Explanation of the person file from the Biographical Portal of the Netherlands

Dr. Rik Hoekstra
Biografisch Portaal
Huygens ING – KNAW
July-October 2012

The File
This file is an extract of the data of the Biographical Portal of the Netherlands (www.biografischportaal.nl). The Biographical Portal is "an online collection of reference works and data sets currently scattered over the internet, containing biographical information on notable persons in Dutch history, from the earliest times until up to present," with data about people who lived or were active in the present-day boundaries of the Netherlands, including its colonies. In principle, all data about the people come from the collections contained in the Portal, although in some case there have been made some corrections. This is especially the case where there were conflicting data from different sources, or when data were not available in structured form and had to be entered by hand on basis of the text of the resources. For a list of resources contained in the Biographical Portal see http://www.biografischportaal.nl/en/about/about/collecties.

The Biographical Portal is a work in progress, to which new sources are added on a regular basis. At present, it contains 80.420 unique persons, described in 125.438 biographies (some persons are described in more than one source). The file is prepared from a snapshot from May 2012. It contains data about 13.308 persons, filtered out from the total collection .

The file consists of a lists of persons with their
- year of birth,
- year of death
- age
- sex
- category in the Biographical Portal

and some aggregates in separate sheets. No effort was made to make all possible aggregations. This description is meant as a description of the data, its selection and an estimate of its reliability/representativity.

Selection and remarks
For many people, especially those from a more distant past, either no exact dates of birth and death are known, or the sources do not contain them. Moreover, all sources are compiled by humans and contain mistakes of various kinds. Apart from the normal corrections of the Biographical Portal, no effort has been made to correct errors. In the file, only persons have been included, for who a year of birth and a year of death was known and who had been given a category by the portal. Exact birthdays and days of death are not included in the file, as they are either not available or contain many

assumptions and uncertainties. Years of birth and death are not always certain either, but the margin of error is smaller here. It may also be assumed that for a large dataset positive and negative errors cancel each other out.

Obvious errors stemming from typing errors (negative ages, ages larger than 110) were deleted. The file contains people born after 1499 and before 1913. For people from the final years, this may perhaps mean a slight bias in the data for people who died young because (1) people who were born in the first decade of the 20th century and lived longer than 100 years have not yet appeared in biographical sources and (2) even if they were not over 100 years, the biographical sources included always have a time lag for the people included. This effect is probably small, and anyway many people are missing from the selection for other years for various reasons.

While I have not tried to determine which sources the people in this selection come from, it is probable that most of them either come from the sources of the highest quality (that put most effort in finding dates)[1] or are well known and whose life data are well known.

Categories
Categorization in the Portal is currently rough. This means that boundaries of the categories are not always clear and sometimes different categories are combined into one. For example, the 'justice system' (Dutch: '*rechtspraak*') contains both judges and criminals. For an explanation see http://www.biografischportaal.nl/en/faq/#6. No effort has been made to make further distinctions in the categories. People for all categories have been included into the file, with three exceptions. The category of 'plastic arts' (Dutch: *beeldende kunst* - comprising painters, sculptors and graphic artists) was omitted because they were available from other sources. I have left out the categories *radio and tv* and '*sports and leasure*' (*sport en vrije tijd*) as they only contained people from the 19th century.

With regards to the categories, there is another complicating factor, as in the Biography Portal some people were included into more than one category. Therefore, some people appear in several categories. This is the case for at most 10 percent of the population. I have chosen to keep these duplicates because from an age point of view, they only have a small effect on the average ages. Moreover, I judged that the alternative of confining them to just one category would mean a further reaching interference in the selection than leaving things as they were.

Representativity
Coming from compilation of biographical resources, the selection of people in the the Biographical Portal (or its sources) is not random or representative per se. In order to be described in a biography, a life has to have at least some biographical interest and some basic facts should be known. The team of the Biographical Portal has tried to balance the composition of the people as much as possible, but, firstly, it is dependent on the collections it includes and secondly, the Portal is a resource under construction to which new sources are still added. This means that there is a general tendency to

1 These include the *Parlementair Documentatie Centrum*, *Digitale Bibliotheek der Nederlandse Letteren*, *Biographical Dictionary of the Netherlands 1880-2000 and 1780-1830; Online Dictionary of Dutch Women*; *Biografisch Woordenboek van Socialisten en de Arbeidersbeweging* and the *Compendium of office holders and civil servants 1428-1861* and some regional sources

favour the healthy and people that stand out: anonymous victims of war, plague or poverty will not appear in a biographical resource. The population of the Biographical Portal is therefore decided more elitist and more rich and famous than the population at large, even if there are also many people included from the medium strata of society (performing artists, industry, and the selection in this file is by no means indicative for life expectancy of the population in general. However, this selection does make it clear that, contrary to popular belief, not everyone from the 16th through the 19th died at an early age and that there were always people who got older, sometimes a lot older.

It is impossible to answer the question whether the selection of the Biographical Portal is representative for the elite stratum of the population. For those who were active in an occupation (here expressed as categories), there is no obvious reason why there should be differences in age composition between the total population and that contained in this selection. The category where we might expect a difference is that of *nobility and royalty* (Dutch: *adel en vorstenhuis*), where it could be argued that it was 'easier' for a noble person to 'get himself described' in a biography than for someone from another category. However, analysis of the age distribution shows that in this respect there are no big differences between this particular category and the total population of the file.

There is an uneven distribution of people over the ages and over the categories. This is a result from selection criteria in the sources, reflecting contemporary interests or 'objective' criteria. Selection are on basis of reputation. Perhaps these *reputed elites* (term of Victor Karády) also reflect (elite) distribution in reality, but at present we have no means to determine this or even the slightest indication as to what proportion of society the sample represents. Only in the case of the *Parlementair Documentatie Centrum*, **all** ministers and members of parliament after 1815 are represented, making this a very reliable source for that group.

Distribution over the centuries is also uneven. In total, the smallest number of persons is from the sixteenth century and the largest from the 19th century. From the 20th century only 12 years were included, so this does not really count. Per category, distributions per century diverge even more; this should be seen as an artefact of the selection and not as indicative of distributions in reality.

Conclusion
All in all, this file gives a rough indication of ages for certain (mainly) elitist and intermediate groups from the Netherlands in the sixteenth through the nineteenth centuries. It is emphatically **no** indication of the age distribution of the population at large. However, for the categories included (imperfect as they may be) it is the largest sample available, leaving aside all remarks about selection. there is no reason why this sample should not give a representative indication of ages for the elite.