A database for Dutch-Australian migrants – setup

Rik Hoekstra, Marijke van Faassen
Huygens ING

1. What do we have, where we stand

Ia Dutch National Archives – emigant cards

The point of departure is the emigrant registration cards from the Dutch consulates, now available in the Dutch National archives under http://www.gahetna.nl/collectie/index/nt00335. It is analysed at length in NL_AUmigrantdatabase_20140123.docx

The database consists of 50,000+ cards about migrant units. The cards themselves have been scanned and there is a database with core data from the cards. It has 51,506 records.

Core data are:

- id
- last name
- interpositions ('van', 'van der', 'de', 'ter' etc.)
- initials
- date of birth (45381)
- means of transportation to Australia
- name of vessel (ship name or name of airline)
- date of departure (19872)
- date of arrival (43558)
- place of registration/arrival (Australia)
- archive link (inventory number)

The original database consists of a single file available in csv and (my)sql format.

Proposal: include table *as is* into the database

Ib Scans

There is no direct link to the cards; it has to be provided by hand. Images of the cards (2 per card) are organized conform the original 51 card file drawers, that had an alphebetic order, though in handling the cards they have gotten into some disarray. Currently we have made a link between database records and cards for 8 drawers; they are organized in spreadsheets. As there is a one to many relation between the records and the scans (at least 2 cards per records, sometimes more), they should be organized in a linking table.

– Recordnr
- imagenr

Ic Sample for analysis

Intern Wouter Schalekamp has drawn a 1% sample of the cards for further analysis and made a further categorisation of the additional information on the cards and the extend up to which they can be structured. Typologies are elaborated in Schalekamp report. The sample is in a spreadsheet; the organization of the fields may (fields in Dutch with English translation).

- Kaart nr. (card nr)
- PersoonsID (person id)
- Naam (name)
- Geboortedatum (date of birth)
- Type vervoer (type of transport)
- Naam Vervoer (name of vessel)
- Vertrek (departure date)
- Aankomst (arival date)
- Kaartenbak (card file drawer)
- Model (emigrant card model)
- Info op kaart (information on card: long, short)
- Unit (number of people in migrant unit)
- Samenstelling (type of migrant unit – single, couple, family)
- Datum inreis (date of arrival)
- Datum na-/terugreis (date of return)
- Godsdienst (religion, limited list)
- Adres (address in Australia)
- Life course NL (life course in NL, short, long)
- Life course AUS (idemAUS)
- Opvangnetwerk (counseling/reception network)
- Opmerkingen (remarks)
- Genoemd kantoor op kaart (office (consulate) indicated on card)

It is not impossible that the cards contain additional structurable information. The structuring and the categories Schalekamp used proved to be useful and could be used as a starting point for further information analysis in the database.

Proposal: a) use structuring of sample as starting point for further information b) include the sample into the database, perhaps with a marker to indicate that it is the original 1% sample.

II Australian immigration files

IIa Autralian National Archives – RecordSearch

Based on the holdings of the National Archives in Australia. These concern the same people, and are the counterpart of the Dutch emigration archives. The archives are searchable with the site's recordsearch.
Searches with RecordSearch yield result lists in HTML, all with the same overall structure:

- Select [not relevant]
- Series no. [archive series of the record]
- Control symbol [unclear, but often name of the migrant]
- Item title [record description with varying degrees of elaborateness]
- Date range [range of dates for NA accessibility, not useful]
- Digitised item [in html: link to digitised files]
- PDF [idem for pdf version of digitized files]
- Item barcode [main identification number for archive; also used for digital files]
- Format [?, not relevant]

HTML files are tables, that can be turned into comma separated (csv) files easily (and even imported into a spreadsheet program.
Procedure is to

- select a transport vessel name (either ship or aviation company) to narrow down passenger selection,
- save the html results (sometimes many result lists depending of number of passengers travelling onboard a ship in different voyages),
- convert html to csv
- concate separate files into one file per ship (may add up to 20,000 records/individuals)
- match the records with the Dutch emigration cards

An iterview with Marjory Bly made clear that there is no better database available behind the screens than what is available through recordsearch. Changing transciption policies over the years, local differences of National Archives' offices etc have made record description inconsistent. Nonetheless, it has proven possible to match at least one quarter to one third of the Australian records to the Dutch with a crude matching algorithm, that could be improved and fine tuned for better results.

Proposal:

- include the csv files into the database *as is*
- as the files are as yet not complete, include shipnames-files into the database incrementally
- perform matching between the dutch and australian records offline and include results to the database as a linking table.

IIb Australian National Archives – Files

The files of the records from the National Archives are sometimes digitised. Autralian archive

policies open them up for general consultation once they have been digitised. The files contain varied information from the immigration and/or nationalisation files, usually including a pass photo, from one or several family members, the original application form with appendices, often including the medical examination files from the Netherlands and at times letters of recommendation from previous employers.
We have not yet inventarised the possibilities for structuring this information.

Proposal: include the cumulative csv files per vessel into the database.
Links between Dutch and Australian records stay in a separate table
Link the Dutch and Australian records offline and include links incrementally

IIb Passenger lists

The National Archives also contain passenger lists (often also called 'nominal rolls') for the ships per voyage. These can be actracted from the results of the 'ship' searches mentioned above either programmatically or by hand. Passenger lists are often digitized by the archives and contain information about all passengers on a ship. They may be used to help determine (by coparison) whether there are structural gaps in either the Dutch or the Australian migrant records.

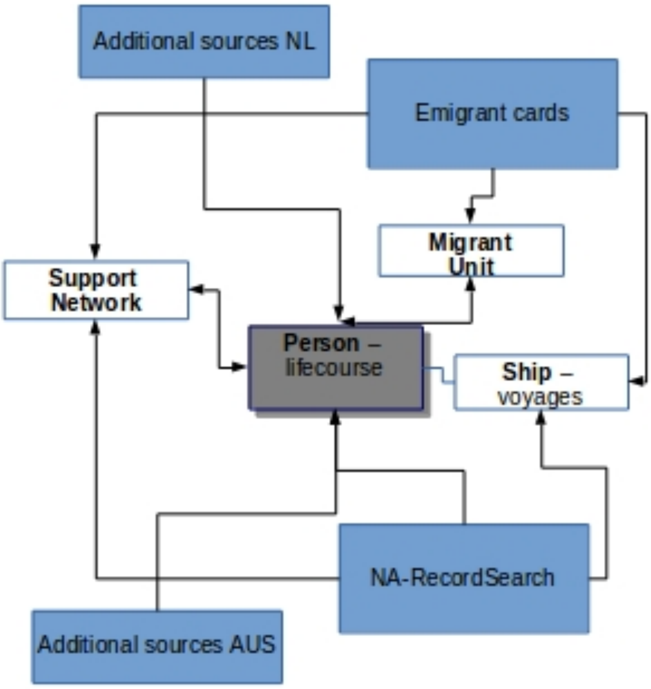
Proposal: download relevant migrant files and link them to the database. Alternatively they could also be linked through a hyperlink to the Australian Archives site (which is more vulnerable to changing address, but also more convenient)

III Other archives and sources

We have examined some other files, as additional sources of information, including the files of

- the Catholic Documentation Center in Nimwegen (containing files from the Catholic Labour Center that selected migrants)
- Image archives in both the Australian Archives and from the Dutch Organisation for Migration in the Dutch National Archives
- …

2. Database design



See schema for conceptual setup

Entities:

*Perso*n – separate entity, initial data come from structurally from emigrants cards, but basic entity is a migrant unit. In database person data taken from primary migrant from unit. A 'person' is primarily a life course following the general model of BioDes, where all events in a person life is modelled as event. For briefness basic data (date of birth, day of death) is are properties of the person. The lifecourse is a table with events, that include the migration itself, but may comprise many more life events, depending on the information available.

Person

- id
- last name
- interpositions ('van', 'van der', 'de', 'ter' etc.)
- initials

Person_description

- person id
- religion
- occupation (though this may also be an event…..)

event

- event id
- event type
- beginning (date)
- beginning_date_type (conform datable format)
- end (date)
- end_date_type (conform datable format)
- place (may be linked to normalized list of placenames)
- description
- source
- remarks

voyage:

- vessel id
- event id

vessel:

- id
- name
- description

Archive document

- id
- document type (emigration card, immigration file ...)[1]

[1] May be specified for card models

- archive link (inventory number)
- remarks

institution:

- id
- type (example: central govt, local govt, church etc)
- name
- country (ex: NL, AUS, INTL}

Additional tables:

Image

- id
- name
- location (for instance: url)

location:

- place id
- lat
- long
- name
- [variants]

linking tables:

archive-images:

- archive document id
- image id

person-document:

- person id
- document id
- unit type (for migration cards; maybe in separate table)

- remarks (for example about variations in the document with the person record)

person-network:

- person id
- person2 id
- relation type (maybe normalize in separate table?)

person-institution

- person id
- institution id
- relation type (employer, assistence, ...)

Elaboration:

Life course

Is a table with events. Emigration consists of two events:

event type: departure

- person id
- event type (here: emigration-departure)
- event description id (linked to voyage)
- start (date): date of departure
- start place: place of departure

event type: arrival

- person id
- event type (here: emigration-arrival)
- event description (link to voyage)
- end (date): date of arrival
- end place: place of registration/arrival (Australia)

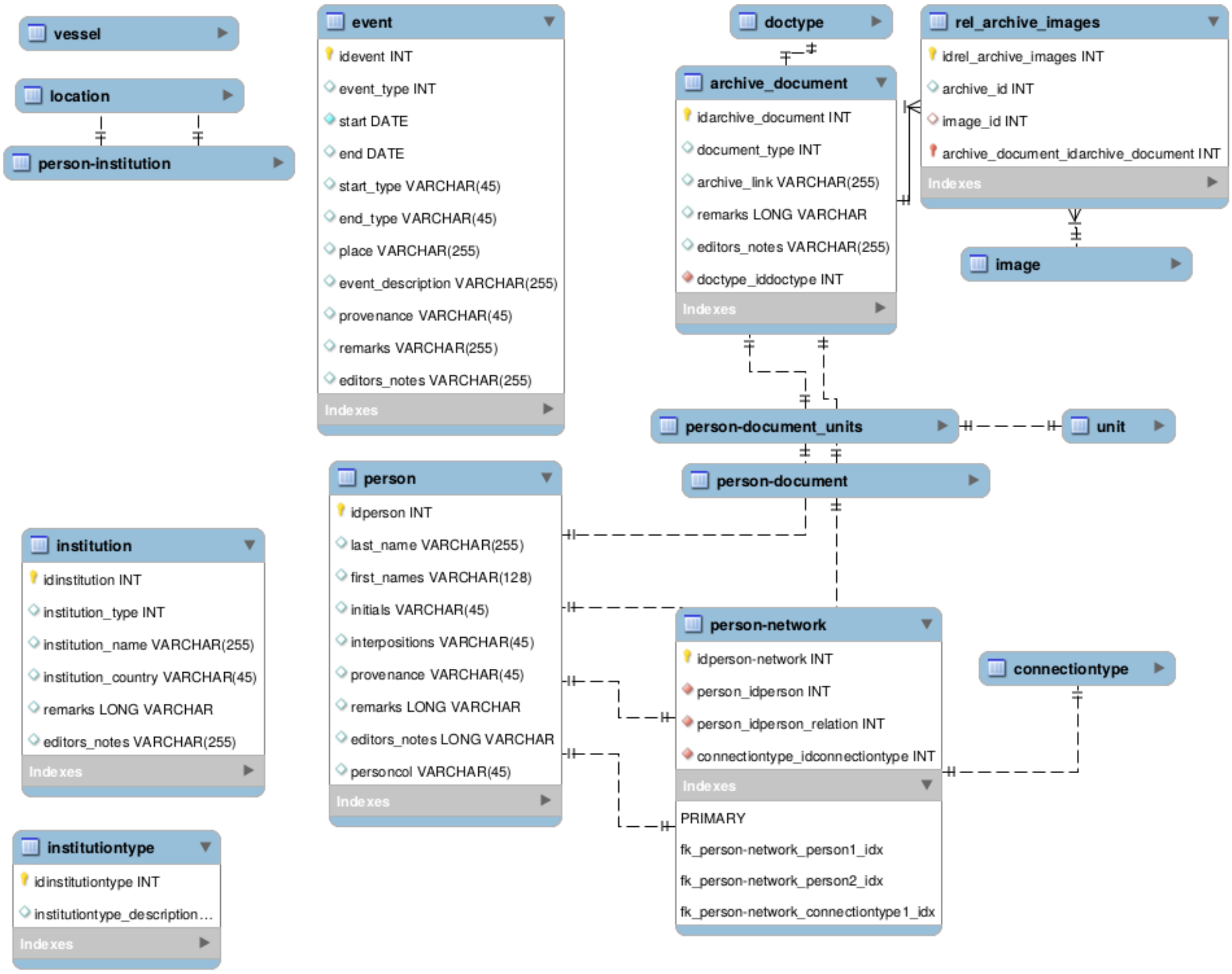
N.B.

1.I have chosen for one person not different variants which might well be true because of

differences in archive records. This being a reference database, the assumption is that differences between different archive records will be recorded in the person-archive relation.
I hesitate whether to make an advanced description type for migrants containing their occupation, religion etc. In the end I decided it would not be a good idea, because nearly all properties should be modeled as event (even religion, especially if this is taken from possibly contradictory archive files)

2. In principle it is a good idea to include a field for (more or less structured) remarks and a second field for (unstructured) editor notes. I have not always included them in the design. The same is true for source/provenance

3. For items like place names or migrant units it may be desirable to normalize types in separate tables.



(unreadable db schema follows)

Normalization of the masterdatabase

Following the design of the database, the data were migrated from the spreadsheet (the *master database*) to a database, using a python script. The script does not fill the database at once but produces sql files with which the tables in the database can be filled.

In fact, the process has three steps:

- normalize the master excel sheet into separate sheets (*normalizer.py*) for
    - persons (*person*)
    - vessels (*vessel*)
    - voyages (*event*)
    - cards (*card*)
    - links people-voyage (*link)*
    - links vessel-voyage (*link*)

  this python module is able to write separate csv files

- translate this to separate sql files for each table (*xls2sql.py*):
    - sql_ves2sql_gen.sql (*vessel*)
    - sql_evt2sql_gen.sql (*event*)
    - sql_rpe2sql_gen.sql (*link person event*)
    - sql_cp2sql_gen.sql (*card*)
    - sql_vr2sql_gen.sql

  sql was meant for a mysql database. For data structure see above. N.B. no indexes are included in the database, but are important for performance reasons

Linking the cards to the scans is a separate action taken care of in a different python module (xsl2scans.py). This yields links between cards and their scans - usually 2 scans, front and back, but sometimes more when extension cards were used for a single migrant). For workflow reasons, the master sheet was split up for linking cards to scans into separate files per 'box' (*bak*).
Each box corresponds to a directory with images. Each box contains approximately 2000 images (or 1000 cards). In the master database there is a field for box number, but it does not quite coincide with the actual directories, that were created in the course of the National Archives scanning project.
As the image numbers are renumbered per 'box' directory, referring to the images depends on both box number and image number. Therefore, it has to be clear to which box the sheet refers. Eventually, this has to be included in the database itself, but as long as separate sheets are used per box/directory, it is acceptable to include the box number in the sheet name. A naming scheme has not yet been worked out.

The 'box'-sheets are derived from the master sheet, but may be altered as they are elaborated. Obviously, this is true for the image numbers that are entered but also for mistakes that the people working on the sheets come across along the way and that may be corrected.
Moreover, all boxes contain some 'lost' or 'extra' images, that are not included in the database, or that contains scans of for example business cards. These extra images cannot be included into the

sql database in the same way are the bulk of the images, but should be separated and manually linked in a separated sheet. There is no procedure for this yet.
In the end, the modified data from the box-sheets should replace the original data in the master sheet, in order to keep the corrections of the researchers working on them. We have not yet decided upon the procedure for this either.

The normalizations described above take care of most of the normalization of the original master sheet, with some caveats:

- the emigrant cards describe emigrant units instead of individuals. We realized this all along, but for data normalization this means that the migrant unit and the individuals in it should actually be separated. We have not done that yet, and treat the unit as the name it is described with, which refers to the main migrant in it, often but not always the head of the family. This is no problem at this stage,but requires reconsideration when the database is further elaborated, for instance when the migrants from the Australian National Archive database should be linked
- in some cases, migrants or migrant units appear on different cards, for example when a card was forwarded to another consulate, or when migrants we repatriated and then remigrated. Except for the sample that was elaborated, this has not been included in the database (and at this stage the sample remains out of it), causing no immediate problems. For the sample we devised a procedure in which migrant number (or actually the migrant unit number) may come to refer to different card numbers – in the master database there is a different number for each card *and* each migrant (unit).

  This solution is good enough for keeping track of the differing information about migrants on different cards. However, the personal information on the migrant (or actually, the migrant unit, see above), is kept as personal information. Eventually, the personal information should be kept with the card, as there may be hard to solve discrepancies between different card. However, we should also devise a way to decide which information about migrants should be considered canonical (or preferred) information and as the personal data. When actual personal files are included, we should use that information, but if not there is no obvious to decide that. However, it may also be acceptable to explicitly decide that differing information should be solved anyway and that contradicting information in the database is not acceptable.
- We have included only links from cards to (one, two or more) scans, but refrained from describing the scans themselves in the database as this appeared superfluous.