УДК 930:004

ACCESS TO CENSUS MICRODATA AND AGGREGATES[1]

Thorvaldsen G.

Norwegian Historical Data Centre, University of Tromsø, 9037 Tromsø, Norway, gunnar.thorvaldsen@uit.no

With the digital computer came a revolution in access to census data in many formats. Many statistical agencies have made their statistical publications from censuses available as scanned images on CD-roms or via the Internet. Increasingly, however, researchers prefer to compute their own aggregates from microdata on the individual level, since this provides more fine-grained statistics, can be easier to compare over time and avoids ecological fallacies. Fortunately, most archives have understood that it is sound source protection to spread copies and most statistical agencies have understood that their own economic interests should not block researchers' access to the microdata. All together, these databases contain in excess of two billion individual records, which can be analyzed online or more thoroughly after downloading selected parts. The optimal solution is to link census microdata with vital registers and thus create population registers.

Key words: census, microdata, Internet, record linkage, population registers.

Until the 1960s, the only form of access to information from censuses was on paper or microfilm, the latter relevant for copying census manuscripts, books being the normal way for each statistical bureau to publish the aggregates. Main variables were also available in comparative international volumes, made available by Mitchell since 1978[1]. With the digital computer came a revolution in access to census data in many formats. The simplest to use are the search engines popular among genealogists for finding information about individuals; these are also used in academic research to identify groups of persons under study. Such research will usually cater for statistical information, including the scanned versions of the above-mentioned volumes with aggregates, which are now more easily available via the Internet than the book versions in libraries. Digital aggregate data was first collected as variables about administrative units such as municipalities, often combining census data with information from other sources such as vital registers. The many border changes of these units create problems if studying development over time, however.

This is one of the reasons why access to individual level census data has become so widespread for statistical research both historically and for studies of the present-day situation. When such materials are younger than

[1]The work was supported by the RSF, grant No. 16-18-10105 "Ethno-religious and demographic dynamics in mountainous Eurasia around 1900. A comparison of the Urals and Scandinavia".

72 to 100 years – depending on country – they must be anonymized by removing names, birthdates and addresses, and de-identified by making it impossible to identify anybody indirectly. Older, historical censuses can be used as complete versions, however, and in this sense the work to spread nominative census taking from the mid-19th century still carries fruit. The digital versions have been encoded so that accidental differences in spelling etc are removed, variables are constructed for the study of family structure, the formats are harmonized with respect to the coding of variable values and the record structure is as similar as possible [2]. This work has been carried furthest by the Minnesota Population Center in order to facilitate comparative research across time and space. Here all the surviving nominative US censuses from 1850 to the present have been made available either as full-count or representative samples, with or without identifying information in the IPUMS – International Public Use Microdata Samples [3]. Another major database is the IPUMS-International with census samples of households from countries on all continents after the 1950s. Among the exceptions are countries where the statistical agencies have not yet understood that the significant advantages of the microdata versions far outweigh the unlikely risk that users can identify anybody in the samples and anyway must sign a declaration of non-disclosure. Thus, the problem is not differences created by the lawmakers, but rather that similar laws of statistics are interpreted differently. The third major integrated database is the NAPP – North Atlantic Population Project with full-count censuses covering North America, Scandinavia and the UK from 1801 to 1910 [4]. Such compilation of joint resources enable the transcription or transfer of an increasing number of censuses into full-count digital sources rather than samples, enabling also the detailed study of small population groups. All together, these databases contain in excess of two billion individual records, which can be analyzed online or more thoroughly after downloading selected parts [5].

In addition to earlier in-house transcriptions, the Minnesota Population Center collects its census materials in cooperation with census bureaus, genealogical companies and population researchers in most countries of the world. Among the partners on the historical censuses, mainly from the 19th century, this book has focused on North America and Scandinavia. Like in the US, the start was made by researchers who turned source materials into individual level census data with specific research questions in mind. Increasingly, the projects turned into specific infrastructure undertakings focusing on the transcription, encoding and standardizing of census materials. These databases are then built with an eye to making historical census data available both for their own research and for other researchers. In Canada, representative samples or full-count versions of all the censuses from 1851 are now available, either for downloading or analysis via specially protected networks in so-called data enclaves. Norway has put five full-count nomina-

tive censuses online for the period 1801 to 1910 while those from 1960 onwards are only available as deidentified files sold by Statistics Norway to bona fide researchers. In a project to build a historical population register from 1801 onwards, work is now underway on scanning the remaining paper-based censuses in order to have them transcribed for integration through record linkage together with the nominative vital records [6].

References

1. *Mitchell B. R.* International historical statistics: Europe 1750–1993 / B.R. Mitchell. London: Macmillan, 1998.
2. *Thorvaldsen G.* Номинативные источники в контексте всемирной истории переписей: Россия и Запад // Известия Уральского федерального университета. Серия 2: Гуманитарные науки. 2016. № 18(3).С. 9–29.
3. *McCaa R.* Thanks to 70 years of Inter American Statistical cooperation, the world's largest integrated census microdata dissemination site www.ipums.org/international // Estadistica, 2013. № 65(184). P. 31–45.
4. *Roberts E.* The North Atlantic Population Project: An Overview // Historical Methods, 2003. № 36(2). P. 80–88.
5. *Ruggles S.* Big Microdata for Population Research // Demography, 2014. № 51(1). P. 287–297.
6. *Thorvaldsen G.* Using NAPP Census Data to Construct the Historical Population Register for Norway. HistoricalMethods. 2011. 44(1). Pp. 37–47.

ДОСТУП К МИКРОДАННЫМ ПЕРЕПИСЕЙ И АГРЕГИРОВАННЫМ ДАННЫМ

Торвальдсен Г.

Норвежский исторический центр, Университет Тромсе,
9037 Тромсе, Норвегия, gunnar.thorvaldsen@uit.no

В эпоху цифровых технологий произошла и революция в доступе к данным переписей населения в различных форматах. Многие статистические агентства опубликовали данные переписей в виде отсканированных изображений на Cd-дисках или в Интернете. Однако все чаще исследователи предпочитают рассчитывать собственные совокупности микроданных на индивидуальном уровне, так как это обеспечивает более точную статистику и дает возможность сравнить динамику и избежать «экологических заблуждений». Большинство архивов признали, что распространение копий способствует защите первоисточника; а большинство статистических агентств – что их экономические интересы не должны ограничивать доступ исследователей к микроданным. В общей сложности эти базы данных содержат свыше двух млрд отдельных записей, которые могут быть проанализированы как в режиме онлайн, так и более тщательно после скачивания выбранных частей. Оптимальным решением является связывание микроданных переписей с актами гражданского состояния и создание таким образом регистров населения.

Ключевые слова: переписи, микроданные, Интернет, связывание данных, регистры населения.