A Practical Introduction to Dataset Merging

Kamila Klauzinska

Independent Scholar, Krakow/Lodz, Poland
Email: kamila.klauzinska@poczta.fm

The development of sophisticated tools designed to combine and compile –to merge- family data from different sources and databases, from different repositories, is currently viewed as a major objective in the virtual reconstruction of past Jewish families and communities. An overview of some of the benefits and complexity involved in database merging is presented here by means of specific examples.

I. Introduction

It is hard to imagine the history of Poland without the historical and intellectual contributions of Polish Jews. The catastrophe of the Holocaust irreversibly changed the ethnic fabric of Poland. In their time, Jews had constituted the largest minority in the country, with a illustrious history going back over one thousand years. Over three million of them were murdered at the hands of the Nazis, and in 1968 many of those remaining departed, adding to the immeasurable and irrevocable damage done to the community by the Shoah. Today, virtually all that remains of the former Jewish population of the country are material artifacts and the many documents stored in the national and municipal archives, registries, museums and libraries of Poland.

The main purpose of contemporary documenting is to rescue memories of the residents of former *shtetls* (towns where Jews lived in significant numbers), to reassemble broken family trees, reconstitute the structure of lost communities, and preserve this knowledge for future generations.

Since the 1980s, Jewish genealogy has developed rapidly. Whereas in Poland, the study has been treated marginally as a domain supporting history, anthropology and sociology, in the world at large there has been a genealogical boom, as carried out both by family historians and social scientists (Klauzinska, 2007). A number of solutions to difficult genealogical problems are increasingly being proposed by researchers in hard-science fields (biologists, mathematicians, physicists, statisticians, computer scientists, and so on)(Wagner, 2006). Computer technology and the Internet have enabled programs to create intricate genealogical trees and specialized databases.

Genealogical studies are therefore becoming increasingly complex and multi-faceted (Mokotoff & Amdur-Sack, 2004; Wagner, 2006; Jones, 2007; Lamdan, 2009; Herszkovitz, 2011). More databases are being indexed, more Jewish cemeteries are being surveyed and cataloged, and as a consequence the need arises for a systematic comparison and merging of these databases. This is a process that, among other things, helps the recreation not only of individual family trees but also kinship groups and indeed communities. The dataset merging project using the Jewish cemetery of Zdunska Wola as a pilot falls into the latter category and concrete examples from that project are used in the present paper.

II. Early Motivating Example

From the early 2000s, Prof. H. Daniel Wagner (Weizmann Institute of science) and the author of the present paper have been engaged in a project to document the graves in the Jewish cemetery of Zdunska Wola, in central Poland (Klauzinska, 2009). In September 2004, we were documenting one of the eleven sections of the cemetery. One headstone (*matzeva*) looked familiar, as its shape and symbol resembled those on a *matzeva* shown in an old black and white photograph taken in the cemetery. The large matzeva on the old photo had a plaque with a detailed text, whereas this plaque had disappeared from the stone we had found. The *matzeva* was now laying on the ground, broken into several pieces, with only the symbol left (which was identical to the one on the old photograph), but without any text. Fortunately, the name of the deceased was discovered, carved on the partially buried bottom side, which confirmed that the *matzeva*, indeed, was the same as the one on the old picture.

Following this event, we decided to encourage the residents of Zdunska Wola to send us any old photographs of headstones from the cemetery in their possession. Our purpose was to compare them with the current tombstones lacking inscriptions, and indeed we were eventually successful in establishing family data for a number of tombstones.

An index was created for all the tombstones we had documented. All told, there were 3,505 *matzevot,* not all of which bore complete data. In many cases we had to compare the *matzeva* data with information from the birth and death records („metrical records"), a time consuming activity. This led us to create a computer program that would search, compare and integrate data from different sources.

III. Basic Principles of Database Merging

While the technical implementation of a merging task is somewhat involved, its principle is quite simple (Wagner, 2008). By merging two or more datasets for a given individual, we may gain additional previously unknown information about that individual. For example, we have a birth certificate for Yisrael MOSZKOWICZ, and we want to identify the correct death certificate of Yisrael MOSZKOWICZ. Indeed, not unfrequently we can find several death certificates pertaining to different individuals all named Yisrael MOSZKOWICZ. Identifying the correct individual requires sometimes careful comparison of the data because the information may not have been accurately recorded by the clerk, or may not have been accurately known by the person who reported the event, or because of spelling mistakes (including on matzevot). It is important to realize that different data sets usually contain complementary information (even if only slightly) for the same individual, which is why the merging concept is important. For example, this is the information at our disposal in birth and death records:

A. Birth record, which normally includes the father's name, his occupation, the mother's name [with sometimes her maiden name], date and place of birth.

B. Death record, which normally includes the father's name [with sometimes his occupation], the mother's name [with sometimes her maiden name], the age of the deceased [sometimes the exact date and place of death]; and occasionally the names of surviving relatives of the deceased – wife or husband, children.

The more information is available about a given individual in the respective datasets, the more valuable and reliable the data merging. Before demonstrating the power of dataset merging using the pilot study of Zdunska Wola, I wish to point out that significant datasets have survived for this town, as follows:

- 35,000 metrical records from the years 1808-1942
- 28 Books of Permanent Residents (Ksiegi Ludności Stałej), containing over 3,500 records of Jewish families [now extracted and held in the author's private archives]
- Applications for identification cards for the years 1930-1934, among which are nearly 600 Jewish applicants, including portraits photo [private archives of the author]
- A collection of 3,505 data for individuals buried at the Jewish cemetery in Zdunska Wola, including photos of the *matzevot* [submitted in the Spring of 2012 to the JewishGen Online Worldwide Burial Registry (JOWBR)]
- A list of 91 Jews, who returned to Zdunska Wola in 1946 [private archives of the author]

- A list of 445 individuals who paid the tithe tax in 1907 [private archives of the author]
- 1,083 entries in the Polish Business Directory from 1929
- A list of 2,300 necrology records in the „Yizkor Book" (Memorial Book) of the Jewish community of Zdunska Wola
- Over 300 records of Jewish families in the Residents Register [Rejestr Mieszkańców] [private archives of the author]
- Local newspapers – for instance *KALISZANIN i GAZETA KALISKA*, that contain pertinent information concerning the town's history and its residents, including the Jewish community [private archives of the author].

For the pilot project, the metrical death records and the recently acquired Jewish cemetery data were compared and merged. There was a very good reason for this specific selection, namely, the need to clarify the identity of the deceased individual. Indeed, according to the orthodox Jewish tradition, no family names (surnames) appear on the *matzevot*, a practice which was widely observed in Zdunska Wola. Only 629 (18%) of the tombstones bear family names, while 2,876 (82%) have no family names. Moreover, not all tombstones have survived in their original state, many being broken and apparently unidentifiable

Manual merging of the datasets allowed the identification of another 1,541 family names, or an additional 54%. Adding these to the original 629, a significant total of 2,170 tombstones could be identified by their surnames, thus 62% of the total (compared to the initial 18%)!

One serious drawback of the manual merging process is of course that it consumes a great deal of time. However, it allowed us to improve and optimize the performance of the computer program. There were two successive steps as follows: (1) A death records Excel database was created by manual extraction and indexing, for the years selected to test the feasibility of the project, which included 2,137 entries arranged specifically for the program in Polish and English (some of the records were translated from the Russian); (2) A computer program was devised to automatically perform data merging between the data from the death records and the cemetery data previously recorded from the *matzevot.* The documentation of the cemetery data took six years of effort with the invaluable assistance of many local volunteers (Klauzinska, 2009).

IV. Computerized merging

As mentioned above, each metrical death record includes the following data: surname, given name(s), record number, year of record, type of record [in this case D for death], date of death, age of the deceased at death, father's name, mother's name, name and family name of spouse (if alive), and sometimes civil status (widower, widow). This set of data was to be compared and merged with the cemetery listing, which comprised 3505 burials/graves from 11 sections, for the (approximate) period 1828-1940. The data for each burial included the following, when available: surname, given name, grave location (section and position in section), civil death date (manually translated from the Hebrew death date), Hebrew death date, father name, guessed (by manual merging) corresponding metrical data, spouse name, comments.

As seen, the data from both data sets include some overlapping information (sometimes with spelling differences in names, and accuracy in dates and so on), as well as non-overlapping information. For example, and most importantly, the surname of the deceased often does not figure on a grave. Similarly, the mother's name almost never figures on a gravestone. These however always figure in the metrical data. On the other hand, the Hebrew date figures on the grave but not in the metrical data; other more personal information also sometimes figures on graves but not in the metrical data, such as revealing symbols, nicknames, circumstances of death (in a fire for example), and so on.

This increase of information about an individual is the fundamental motivation for the merging procedure developed here. In particular, the most important task for descendants is often to identify the correct grave of an ancestor, which is usually difficult without a surname on the tombstone: details that appear both on the grave and in the metrical data (the first name, date of death, and the father's name) usually lead to the correct assignment of a surname on a grave and thus to the sought after identification of the grave. Merging with the metrical listing is therefore a necessity, even though it is often unavoidably ambiguous (dates on both data sets do not exactly fit, father name is absent because the grave is broken, etc).

Two software packages were used, Turbo Delphi and InterBase 2007, which enabled implementation of the following tasks:

i. Tables were constructed (T_METRICAL and T_OTHER_SURNAME_MET), which store the data in a format more appropriate than the original EXCEL files.

ii. A program was designed and implemented for automatic transfer of data from the metrical books (in EXCEL tables) to an InterBase table (source code available on demand).

iii. A table T_CEMETERY was prepared.

iv. The main program (source code, compilation, screenview) was constructed, using InterBase.

v. A small program was written to convert the age at death into a mathematical number.

vi. A SOUNDEX program was created based on the Daitch-Mokotoff algorithm.

Upon running the merging program, the opening window appears as in Figure 1.

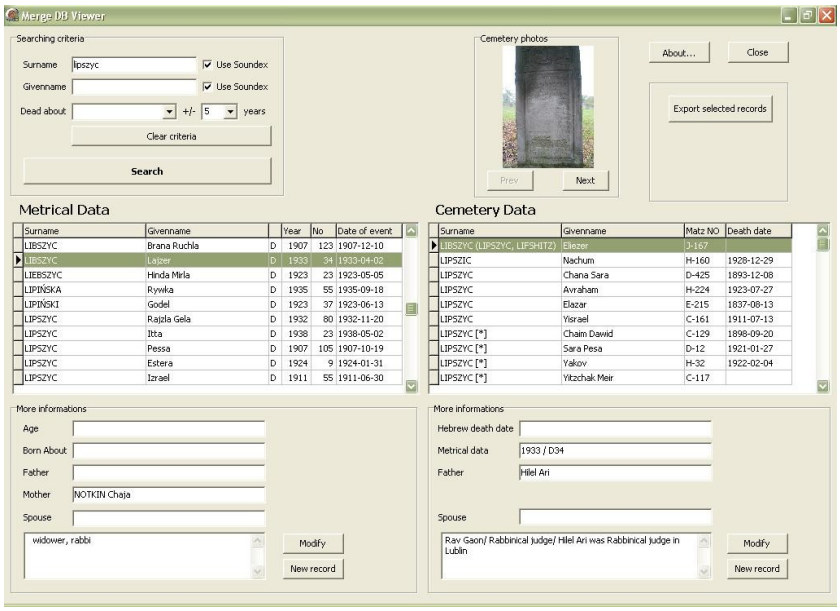

Figure 1: Opening window of the Merge.exe program. Metrical data at left, cemetery data (including picture) at right. A search is performed by inputting a surname at the top left side of the page.

The Merge program is very flexible and may be augmented by including additional database types (an example is given in Figure 2), among them:

1. Identity Card applications (*Podania o dowód osobisty*) – which include a (sometimes rare of even unique) photograph of an individual.
2. Confirmations of Polish citizenship and identity (*Potwierdzenie obywatelstwa polskiego i tożsamości*).
3. Inhabitants' register (*Rejestr mieszkańców*) – which replaced the books of permanent residents in 1931.
4. Supplements to the books of permanent residents (*Dowody do ksiąg ludności stałej*) – a potentially rich source of knowledge on the family stories (Klauzinska, 2011a). The number of such surviving supplements is not known.

5. Residents‘ cards (*Karty meldunkowe*). There is an large number of well preserved cards of Jewish residents in the Historical Museum of Zdunska Wola, awaiting classification.
6. House residents‘ book (*Księgi meldunkowe domowe*).
7. Passport applications to Palestine (P*aszporty emigracyjne do Palestyny*).

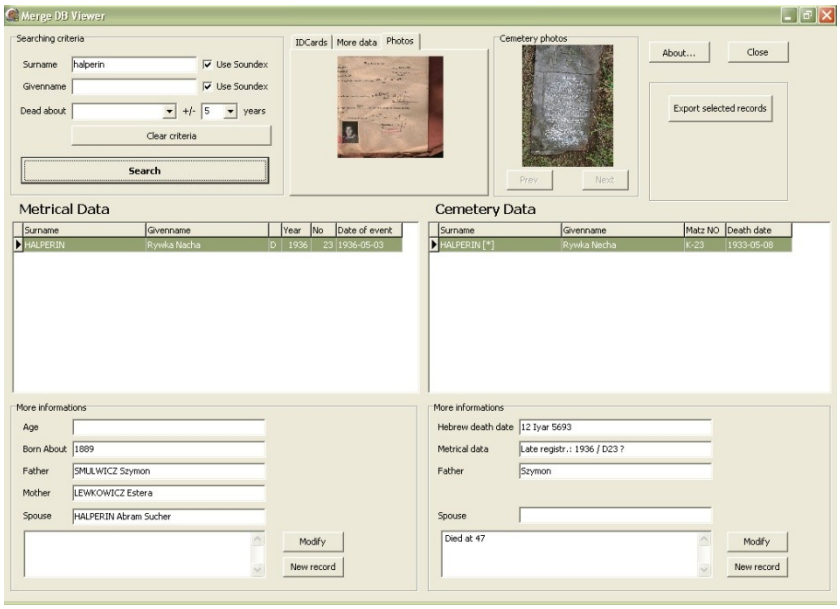

Figure 2: Opening window of the augmented version of the Merge.exe program. Metrical data at left, cemetery data (including picture) at right, ID cards at the center and top of the page.

V. Practical Examples

The usefulness and power of the merging process can simply and convincingly be demonstrated by means of several remarkable examples, as follows.

Example I:

The following partial text appears on a fragment of a tombstone in the Zdunska Wola cemetery (Figure 3): *„(…)shon [?], son of Moshe, died 2 Tamuz 5665* (according to the Hebrew calendar). Can we identify the individual recorded on this tombstone?

Figure 3: Fragment of a broken tombstone, with partial text only.

It is a simple task to convert the Hebrew date into standard dating (*http://stevemorse.org/jcal/tombstone.html*): 2 Tamuz 5665 is 5 July 1905. In the Zdunska Wola metrical death records, we only find a single (luckily) 1905 death record with the name Gershon, according to which Gershon, son of Moshe, died June 21, 1905, at age 20. Since in 1905 Zdunska Wola was located in the Russian partition of Poland, Russian clerks used dates according to the Julian calendar, which in 1905 differed by 13 days from the Gregorian calendar: June 21 in the Julian calendar is July 5 in the Civil (Gregorian calendar) in use today, which exactly fits the Hebrew date on the tombstone.

There is something more about Gershon, which I discovered by careful examination of the cemetery and metrical death records: From the cemetery listings, on July 5 1905, two more people died whose father's name was Moshe and, from the metrical data, a unusually high number of people (eight) died on that day, with their deaths recorded in sequence. Could it be that a unique, likely dramatic, event took place in Zdunska Wola?

To find out, I set out to examine the „*Gazeta Kaliska*" newspaper which since 1898 had been published daily. In the of July 7, 1905 issue, I discovered a long article about a fire that took place in Zdunska Wola, with the names of all the victims. Gershon WINTER, only 12 years old (even though the death record of Gershon states that his age is 20), was among the victims. We cannot be sure of his real age, but we do know that Gershon was born in Widawa, lived on Laska Street, and that a fire erupted in the post office located at that same address. A total of 16 people died, eight Catholics and eight Jews. Amongst them were Gershon WINTER, his siblings Ryfka Fajga and Itta (the metrical records also includes the 1905 death of a Chil Majer WINTER), the pregnant Matla RZESZOWSKA from Klodawa, and most probably her husband Lajzer Dawid RZESZOWSKI. The article stated that Gershon's father, Moshe, a carpenter, was a widower.

Gershon's burial and that of the other victims took place, according to the Jewish tradition, within the next 24 hours, on July 6 at 2 p.m. (the fire broke out at 2 a.m. on the night of July 5-6, 1905). Jews and non-Jews took part in the burial. At 7 p.m. there was another burial, this time of the Catholics victims, with many

Jews attending. After the funeral, the Jews, together with the Catholics, raised funds on behalf of Tadeusz BIEGANSKI, whose wife and four children died in the fire.

This beautiful, if sad, lost moment in the history of Zdunska Wola was revealed and brought back to life, as it were, by that small fragment of stone from the Jewish cemetery, and with the aid of merged datasets.

Example II:

A second example is that of the tombstone of Rywka Nacha, daughter of Shimon, died 12 Iyar 5693 (Figure 4), which corresponds to 8 May 1933. Amongst the ~7,000 death records, nearly 130 records could be found for the names Rywka and Rywka Nacha, among them only two death records for Rywka Nacha. The first one is from 1889 and the second from 1936. Obviously, the 1936 record seemed more appropriate, and stated „Rywka Nacha HALPERIN, daughter of Shimon SZMULEWICZ and Ester LEWKOWICZ". (Death records were occasionally created later than the actual data of event, which generates obvious search difficulties which should also be dealt with by the merging procedure). In this particular case, however, we had read the date on the tombstone incorrectly, due to the very poor condition of several letters. Instead of a *vav*, the last letter in the date had been read and recorded as a *gimel*, thus yielding the (mistaken) year 1933, instead of 1936.

Figure 4: Restored tombstone of Rywka Necha HALPERIN née SZMULEWICZ.

Merging the cemetery and death records of Rywka Nacha with a number of other sources led to much additional information, including her 1913 marriage record with Abram Sucher HALPERIN who, according to the 1929 Business Directory, was a grocery salesman on Juliusza Street. More information appears in the Books of Permanent Residents (located in the National Archives in Sieradz), under the misspelled name of HARPEN–SZMULEWICZ. However, for Rywka Necha's elder son, Alter (who passed away in 2008 at the age of 84), and his own

children, the most emotional finding was the discovery of Rywka Necha's 1933 Application for an Identity Card (Sieradz Archives), in which Alter finally found a picture of his mother's: this is her only surviving photograph, as Alter, who had survived the Shoah, had no picture of his dead parents.

Example III:

This last example brings us back to something mentioned at the beginning of this paper: old photographs from the cemetery. In the „*Yizkor Book*" for Zdunska Wola one finds a photograph showing in the distance, on the left, the back of a tall tombstone with the name „Mirel BIRMA" (Figure 5).

Two questions arise: (1) Can this tombstone still be found today in the cemetery ? Indeed, over the years many tombstones have disappeared from the Jewish cemetery, often to be used for walkways and buildings by locals; (2) Can we determine with precision relevant missing data such as the date of death, full family name, and so on?

A manual comparison was first carried out, and records from 1808 – 1942 showed the name Mirel appearing 25 times, while Mirla occured 201 times. However, in searching for BIRMA(N) or BYRMA(N), a single death record appeared, that of Mirla BYRMAN who died in 1911.

Turning to the computerized merging program, seven metrical records and 12 cemetery records appeared for the name Mirel/Mirla (remember that for the time being, the database of metrical records in the program includes only a limited number of years; were it to comprise the full 35,000 metrical records, the number of occurences of the name Mirel would certainly be higher). The comparison proved successful: the *matzeva* still exists in the cemetery, (tombstone A463), and Mirel was the 82 year old widow of Moshek Gersh (Hersh/Hirsh) GOLDBART. The *matzeva* A463 identified by the computer was lying on the ground (Figure 6) and after it was turned over, the inscription on the back was found to be similar to that in the old photograph: Mirel BIRMA(N)!

Figure 5: Old photograph from the Zdunska Wola Yizkor Book (p. 583).

Figure 6: Tombstone A463 found in the Jewish cemetery of Zdunska Wola. Front (left) and back (right) sides.

VI. Conclusions and Pointers for Future Work

Several challenges and proposals for improvement must be addressed in future merging software. Some of these are listed as follows:

1. The merging of incomplete data, such as when a *matzeva* provides only a fragment of text (names, dates), as in the first example above.
2. Names spelled differently, for example in different languages, which a soundex system would not be able to catch. For example Zvi in Hebrew

(which is „Cwi“ in Polish) is the same as Hirsh in Yiddish; Zippora (Cypora) in Hebrew is Fajga in Yiddish and so on.

3. The problem of late registrations of metrical records (especially birth records).
4. Inconsistencies in the ways the date is written. Sometimes the date is written in Hebrew according to the Jewish religious calendar, as on *matzevot*. Metrical records have Julian or Gregorian dates, or Julian and Gregorian dates side by side.
5. The issue of rating and ranking results from a program run – indicating the person according to the highest probability (by reading and merging other data from different databases and, for instance, matching the father's name, which is an important element in identification, appearing on the vast majority of *matzevot*.).
6. Keywords search capability – for example, by searching for "fire", we should easily pull up the four *matzevot* at the cemetery which relate to the fire of July 5, 1905.
7. Including all available information from tombstone epitaphs. Some of these include extensive genealogical information, especially in the case of rabbinical families, see an example in Figure 7. In Zdunska Wola there are 44 *matzevot* with genealogical information of this sort, thus well beyond just the name of the father of the deceased. Wodzinski (1996) mentions that there are well known *matzevot* of *tzadikim* with at times genealogical data back to seven generations.

Figure 7: The matzeva of Yakov Moshe [TAUB] in the cemetery of Zdunska Wola. Excerpts from the epitaph: *Yakov Moshe died on November 4, 1902. He was the son of Menachem David, and, grandson of Rabbi Yehezkiel [TAUB] from the town of Kozmir (Kazimierz Dolny), and grandson of Rabbi Itzhak Meir from the town of Ger (Góra Kalwaria)*.

Clearly, manual checks will still be necessary, but a well-designed computer program can greatly narrow the range of validation of data of interest. Now already, thanks to the current program, family descendants are able to find tombstones of their relatives faster and with greater probability than ever before. It is now possible to (re)create the genealogical trees of families belonging to the same Jewish community in any given town, and eventually to arrive at a „communal/regional genealogical tree“.

One last point, of importance: rare and very interesting genealogical sources will continue to appear from searches in archives, repositories, and local and national museums. As an example, books in Hebrew were recently discovered in the Historical Museum of Zdunska Wola. One of them is ‚*Dvarim*‘ (*Deuteronomy*), the fifth book of the Hebrew Bible (*Tanach*), which includes a stamp with the name of its owner (Figure 8). Another is ‚*Bereshit*‘ (Genesis), the first book of the *Tanach*, which includes a handwritten list of births of the owner’s four sons (Figure 9). Handwritten genealogical information in Hebrew books are not unusual: as a tradition the eldest member of family – the father or grandfather- inserted this kind of information on the inside cover of a prayer book such a Mahzor, Siddur, Bereshit, etc, or in private notebooks (Klauzinska, PhD thesis, 2011b).

Figure 8: *Sefer Devarim*, Archive of the Historical Museum of Zdunska Wola. Three red stamps (in Hebrew, Polish and Russian) reveal the owner’s name: Symon-Symcha Kronman. From database merging we learn that Szymon Kronman was born in 1898 in Zdunska Wola. The Kronman family was researched for many years by Martin Kronman from Syracuse (New York, USA). Martin passed away in 2009 and never saw this book which had remained inaccessible for the last 20 years in the local Museum archives.

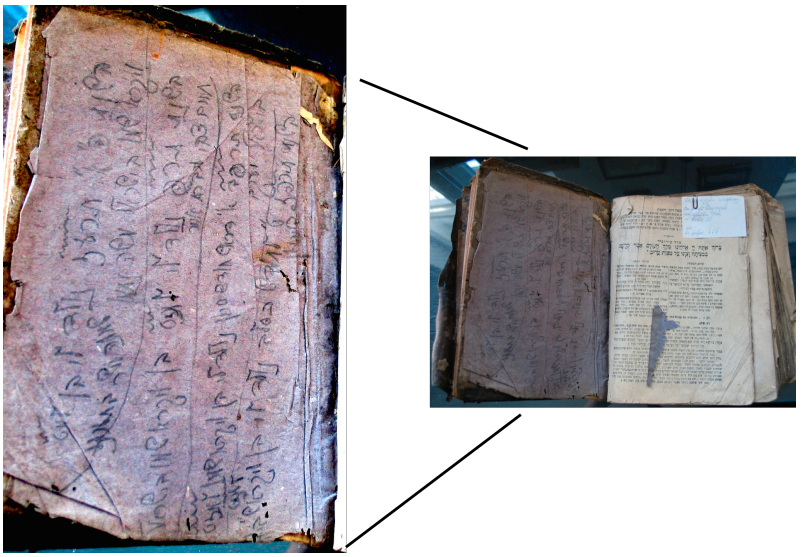

Figure 9 - *Sefer Bereshit*, Archive of the Historical Museum of Zdunska Wola. The inside cover includes a handwritten list of births of the owner's four sons.

Acknowledgments

This work was supported in part by a grant from the International Institute of Jewish Genealogy, which is gratefully acknowledged. Thanks are due to Jakub Zajdel for his assistance with the computer software.

REFERENCES

Hershkovitz, Arnon, 2011, "Leveraging Genealogy as an Academic Discipline", Avotaynu, xxvii, 3 (Fall issue), 18-24

Jones, Thomas W., 2007, "International Institute: A Breakthrough for Academic Genealogy?", Association of Professional Genealogy Quarterly, xxii, 1

Klauzinska K., (2007), „A Modern Approach to the Genealogy of Polish Jews: Zdunska Wola as a Test Case", Scripta Judaica Cracoviensia, vol. 5, 39-51

Klauzinska K., 2009, "Colored Tombstones in the Jewish Cemetery in Zdunska Wola", Ars Judaica, vol. 5, 121-128.

Klauzinska K., 2011a, „Księgi ludności stałej jako źródło do badań nad Żydami sosnowieckimi", Żydzi na Górnym Śląsku i w Zagłębiu Dąbrowskim. Historia. Kultura. Zagadnienia Konserwatorskie. Kraków, 67-74

Klauzinska K., 2011b, „Modern genealogy of Polish Jews", PhD dissertation, manuscript submitted to the Jagiellonian University in Cracow, 284

Lamdan, N. 2009, "Jewish Genealogy: Moving Towards Recognition as a Sub-branch of Jewish Studies", Avotaynu, vol. 25, No 2 (Summer issue), 3-8

Mokotoff, Gary & Sack, Sallyann Amdur, 2004), "The Next Step: Jewish Genealogy Goes Academic", Avotaynu, xx, 3 (Fall issue), 3-4

Wagner H.D., 2006, „Genealogy as an Academic Discipline", Avotaynu, vol. 22, No 1, (Spring issue), 3-11

Wagner H.D., 2008, „Tombstone identification through database merging", Avotaynu, vol. XXIV, (Spring issue), 8-10

Wodziński M., 1996, Hebrajskie inskrypcje na Śląsku XIII-XVIII wieku, Wrocław, 98

BIOGRAPHY

Kamila Klauzinska has a PhD from the Department of Jewish Studies, the Jagiellonian University in Krakow, Poland. Her PhD dissertation focused on Modern Jewish genealogy. Previously, she earned a M.A. degree in Ethnology from the Faculty of Philosophy and History, the University of Lodz, Poland. Her M.A. thesis focused on the Jewish cemetery in Zdunska Wola. Dr Klauzinska was a Visiting Scholar in a number of prominent institutions, including the Russian, East European, and Eurasian Center at the University of Illinois at Urbana-Champaign, USA (2011). She was awarded a number of prestigious scholarships, among them a Jagiellonian University Scholarship (Krakow, 2006-2010), a Rothschild Foundation Grant (2008) and a grant from the International Institute for Jewish Genealogy in Jerusalem, (2007-2008).

As a researcher and consultant, Kamila Klauzinska specializes in Jewish metrical data from large numbers of sources stored in the Polish Archives and Registry Office. She was co-leader of the Photographic and Topographic Census Project in the Jewish Cemetery of Zdunska Wola (2002-2008). She cooperated with the Museum of the History of Polish Jews in Warsaw – "Virtual Shtetl" project (2008-2009).