מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

H. Daniel Wagner

Genealogical Database merging

A tool for the virtual reconstitution of vanished Jewish Communities

The 15th World Congress of Jewish Studies

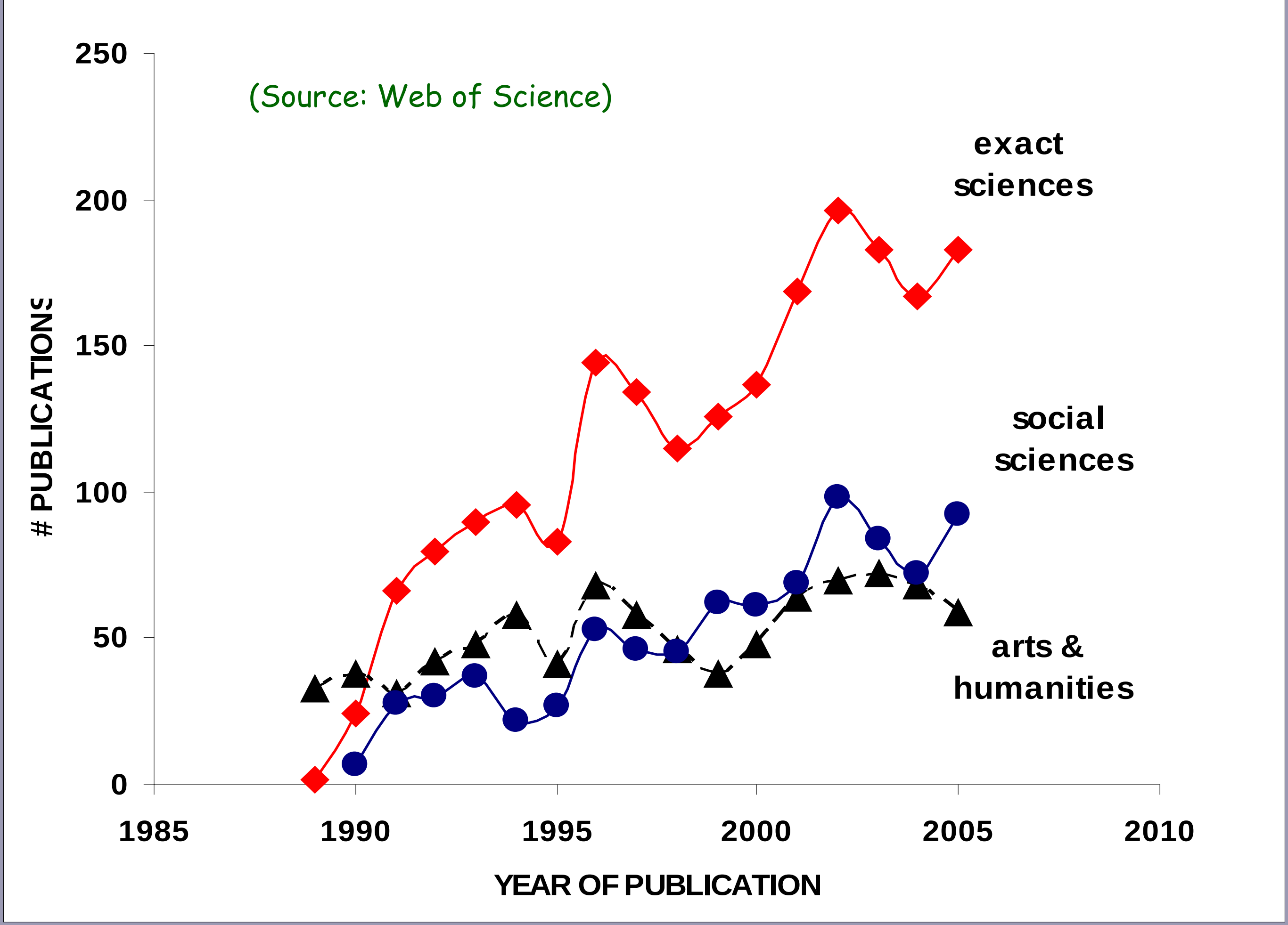
2-6 August 2009, Jerusalem, ISRAEL

The Fifteenth
World Congress
of Jewish Studies

Lecture contents

- Genealogy – A discipline in transition
- What is ‘Virtual Reconstitution’?
- Zdunska Wola as a pilot project
- A tool for ‘Virtual Reconstruction’: Database Merging
- Problems to resolve and simple examples
- Conclusions and recommended future work

Genealogy - A discipline in transition



H.D. Wagner, *Genealogy as an academic discipline*, **Avotaynu (2006)**

Genealogy - A discipline in transition

TOOLS AND INTERESTING PROBLEMS FROM THE EXACT SCIENCES:

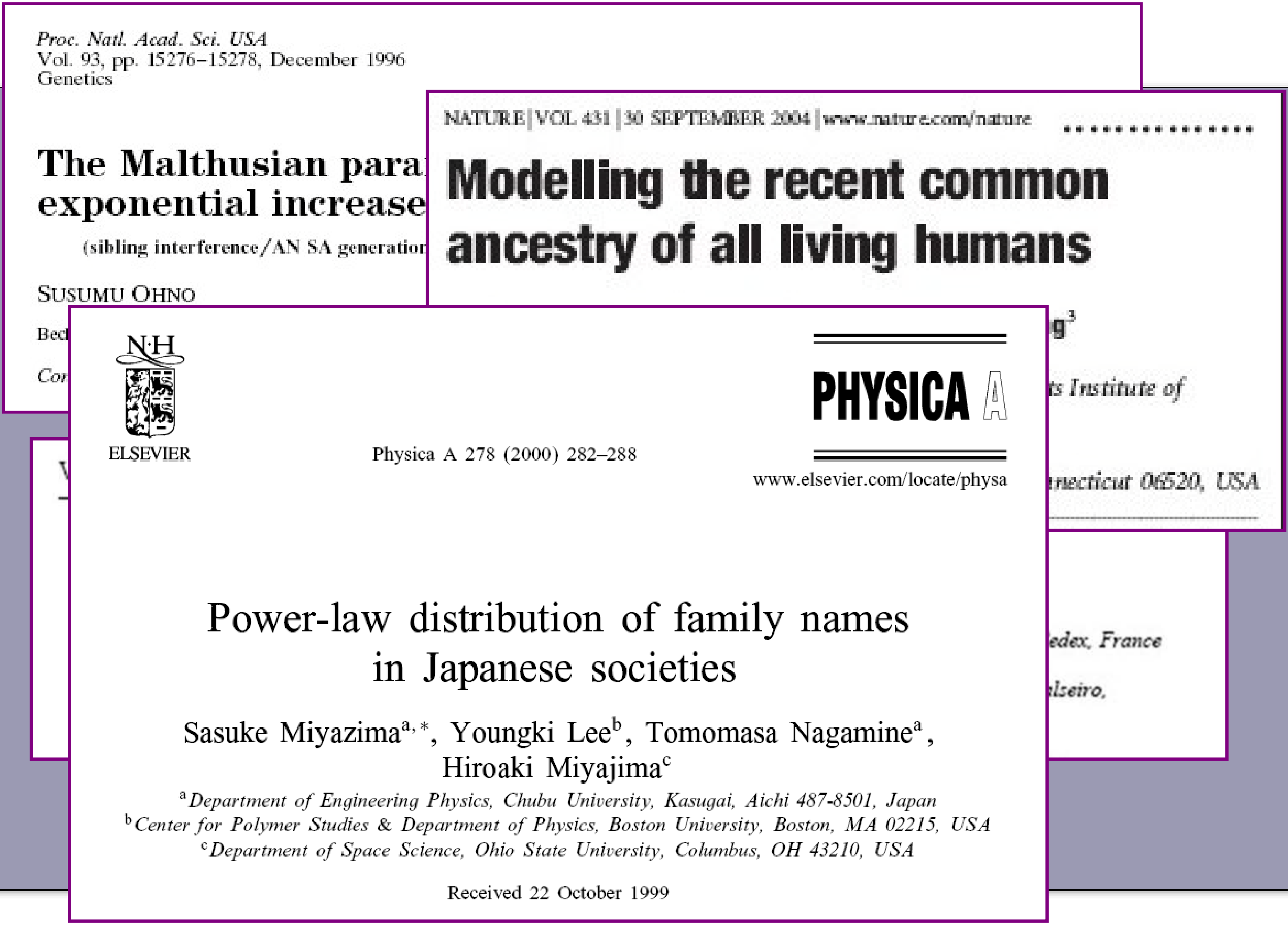
- Mathematics & Statistics – ‘Perturbations’ in family trees due to ‘tribal/village/royal etc. confinement effects’ leading to intermarriages

- Statistical Physics – Study of the size and geographical distribution of migratory movements (or stability of surnames) using annual telephone directories, leading to universal scaling laws (as in physics)

- Molecular Biology – DNA studies yield insights into the origins of human groups, the transmission of genetic diseases, the solution of historical and genealogical debates, problems of forensic nature, etc.

- Computer Science – Infinite repositories for databases, data retrieval is instantaneous, pure research tools (specific search engines, improved soundexes, database merging, etc)

H.D. Wagner, *Genealogy as an academic discipline*, **Avotaynu (2006)**



Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 15276–15278, December 1996
Genetics

The Malthusian para
exponential increase

(sibling interference/AN SA generatio

Susumu Ohno

Bec

Cor

NATURE | VOL 431 | 30 SEPTEMBER 2004 | www.nature.com/nature

Modelling the recent common ancestry of all living humans

g[3]

ts Institute of

necticut 06520, USA

V

edex, France

lseiro,

ELSEVIER

Physica A 278 (2000) 282–288

PHYSICA A

www.elsevier.com/locate/physa

Power-law distribution of family names in Japanese societies

Sasuke Miyazima[a,*], Youngki Lee[b], Tomomasa Nagamine[a], Hiroaki Miyajima[c]

[a] *Department of Engineering Physics, Chubu University, Kasugai, Aichi 487-8501, Japan*
[b] *Center for Polymer Studies & Department of Physics, Boston University, Boston, MA 02215, USA*
[c] *Department of Space Science, Ohio State University, Columbus, OH 43210, USA*

Received 22 October 1999

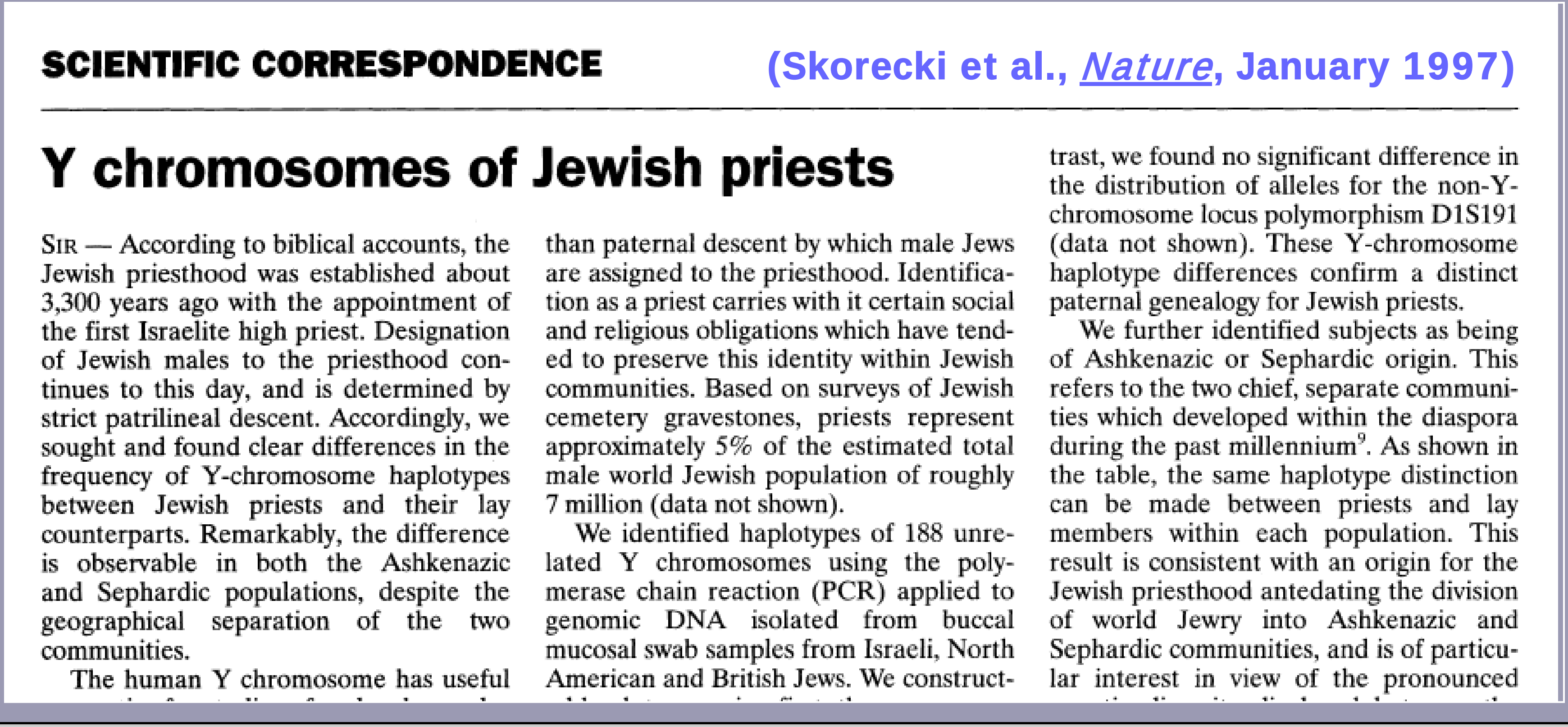
MOLECULAR BIOLOGY

Modern genetic research provides new insights into the ways we are related to each other

Two powerful tools:

1. the Y-chromosome (transmitted from father to son without alteration)
2. Mitochondrial DNA, inherited from mother

Significant differences are found between the Y-chromosomes of Cohanim and all other Jews. The origin (coalescence) of Cohen chromosomes may be traced to 106 generations back (106 x 25 = 2,650, 106 x 30 = 3,180) years ago (between the Exodus and destruction of first temple), with very small differences between Sepharadim and Ashkenazim.



SCIENTIFIC CORRESPONDENCE

(Skorecki et al., *Nature*, January 1997)

Y chromosomes of Jewish priests

SIR — According to biblical accounts, the Jewish priesthood was established about 3,300 years ago with the appointment of the first Israelite high priest. Designation of Jewish males to the priesthood continues to this day, and is determined by strict patrilineal descent. Accordingly, we sought and found clear differences in the frequency of Y-chromosome haplotypes between Jewish priests and their lay counterparts. Remarkably, the difference is observable in both the Ashkenazic and Sephardic populations, despite the geographical separation of the two communities.

The human Y chromosome has useful

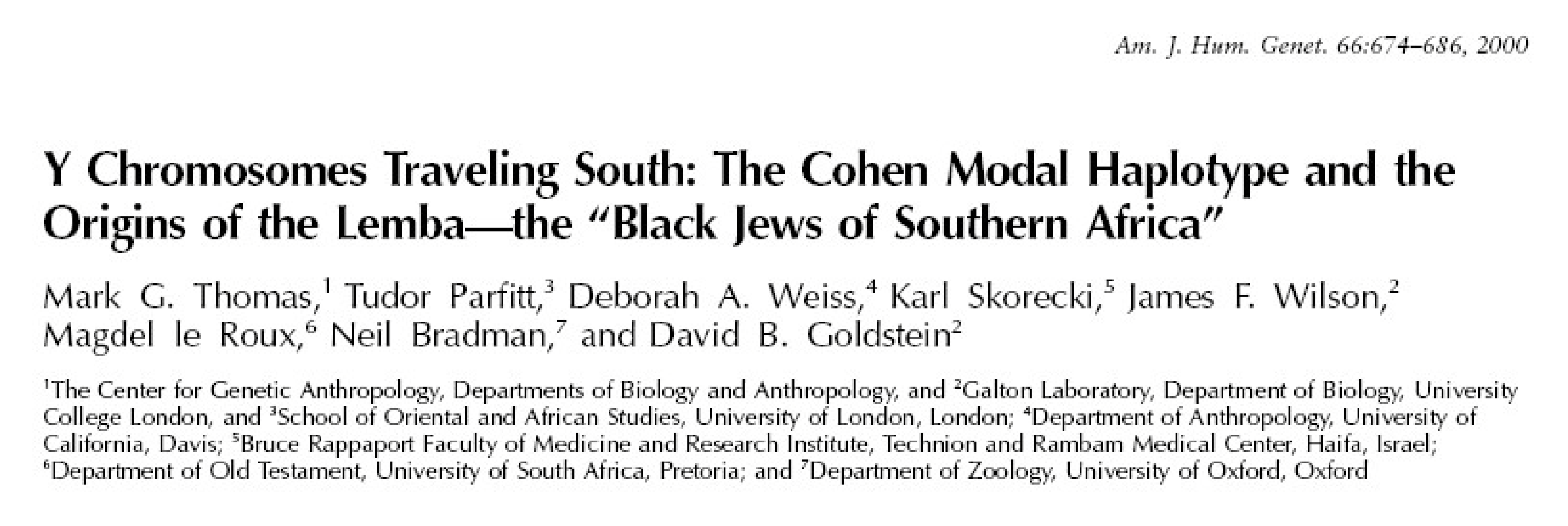
than paternal descent by which male Jews are assigned to the priesthood. Identification as a priest carries with it certain social and religious obligations which have tended to preserve this identity within Jewish communities. Based on surveys of Jewish cemetery gravestones, priests represent approximately 5% of the estimated total male world Jewish population of roughly 7 million (data not shown).

We identified haplotypes of 188 unrelated Y chromosomes using the polymerase chain reaction (PCR) applied to genomic DNA isolated from buccal mucosal swab samples from Israeli, North American and British Jews. We construct-

trast, we found no significant difference in the distribution of alleles for the non-Y-chromosome locus polymorphism D1S191 (data not shown). These Y-chromosome haplotype differences confirm a distinct paternal genealogy for Jewish priests.

We further identified subjects as being of Ashkenazic or Sephardic origin. This refers to the two chief, separate communities which developed within the diaspora during the past millennium[9]. As shown in the table, the same haplotype distinction can be made between priests and lay members within each population. This result is consistent with an origin for the Jewish priesthood antedating the division of world Jewry into Ashkenazic and Sephardic communities, and is of particular interest in view of the pronounced

The LEMBA (South of Africa) and the Samaritans do carry the Y-chromosome type assigned to Cohanim !



Am. J. Hum. Genet. 66:674–686, 2000

Y Chromosomes Traveling South: The Cohen Modal Haplotype and the Origins of the Lemba—the "Black Jews of Southern Africa"

Mark G. Thomas,[1] Tudor Parfitt,[3] Deborah A. Weiss,[4] Karl Skorecki,[5] James F. Wilson,[2] Magdel le Roux,[6] Neil Bradman,[7] and David B. Goldstein[2]

[1]The Center for Genetic Anthropology, Departments of Biology and Anthropology, and [2]Galton Laboratory, Department of Biology, University College London, and [3]School of Oriental and African Studies, University of London, London; [4]Department of Anthropology, University of California, Davis; [5]Bruce Rappaport Faculty of Medicine and Research Institute, Technion and Rambam Medical Center, Haifa, Israel; [6]Department of Old Testament, University of South Africa, Pretoria; and [7]Department of Zoology, University of Oxford, Oxford

The focus

Database merging

A tool for

(i) assembling a better picture of one’s ancestor

(ii) the virtual rebuilding of vanished Jewish Communities

H.D. Wagner, <u>Roots-Key</u> (2007)
H.D. Wagner, <u>Avotaynu</u> (2008)

Why do this ?

To perpetuate the memory of a lost ancestor

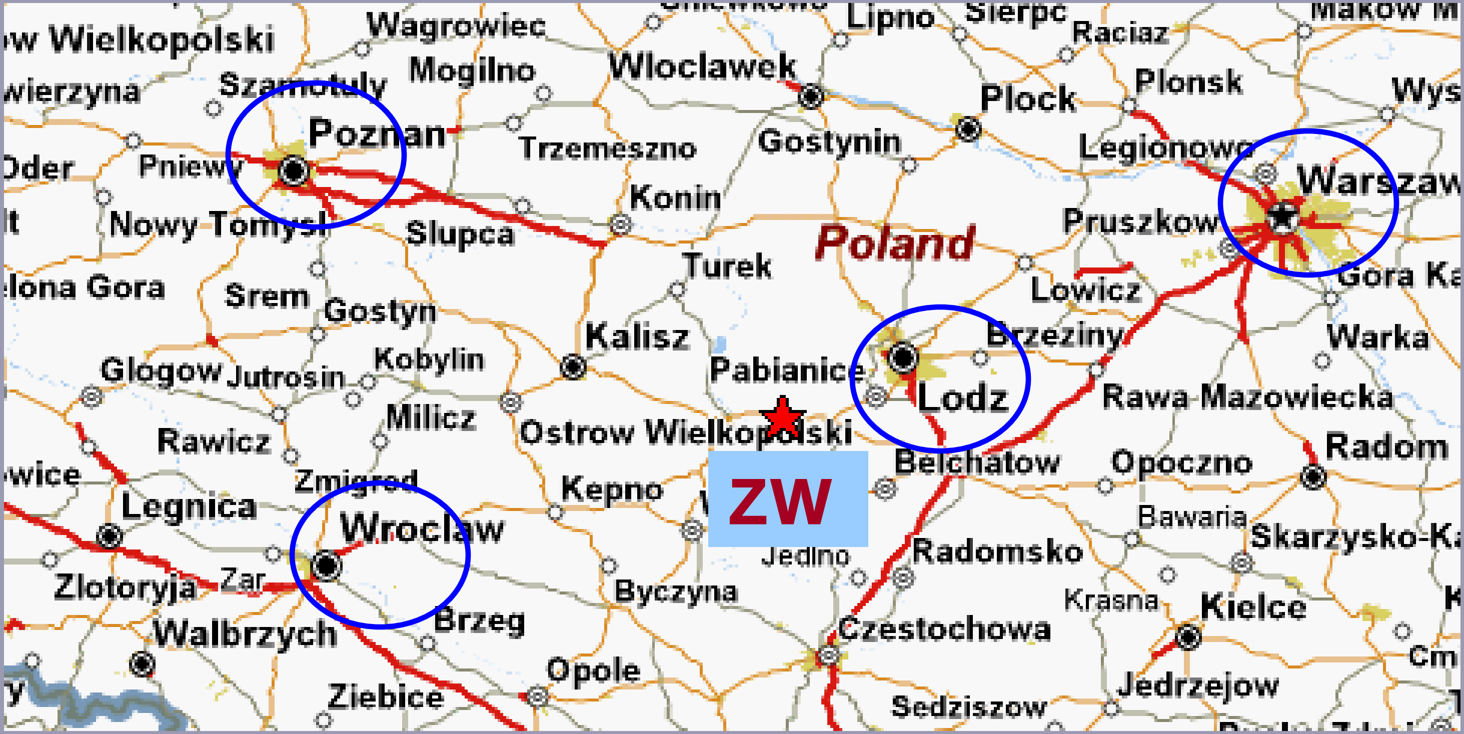
To perpetuate the collective memory of a lost community

To create a significant act of remembrance

To signify to the world what has been lost to humanity

Zdunska Wola as a pilot project

Where is it?



Zdunska Wola as a pilot project

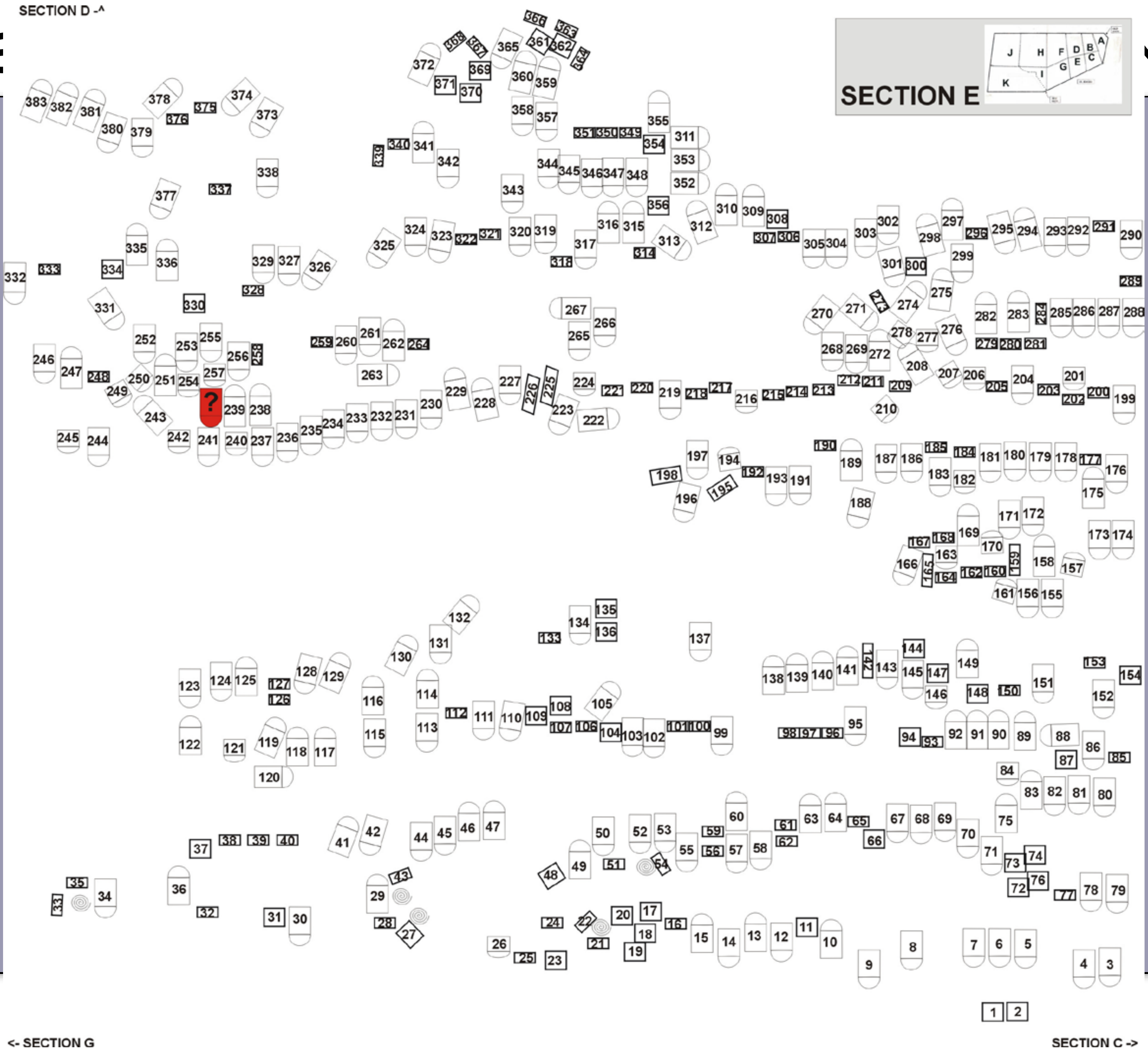
RECENTLY COMPUTERIZED:

- 32,000 B/M/D metrical data (1808-1942, at USC & Lodz archives)

- 3,500 entries for Jewish families in the 28 Books of Permanent Residents, or KLS (up to 1931).

- 3,505 tombstones in the Jewish cemetery (1828-1940) (including photographs, exact locations).

CRE N



Zdunska Wola as a pilot project

- 500 applications by Jews for identity cards (1930-1934) (including photographs)
- 2,300 entries from the ZW Yizkor Book necrologies
- 1,300 names on the memorial monument at the Trumpeldor cemetery in Tel-Aviv
- Thousands of Pages of Testimony (PoT) at Yad Vashem
- 1,100 surnames from the 1929 Polish Business Directory

Database Merging

- The information about every individual, contained in currently separate databases, is to be channeled into a single database

- This ‘unification process’ is called DATABASE INTEGRATION/MERGING

- Non-trivial !

Not trivial because...

- The name of an individual may have been registered with different spellings in different databases

- Spelling problems appear when the databases were created in different languages (Polish/Russian/Yiddish...)

- Birth dates of a given individual are often different in various databases (also: Julian vs Gregorian vs Jewish calendars)

- Often there are no surnames on Jewish tombstones and there may be 5 different Abraham ben Yakov in a given year

Not trivial because...

- The birth metrical record appears as Efraim Yehuda whereas the name on the tombstone is Fiszel Lajb...

- Yitzchak Majer on the tombstone, but only Meyer in the D data

All of this means that merging criteria of ‘identicalness’ must be defined as accurately as possible, with an assigned probability level. [*Some commercial genealogy packages do this already by prompting the user regarding possible matches for 2 individuals who seem to be the same person*.]

SOUNDEX: A single code for names that sound the same

Ehud OLMERT or ULMERT or ULMART etc...

Gideon KOUTS or KUC or KUTZ etc...

Shimon PERES – PERETS – PEREC – PEREZ etc...

Concept patented in 1918

Used by National Archives to organize US Federal Census data 1880-1920.

Problem with NARA Soundex: ZILBER (Z416) ≠ SILBER (S416) !

1985: The Daitch-Mokotoff Soundex

D-M (Zilber) = D-M (SILBER) = 487900

SOUNDEX

The Daitch-Mokotoff Soundex System is not perfect:

Looking for the D-M code of ZILBER, you will find SZLEIFER (a false hit) with the same code!

2008 – The Beider-Morse Phonetic Matching algorithm

More problems

- The deceased may have been registered in the D metrical data several years after the (true) date figuring on the tombstone.

- How to deal with families in which a sudden change of surname occurs (In my family: PIETRKOWSKI suddenly became MANOWICZ !)?

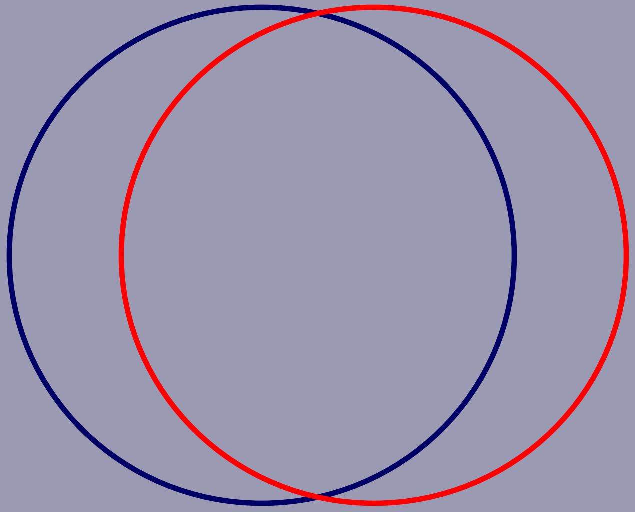
And yet more problems...

- Assume 2 people with the same surname are present in a town, but no formal connection exists (no documentation). What is the probability that they indeed belong to the same family?

Possible clues:
(1) The probability of belonging to the same family is higher if the surname is rare in that town (KUMEC in Konskie).
(2) The probability of belonging to the same family is higher if the children in both group bear a similar first name, possibly pointing to a common grandparent.

Database merging

Example #1: Metrical death + birth records

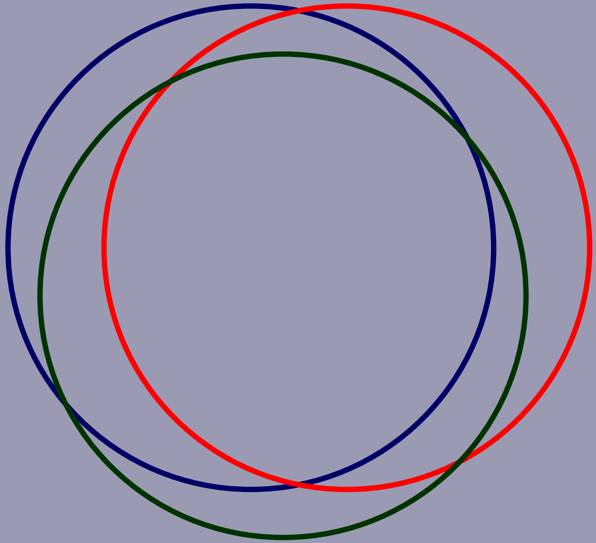


Birth records usually include also the names, ages and occupations of the parents

Death records, on the other hand, usually include age at death and often identify surviving family members.

Database merging

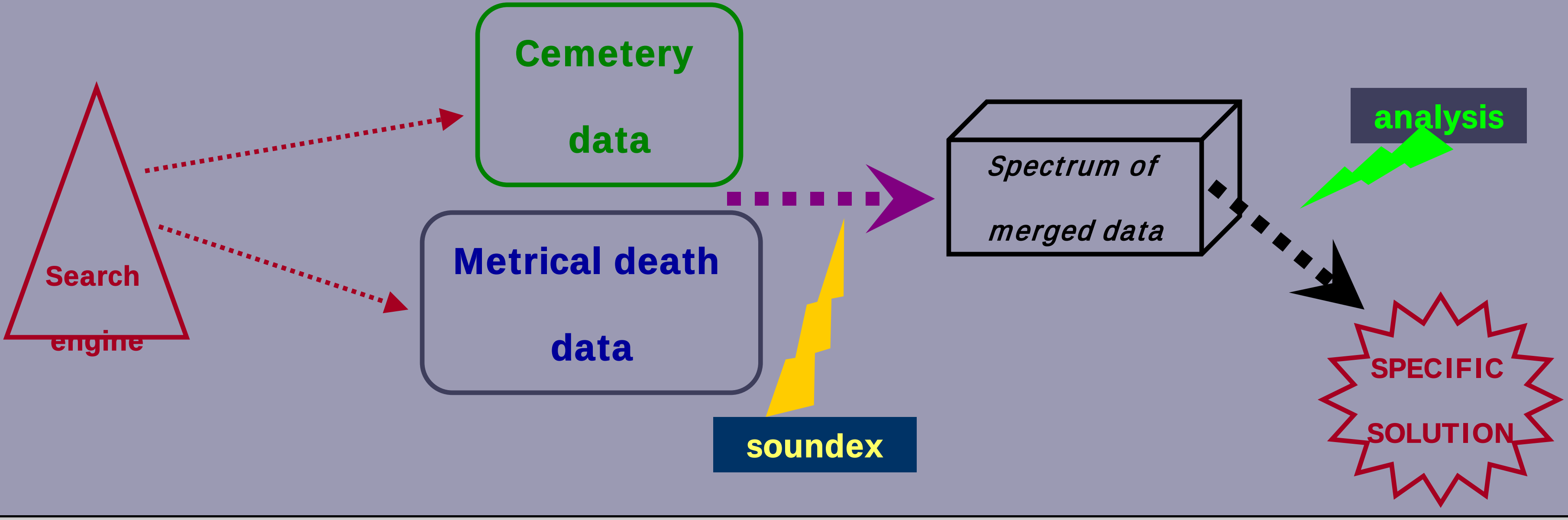
Example #2: death records + birth records + cemetery records



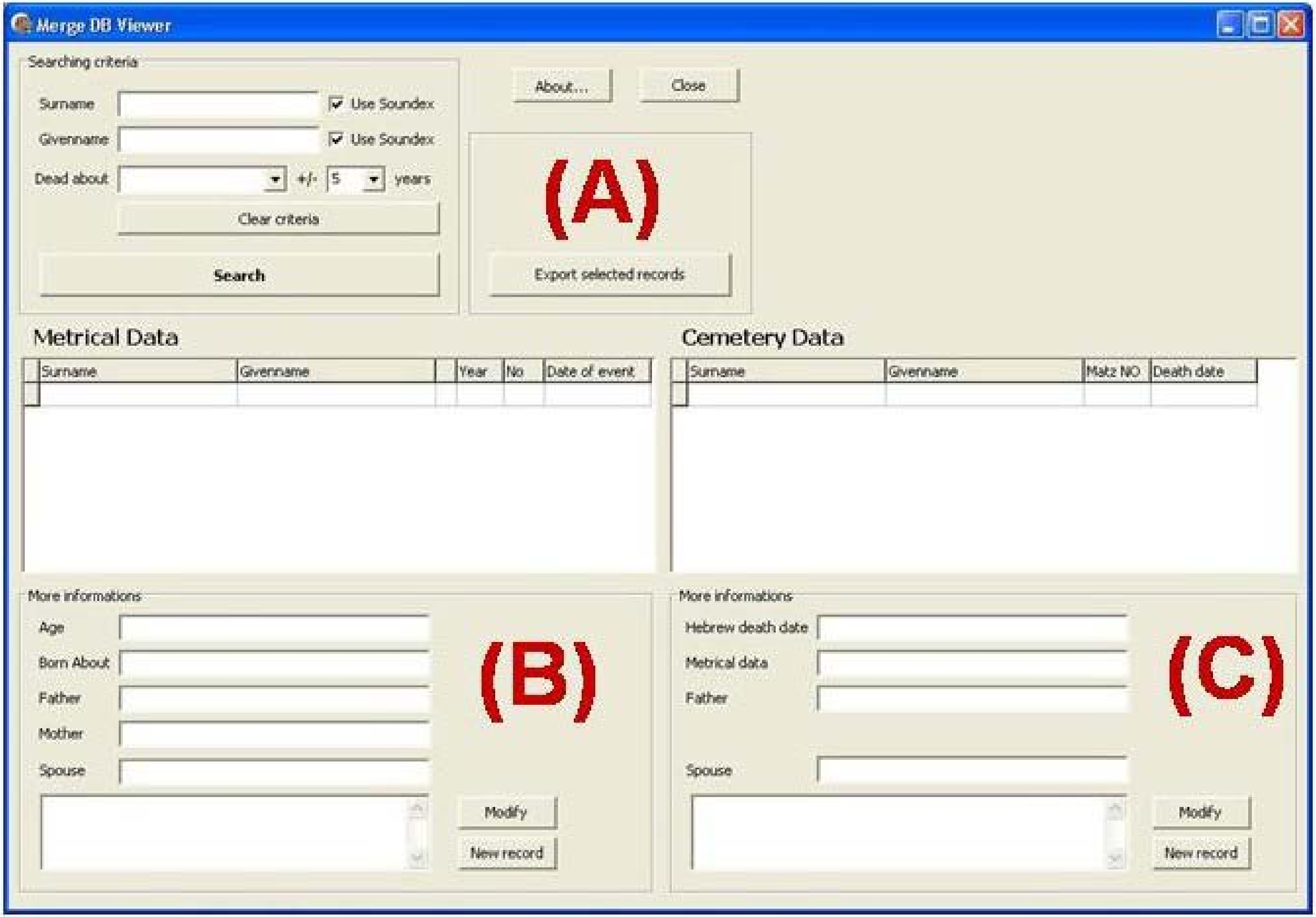
Software for Database merging

- Phase I – Creation of metrical death DB and cemetery DB for Zdunska Wola (Excel DBs with the right format).

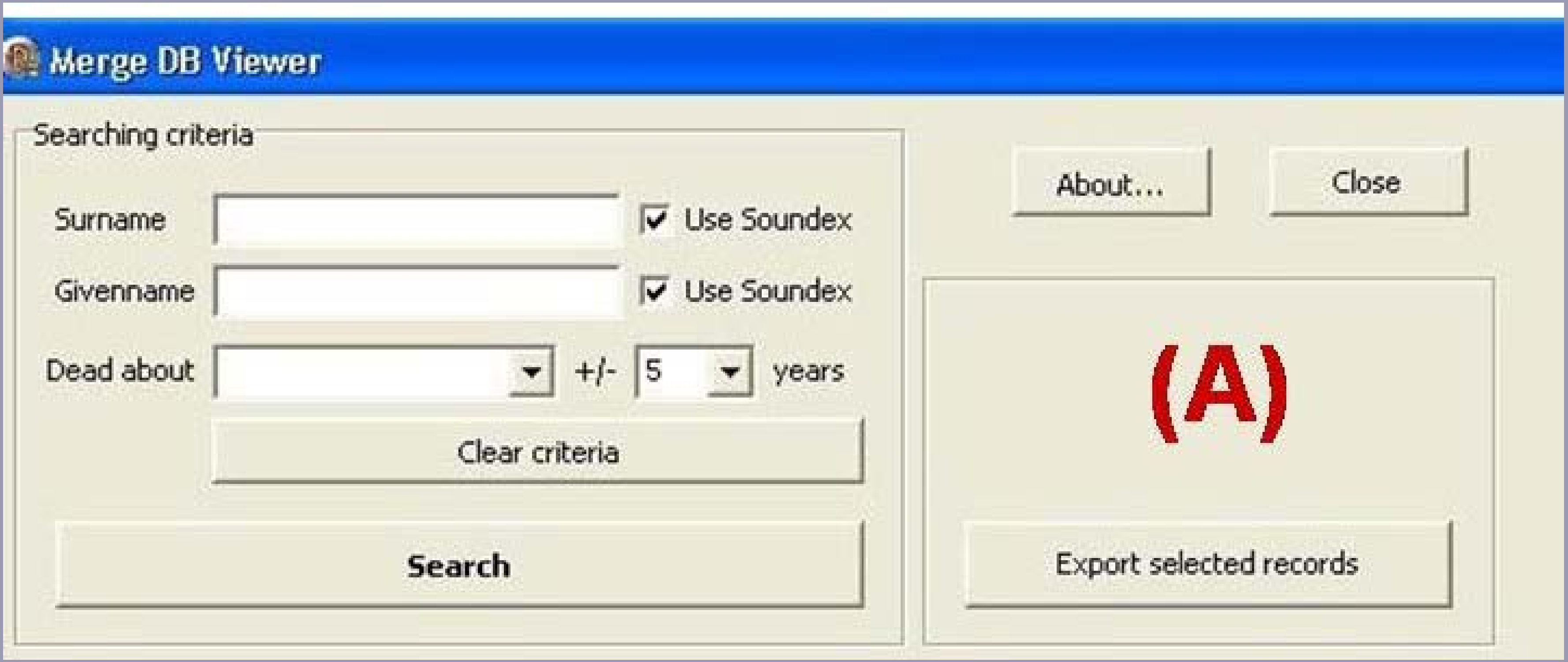
- Phase II – Computerized merging algorithm (including D-M soundex)



Software for Database merging



Software for Database merging



Merge DB Viewer

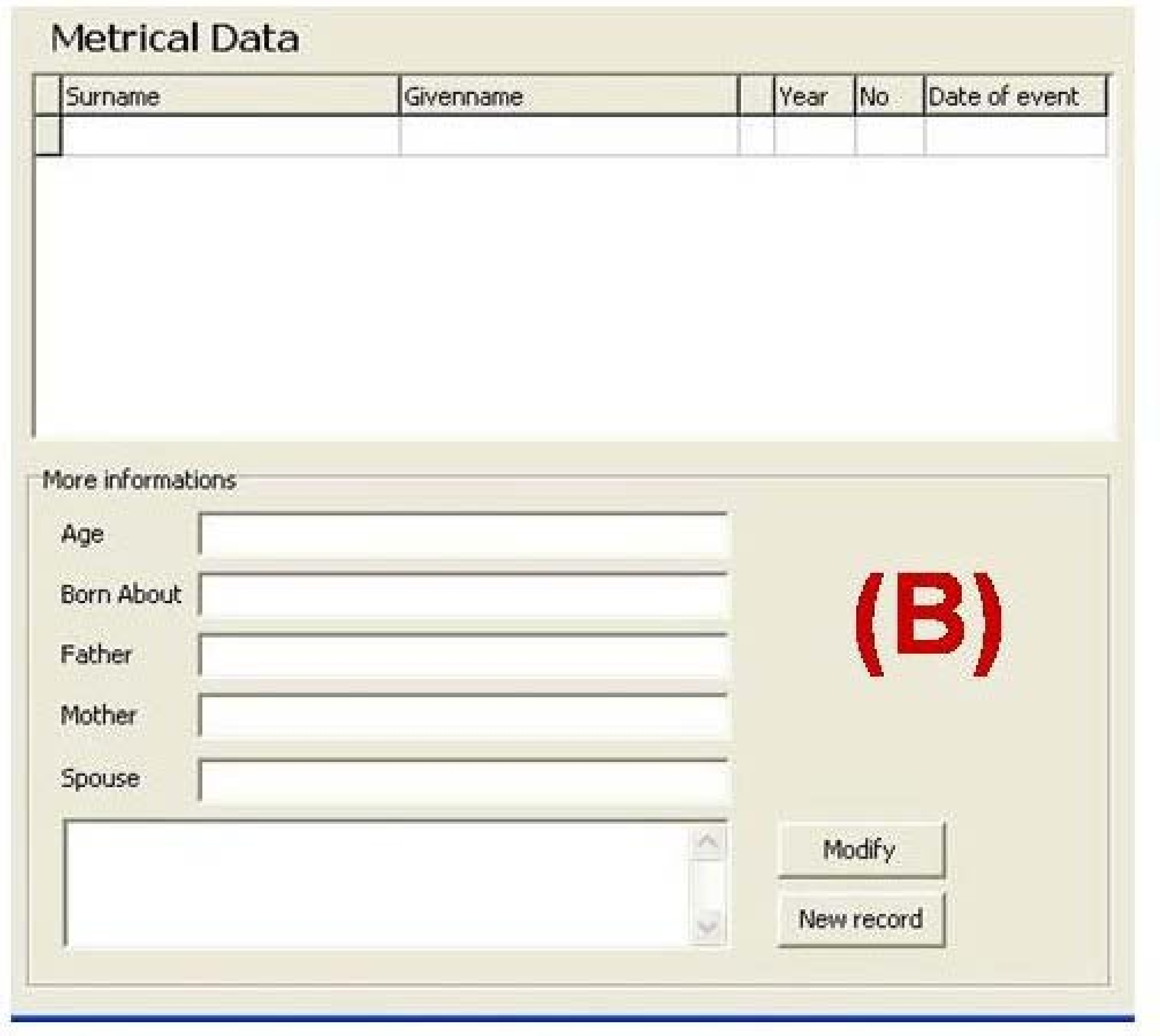
Searching criteria

Surname ☑ Use Soundex

Givenname ☑ Use Soundex

Dead about +/- 5 years

Software for Database merging



Metrical Data

	Surname	Givenname		Year	No	Date of event

More informations

Age

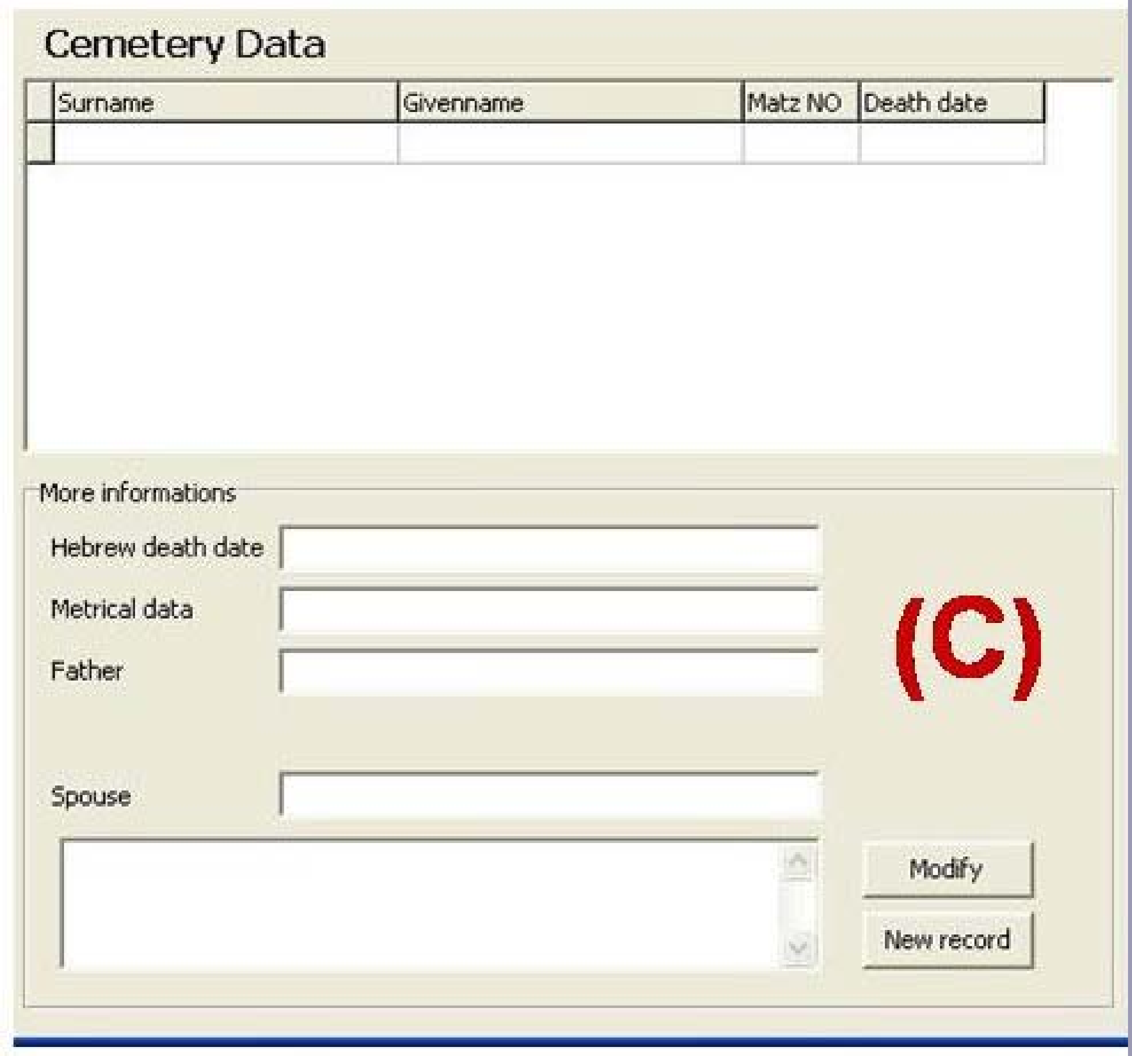
Born About

Father

Mother

Spouse

Software for Database merging



Cemetery Data

Surname	Givenname	Matz NO	Death date

More informations

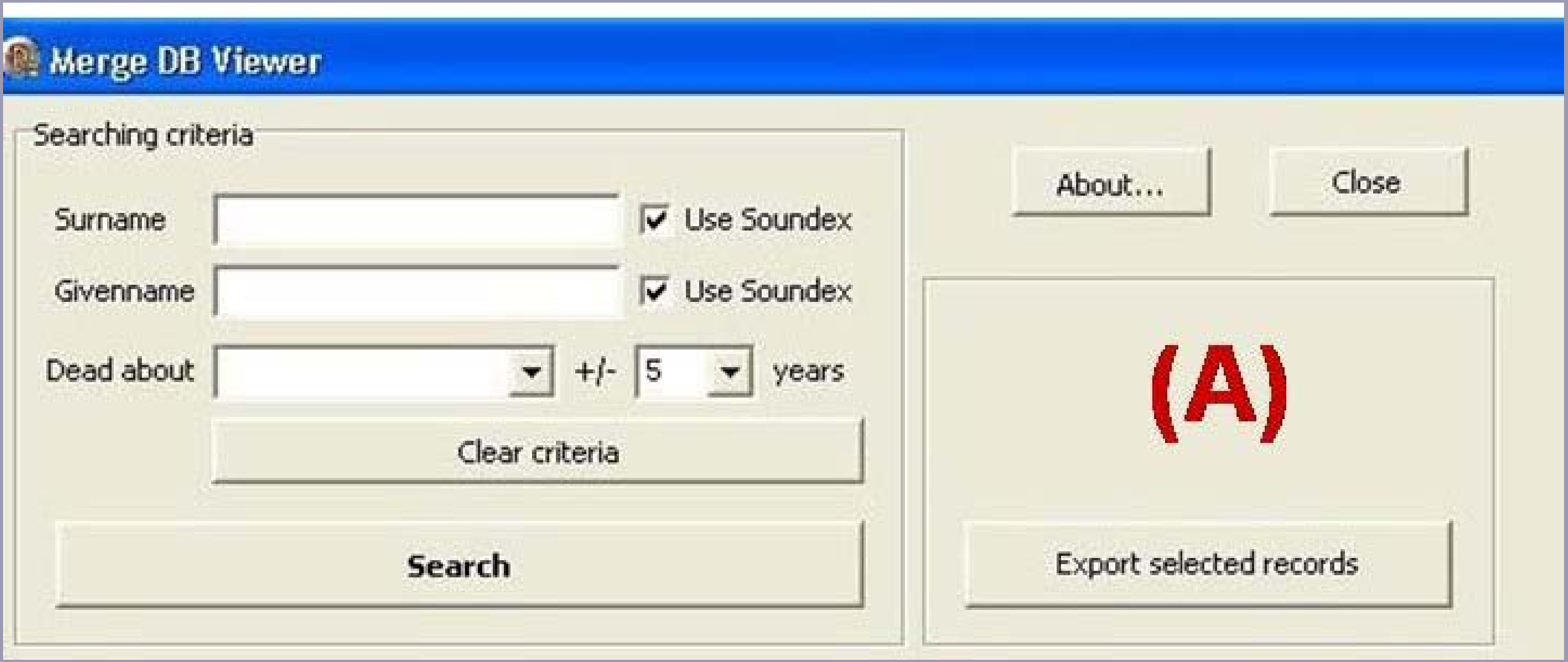
Hebrew death date

Metrical data

Father

Spouse

Software for Database merging



Merge DB Viewer

Searching criteria

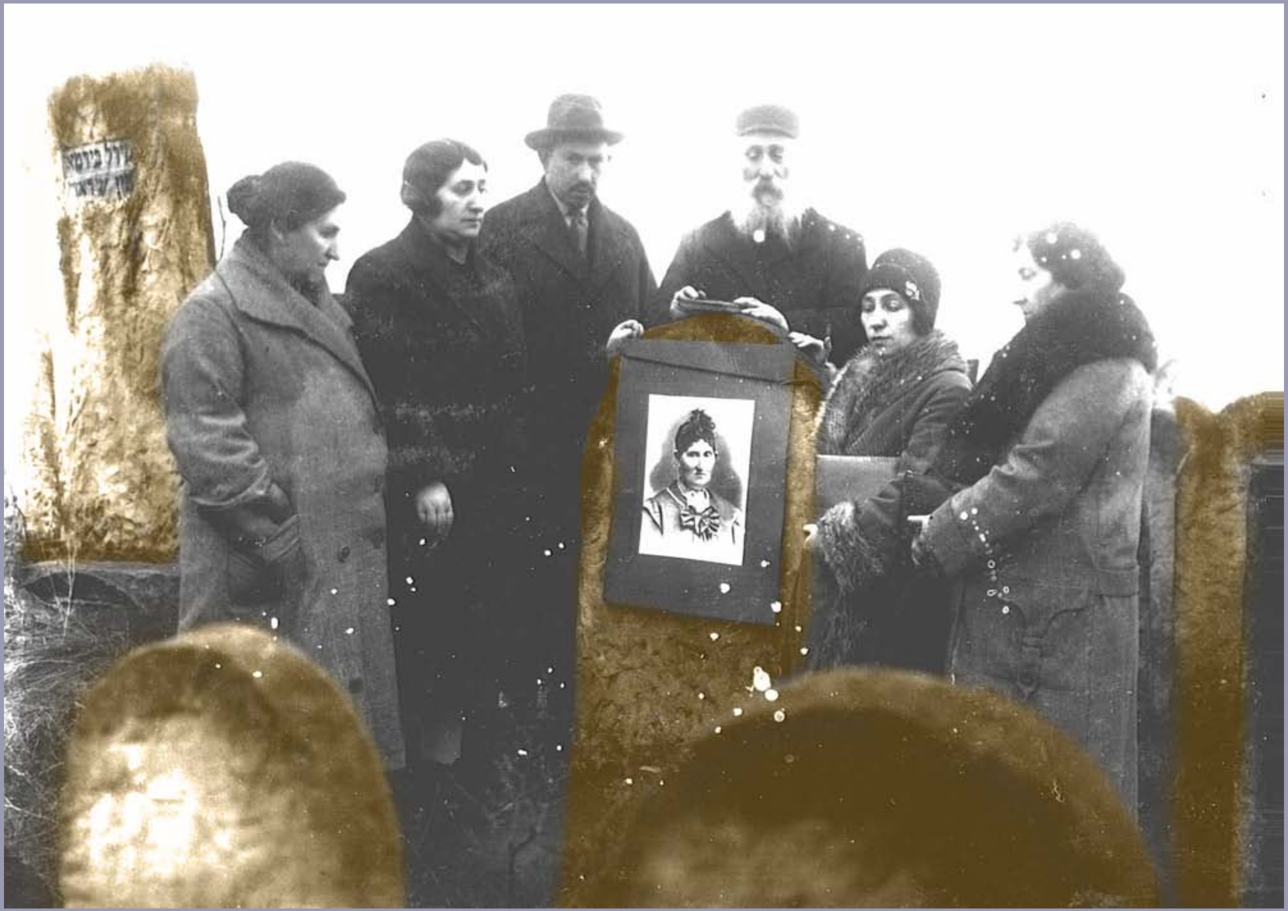
Surname ☑ Use Soundex

Givenname ☑ Use Soundex

Dead about +/- 5 years

Illustrative example

An old picture found in the Yizkor Book of Zdunska Wola:



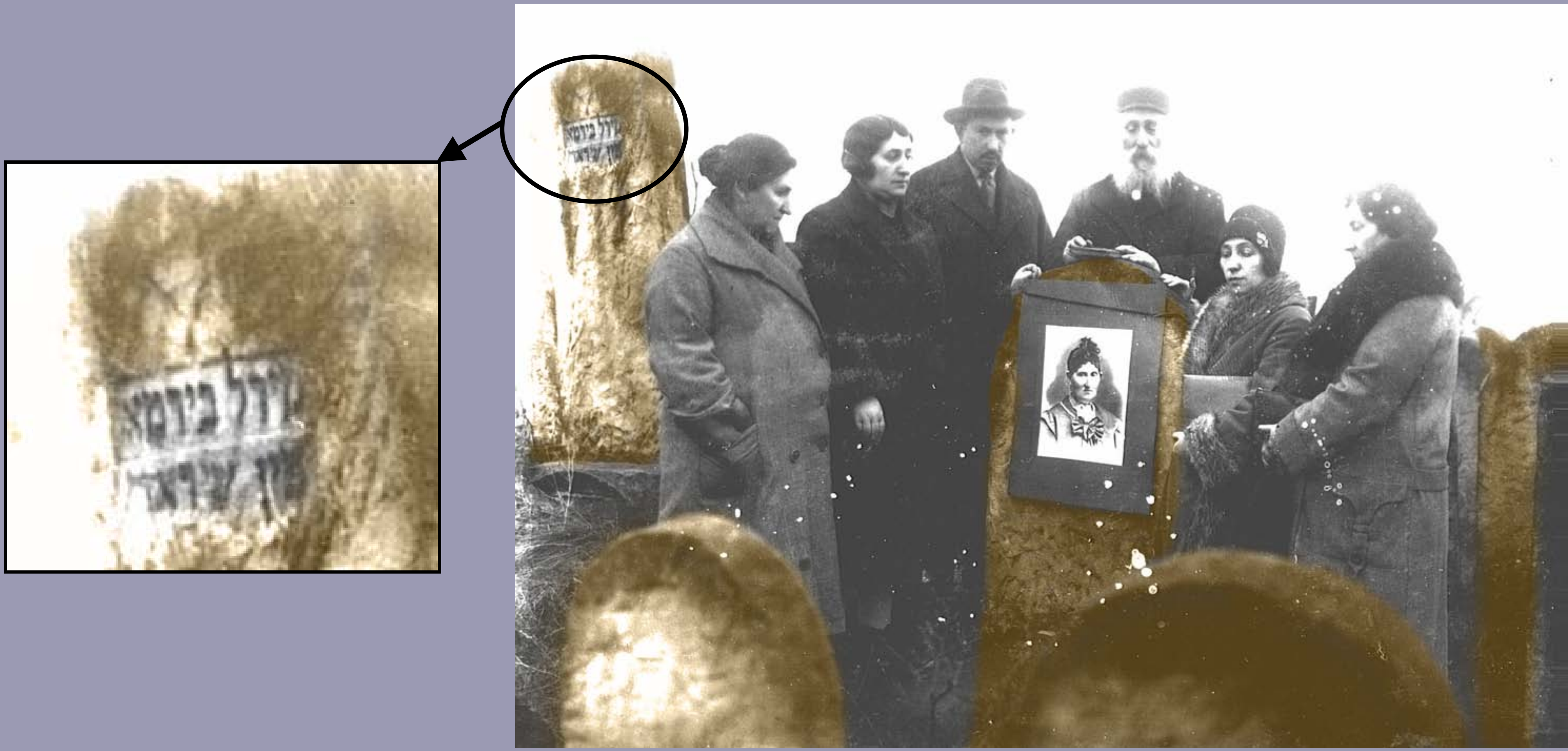
Illustrative example

QUESTIONS:

1. Does any of these stones on the picture still exist in the cemetery?

2. If yes, can we fully identify the deceased (full name, date of death etc)

Illustrative example



Illustrative example

MANUAL MERGING:

1. 1808-1942: 25 Mirel, 201 Mirla

2. Searching for BIRMA(N)/BYRMA(N): Mirla BYRMAN, died 1911, record #84

Illustrative example

COMPUTER MERGING:

_______ Genealogy Merge Data Base _______

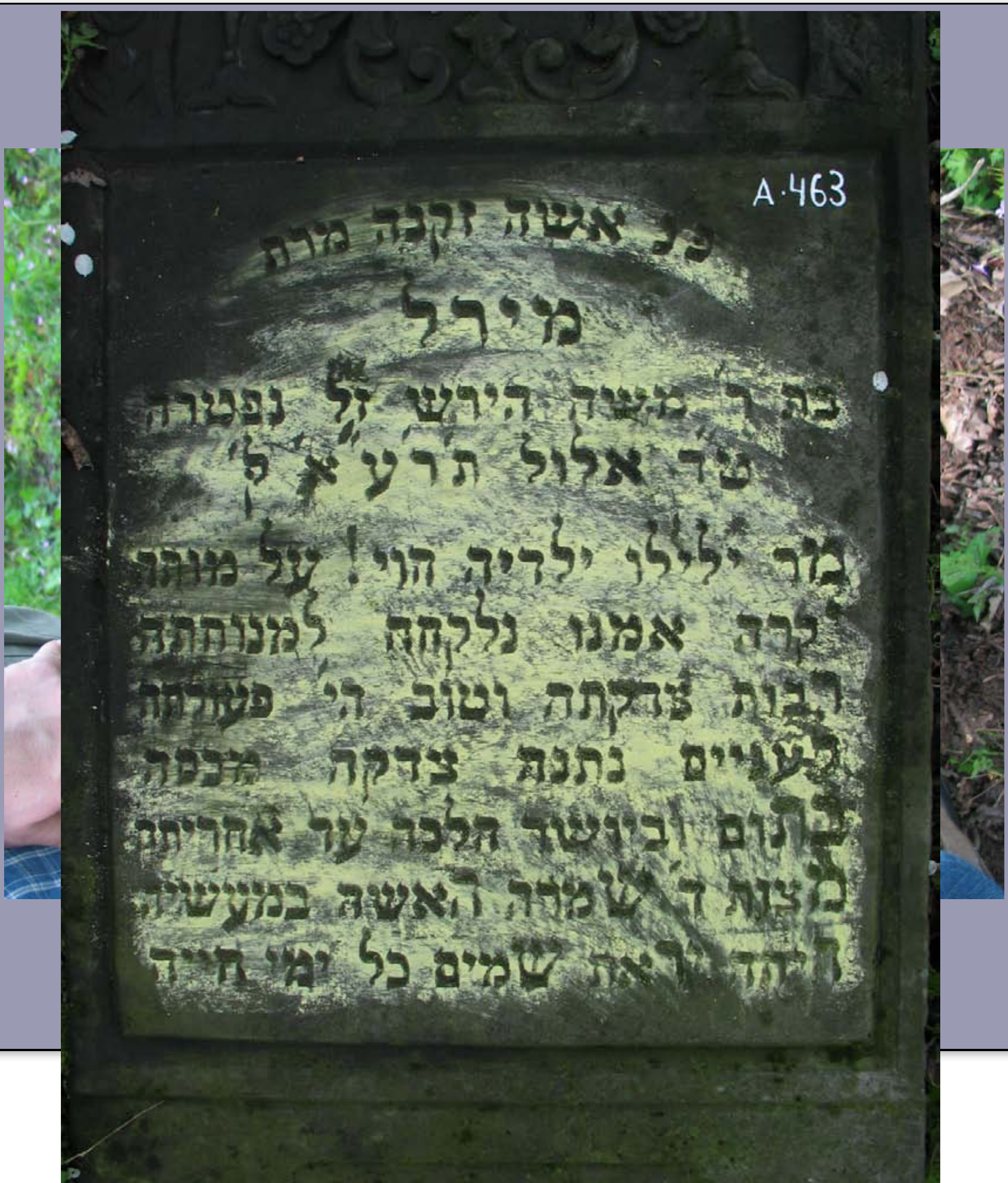
Metrical Data

Name........... BYRMAN Mirla
Act............ D 1911 No. 84
Date of event.. 1911-08-24
Born about.....
Father......... GOLDBART Moszek Gersz
Mother.........
Spouse.........
Comment: from Sieradz widow

Cemetery Data

Name........... BYRMAN [*] ?? Mirel
Death date......1911-09-06
Heb. death date... 13 Elul 5671
Tombstone No....A-463
Father......... Mosze Hirsz
Spouse.........
Comment: old woman

Illustrative example

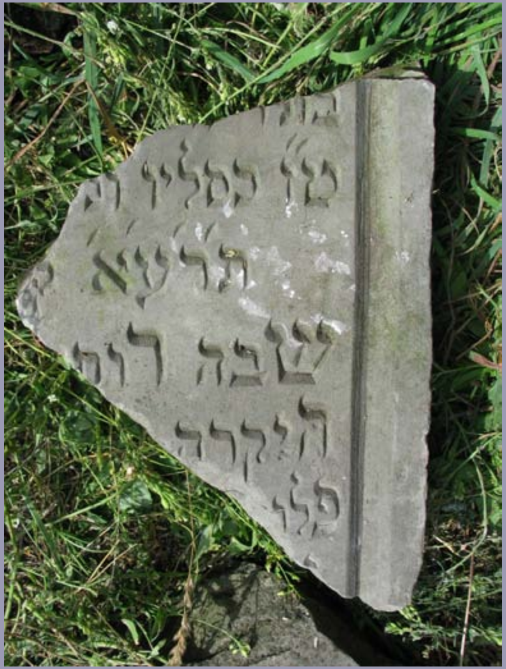


A·463

כל אשה זקנה מרת
מירל
בת ר' משה הירש ז"ל נפטרה
ט'ד אלול תרע"א ל'
מר ילילו ילדיה הוי! על מותה
יקרה אמנו נלקחה למנוחתה
רבות צדקתה וטוב הי' פעולתה
לעניים נתנה צדקה הרבה
בתום וביושר הלכה עד אחריתה
מצות ד' שמרה האשה במעשיה
היתה יראת שמים כל ימי חיה

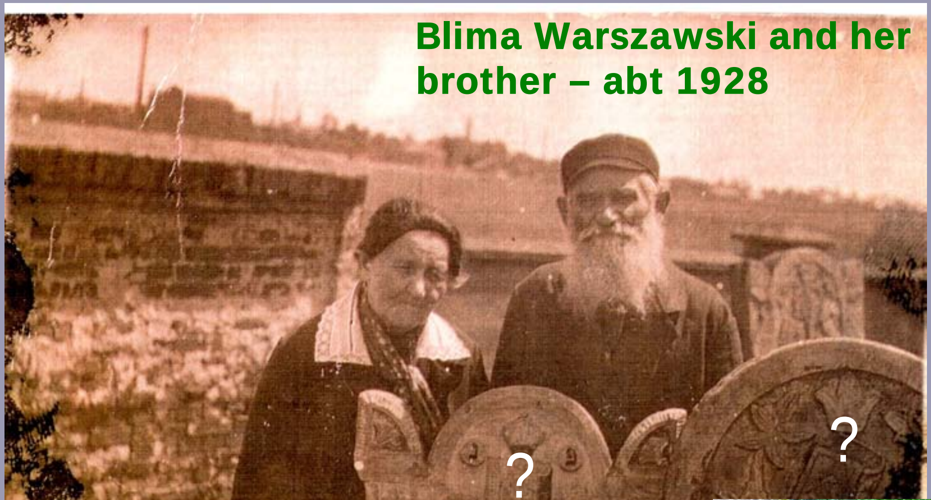
More merging cases

To assign the correct name to a small fragment of tombstone



16 December 1910
A9 - Sura Perla BERKOWICZ

The most important task for descendants is often to identify the grave of an ancestor (from an old pic for example), which usually is difficult without a surname on the tombstone


Blima Warszawski and her brother – abt 1928

A585 – Mordechai WARSZAWSKI

A568 - Chaim

A SIGNIFICANT RESULT

- 3,505 graves in the cemetery of Zdunska Wola
- Only 629 have surnames (18%)
- As a result of merging with metrical death DB:
2170 graves with surnames (62%) !

Conclusions and recommended future work

- Pilot studies help identify various problems arising in merging of Jewish data sets
- Merging software should eventually include more than 2 DBs (passport/ID applications with photos, Yizkor book, Kahal lists etc)
- Include the new Morse-Beider soundex
- Expand software to create ‘restricted’ family trees (thus, for each surname)
- Expand software to integrate ‘restricted’ family trees into ‘connected’ family trees

Conclusions and recommended future work

- Additional complexity is expected when merging entire family trees, but the reward may be exceptionally great:

  (i) the linking of different trees into a *shtetl* ‘forest’, then into a regional ‘forest’

  (ii) the discovery of new family branches due to a second (previously unknown) marriage, etc.

Acknowledgments

The International Institute of Jewish Genealogy

Jakub Zajdel (software creation)

Kamila Klauzinska (metrical database)

SEPTEMBER 2005 – 180th ANNIVERSARY OF ZDUNSKA WOLA

(a different kind of ‘merging’)