**Notes on the Enigma schema**

**Copyright and disclaimer**

Please read the separate “Copyright and disclaimer” Notice.

**Background**

I first began developing this database in the mid-1990s as a way to keep track of HK-listed companies in which I was interested, storing due diligence data on their directors and advisors and my notes, often along the lines of “This company is crap, avoid”. Those notes have of course been removed from the public dataset.

After I founded Webb-site.com in 1998 (the same year as Google, BTW), I decided to complete the dataset, using all available archives of listed company documents back to 1990, and to make it available for free on Webb-site as a public service.

We have continued to maintain that dataset up to the cut-off of 31-Mar-2025, a 35 year run. I’ve spent a considerable amount paying a full-time assistant to review daily disclosures on appointments and removals of directors, supervisors, auditors and financial advisors under the Listing Rules and Takeovers Code, as well as annual reports for corporate advisors, and Government Gazette Notices and press releases for appointments to over 800 statutory and advisory bodies, amongst other things.

One day soon, Artificial Intelligence may be able to read all these documents and do as well as a human, but there are challenges – how do you know which “Chan Chi Keung” or “Zhou Li” has been appointed to a board? We always run the risk of either double-identity or mistaken identity in our database (if in doubt, create a new person).

When the database was just for my own use, it was stored internally in a Microsoft Access file. That clearly doesn’t scale, so I moved it all to MySQL as a back end, but we still use Access for editing on the Local Area Network (LAN) in a front-end application called “Enigma”. For the avoidance of doubt, my use of that name long predates the “Enigma Network” which I published in 2017 and there is no connection between them. I just like the name – ever since I read as a child about Alan Turing’s wartime work to crack the German Enigma code. In a way, I regard the entire stock market as a puzzle.

I have also developed a set of web pages (in classic ASP) which we can use to edit the database when not on the LAN, obviously with authentication required. This allows multiple editors to work remotely. It also enabled collaboration with various academics who worked to collect detailed ownership structures and data on board committees. Both of those projects are dormant, but you could take them forward.

The authentication system is explained in a separate document.

When I set up the CCASS Analysis system in 2008, I knew that this was going to be a large and quite separate dataset, so this is contained in a separate schema (see “Notes on the CCASS schema”). Somewhat arbitrarily, that schema also contains the quotes database for stock prices.

A lot of the datasets in Enigma are maintained automatically by code I wrote in VB.Net. For example, when the Immigration Department began publishing daily border movements during COVID, I added a dataset for that and code which still runs daily. These programs of course depend on the data source (often a web page) not changing its format, in which case you will need to update the code.

In another example, in the wake of the public protests in 2019, there was increased public interest in our justice system, so I found a data source on people in custody (whether or not convicted) and coded that collection into a new dataset which revealed that a record share of the prison population had not been convicted. We collect that every quarter.

I hope that my work inspires people more skilled than I am in modern database management to uncover more public data and add datasets in any area that interests them, making data available to the public in user-friendly form.

I also hope that these IT-skilled people will find ways to help those who lack such skills but want to collect data, particularly in the activist/NGO space. In any event, AI-driven coding is likely to make that easier over time.

I won’t try to document all of the Enigma tables here. You will find some guidance in the “comments” field of the MySQL tables and columns.

**The MySQL data dump**

The enigma schema is partly driven by MySQL Triggers on tables. I have therefore broken the dump into 3 files: Structure, Data and Triggers. Restore them in that order (with Triggers last), so that the Triggers do not modify the data as you load them.

**Naming conventions**

All names of people and organisations must be unique, so that when a user searches, they aren’t presented with a list of identical names and there is at least some clue as to which one they should click.

***Organisations***

For organisations with the same name, we append either their domicile (2 or 3-letter code), date of incorporation or date of dissolution if known. Our code handles the differentiation.

The organisations table has a column “namehash” maintained by a MySQL Trigger. This calls the orgHash function, which creates an MD5 hash of a normalised version of the name, created by the normOrg function. Amongst other things, the namehash is used by our SFC.vb to find matches when a new organisation gets a licence, and by HKLawSoc.vb to find matches for law firms and employers.

***People***

For people, the name pair (name1, name2) must be unique. Name2 can be null – for example some Burmese or Indonesian names are only one word. For people with the same name, we have code which runs on data entry (either via Access or the web) and extends names when necessary, either with a unique identifier (such as an SFCID) or a known date of birth (which may be approximate to the nearest month or year).

For people with both English and Chinese given names, we put their English name first wherever possible, for example “Chan, David Chi Keung” rather than “Chan, Chi Keung David”. This helps to reduce double-identities in the database, but they happen.

For females who are married and have taken their spouse’s family name in addition to their own, we put the maiden family name last. For example, if “Cheng, Yuet Ngor” marries a Mr Lam and takes his name, then her name would be “Lam Cheng, Yuet Ngor”, not “Cheng Lam, Yuet Ngor”.

The people table contains a simplified version of each name (dn1, dn2) which is not unique and allows for quick searching as well as being used by our code when auto-differentiating names. This is maintained automatically by MySQL Triggers which strip off the extensions. There is also a 5-letter “namestem” field in the people table which is maintained by MySQL triggers and is used by the Access front-end on the LAN when typing a name into the people form, otherwise the search would have to combine 2 columns.

On Webb-site, the dbeditor folder contains a full suite of editing pages which handle all this, as does the code in the Access front-end. When we find a duplicate person, there are forms to combine the records, as long as the data (such as domicile, organisation type, or date of birth) are not inconsistent. On combination, our code handles all the consequent amendments to other tables, of which there are many! There is also a “mergedpersons” table so that URLs with a removed personID are forwarded to the merged personID.

**Corporate officers**

These are contained in the directorships table. I never changed its name, but now of course it includes supervisors, senior managers and even a few employees, UK/Irish hereditary peerages (just for fun), Bauhinia awards, judges, statutory boards, councils etc.

*Rules*

* A person cannot have more than one directorship in a company at a time, so don’t enter data that conflicts.
* We use the right-open interval convention for time periods, so the period includes the “from” date but excludes the “until” date, which is the first date on which a person no longer holds a position. So when a person changes titles on a board (e.g. from Finance Director to CEO) then the resDate from the first role is the apptDate for the new role. They don’t take a day off in between!
* Each entry has a positionID corresponding to the positions table.
* The positions table contains all the known position types – we occasionally add to this when someone has a weird position – for example, the latest position was INED and Alternate Chairman. Generally, the sequence of titles in the “posShort” and “posLong” descriptions is by seniority and then whether they are a Founder or Co-founder. By default, a CEO or President is assumed to be a director, otherwise use the one with a “(Non-dir)” extension.
* The status column in the positions table records whether a position is Executive, Non-executive, Independent Non-Executive or Unknown. This is used in Webb-site to show analysis of INEDs, for example. See the status table.
* The rank table records the level of the position, ranging from the Supervisory (0), Main Board (1) and so on. See the rank table. Beware that “rank” is now a reserved word in MySQL, so you’ll have to quote it when using the column name.
* Don’t change data for people with a position ID of 394 (“Rep”) or 395 ( “RO”) as these are SFC licensee positions computed automatically by our SFC.vb code when it checks the SFC licensee database daily.
* Some positions are maintained by our Law Society collection code HKlawSoc.vb. These are assigned an “LSrole” in the positions table.
* All directors and members of UK companies/ limited partnerships are auto-maintained by our UKCH.vb module, which runs continuously against an API – if only HK’s Companies Registry was so transparent!

**Adviserships**

The adviserships table includes 2 types of advisers – a continuing role (for example, auditor) or a one-time role (for example, an Independent Financial Adviser on a deal). The latter does not have a removal date.

The advisers table contains a list of organisations which have been advisers.

The roles table contains a list of roles of advisers and whether they are one-time or continuing.

For advisers taken from annual reports, we use the date on which the full-year results are published, which is usually also the date printed at the end of the “Directors’ Report” in the annual report. So a bank which first appears/disappears from a company’s report is appointed/removed on that date.

For deals, for IFAs appointed under the Takeovers Code, we take the appointment date as announced, because the role begins immediately. For other roles in deals, we only give credit on the date of the shareholder circular which contains their advice (such as a Listing Rules IFA or a Reporting Accountant). If an IFA is working under both the Takeovers Code and the Listing Rules, then they get 1 credit for each role.

**Ownership analysis**

This was a project by a team of academics, originally out of HKU, to analyse the detailed ownership structure (including control and economic interests) in HK-listed companies before and after certain key changes in the Listing Rules which required the number of INEDs to increase from 2 to 3 (from 30-Sep-2004), and then later for the proportion of INEDs to be at least 1/3 of the board (from 31-Dec-2012). What kind of ownership profiles were ahead of the curve, and what kind made changes at the last minute?

The team went carefully through thousands of [disclosures of interest](https://di.hkex.com.hk/di/NSSrchMethod.aspx?src=MAIN&lang=EN&in=1&) and annual reports, noting the layers of ownership via partly-owned entities, trusts, partnerships and so on. I guided them on the interpretation of the byzantine forms.

I wrote code to compute the economic interests of the ultimate controllers, which often differ from their voting power because of pyramid ownership structures. I wrote that code years ago in Visual Basic for Applications (VBA) in Access. Unfortunately, I never got around to converting it to VB.net after I got cancer, so you will find it in the “HKU” module of the Enigma access file. The top routine is profileSnaps, which takes hours to run, as it has to iterate up and down the ownership tree using a recursive algorithm.

Personally, I didn’t expect the researchers to find much, given that most INEDs are nominated and elected by controlling shareholders (who are often also the Executive Directors) whom the INEDs are supposed to monitor; and given that very few HK-listed companies don’t have a controlling shareholder with a stake of 30% or more. See this [presentation](https://webb-site.com/codocs/DMW171110.pdf).

The tighter the control, the less likely to me that the company would comply early and actually take INEDs seriously – very few companies do. Anyway, this project did result in a large dataset which we retain in the database, most of which did not appear on Webb-site because we decided not to publish outdated shareholdings in listed companies – but we do display holding structures of intermediate companies.

You might want to take this forward with more regular and recent data collection.

The relevant tables are:

* ownerprof (the ownership profile of each company on each date, computed in VBA from the ownerstks table)
* ownerstks (ultimate controllers of beneficial shareholdings, computed in VBA from the ownership tree, with their economic interest).
* ownertype
* sholdings (which contains a lot of other shareholding records that I’ve added over the years).

In the Webb-site dbeditor folder are pages related to this project:

* holding.asp
* holdinghist.asp
* ownerdets.asp

**Committee Analysis**

Another project, by a related team, was to analyse the structure and attendance of Board Committees (such as the Audit Committee, Nomination Committee and Remuneration Committee), using data from annual reports. This was challenging as companies adopted different interpretations of the rules for reporting attendance when someone joined or left during a year, or even different interpretation of a “year” – was it the financial year, or the period between annual reports or results? In my view, this was due to poorly-defined Listing Rules – the regulators really don’t think things through.

These data never appeared on Webb-site as they were somewhat incomplete. You may wish to take them forward. The relevant tables are:

* comeets
* comex
* compos
* coms

In the Webb-site dbeditor folder are pages related to this project:

* coms.asp, coms.js
* comeets.asp, compos.asp and comattend.asp – these 3 are response pages to AJAX calls from coms.js

For both of these projects, I built a logging system for the data-entry team to record their progress and outstanding problems. These are contained in the snaplogs table. They are displayed in a page in the dbeditor folder:

* admlogs.asp
* snaplog.asp

**Content Management system**

I developed my own content management system for the articles and reports on Webb-site.com. Often these are brief reports of announcements by third parties or of judgments rather than full-form editorial content. The stories table contains the details of each story including its source, title, story date-time, publication date-time and summary.

The source names are in the sources table.

If sourceID=1 (Webb-site.com) then the URL is just the name of a page in the articles folder on Webb-site. There is an include-file in the top of each article (topart.asp) which looks up the page in the stories table and then produces the summary, title and date. There is an include-file at the bottom of each article (tailart.asp) which produces a list of any people, organisations and topics tagged in the article.

Each story can have a second URL with text. If either URL is to a supporting file on Webb-site then the URL begins with “..\” which tells our pages that it is an internal link. For example, radio shows are in the “..\audio” folder (including my old Backchat appearances), and documents are in the “..\codocs\” folder.

There is a separate table, sfcnews, which contains data collected from SFC announcements, including the English and Chinese content. That is so that we can archive them as originally published and not as subsequently amended, withdrawn or redacted. We used to do that by creating PDFs but a few years back the SFC started making a JSON package available inside the news page, so we use that. The contents of the press release appear in a letterbox in the “artlinks.asp” page with our links to relevant topics, people and organisations.

For other sources, if they are regulatory or law enforcement then we usually convert the announcement to PDF and put it in codocs. If we are linking to a copyright publication such as a media website then of course, we cannot keep our own copy. This does of course result in broken links if the media site gets shut down (as many have) or reorganised. Orwellian memory holes are everywhere in HK, but you might find original copies on <https://archive.org/>.

The people and organisations tags for each story are contained in the personstories table.

The topic tags for each story are contained in the storytags table.

The tags are in the categories table. The tags are in a hierarchical structure contained in the cattree table. This tree is displayed in the default.asp page in the articles folder.

When we want to publish a story, we either use the Story form in the Access application, or the corresponding page in the Webb-site dbeditor folder, story.asp. This of course requires authentication; all stories in the past were published by me, but there’s no reason why a team of editors could not use this to manage a more frequent publication system.

When fetching a story from the SFC, just pick SFC as the source, enter the refNo parameter for the URL (without the rest of the URL: for example, 25PR49 is the 49th press release in 2025), then hit “Fetch SFC PR” and our code will work its magic.

**SFC licensee database**

We run SFC.vb daily to collect the latest licence data for SFC-licensed individuals and corporations. We do not collect data on those licensed under HKMA, because it has historically used multiple identifiers for the same individual and also they do not use the same identifiers as the SFC, making it impossible to track people as they move between the two licensing regimes.

Tracking who’s who in HK would be a lot easier if all public bodies just used HKIDs or (for those without an HKID) passport numbers and put them on the register. [Your HKID is not a password](https://webb-site.com/articles/identity.asp) – it is an identifier, not an authenticator. Also, HK badly needs a single financial services regulator for those working in banks, brokers, MPF and insurance (including agents), to provide a consistent approach and prevent regulatory arbitrage. Prudential regulation of banks and insurers (for capital adequacy) is a separate matter that should be handled by a dedicated regulator.

Our software analyses the separate “activity” licenses of people at each firm. If they are both a Responsible Officer in one activity and a Representative in another, then RO outranks Rep, so we record them as an RO in the directorships table.

**Law Society database**

We run HKLawSoc.vb daily to collect the public data on the Law Society register. As there is some lag between an appointment or departure at a law firm and the LS publishing the appointment or removing them, we only display data to the nearest month. Sometimes (particularly around year-end) people disappear from the register and reappear after they have paid their membership fee, so they will drop in and out of our records. Not our problem.

We also track those who are working in private practice at employers (the SFC and HKEX have lots of lawyers, for example). This does require some manual intervention to link the personID of the lsemps table to the personID of organisations, so you will need to keep any eye on that. We don’t link employers if they are just “C/O” (Care Of) a law firm as these are almost always overseas law firms operating under a single brand, so it would be confusing.

**The public interest in the SFC and Law Society databases**

The LS and SFC tracking is undoubtedly useful to the public in several ways:

* Potential clients need to know whether the firm has a high turnover of staff, or long-term staff who can be relied upon.
* Head-hunters need to know which firms are shrinking/growing and potentially have staff to lose or need new staff.
* Potential employers can look at how many and which past firms a person has been through, to cross-check CVs.
* Firms like to have bragging rights about how many staff they employ (see the rankings pages on Webb-site), and so on.
* Journalists at leading media have used our database in the past to analyse trends in the financial markets – are foreign-owned firms shrinking? Are mainland-owned firms growing?

DMW

2025-04-30