN-version of the first database version (1991), published by ChessBase.

hhdbII.pgn — PGN-version of the second database version (2000), also published by ChessBase as *Endgame Study Database 2000*.

xref.xls — A Microsoft Excel file documenting correspondences between *hhdbII.pgn* and *hhdbIII.pgn*.

These last three files are not clearly of any major interest for ordinary users.

One undocumented file also appears on the CD:

release.pdf — this file is not mentioned in the *readme.pdf* document: it appears to contain a description of the changes that have been made since the first release of the database. (I am using release 1, dated 2005-05-13.

By now it is probably clear that further software is required to use this CD:

The PDF files will require Adobe Reader 4.0 or later for reading (or equivalent software). And anyone wanting to read the *xref.xls* file will need Excel Reader or equivalent. Both Adobe Reader and Excel Reader can be found free on the net, but it would have been nice if references to the relevant download pages had been included to help the non-Internet-savvy user over this initial hurdle. A user without any Internet connection at all is left stranded, which is quite unfortunate, as the information in *readme.pdf* really must be available for everyone. It would have been better to make this a pure text file.

The database itself, of course, requires some kind of PGN-processing software for use, and it is again assumed that the user knows where to find this. As will be noted below, the actual use of the database relies entirely on the capabilities of this software, and so I think it would have been helpful to give the user some suggestions as to what would be useful and not.

The only software mentioned is CQL, Chess Query Language. In this case, a URL has been provided to the CQL web site.

My own fairly limited experience suggests that ChessBase or some similar database software is probably the best to use in order to get the kind of reporting and searching capabilities a user will want. Chessbase Reader (which I found on the CD of a relatively recent copy of *Chessbase Magazine*) also seems as if it would work well, but chess-playing programs such as Fritz seem too limited for anything but the most trivial searches: something more is required

to use the database to the fullest. The free Chessbase Light is not useful, as it cannot handle databases of more than 8000 'games'.

The Database

Once you have successfully imported or opened the PGN-database to your database software, what next? This is actually quite a difficult question to answer, mainly because there is no documentation on what is actually *in* the database, or what the database is intended to be useful for. This is an unfortunate omission.

Here are some possible uses, but in the absence of authoritative information they should not be considered to be anything but guesses on my part:

- What studies have N.N. published? Where?
- Has N.N. co-authored any studies?
- Who has composed most endgame studies?
- Is a given position *P* a published study?
- What studies were published in source *S*?
- What studies have won first prize in tourneys?
- Has a certain study been anticipated or cooked?
- What studies rely on en passant capture?

However, it must be noted that although the database contains some or all of this information, the PGN-software you actually are using may not necessarily be capable enough to to find it. And if the user is looking for name-related information, it may even be that the database information doesn't follow the PGN format specification well enough for standard searches to work as expected. For more details of this, see below.

As the database relies on the PGN-software to provide the actual search capability, the user is in a bit of a bind: there's nothing that says how the database is intended to be used, and there's nothing that says what PGN-processing software actually is capable of allowing such use. In other words: the PGN database contains some unspecified information, but it's up to the user to find the right software to be able to use it as he wishes.

The only area in which the database can be evaluated that doesn't at the same time involve some particular computer program seems to be its actual format, the contents, and the information it provides. In order to do a proper evaluation, then, it is clearly useful to know that the database actually contains, but no detailed description is present on the CD.

Some ideas can be found by the codes documented in the *readme.pdf* file: there is for example a code '{pl}' that indicates that the study has been anticipated to some (undocumented) extent. This suggests that one purpose of the database may be to help detecting such anticipations as far as they are not

already detected by simple search functions. As long as the source and degree of anticipation is not described, it is difficult to know how far this information can be trusted: is it based on position or on content? If the latter, who made the decision? It is possible to make guesses, of course – but the answers should really have been included.

While examining anticipated studies, I encounter another minor problem: there is no good way to identify a particular study in the database. The database I use provides a number (*e.g.*, 2745, a study by Moisovski), but that number is not guaranteed to refer to the same study stored in another database. When I try *Chess Informant Expert,* the number is 2744, and in one of my personal tools, the original order is not retained at all, so any sequence number is useless.

Back to possible purpose of the database. Correctness is a second possibility after anticipation: there are also codes to indicate cooks and duals of various degrees, although all have not been documented. Typically the solution contains a variation with comment such as 'cook JU' or 'cook MG', yet there's nothing that says what those codes mean (not even the *readme.pdf* file), or what criteria have been used to establish cookedness. In a database that covers studies from more than three centuries it would be somewhat odd if modern standards were applied also to the very oldest entries.

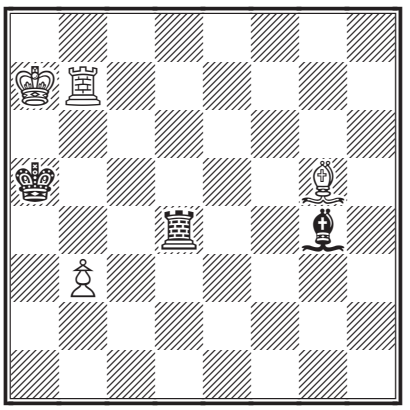
The publisher estimates that the database contains approximately 80% of all studies published. I'm not capable of evaluating completeness well, but a more-or-less-random test is fairly easy enough to make. The randomness is fairly low: my personal interests are firmly on the far side of the year 1900, and so most studies I have collected in any form are comparatively old.

Of 187 studies tested from sources between 1858 and 1927, I find 41 not present in HHDB (*i.e.* with a identical position or mirrored, rotated or translated). This is roughly 20%, though it must be noted that the real percentage is probably lower, as I expect around 5% of my test studies to have been mistyped.

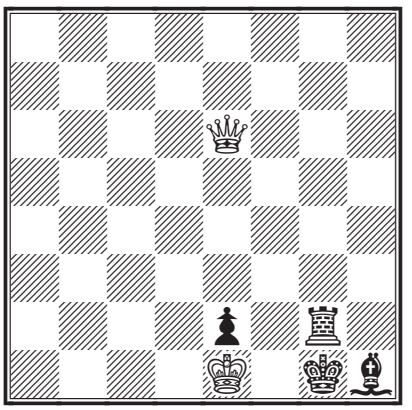
For the studies actually present, my source information mainly agrees with HHDB III. Quite often I have noted the same source, and in most of the other cases, HHDB III refers to an older source than I have documented. In just a few cases, it looks as if I may have the better source information. Thus, I have no complaints as to completeness.

The reverse evaluation is less clear: is there something the database absolutely should not contain but still is present? In the absence of any documented purpose, I can only guess that since it is a endgame study database, I probably shouldn't expect to find chess problems in it. But when I correlate the positions in the database with a smallish problem collection of my own, I do find a number of chess problems.

For instance, problem 310 in the Russ miniature book (a 6-mover by T. Nissl) is included as a win-study by unknown author from unknown source. Same thing with problem 281 from the same book (a 5-mover by Galitzky) (see dia-



Nissl: Mate in 6 moves



Galitzky: Mate in 5 moves

grams). A quick search for other possible material by Nissl reveals a position by 'T. Nisse' from *Akademische Monatshefte für Schach* that also looks much more like a problem than a endgame study.

Including problems makes a certain sense in a database intended for finding anticipations: more-movers are closely related to studies, and as there is some overlap between the fields, studies could perhaps be 'actively anticipated' from such problems.

However, when such database entries are tagged as being faulty in various degree, it raises further and more difficult questions as to their presence. When regarded as studies most problems are most likely faulty, but it is not clear why they should be regarded as studies in the first place.

The absence of source and author information in these two cases rather suggests that some of this information may have been collected by less formal methods than the rest of the material.

Continued checks up to the year 1900 produced another half-dozen of studies that seem likely to have been originally published as more-mover problems. The small number of finds may indicate that the database doesn't contain very many of these, but it might also be related to the relatively small sample of problems (about 7000) against which I checked the study database.

This suggests that some entries in the database are not really endgame studies. If this is deliberate, it really should merit a word of warning in the introduction: a user should not be allowed to assume that everything in the database are studies.

The presence of studies from unknown sources by unknown authors in the database also raises some questions. They are not useful for finding anticipations, as they contain no source information, and they are not obviously useful for any other purpose. There is just a handful of them, yet their presence is difficult to explain.

Another possible evaluation regards name correctness and consistency: are names of study composers spelled consistently, especially names that have been transcribed from Cyrillic scripts? This is not entirely easy to evaluate, but it should at least be possible to look for possible inconsistencies. I find, for instance, among the names beginning with 'A', one single instance of 'Afansiev=G' together with 115 of 'Afansjev=G': this may indicate a misspelling. (When I check this name against Jeremy Gaige's *Chess Personalia,* I find that Gaige prefers the transcription 'Afansiev', which raises further questions about what transcription system, if any, is used in the database.) It is less easy to decide if the single 'Afonin=A' may be a misspelling for 'Afonin=S' of which there are nine, as *Chess Personalia* has no entry for anyone of the name Afonin. Indeed, the publisher is probably one of the relatively few people who can decide such matters. The presence of 'Alekseev=V' and 'Aleksejev=A' suggests that there may be inconsistencies in transcription methods as well that need to be taken into account by the user. I also find 'Akerblom=A', which I would have preferred to see as 'Åkerblom=A', which raises another question about handling of unusual accented Latin letters: a very quick check suggests such accents may simply have been dropped.

A few words on the problem of name consistency, and the tradeoffs that have been made, added to the *readme.pdf* file, would do much to prepare the user for this problem. Name unification is a problem in most chess databases, and so it should not really come as a surprise that it is present here as well.

The PGN Data

(*This section is comparatively technical, and may safely be skipped. But in order to understand exactly what the database contains, it is necessary to understand some of the problems involved in shoehorning a endgame study database into a file format intended for chess games.*)

The database is distributed in PGN form. PGN (Portable Game Notation) is a data exchange file format developed by Steven J. Edwards around 1991, and which has since become very popular for exchange of chess game scores. Most modern chess database programs have some degree of PGN support.

PGN defines two major groups of formats: import format and export format. The import format is intended for data produced manually, and is therefore rather forgiving in nature. The export format is more strictly defined,as it is intended to be used for archive storage and data exchange between different computer programs. Here I assume that only the export format is of interest for the examination of the database.

The PGN specification stipulates that files intended for exchange purposes shall use single LF (line feed) characters as end-of-line indicators. The database file uses CR+LF, and although this is unlikely to cause any problems in practice, it is nevertheless a possible source of problems.

PGN also stipulates that lines in data files should not be longer than 80 characters. As far as I can find, the maximum line-length in the database is 2730

characters, which is yet another possible area for problems: some reader programs may refuse such excessively long lines.

PGN was designed for exchanging information about chess games, *not* endgame studies. Chess games have players on two sides, and are usually played at a date and at events of various kinds, while endgame studies are by one or more composers, and have been published and possibly even won major awards. In other words, endgame studies come with information that are not obviously possible to store in a PGN database.

A number of tradeoffs have been made to make the whole thing work. These can cause problems, so it is necessary to know what they are, and how they affect the user.

The first entry of the *hhdbIII.pgn* file looks as follows

```
[Event "1.c ja13 te03"]
[Site "?"]
[Date "2005.??.??"]
[Round "?"]
[White "Bocharov=J"]
[Black "[=0003.31h4b5"]
[Result "1/2-1/2"]
[SetUp "1"]
[FEN "7n/2p4P/8/1k1P4/3P3K/8/8/8 w - - 0 1"
[PlyCount "13"]
[EventDate "2005.??.??"]

1. Kg5 Kc4 2. d6 $1 cxd6 3. d5 (3. Kf6 $2 Kd5 4. Kg7 Ke6 5. Kxh8 Kf7) 3... Kd4
4. Kf5 Kc4 5. Ke6 Kc5 6. Ke7 Kxd5 7. Kf6 1/2-1/2
```

and represents a endgame study along with its solution.

The entry consists of two parts: the first contains a number of lines bracketed in [], followed by an empty line, and the second part consists of two lines containing a number of chess moves.

In the first part, the first seven entries (from Event to Result), in PGN terminology known as ‘tags’, are required by the PGN format, and must be present in this exact order. The following four tags (SetUp, FEN, PlyCount and EventDate) are optional, but if they are present they must be in alphabetical order.

Although it is not very likely that a PGN-processing program will refuse to accept the data because that order is not followed, it must be noted that there is a risk it may happen, and that any program that does refuse to accept the information because ‘SetUp’ doesn’t follow ‘PlyCount’ does so with the full support of the PGN format specification.

What information is found in these tags? If this was a standard PGN file, the PGN format specification itself defines the contents (with some omissions), but as the HHDB database does not contain chess game scores, further documentation is necessary.

Unfortunately, there is no such information on the CD, so the user is left to his own guesses and assumption. It is usually possible to make a relatively sound guess, but it should not be considered acceptable to leave such vital information undocumented. The explanations below are my guesses for those cases where I feel I can reasonably safely make them.

```
[Event "1.c ja13 te03"]
```

Award or source information. *ja13* and *te03* are codes that are documented in the *codes.txt* file. Other codes may appear here (indications of cooks, duals, corrections etc. — these are documented in the *readme.pdf* file, except for the symbol '#' which occurs quite frequently.

The award hierarchy is not documented. Users who have seen one or two tourney reports will have no problems in figuring out what '1.p' or '1.hm' means. '1.c' may be more of a puzzler, especially in the apparent absence of any '1.m' . Less knowledgeable users will have to go elsewhere to interpret these codes. (I guess that p = prize, hm = honorable mention, c = commendation and m = mention.) As these are English names, non-English awards must have been translated somehow, and again it would have been useful to know how this translation has been performed, especially if the user wants to know what the original source said. Is the German 'ehrende Ehrwähnung' the same as 'hm', for example? Can the reverse translation be performed reliably?

```
[Site "?"]
```

Required by PGN, but not used by the database.

```
[Date "2005.??.??"]
```

Not documented, so its relation to EventDate is not explained. As far as I can make out, they are identical, which prompts the question of why the EventDate information is included at all.

Presumably the year of the award (when one is specified — although that concept does not seem to be quite well-defined either) or the date of the publication (in case of a plain source reference), though its not clear what it means when both are indicated in the Event tag.

```
[Round "?"]
```

This field seems to have been used for some purpose, as there are 129 occurrences of [Round "1"] and [Round "2"] respectively, but again there's nothing to explain if or how this is significant.

```
[White "Bocharov=J"]
```

The composer(s) name — presumably the '=' is used as the equivalent of a ',' between family name and given name. Multiple composers are given as:

```
[White "Van_Essen=M Wohl=A Afek=Y"]
```

with a space character separating the names. (In those cases where names actually do contain spaces, they has been replaced with an underline, *e.g.* 'Van_der_Heijden=H'). PGN, however, requires personal names to be on the form "Bocharov, J", with a space following the comma, and also stipulates that multiple names should be separated by a colon, thus:

```
[White "Van Essen, M:Wohl, A:Afek, Y"]
```

This means that the HHDB III conventions will cause problems for PGN-readers that actually follow the PGN specification: such software will not be able to find joint compositions, as there is no colon to indicate the presence of more than one name, and may have problem find all compositions by 'Wohl, A', say. It may be possible to make some software-specific search, though this calls for very detailed knowledge about '=' and '_' on the part of the user, but there's no guarantee it will work. This refusal or inability to follow PGN format specification is unfortunate.

```
[Black "[=0003.31h4b5"]
```

Extended GBR-code for searching on material and stipulation. This code is explained in the *readme.pdf* file, and although there should perhaps have been an additional note that the information is stored in the name field for Black, it does not seem likely to be misinterpreted or mistaken for anything else.

```
[Result "1/2-1/2"]
```

Encodes the stipulation: "1-0" is win, and "1/2-1/2" draws.

```
[FEN "7n/2p4P/8/1k1P4/3P3K/8/8/8 w - - 0 1"]
[SetUp "1"
```

These two tags together indicate that the 'game' encoded should start from the specified position, rather than the standard array. In ordinary PGN files they are rather rare, and used only for game fragments or games at odds.

The FEN tag contains information that is relevant for a game, but not necessesarily for a study. So for instance can information about castling status (what castling opportunities remain) and en passant capture be found here, along with what side is to move.

Castling and en passant status are not explicitly stated in a study, so this information must have been added in some way: what method has been used? Again, there is no information available.

One method would be to inspect the printed solution, and add whatever castling rights it needs: if the solution relies on castling moves, the corresponding status is added. This is relatively straightforward, but has the downside that cooks (unintended solutions) involving castling may go undetected.

Another method would be to add castling status whenever kings and rooks are in original positions. This, on the other hand, may possibly be wrong for very old studies, where I believe castling is considered illegal, unless it can be

proved that it is legal.

If the information in the database is handed over to an endgame analysis program that is capable of handling castling correctly, the presence and absence of these rights is going to be interpreted literally, and so the program may not make a correct analysis, and may fail to find cooks, or find cooks where none exist. This could be a problem the user need to be aware of.

This is one of the consequences of shoehorning an endgame database into a chess database: not everything fits neatly. This interplay between real PGN information and added information need to be taken into account, and should preferrably have been documented.

```
[PlyCount "13"]
```

I expect that this indicated number of half-moves in the solution, as the PGN standard requires. Be prepared to find studies with PlyCount as high as 551 (a sea-snake by Blathy).

```
[EventDate "2005.??.??"]
```

As already has been noted above, it is not clear if this tag provides any information over and above that already encoded in the Date tag.

```
1. Kg5 Kc4 2. d6 $1 cxd6 3. d5 (3. Kf6 $2 Kd5 4. Kg7 Ke6 5. Kxh8
Kf7) 3... Kd4
4. Kf5 Kc4 5. Ke6 Kc5 6. Ke7 Kxd5 7. Kf6 1/2-1/2
```

The second part of the entry, after the empty line, contains the solution (here the first line is too long to fit the column width, and has been split over two lines), and is in some cases the solution is extended with notes. Some of those I find in the database are '{main}', '{eg}', '{or}', and '{cook MG}'. The last probably indicates some kind of fault, but as this is not documented — does MG indicate the type of the error, the person who reported it, or something else entirely? — it is left to the user to make sense of.

The source of the solution is not obvious: in the absence of any direct statement it may not be safe to assume it is the author's full solution, and possibly not even it is from the same source as the diagram. That is, the presence or absence of a particular variation may not be of any particular significance.

The PGN specification places some rather tough requirements on this information, such as requiring files to be as compact as possible (see section 8.2.1 of the specification) but again it must be noted that few programs insist on strict interpretation of such requirements, and even fewer report when they do find something that may indicate a problem. The user is probably well advised to verify that 67691 studies indeed have been recognized by the software used — anything else may indicate problems in the PGN interpretation.

For instance, in a study of Kondratjev, the following fragment of the solution can be found:

```
6. Nf4 (6. Ng5 {{or}}) 6... c1=Q
```

A {}-pair indicates a comment, and that comment ends where the first '}' occurs: nested comments are not possible. In this particular example, '{{or}' would be such a comment, immediately followed by a '}', which is not permitted according to PGN. This should, strictly speaking, produce a warning about a bad, possibly damaged PGN file, but it might be passed over in silence while the 'game' is dropped as formally faulty, or perhaps by ignoring the extraneous '}' entirely. The presence of such non-PGN data in the database is rather disturbing: it should not have been possible to produce a non-compliant PGN file at all.

There are a few tools for checking for PGN datafile correctness, and although they are mainly targeted to game scores, they can occasionally be useful also for other types of files. The comment problem described above was the only problem directly reported by the *pgntrim5* program — but as it proved to have dropped around 1000 studies during processing, it should perhaps not be safe to assume that it is the only error present.

The PGN-related problems mentioned above are with a very high degree of probability problems in the computer programs used to create the database: they don't follow the PGN specification as well as they should, and any shortcomings inherent in those programs will then be reflected in the database file. That indirectly suggests that there may be other errors present in the files.

Further technical notes

The CD is a ISO9660 standard Mode 1 Data CD, using the Microsoft Joliet extensions, and should not give any problem to read with a reasonably modern CD reader and software.

A closer examination of the CD shows that none of the Joliet extensions, such as long file names or extra characters in file names, are actually used: it would probably have been possible to increase data portability slightly by using a pure ISO CD format.

Summary

There are some shortcomings in the database, some of them quite severe. Most of the problems are related to the PGN file format, which a) isn't designed for endgame composutions in the first place, and b) is not followed as strictly as it should be. They may lead to PGN-processing software ignoring parts of the database contents, and in other cases, that searches for composer names do not work as expected.

The documentation also has a number of weak points, the most important of which is the lack of information as to the purpose and contents of the database: how is it intended to be used? what information is included? Without knowing that, it is quite difficult to decide if the presence of chess problems

among the studies is a fault or a feature. The lack of information about name representation problems may also be a problem a user finds it difficult to get past.

Apart from such glitches, I have not had any important problems in using the data with ChessBase Reader, and a couple of personal PGN utilities. The knowledge of their presence, however, makes it somewhat difficult to rely on search results from the database.

It is, however, fairly obvious that using PGN as exchange format is not an ideal solution. It's probably the best that can be expected, in the absence of an data exchange format tailored specifically to the purpose, and database software that 'knows' about endgame studies.

Source References

Gaige, Jeremy: *Chess Personalia: a biobibliography*
Jefferson, NC : McFarland & Co., Inc., 1987

Edwards, Steven J.: *Portable Game Notation Specification and Implementation Guide* (Revised: 1994.03.12)
see *http://en.wikipedia.org/wiki/Portable_Game_Notation*

Russ: *Miniature Chess Problems from Many Countries,* 2nd ed.
London : Unwin Paperbacks, 1987