A Methodology for Error Detection and Correction of Jewish Names in Digitized Genealogical Records

Jean-Pierre Stroweis

Independent scholar, Jerusalem, Israel
E-mail: stroweis@zahav.net.il

This paper proposes three algorithms for error detection and error correction of Jewish names found in digitized genealogical records such as civil records. Error detection is performed by matching names against reliable dictionaries of Jewish names for a given geographical area, using both the Daitch-Mokotoff Soundex and the Beider-Morse Phonetic Matching. Error correction is made possible by exploiting the inherent redundancy embedded in the various records pertaining to the same family cell. The combination of these techniques has been applied to large sets of civil records of Jews from two Polish towns, detecting an unexpectedly large number of errors and in most cases suggesting corrections, even before consulting again the original records. The methodology has significant side-effect benefits, such as inferring unregistered information and tracking variants of given names and surnames. The method can be used as a quality control tool for genealogical databases published on the web.

"Genealogy information found on the Internet should never be trusted."[1]

Errare humanum est, perseverare diabolicum.[2]

I. Presentation of the Problem

The Internet has become a major tool for genealogy research. New repositories are published on the web daily, granting genealogists the largest insight ever to their ancestors' existence with just a few clicks. This instantaneous availability raises the question of the reliability of the information obtained from the Internet. We rely a great deal on the precision of those who indexed and placed these sources on the web. The accuracy and completeness of the online

[1] Cited by Dick Eastman's Online Genealogy Newsletter, March 2011, http://tinyurl.com/bxyunq8

[2] To err is human, to persist is of the devil.

repositories is very uneven. Providers tend to publish as fast as possible, often at the expense of the quality.

For example, the major genealogical event of 2012 is no doubt the release and indexing of the 1940 United States Federal Census[3]. Not that everyone has relatives in America, but the size of this repository and the pace of its indexing are unprecedented. 3.8 million scanned pictures of census pages have been made public on the Internet on April 2, 2012. Census data for 132 million persons were then turned into an online searchable database within just 120 days at a staggering pace of 1,200,000 individuals indexed per day. Some 120,000 volunteers were drafted to perform data entry on a daily basis, as many as census takers in 1940. Indexing was completed and database made ready during August 2012. Could this incredible pace have any impact on the quality of the database? Could the proverb "More haste, less speed" apply here? Possibly. Quality control was usually minimal; on selected cases, two indexing teams worked in parallel on the same set, and when they did not match, an arbitrator made a final decision. Unfortunately, this resource-consuming approach cannot be applied globally.

Providers of large-scale population enumeration are aware of the dangers. The superintendent of the 1870 US census already wrote, 145 years ago: "It is exceedingly undesirable to bring anything into the census which is not thoroughly trustworthy; such material always does more than discredit the work; it accomplishes a great deal of positive harm. It is exceedingly undesirable to bring anything into the census which is not thoroughly trustworthy; such material always does more than discredit the work; it accomplishes a great deal of positive harm."(Walker, 1878)

Techniques for assessing errors and improving the quality of enumeration include re-enumeration, comparison of successive censuses, matching names in sets of individual records, checks of internal consistency, checks against independent aggregates, and post-enumeration sample surveys (Walker, 1878). Despite their existence, these techniques have not yet been systematically applied to digitally indexed repositories. There are no standard acceptance procedures and very little has been published to assess the quality of searchable databases. If volunteers have plenty of good will, that does not make always their work perfect, alas. Mormons volunteers indexed the Ellis Island immigration manifests. This database has been extremely valuable to trace immigration to the USA, despite the large number of errors it contains. Essential sources (such as Ellis Island Passenger Lists and Yad Vashem Pages of Testimonies) known to be inaccurate only offer an online tool to submit corrections. The problem becomes critical for databases that do not provide access to scanned images of the original source.

[3] https://familysearch.org/1940census

This paper proposes a systematic method for error detection and correction of Jewish genealogy databases. The goal is to define an automatic or mostly-automatic tool that could serve for quality control. As a side benefit, it offers a tool to measure the reliability of a genealogy database. The method was applied to late 19th century / early 20th handwritten century civil records. These primary records contain rich genealogical data and are supposed to offer the highest level of reliability, i.e. a low error rate for primary errors, as they are maintained by professional clerks and local administrations, following national standards and regulations. We chose two sets of Jewish civil records from two Polish towns, Staszów, and Ostrów Mazowiecka (where births, marriages and deaths of Jews were registered on separate books).

II. Classification of the Errors

What are the various kinds of errors, their causes and their impact? We distinguish **primary** and **secondary** errors. Primary errors were made at the creation of the records. They include:

- Under-enumeration, when a record or an element was omitted,
- Over-enumeration, when a person was included more than once in the records,
- Misreporting, when the recorded information is not faithful to the facts.

Among the reasons for primary errors:

- Evasion from registration (e.g. to avoid military draft),
- Inaccuracy of the reporters (e.g. inconsistent usage of double given names; age piling, a trend to round the ages up or down to the nearest number that ends in zero a five),
- Delayed registrations,
- Intentional mis-reporting and identity theft,
- Mis-registration by the administration in charge of the recording.

Secondary errors are introduced, un-intentionally, by the digitizing process. They include skipping, mis-interpretation, mis-transliteration, mis-typing and mis-formatting. There are many causes for secondary errors. Digitization of civil records is a labor-intensive, error-prone process. Old, handwritten civil records often are hard to read because of damaged paper, faded ink, ancient characters, idiosyncratic handwriting, or poor photocopy or microfilm quality. Errors may occur during human interpretation, transliteration, optical character recognition or data entry. The risks of secondary errors rise when someone

unfamiliar with the languages and the names on the documents is in charge of the data entry.

Secondary errors must be eliminated. They result from the digitization and they are simply nuisances. Primary errors are either the result of unreliable recording or intentional changes in the submitters' declaration, in which case they should draw the genealogist's attention. Soundex-aware search engines do not correct errors, but they offer a convenient way to access potentially matching records that have slight variations in spelling. However, when names are too much distorted or too transformed, soundex cannot help. Therefore, it is not necessary to catch all the spelling differences, it is sufficient to correct the distortions that modify the soundex of the original names and cause the search engines to skip over the relevant records.

Database errors are impediments on the way to reconstruct a family history, and they affect the reliability of the research. Then they are propagated and replicated all over, like computer viruses, making it even more difficult to find and fix them. Such errors must, therefore, be detected as early as possible.

III. The Proposed Methodology for Detecting Errors

Best practices for detecting basic data entry errors in digital repositories include:

- Usage of standards for transliteration, names of places, dates, acronyms, upper case characters, etc.
- Review spreadsheet tables, column-by-column, to catch typographic errors, reject invalid or out-of-range values, remove extra or superfluous characters, etc.
- Verify that the internal rules and implicit redundancy in the original records are not broken by the data entry. Here are two examples: final Hebrew letters [ך, ם, ן, ף, ץ] are only expected at the end of Hebrew words; all members of a family cell listed in a census should share the same address.

As online repositories are searchable per name, errors on names are the most critical. If a name in the digital repository is incorrect, the user search is likely to miss that record. So, the error detection operation assesses the accuracy of the given names and surnames in a digital repository. That is, it identifies the most likely correct names and the most likely incorrect names suspected of being erroneous. Preliminary steps will handle separately double names (e.g. for "Majer Dawid", treat "Majer" and "Dawid" apart), eliminate language declension suffixes (e.g., use Mordko instead of Mordkowna) and switch surnames to their masculine form, that is, Lipski instead of Lipska.

Then, it is useful to perform a frequency analysis on the data set in order to count how many times each given name or surname appears in the data set. Frequent names are more likely to be correct, while rare names may reflect errors.

Next step consists to verify if a name/surname is a known Jewish name/surname for the given geographical area and historical period. Fortunately, several valuable dictionaries of Jewish names from Central and Eastern Europe are available to consult, including the Handbook of Ashkenazic Given Names and their Variants, the Dictionary of Jewish Surnames from the Russian Empire (DJSRE), the Dictionary of Jewish Surnames from the Kingdom of Poland (DJSKP), the Dictionary of Jewish Surnames from Galicia (DJSG), all by Alexander Beider, and the Dictionary of German Jewish Surnames (DGJS) by Lars Menk.[4]

If the exact spelling of a particular name or surname appears in the relevant dictionaries, it should generally be accepted as correct[5] : the dictionaries confirmed the existence and the spelling of the name. Steve Morse has a surname reference web tool that does exactly that[6].

Otherwise, should we accept the names not found in the dictionaries?

Names that sound exactly like dictionary names should also be accepted, even if they are spelled different. For this purpose, the Beider-Morse phonetic matching (BMPM)(Beider & Morse, 2008) serves as a robust and strict criterion for sound similarity. Practically, are considered as correct those names whose BMPM value is identical to the BMPM value of a dictionary entry. For example, the surnames Cereśnia and Mężyński are not in DJSKP, but they share the same BMPM value as Czereśnia and Menczyński, names that are in the DJSKP dictionary. So Cereśnia and Mężyński are accepted.

What about other names? As many names are not matching with BMPM, the next criterion is the Daitch-Mokotoff soundex[7] (DM), which also aims to produce the same soundex value for names that sound similar. For example, the alternate spellings Chasonow and Khazanov share the same DM code (546700). As the DM soundex ignores most vowels and as it is usually restricted to six digits, a DM soundex will match names not always matched by the BMPM. However, DM soundex match may also occur between unrelated names (false positives). First example, the distinct surnames Cyrkman and Zorgman share the same DM code (495660); second example, the surnames Rajchensztajn and Rozensztengien share the same six-digit DM code (946436), while an extension to eight-digit of the DM code would produce different codes (94643600 versus 94643656).

[4] All these dictionaries were published by Avotaynu.

[5] Unless inconsistencies are uncovered during the correction stage.

[6] http://www.stevemorse.org/phonetics/beider.php

[7] http://www.avotaynu.com/soundex.html

Note that while Mężyński and Menczyński are BMPM equivalents, they have different DM codes (646450 versus 664645) because of the Polish letter ę (e ogonek), which sounds like “en” or “em.” Although not an inherent limitation, the DM soundex lacks explicit rules for diacritic signs; thus, most search engines ignore them when computing the DM code.

So, DM soundex match will be used with caution:

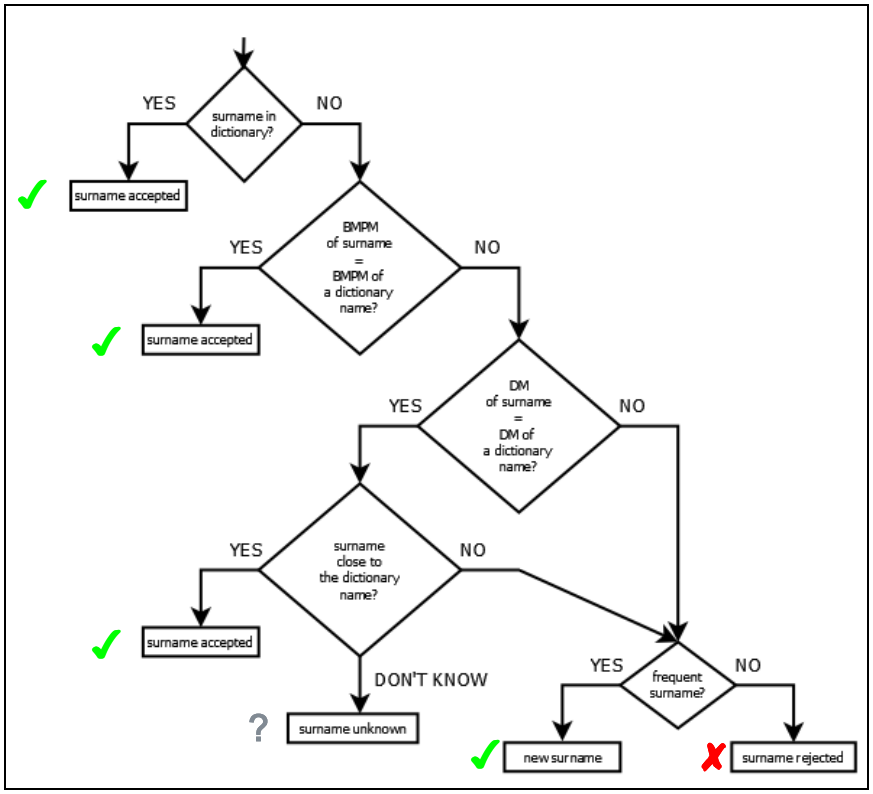
- Names whose DM code is not the DM code of any relevant dictionary entry are most likely errors, unless they are new names not yet listed in the dictionary.
- Names whose DM code is identical to the DM code of a dictionary entry are most likely acceptable variants or distortions from existing surnames, unless the DM match is coincidental in which case the name must be rejected.

In either case, the DM soundex match does not provide a clear-cut decision. Additional criteria are necessary to confirm or infirm the result of the DM match. They include:

- The frequency of the surname in the repository,
- The frequency of the surname in other repositories, in particular from adjacent towns.
- The match of the “approximate” BMPM value.

If a name in this category occurs in a large number of records, it likely is a new name to add to the surname dictionary. A further search on the Internet—for example, on the JRI-Poland database, particularly a search on records of nearby towns—may confirm or invalidate this hypothesis.

The above algorithm for error detection is best represented by the following organization chart:



Here are a few examples:

- The data entry name Brajdak has no dictionary entry with an equivalent BMPM value. Its DM soundex value is 793500, just like many entries in DJSKP, such as Brajdyk, Frajtag and Wartecki. This correspondence indicates that Brajdak is a spelling variant of the DJSKP surname dictionary entry Brajdyk .
- Dizenhaus is not in any dictionary, but it shares its BMPM value with the DJKSP entry Dyzenhauz, so it is accepted.
- Blugant has the DM soundex value 785630, which is found in DJSKP with Wolgiemut and in DJSG with Pelikant, but these obviously are false positive coincidences. Blugant has no close dictionary entry; hence, it is a suspect name.
- Zajdelwar shares its DM soundex value 438790 with Szydlower (in DJSKP) and Jodliwer (in DJSG), two other false positives. Zajdelwar also is a suspect name.
- Pantyrer DM soundex value is 763990, just like Wanderer (in both DJSKP and DJSG), which makes it a suspect name. However, as this surname is found on 35 records in the sample repository, it is most likely a correct, new name not yet registered in the dictionaries.

IV. Approaches to Correcting Errors

Once we identify suspect names, how might we check and correct them? Two methods are possible: one is a speculative correction, basically an educated guess; the other is a demonstrable correction, deducting the name from other related records.

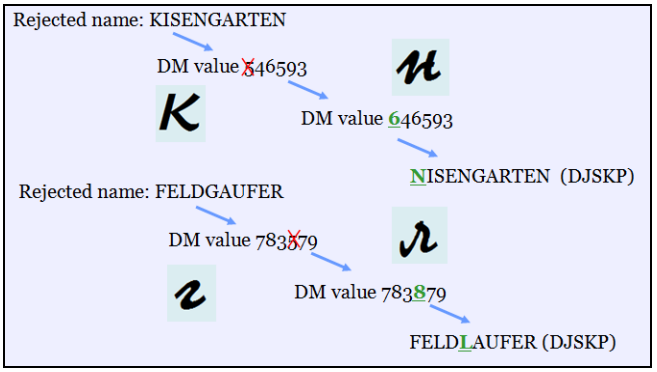
Speculative correction

If the DM soundex code of a suspect name is not the code of a dictionary entry, we may search for DM soundex codes (in the relevant dictionaries) that are as close as possible to the suspect name. Computer science literature abounds in metrics that estimate the distance between two strings of characters, i.e. tools that establish a measure of the degree to which two strings are close or not; the closer the strings, the smaller their distance. The Levenshtein distance[8] between two strings is defined as the minimum number of edit steps needed to transform one string into the other, with the allowable edit operations being insertion, deletion, or substitution of a single character.

We may apply this distance between two DM codes. For example, if we believe that the suspect name Birnshtuck is a misinterpretation of the dictionary name Binsztok, we may ask how similar are their DM values (796435 and 764350 respectively)? The Levenshtein distance between these DM codes is two edit steps, because, in order to pass from string '796435' to string '764350', we first need to delete the digit '9' then append the digit '0'. Calculating the distance between two DM values is a measure of how closely these soundex values sound alike, whereas calculating the distance between the surnames themselves is a measure of how close their spellings are. Measuring distances between DM codes is more efficient here, as it ignores changes in vowels and spelling differences that have no or little impact on the name sound (e.g., SH instead of SZ, K instead of CK).

Let us now examine the substitution of a single consonant. In this common case example, a suspect name differs from the real name by one consonant only, that is, the DM soundex code of the suspect name is likely to differ only by a single digit from the DM soundex code of the real name in the dictionary, resulting in the smallest Levenshtein distance of one single edit step. This situation appears frequently. For example, in handwritten Cyrillic, the letter К ("K") often is confused with the letter Н ("N"), and the letter Г ("G") often is confused with the letter Л ("L"), resulting in the suspect name Kisengarten (DM soundex value 546593) instead of the real name Nisengarten (soundex value 646593), or the suspect name Feldgaufer (soundex value 783579) instead of the real name Feldlaufer (soundex value 783879). The substitution process is presented in the following chart:

[8] http://en.wikipedia.org/wiki/Levenshtein_distance



Because each digit may have as many as nine alternatives, a maximum of 9 x 6 (the number of digits in a DM soundex code) = 54; DM soundex codes must be checked per suspect name - assuming that the error resulted simply from the substitution of one consonant. We may search correct names among nearby DM codes and try to find a correction to our erroneous name.

Such a systematic procedure produces only hypotheses —best guesses— or what we may call speculative correction. This method is simple and always applicable.

Demonstrable correction

Error correction using deductive steps is built upon redundancy. The idea is to compare the record containing a suspect name with other related records that incorporate the correct spelling of this name. During the 19th- and early 20th-century, European Jewish families often had many children, so it is extremely valuable to trace and compare the records of siblings. For example, given the suspect surname of a new-born, searching the records (from the same town) that share the same maiden name of the mother is likely to reveal the birth or death records of siblings, and/or the marriage record of the parents. After eliminating records that are unrelated to this family group, we usually find alternatives for the suspect surname. Following are two examples :

Example 1

A suspect surname, Bulman, is found in the birth record of Fajga, daughter of Majer Bulman and Marja Dwojra Urfajg. Searching for other records with the mother's maiden name of Urfajg yields:

- Birth record of Abuś, son of Majer Bulwa and Marya Urfajg,
- Birth record of Chaim, son of Majer Bulwa and Marja Dwojra Urfajg,
- Marriage record of Sana, son of Majer Bulwa and Marjem Urfajg (to Ajdla Flajszakier).

Thus, from the preponderance of evidence in these records, Bulwa, and not Bulman, may be assumed to be the correct surname.

Example 2

The data entry birth record of Menasze Kohn indicates that he is the son of Bencjon Kohn and Chaja Dyndkorn. The name Dyndkorn is suspect; no surname dictionary entry shares its BMPM value or its DM value (363596). A search through the other records with the father's surname Kohn (i.e., search matching records that share the spouse's surname) yields:

- Birth record of Pinkwas, son of Bencjon Kohn and Chaja Ryndhorn,
- Birth record of Abraham Samuel, son of Bencjon Kohn and Chaja Ryndhar,
- Birth and death records of Faiga Blima, daughter of Bencjon Kohn and Chaja Ryndhorn.

We may conclude that Dyndkorn is a data entry error created instead of Ryndhorn, a name that is listed in DJSKP. The two DM values differ by a single digit (363596 versus 963596). Possibly the spelling Ryndhar is an error introduced by the clerk at the creation of the birth record for Abraham Samuel.

In these examples, an error in a surname was detected and then corrected with a high level of confidence by correlating the data in several records. Further consultation of the town registries will assess if the deduction is well-founded, and should determine also whether the error is a primary or secondary error. Error correction based on redundancy is the result of a demonstrable, deductive reasoning; therefore, it is superior to a speculative correction. It works, however, only if the data entry includes the parents' names, and the data set includes other records for the same family with which to compare.

Because of the limitations inherent in each method, we recommend using a combination of both correction techniques.

V. Additional Benefits

Use of the comparison process described above can enhance the quality and the completeness of the digitized data, allowing the inference of details not even recorded in the original records.

To proceed, we merge all the birth, marriage and death records in a town into a single, four-column table, including only the parents' given names and surnames and intentionally ignoring the given names of newborns and deceased individuals, in order to focus on the name correlation issues:

Father's Given Name	Father's Surname	Mother's Given Name	Mother's Maiden Name
Abram	KUPFERBERG	Fajga	GOLDFLUS
Abram	KUPFERBERG	Fajga	GOLDFLUS
Abram	KUPFERBERG	Fajga	KUPFERBERG
Abram	KUPFERBERG	Fajga	KUPFERBERG
Abram	KUPFERBERG	Fajga	KUPFERBERG
Abram	ROZENTRAUB	Fajga	RAJNSZTAJN
Abram	ROZENTRAUB	Fajga	RAJNSZTAJN
Abram	WEJGMAN	Fajga	WEJGMAN
Abram	WAJNBERG	Fajga Brandla	Not mentioned
Abram	WAJNBERG	Fajga Brandla	BERMAN
Abram	WAJNBERG	Fajga Brandla	BERMAN

We can now sort this parents table according to various orders, depending on the aspect to examine: typically, we sort according to three columns (considered as correct) as we look for missing information or inconsistencies in the fourth column.

Completion of missing details

Mothers' maiden names are not always listed in the original records, or sometimes the mother's maiden name as recorded is similar to the father's surname, which further adds to the confusion. In both cases, correlating records with the same father's given name, father's surname and mother's given name, allows a researcher to deduct the missing mother's maiden name. In such a case, we sort the table first by father's given name, then by mother's given name, then by father's surname. In the above table, we notice that, in three records, the maiden name of Fajga, spouse of Abram Kupferberg, was recorded also as Kupferberg, but twice it is reported as Goldflus. It is not recommended to modify the data, but one may indicate that Goldflus is an inferred maiden name. On another record, the maiden name of Fajga Brandla, spouse of Abram Wajnberg, was not mentioned, but we have good reasons to believe it is Berman.

Variations in surnames

Some inconsistencies are not revealed by the error detection phase, i.e. when the variations in the parents' surnames occur among names found in the dictionaries. In this case, the table must be first sorted by father's given name, then by mother's given name (assumed both correct). For example, in three birth records that occurred in a 10-year period, the father is registered as Anczel Wurcelman and the mother as Basia. The mother's maiden name is recorded once as Eliasiewicz, once as Elasiewicz and once as Lisiewicz. This may be simply another data entry (secondary) error—or an intentional change.

Father's Given Name	Father's Surname	Mother's Given Name	Mother's Maiden Name
Anczel	WURCELMAN	Basia	ELIASIEWICZ
Anczel	WURCELMAN	Basia	ELASIEWICZ
Anczel	WURCELMAN	Basia	LISIEWICZ

Similarly, by browsing the table, we detected that Kirsz is sometimes reported as Kirszner, Bergier as Bergman; Kupferberg as Kupferberger; Tencza as Tencer; and Mandelman as Mandel or Mandelbaum. These variations seem to be primary errors; that is, errors that originated at the creation of the record and not an artifact of the data entry.

Correction of formerly accepted names

Icek Diament is recorded as the father in five births over a period of ten years; for two births, the mother is recorded as Gitla Cynamon; in two other births, the mother is recorded as Sura Gitla Cynamon; and in the fifth birth, the mother's name is Sura Gitla Cymerman. Cymerman is likely a primary error, subject to further verification in the original records. On other cases, this could reveal variations in surnames over time.

Father's Given Name	Father's Surname	Mother's Given Name	Mother's Maiden Name
Icek	DIAMENT	Gitla	CYNAMON
Icek	DIAMENT	Gitla	CYNAMON
Icek	DIAMENT	Gitla	CYNAMON
Icek	DIAMENT	Sura Gitla	CYMERMAN
Icek	DIAMENT	Sura Gitla	CYNAMON
Icek	DIAMENT	Sura Gitla	CYNAMON

In the following example, the surname Pinkowska was not questioned during the error detection phase, because it is a dictionary name. But it seems that the real maiden name in this family cell should be Pinczowska.

Father's Given Name	Father's Surname	Mother's Given Name	Mother's Maiden Name
Hercak	SZMAJSER	Blima Laja	PINKOWSKA
Hercka	SZMAJSER	Blima Laja	PINCEWSKA
Hercyk	SZMAJSER	Blima Laja	PINCZOWSKA
Hercyk	SZMAJSER	Blima Laja	PINCZOWSKA
Hercyk	SZMAJSER	Blima Laja	PINCZOWSKA
Hercyk	SZMAJSER	Blima Laja	PINCZOWSKA

Given Names

The methods for correcting surnames also apply to the detection and correction of given names. Indeed, many errors in given names appear in the vital records. Polish civil records, for example, are less consistent when it comes to given names than they are with surnames (and even less consistent with ages). Individuals sometimes are identified by a single given name (Brucha), sometimes by a double given name (Brucha Rajzla) and sometimes by the double name in reverse order (Rajzla Brucha). Variations are frequent. For example, Eta instead of Etka, maybe a diminutive or affectionate form. Nonetheless, obvious errors may be easily detected by simple inspection, for example, Chuwa Lejb instead of Chuna Lejb. To search for inconsistencies in given names, first sort the parents table by father's surname, then by mother's maiden name (i.e., the surnames are considered correct for the purpose of studying variations in given names).

VI. Evaluating the Methodology

Analysis of digitized civil records

The author applied the error detection and correction method to the Jewish civil records of two Polish towns, Staszów, and Ostrów Mazowiecka. For Staszów, the digitized repository contains 7,439 births, marriages and deaths between 1885 and 1904. For Ostrów Mazowiecka, it includes 8,599 civil acts for the same period.

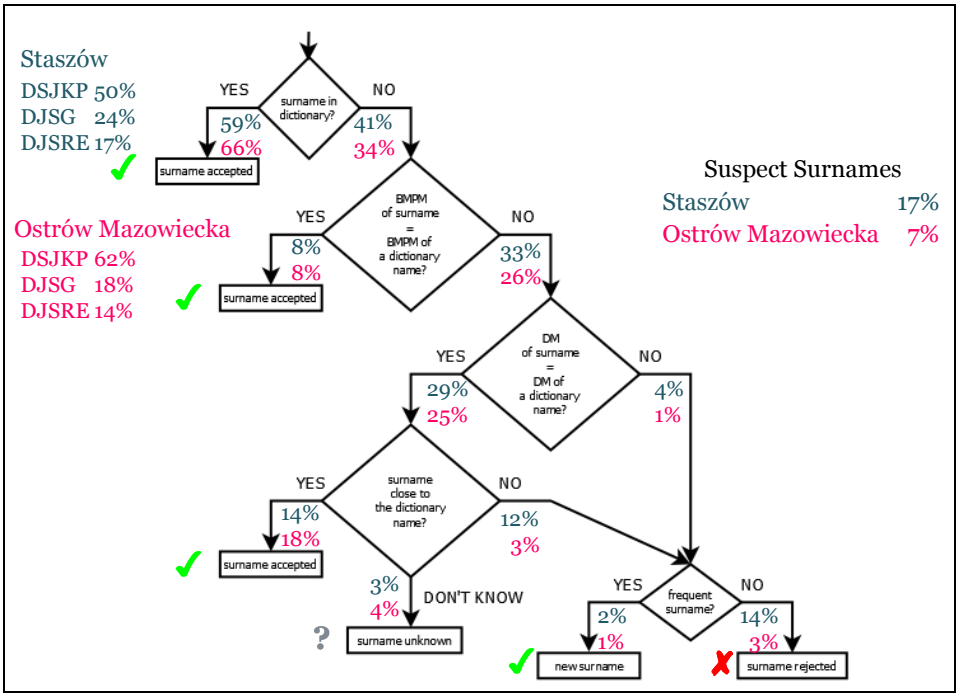
Located now in Kielce province, Staszów once belonged to the Sandomierz district of Radom gubernia section in Congress Poland. But it is only 12 miles (20 km) north of the Visła (Vistula) River, which served as the political and linguistic border between the Kingdom of Poland and the Austrian province of Galicia. Staszów also was under Austrian control for a short period. It is likely, therefore, that some of the family surnames in Staszów came from Galicia and were Germanized names. Thus, it was useful to check Staszów surnames acquired from civil records against the names in both DJSKP and DJSG. We also found Staszów names in DJSRE.

Ostrów Mazowiecka was located in the Łomża guberniya of Congress Poland, around 80 miles distant (130 km) from contemporary Belarus. It was at a linguistic boundary between the northern (Lithuanian) Yiddish dialect and the southern (Polish/Galitzianer) Yiddish dialect. From the town records, the history of the town, and genealogical research, indications are that families in the town had connections to Lithuania and Belarus, among other places. For these reasons, Ostrów Mazowiecka surnames were checked against three dictionaries, the DJSKP, DJSG and DJSRE.

The author indexed some of Staszów's civil records for Jewish Records Indexing-Poland (JRI-P) and is familiar with that town's surnames. By contrast, he

had no previous experience with the surnames of Ostrów Mazowiecka, one of the best-documented Polish Jewish populations.

The results of the error detection are presented below:



More than half of the digitized surnames had an exact match with a DJSKP (Kingdom of Poland) dictionary entry. Other surnames matched the Galician or the Russian Empire dictionary. Another 8% of the digitized surnames match the BMPM of a dictionary entry. The fate of around 30% of the surnames is determined by the DM and additional criteria.

The results of the error correction are summarized in the following Table. As mentioned earlier, this stage also detects and corrects errors on formerly accepted names, raising the overall error rates.

	Staszów				Ostrów Mazowiecka			
Distribution of surnames	Qty	%	Accept	Reject	Qty	%	Accept	Reject
Exact spelling or BMPM match	1321	63.4%	1321		1999	66.3%	1999	
Good DM match or variant spellings	316	15.2%	316		737	24.5%	737	
New names	32	1.5%	32		19	0.6%	19	
Suspect and corrected names	253	12.1%		253	82	2.7%		82
Corrected formerly accepted names	45	2.2%		45	10	0.3%		10
Suspect names with speculative fixes	53	2.5%		53	27	0.9%		27
Suspect names, not corrected	65	3.1%		65	139	4.6%		139
Total	2085	100%	1669	416	3013	100%	2755	258
			80%	**20%**			**91%**	**9%**

Only 63 percent of the Staszów names and 74 percent of the Ostrów Mazowiecka surnames are recognizable immediately (i.e. names found in the dictionaries or whose BMPM code matches the BMPM code of a dictionary entry). After reviewing the other names, 80 percent (Staszów) and 91 percent (Ostrów Mazowiecka) of the surnames are considered correct, still leaving a high percentage (20 percent) of rejected surnames in Staszów records (appearing in 5.6% of the records). The records for Ostrów Mazowiecka are much cleaner, but even here 9 percent of the surnames (which appear in 3 percent of the records) are rejected.

For each town, we deduced corrections with a high level of confidence (using redundancy), for about half of the suspect surnames (253 surnames in Staszów, 82 surnames in Ostrów Mazowiecka). For example, we suggest Rajfer instead of Gajfer; Ingbir instead of Igler and Jugbir; Wurcelman instead of Kurcelman; Kucharska instead of Nuchowska.

In both towns, the procedure detected a few new names (or new variants) not yet in the dictionaries and we submitted these names to Alexander Beider for review. For Staszów these names are: Ajlmacher, Ancman, Apolet, Breze, Chałupnik, Fajerholc, Grosnacht, Grynewize, Gula, Hisencwajg Jekowicz, Kanias, Kopsztajn, Krozman, Kupferber, Kuflewicz, Pantyrer, Pokoik, Rauszer, Rozenburszt, Szajbman, Sznold, Szyndelkauf, Tajerholc, Tanchemowicz, Urfajg, Urmanowicz, Wajnbirer, Wajnperl, Wajschan, Zajdenwar, Zandlicht, Zylberband, Zondlicht and Zylberband. Beider validated all "new" Staszów names, except Hisencwajg, Tajerholc and Zondlicht, considered as spelling errors for Nisencwajg, Fajerholc and Zandlicht, respectively. Beider already had seen the names Pantyrer and Sznold but did not include them in the DJSKP book, because he was not certain they existed. Thanks to Staszów records, he has now concluded they that, indeed, are correct surnames.

For Ostrów Mazowiecka, the new names are: Berengolc, Dergel, Etmański, Golcman, Lasinowicz, Lejtmańska, Liwsza, Nozdierek, Perlamuter, Rajbuda,

Slowatys, Szafit, Szarmacher, Trembliner, Wapno, Wielgolewski and Zegadło. Beider validated all "new" Ostrów Mazowiecka names, except Lasinowicz, Lejtmanska, Liwsza, and Perlamuter, which he believes may incorporate spelling errors. Beider added that the surnames Berengolc and Golcman, indeed, existed despite their peculiar spelling, which resulted, he believes, from a transliteration from the Russian to Polish (made by people who prepared the extracts from civil records originally written in Russian), which in turn produced a hybrid spelling. The "g" comes from Russian (instead of the Polish "h"), while the "c" is the mark of a Polish name. In certain cases (including this one), BMPM does not recognize such names, because treating hybrid names across languages would introduce too many false positives.

After suggesting corrections for half of the suspect surnames, the method provided no good answer for the rest, leaving uncertainty about a disturbing 3.1 percent (Staszów) and 4.6 percent (Ostrów Mazowiecka) of the surnames. For Staszów, here is a sample list of the still suspect surnames :

Akt, Albuel, Alkierski, Atis, Baghadi, Bajger, Balman, Baslaf, Bauwald, Beger, Bejkman, Benska, Birem, Bjankiewicz, Blauenfeld, Blie, Blitentajl, Blusz, Blusztaut, Blusztok, Borelbajm, Boteman, Branat, Brusztajn, Brycka, Bukman, Buksznal, Bulmer, Burkow, Busko, Cecelewicz, Cielewicz, Ciete, Cikowicz, Copen, Cynkar, Cypun, Dyzensztraub, Efektor, Episbaum, Erao, Eudenman, Euher, Fajchman, Fajdelbaum, Fajsztam, Findberg, Finkelsztaut, Fiszenfarb, Flauerman, Forszel, Frajmil, Friszenman, Fugier, Fuker and Fursztajn.

For Ostrów Mazowiecka the still suspect names are (partial list):

Mael, Mejmark, Migron, Mile, Mincber, Molotowska, Moref, Mosiąc, Motlowna, Mysien, Napiętka, Nelrin, Niedzielski, Nowach, Pętnicka, Pientrza, Poniaprya, Pukawka, Purwicz, Rajczykow, Reblińska, Remblinkier, Rolińska, Ronsztejn, Rytolc, Sczesak, Sejtla, Serek, Siberman, Siwowicz, Skiersz, Sławska, Słynowicz, Srenia, Sywowicz, Szeljfer, Szlubów, Szmanowicz, Szrabkowicz, Sztabinowicz, Szwarcyger, Szyckowicz, Tachla, Totub, Tumkenlank, Wąsiak, Wetrow, Wiklowska, Wilgolar, Woldchowna, Wybykow, Zajfler, Zołać, Zubnin.

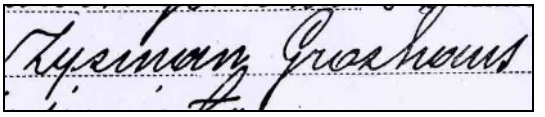
The above lists of most likely non-existent, erroneous surnames in the databases illustrate the difficulty of making educated decisions. Comparisons with online records covering other periods of the town and/or with neighboring towns might help decide about each name, one by one. In any case, the original records should be consulted again.

Success story

The return on investment on this methodology came on one recent case study. The surname in a Staszów birth record had been digitized as Proshaus, a name we suspected to be an error. Our proposed correction, correlating with other records in town, was Groshaus. We ordered the original record and found it

to be indeed for the Groshaus family. The error occurred during data entry as the majuscule letter of the name could be easily interpreted as P instead of G:

But another instance of the surname in the same record (the father name, Zysman Groshaus) confirmed that the name indeed started with a G.



This reading error modified significantly the soundex code of the original name and was fixed by the substitution of a consonant.

VII. Implementation of the Methodology

Name error correction based on redundancy requires full extract indexing: parents' given names and surnames are necessary for comparing records. Moreover, indexing parents' names typically resolves many errors. The error detection and correction methodology is useful only if it can be implemented automatically, or at least if significant steps can be done automatically. The Daitch-Mokotoff soundex codes and Beider-Morse phonetic matching codes were generated for all surnames of both towns and from each surname dictionary thanks to Steve Morse batch one-step tools tailored to process hundreds of names in one time[9] as exemplified in this sample direct correspondence table:

[9] Steve Morse one-step tools for generating DM codes and BMPM codes in batch are respectively located at http://stevemorse.org/census/soundexbatch.html and at http://stevemorse.org/phonetics/phoneticbatch.html.

Surnames	DM Soundex	BMPM
ADLER	38900	adler
AGATER	53900	agater
AJCHENBAUM	056760 046760	ajxenbaum
AJCHENCWAJG	056575 046575 056475 046475	ajxentsvajk
AJCHENHOLC	056585 046585 056584 046584	ajxenholts
AJDELBERG	38795	ajdelberk
AJZENSTAJN	46436	ajzenStajn

Then, using Excel, the author deduced the reverse correspondance tables to list all the dictionary entries that share a given DM code or BMPM value, as in this sample:

DM Soundex	Surnames
38674	ADLENOWSKI
38690	ITALIANER
38745	UDALEWSKI
38760	ADELBAUM AJDELBAUM EDELBAUM
38765	ADELFANG ADLIWANKIN
38784	AJDELFELS EDELFELS
38785	ADLEFLIGIER
38795	ADALBERG ADELBERG AJDELBERG EDELBERG EJDELBERG

Finally, comparison of the surnames encountered during data entry with the dictionary entries and their DM and BMPM values is possible with Excel built-in functions.

VIII. Conclusion

We have presented guidelines for detecting and correcting name errors in genealogical data sets such as civil records from geographical regions covered by a comprehensive dictionary of surnames. This method is best applicable to homogeneous Jewish populations with little incoming migration, such as 19th-century Jewish communities in small and mid-size towns, and probably is less successful in larger cities with heterogeneous populations and Jewish names from various origins. Examination of data sets of surnames from two towns revealed a significant number of errors in digitized civil records. This method offers a strategy to identify potential errors and inconsistencies, to suggest corrections before reviewing the original records and to evaluate the reliability of a given data set. Some cases still require manual treatment. The process is partially automatic and

can be automated further via dedicated software. Future research may generate additional tools to address the unresolved cases .

This study, begun as a preliminary step for the International Institute for Jewish Genealogy (IIJG) Reconstruction of Destroyed Jewish Communities project, demonstrates that quality improvement of individual data should be a prerequisite before merging several data sets together. When dictionaries of Jewish surnames for locations such as Bulgaria, Greece, Hungary, The Netherlands and Romania are published, the error detection and correction methodology will be applicable to more geographical regions.

IX. Acknowledgments

Several experts contributed to this study. Alexander Beider gave valuable guidance and feedback on surnames. Stanley Diamond provided sample civil records from Jewish Records Indexing-Poland. Avotaynu publisher Gary Mokotoff supplied the lists of surnames from the surname dictionaries. Steve Morse lent support with his one-step tools. An early version of this study was published in Avotaynu (Stroweis, 2011).

REFERENCES

Beider A. and Morse S. 2008. “Beider-Morse Phonetic Matching: An Alternative to Soundex with Fewer False Hits,” Avotaynu Vol. XXIV, No. 2, Summer 2008, pp. 12-18; also at http://stevemorse.org/phonetics/bmpm.htm

Stroweis, J.-P., Avotaynu Fall 2011. Vol. XXVII, No. 3, pp. 4-11.

Walker, Francis A. 1878. "Interview of the select committees of the Senate of the United States and of the House of Representatives to make provision for taking the tenth census." 45th Cong., 2d sess. S. Misc. Doc. 26, cited by Richard H. Steckel “The Quality of Census Data for Historical Inquiry: A Research Agenda”, Social Science History, Vol. 15, No. 4 (Winter, 1991), pp. 583-584.

BIOGRAPHY

Jean-Pierre Stroweis is a software engineer born in Paris, France. A past president of the Israel Genealogical Society, he co-chaired the 2004 IAJGS conference in Jerusalem. He is a member of the academic committee of the International Institute for Jewish Genealogy and Staszów town leader for Jewish Records Indexing - Poland.