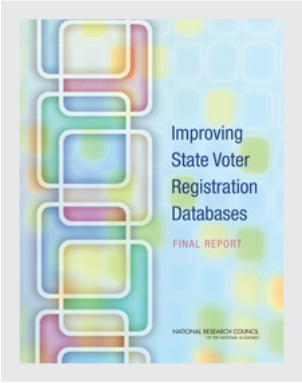
THE NATIONAL ACADEMIES PRESS

This PDF is available at http://nap.edu/12788

SHARE



Improving State Voter Registration Databases: Final Report

DETAILS

128 pages | 8.5 x 11 | PAPERBACK
ISBN 978-0-309-14621-0 | DOI 10.17226/12788

BUY THIS BOOK

FIND RELATED TITLES

AUTHORS

Committee on State Voter Registration Databases; Computer Science and Telecommunications Board; Division on Engineering and Physical Sciences; National Research Council

Visit the National Academies Press at NAP.edu and login or register to get:

– Access to free PDF downloads of thousands of scientific reports
– 10% off the price of print titles
– Email or social media notifications of new titles related to your interests
– Special offers and discounts

Distribution, posting, or copying of this PDF is strictly prohibited without written permission of the National Academies Press. (Request Permission) Unless otherwise indicated, all materials in this PDF are copyrighted by the National Academy of Sciences.

Copyright © National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases

FINAL REPORT

Committee on State Voter Registration Databases

Computer Science and Telecommunications Board

Division on Engineering and Physical Sciences

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

Copyright National Academy of Sciences. All rights reserved.

THE NATIONAL ACADEMIES PRESS **500 Fifth Street, N.W.** **Washington, DC 20001**

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This project was supported by Contract No. N07PC10354 between the National Academy of Sciences and the U.S. Election Assistance Commission. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the organizations or agencies that provided support for this project.

International Standard Book Number-13: 978-0-309-14621-0
International Standard Book Number-10: 0-309-14621-6

Additional copies of this report are available from:

The National Academies Press
500 Fifth Street, N.W., Lockbox 285
Washington, DC 20055
(800) 624-6242
(202) 334-3313 (in the Washington metropolitan area)
Internet: http://www.nap.edu

Copyright 2010 by the National Academy of Sciences. All rights reserved.

Printed in the United States of America

Copyright National Academy of Sciences. All rights reserved.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

Copyright National Academy of Sciences. All rights reserved.

Copyright National Academy of Sciences. All rights reserved.

COMMITTEE ON STATE VOTER REGISTRATION DATABASES

FRANCES ULMER, University of Alaska, Anchorage, *Co-chair*
OLENE WALKER, State of Utah (retired), *Co-chair*
RAKESH AGRAWAL, Microsoft Corporation
R. MICHAEL ALVAREZ, California Institute of Technology
GARY W. COX, University of California, San Diego
PAULA HAWTHORN, Independent Consultant
SARAH BALL JOHNSON, Kentucky State Board of Elections
JEFF JONAS, IBM Corporation
DENISE LAMB, County of Santa Fe, New Mexico
JOHN LINDBACK, Pew Center on the States
BRUCE McPHERSON, State of California (retired)
WENDY NOREN, Boone County Clerk's Office
WILLIAM WINKLER, U.S. Census Bureau
REBECCA N. WRIGHT, Rutgers University

Staff

HERBERT S. LIN, Study Director
KRISTEN BATCH, Associate Program Officer (through August 2008)
ENITA WILLIAMS, Associate Program Officer (since September 2008)
MORGAN MOTTO, Program Associate (through April 2009)
ERIC WHITAKER, Senior Program Assistant (since May 2009)

Copyright National Academy of Sciences. All rights reserved.

COMPUTER SCIENCE AND TELECOMMUNICATIONS BOARD

ROBERT F. SPROULL, Sun Microsystems, Inc., *Chair*
PRITHVIRAJ BANERJEE, Hewlett Packard Company
WILLIAM J. DALLY, NVIDIA Corporation and Stanford University
DEBORAH ESTRIN, University of California
KEVIN KAHN, Intel Corporation, Hillsboro
JAMES KAJIYA, Microsoft Corporation
JOHN E. KELLY, III, IBM
JON M. KLEINBERG, Cornell University
WILLIAM H. PRESS, University of Texas
PRABHAKAR RAGHAVAN, Yahoo! Research
DAVID E. SHAW, Columbia University
ALFRED Z. SPECTOR, Google, Inc.
PETER SZOLOVITS, Massachusetts Institute of Technology
PETER J. WEINBERGER, Google, Inc.

Staff

JON EISENBERG, Director
RENEE HAWKINS, Financial and Administrative Manager
HERBERT S. LIN, Chief Scientist, CSTB
LYNETTE I. MILLETT, Senior Program Officer
NANCY GILLIS, Program Officer
ENITA A. WILLIAMS, Associate Program Officer
VIRGINIA BACON TALATI, Program Associate
SHENAE BRADLEY, Senior Program Assistant
ERIC WHITAKER, Senior Program Assistant

For more information on CSTB, see its Web site at http://www.cstb.org, write to CSTB, National Research Council, 500 Fifth Street, N.W., Washington, DC 20001, call (202) 334-2605, or e-mail the CSTB at cstb@nas.edu.

Copyright National Academy of Sciences. All rights reserved.

Preface

In late 2006, the National Research Council (NRC) convened the Committee on State Voter Registration Databases. Supported by the U.S. Election Assistance Commission (EAC), the committee was charged with organizing a series of workshops and the preparation of an interim report addressing challenges in implementing and maintaining state voter registration databases and providing advice to the states on how to evolve and maintain these databases in order to share information with other states securely and accurately in fulfillment of the Help America Vote Act of 2002. Specifically, the EAC asked the NRC to convene a number of workshops among state policy officials and information technology experts and Academy-selected technology experts on specific topics of interest related to state voter registration databases, to prepare an interim report drawing on these workshops that describes challenges in implementing and maintaining state voter registration databases, and to provide a final report to the EAC on a plan for achieving database interoperability. This plan would provide advice aimed at assisting the states in maintaining statewide voter registration databases that are capable of sharing information with other intrastate and federal databases, as well as across state lines, securely and accurately and address concerns of state technical representatives responsible for database implementation and maintenance.

In April 2008, the committee released its interim report,[1] which outlined various challenges to the deployment of state voter registration databases and described potential solutions to these challenges. These solutions fell into two categories: those that could have been implemented prior to the November 2008 election, and others that would have required a longer timeline for implementation.

This final report builds extensively on that interim report. So that this report can stand by itself, it includes nearly all of the material from the interim report, though in some places, that material has been revised to clarify the committee's intent. In other places, material has been reorganized. This final report repeats all of the recommendations provided in the interim report because there remains a need for those particular recommendations, but it also adds new analytical material and makes a number of new recommendations.

Note that this study was not intended to address all of the issues associated with voter registration.

[1] National Research Council, *State Voter Registration Databases—Immediate Actions and Future Improvements: Interim Report,* The National Academies Press, Washington, D.C., 2008.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Rather, the report focuses on the functioning of state voter registration *databases*, and it does not address other important issues, such as barriers that different groups—minority groups, the poor, voters in the U.S. armed forces and/or serving abroad—face when attempting to register. In addition, although the committee provided specific information on best practices when it could, a comprehensive survey of best practices or compilation of a detailed "how-to manual" related to voter registration databases was beyond the scope of the committee's resources and tasking. Rather, this report is intended to depict some of the problems inherent in acquiring, operating, and maintaining VRDs, and to identify some general approaches to addressing these problems. By implication, the committee believes that the details of how specifically to address these problems are best left to the election officials on the ground who know their systems and operating environments best.

This study was undertaken by a committee of 14 people with a broad range of expertise and backgrounds, including election operations, databases, computer and network security, and political science (see Appendix F)—such a range was necessary to address the topic of state voter registration in all of its organizational, technical, and political complexity. To provide a forum for discussions among state and local voting officials and other experts, to put information on the public record, and to educate the committee, workshops were held in August and November 2007 as part of the information gathering for the interim report. Additional workshops were held in May, July, and December 2008, and in March 2009 to conduct more information gathering. Agendas for all of these workshops are provided in Appendix E.

Note: As this report goes to press (September 2009), the National Association of Secretaries of State (NASS) released a report entitled, *NASS Report: Maintenance of State Voter Registration Lists*. According to the accompanying press release, the report describes laws and procedures of various states related to voter registration and the maintenance of voter registration databases, including verification procedures, address confirmation programs, and removal of names from lists. Unfortunately, this report was not available to the committee in time for it to be helpful in the committee's deliberations. The report can be downloaded at www.nass.org.

The committee thanks all those who participated in its workshops and contributed to its deliberations (Appendix E). The committee also thanks the NRC staff for their work on this report. Herbert Lin provided invaluable and expert assistance to the committee by sorting through comments and suggestions and by drafting the report with the committee's guidance. Kristen Batch and Enita Williams did a masterful job in organizing the workshops that served as the information basis for this report and in preparing the report for review. Jon Eisenberg, director of the Computer Science and Telecommunications Board, worked closely with the Election Assistance Commission throughout this study. Morgan Motto and Eric Whitaker provided administrative support.

Frances Ulmer and Olene Walker, *Co-chairs*
Committee on State Voter Registration Databases

Copyright National Academy of Sciences. All rights reserved.

Acknowledgment of Reviewers

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this report:

Brad Bryant, State of Kansas
Paul DeGregorio, Everyone Counts, Inc.
Morris Fiorina, Jr., Stanford University
Rick Hasen, Loyola Law School
Susan Inman, Little Rock, Arkansas
Ray Martinez III, Rice University
Deirdre Mulligan, University of California, Berkeley
Ion Sancho, Leon County, Florida
Pat Selinger, Los Baros, California

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by Elsa Garmire of Dartmouth University. Appointed by the National Research Council, she was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring committee and the institution.

Copyright National Academy of Sciences. All rights reserved.

Copyright National Academy of Sciences. All rights reserved.

Contents

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Copyright National Academy of Sciences. All rights reserved.

Summary

Voter registration plays a central role in elections in the United States. Today, the states operate under a federal mandate (the Help America Vote Act (HAVA) of 2002) to develop "a single, uniform, official, centralized, interactive computerized statewide voter registration list defined, maintained, and administered at the state level."[1] Each state's database must contain the name and registration information of each legally registered voter in the state, and each legally registered voter is required to be assigned a unique identifier. Election officials must perform regular maintenance regarding the accuracy of the registration lists. In addition, the National Voter Registration Act (NVRA) of 1993 establishes rules under which names may be removed from voter registration lists. (A voter registration list is the list of names contained in a voter registration database, and the terms are often used interchangeably.)

Two basic tasks must be performed for voter registration databases: adding individuals to the voter registration database (VRD) and maintaining the VRD.

- Adding individuals to the VRD requires that an attempt be made to verify the information provided on a first-time voter registration application against the relevant state's department of motor vehicles database of driver's license numbers or the Social Security Administration's database of Social Security numbers. If a nonmatch is found, the election officials in most states will make an attempt to contact the applicant so that he or she can provide additional information, but there is variation in how the states manage the nonmatch. In addition, HAVA Section 303(b) requires that an applicant who cannot be matched against one of these databases be allowed to vote on Election Day provided he or she can present appropriate identification at the polling place.
- Maintaining the VRD is needed to keep voter registration information current and to remove the names of ineligible voters and duplicate registrations from the voter lists. This task requires comparing records within a VRD to other records in order to identify duplicate registrations (usually associated with changes of address or name) and (by law) comparing VRDs to databases of known felons, deceased individuals, and individuals declared mentally incompetent.

[1] Section 303(a)(1)(A) of HAVA.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Databases that are accurate and complete require execution of both tasks. (Accuracy refers to the factual correctness of the data that exist in the database; completeness refers to the presence in the database of all individuals who should be in the database.) These tasks require good data as well as good matching procedures. But in practice, a variety of practical problems arise such as data entry error. In addition, to the best of the committee's knowledge, the matching procedures used by many states have not been subjected to rigorous evaluation or testing.

The VRD also drives the preparation of pollbooks (the list of eligible voters in localities for use at polling places). Additional functionality implemented by many states in their (centralized) voter registration systems—including ballot preparation; signature verification for absentee or mail-in ballots; and management of election workers, polling places, petitions, and requirements for disability access under HAVA—assists the local elections official in conducting an election.

RECOMMENDATIONS

The recommendations of the Committee on State Voter Registration Databases are divided into two categories: actions that can be implemented in a relatively short time frame, and actions that will need more time to implement. The committee also notes that although this report focuses on voter registration databases, such databases are always part of a larger system that includes human beings and institutions. Solutions to technical problems may in some cases also require changes to state election law or regulation and/or to state or local practice and procedures, and should not be regarded as being exclusively about changing computer systems.

The Help America Vote Act provided a substantial one-time infusion of money for states to acquire modern information technology for supporting election administration, including the statewide voter registration systems that have been deployed. However, all experience with information technology suggests that the initial acquisition cost of information technology is a relatively small fraction of its life-cycle costs. Ongoing funding streams will be needed both to maintain VRD systems (and the data they hold) in good operating condition over time and to implement many of the improvements described below.

The short-term recommendations address changes of a nontechnical nature related to (1) education and dissemination of information and (2) administrative processes and procedures. The long-term recommendations address the improvement of data collection and entry; matching procedures; privacy, security, and backup; and database interoperability.

All of these recommendations are directed primarily at election officials (voter registrars) at the state and local/county level, and the legislatures and county commissions that make policy regarding the conduct of elections at the state and local level. In some cases, the Election Assistance Commission has a useful role to play as well in facilitating and promoting their implementation.

Short-Term Recommended Actions—Public Education and Dissemination of Information

S-1: Raise public awareness about the legibility and the completeness of voter registration card information. Jurisdictions could take some or all of the following specific steps:

- Emphasize in the instructions for filling out voter registration forms the importance of legibility and completeness (for example, "Please print all responses; if your answers are illegible, your application may be mis-entered, rejected, or returned to you.").
- Conduct media campaigns emphasizing the importance of legibility and completeness in the information provided on voter registration forms.
- Coordinate with third-party voter registration groups and public service agencies, emphasizing the need for their field volunteers to attend to legibility and completeness as they distribute and/or collect registration materials.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Short-Term Recommended Actions—Administrative Processes and Procedures

S-2: Resubmit alternate match queries if the response returned from the Social Security Administration or department of motor vehicles is a nonmatch.

S-3: Provide human review of all computer-indicated removal decisions.

S-4: Improve the transparency of procedures for adding voters and for list maintenance.

S-5: Use printable fill-in online registration forms.

S-6: Perform empirical testing on the adequacy of processes for adding to and maintaining lists.

S-7: Take steps to find and minimize errors during data entry.

S-8: Allow selected individuals to suppress address information on public disclosures of voter registration status.

S-9: Encourage entities sponsoring voter registration drives to submit voter registration forms in a timely manner to reduce massive influxes at the registration deadline.

S-10: Improve information sharing regarding best practices and lessons learned regarding VRD acquisition, operation, and maintenance.

Long-Term Recommended Actions—Funding

L-1: Provide long-term funding for sustaining voter registration database operations.

Long-Term Recommended Actions—Data Collection and Entry

L-2: Develop and promote public access portals for online checking of voter registration status.

L-3: Allow voters to register and to update missing or incorrect registration information online.

L-4: Encourage/require departments of motor vehicles as well as public assistance and disability service agencies to provide voter registration information electronically.

L-5: Improve the design of voter registration forms.

L-6: Encourage and if possible require departments of motor vehicles, public assistance and disability service agencies, tax assessors, and other public service agencies of state and local government in their communications with the public to remind voters to check and update their information.

L-7: Consider providing tracking tags for voter registration forms to improve administrative processes.

Long-Term Actions—Matching Procedures

L-8: Upgrade the match algorithms and procedures used by election officials, the Social Security Administration, and departments of motor vehicles:

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

- Use automated name rooting (the process through which name equivalents are generated, such as "Bill" and "Will" for "William");
- Use automated name ordering (the process through which permutations of possible name equivalents are generated, such as "Lucia Vega Garcia" being represented as "Lucia Vega," "Lucia Garcia," or "Lucia Vega-Garcia");
- Provide wildcard matching capabilities (capabilities for performing searches on incompletely specified names); and
- Use blocking and string comparators (comparison techniques used to generate a score reflecting degree of similarity rather than a simple "match-or-nonmatch" result).

L-9: Use commonly used unique identifiers for voter identification when available and when necessary privacy safeguards are in place.

L-10: Establish standards or best practices for matching algorithms.

L-11: Use the Social Security Death Master File and STEVE[2] (when deployed) for list maintenance.

L-12: Use third-party data when available to resolve possible matches.

L-13: Develop procedures for handling potential disenfranchisement caused by mistaken removals from voter registration lists.

Long-Term Recommended Actions—Privacy, Security, and Backup

L-14: Implement basic practices for backing up important data.

L-15: Implement basic security measures.

L-16: Take measures to help ensure system accessibility during critical times.

L-17: Consider fair information practices as a point of departure for protecting privacy in voter registration databases.

L-18: Take steps to protect voter privacy when voter registration data are released on a large scale.

L-19: Review appropriate nonelection uses of voter registration data.

Long-Term Recommended Actions—Database Interoperability

L-20: Encourage and if possible require state, local, and federal agencies to cooperate with election officials in providing data to support voter registration.

L-21: Use inexpensive data export functions to facilitate data exchange.

L-22: Develop national standards for data-exchange formats for voter registration databases.

[2] STEVE refers to the State and Territorial Exchange of Vital Events, operated by the National Association for Public Health Statistics and Information Systems. At the time of this writing (fall 2009), STEVE has not been fully deployed.

Copyright National Academy of Sciences. All rights reserved.

1

The Context for Voter Registration

Voter registration plays a central role in U.S. elections. Today, every state except North Dakota[1] operates under a federal mandate (the Help America Vote Act (HAVA) of 2002) to develop "a single, uniform, official, centralized, interactive computerized statewide voter registration list defined, maintained, and administered at the state level."[2] Each state's database must contain the name and registration information of each legally registered voter in the state, and each legally registered voter is assigned a unique identifier. Election officials must perform regular maintenance regarding the accuracy and completeness of the registration lists. In addition, the National Voter Registration Act (NVRA) of 1993 establishes rules under which names may be removed from voter registration lists.

As a registration deadline nears, the processing of voter registration applications can present enormous logistical problems. The reason is the sheer volume of voter registration records that need processing (either new voter registration applications or updates of information for already-registered voters)—and especially in a presidential election year, this volume can be a substantial percentage of the entire voter registration database. Most of these documents typically arrive within a few weeks of a registration deadline and, depending on the registration cutoff in a particular state, that can require around-the-clock data entry up to the last minute so that pollbooks can be printed. In some instances, there have been outstanding documents to be processed even on Election Day, and staff were needed to manage inquiries from polling places from a physical file of registration forms not yet entered into the VRD. Adding to the complexity of the data entry process is the fact that election officials may receive a multiplicity of different unstandardized forms, in the mail, over the Internet, from other state agencies, in person, and of course from third-party groups.

[1] North Dakota does not formally require voter registration as a condition of voting and was exempted from certain provisions of HAVA. For more background information, see www.nd.gov/sos/forms/pdf/votereg.pdf. On the other hand, North Dakota maintains a "central voter file," which contains most of the information that the VRD systems of other states contain, including the voter's complete legal name, complete residential address, complete mailing address, a unique identifier for the individual generated and assigned by the state, and the voting history for the last 4 years. North Dakota's central voter file is used for purposes of "preventing and determining voter fraud, making changes and updates, and generating information, including pollbooks, reports, inquiries, forms, and voter lists." (Chapter 16.1-02, North Dakota Code, available at http://www.legis.nd.gov/cencode/t161c02.pdf.) Thus, many of the issues described in this report regarding VRDs are also likely to be found in North Dakota.

[2] Section 303(a)(1)(A) of HAVA.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Many of the challenges discussed in this report that are faced by election officials in developing effective voter registration databases are ultimately rooted in the fact that election administration is largely a state matter in which the procedures and regulations governing the electoral process for voters are virtually guaranteed to be different from state to state. Whether it is desirable for greater uniformity among states regarding policies and procedures governing election administration is a controversial policy question. Advocates of greater uniformity sometimes argue that it would be more consistent with equal protection requirements. Greater uniformity might also lead to less confusion among voters who move from state to state. On the other hand, greater uniformity—if imposed externally by the federal government—may be seen as negatively impacting state prerogatives.

A more detailed discussion of the background and context for voter registration can be found in Appendix A.

Copyright National Academy of Sciences. All rights reserved.

2

Key Processes for Voter Registration Databases

It is helpful to consider the two basic information management functions of any voter registration database (VRD): adding individuals to the list and maintaining the list.[1] The VRD is also the source of data for pollbooks (the list of eligible voters in localities for use at polling places). Many states have implemented additional functionality to their (centralized) voter registration systems that assists local election officials in conducting an election. Such functionality may include ballot preparation, signature verification for absentee or mail-in ballots (Box 2.1), management of election workers, polling places, petitions, and requirements for disability access under HAVA.

2.1 POSTING NEW VOTER REGISTRATION INFORMATION TO A VOTER REGISTRATION DATABASE

In processing a voter registration application form, the first question to be answered by the election official is whether the applicant is already on the list. Although states handle this process in different ways, one notional way of handling it is that if the person is already in the VRD, the status of the previous registration is changed to "out-of-date" and a pointer added to the new registration. The new registration information must then be added to the VRD, just as it must be if the new registrant is not on the list, except that the registrant is not subjected to the HAVA-mandated verification procedures described below. Alternatively, the database's functionality may allow an update of the voter's registration to reflect the new information regarding address or name.

In those instances in which data are entered in a distributed manner throughout the state, checking to see if the applicant is already in the VRD may occur after the applicant has been added as a new voter. In this case, the new record must be handled as a duplicate of an existing record, each referring to the same person but possibly with different recorded information. (The information might be identical if one person submitted two identical registration forms, as might be the case if he or she had forgotten about one of them or if he or she were unsure of the actual registration status.)

[1] These two general processes—verifying voter registration information and maintaining voter registration lists—are central to the technical and policy dimensions of voter registration databases. Other processes, not covered in this report, are relevant to other requirements and verification procedures covered under Section 303(b) of HAVA.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Box 2.1 Voter Signatures and VRDs

A required component of all voter registration forms known to the committee is an original signature of the registrant. That is, a properly completed voter registration form must include the voter's signature on the physical form itself.

The signature requirement serves two purposes. First, the signature is the voter's certification (under penalties of perjury) that the information provided on the form is true to the best of the voter's knowledge and belief. The signature is thus intended to increase the likelihood that valid information is captured on the form. Second, the signature provides a method for authenticating the identity of the voter at the polling place (usually after the fact). In principle (though often not in practice), a voter's signature when he or she appears at the polling place can be compared to the signature on file if doubts arise about whether the voter is in fact the person who filled out the voter registration form. More commonly, signatures are used in processing absentee and/or mail ballots and for petition verification.

Voter registration databases often integrate an image of voter signatures into their records of registered voters, but to the best of the committee's knowledge, they continue to store original signatures that are captured on paper. Handwriting experts—who may be asked to judge whether two signatures are sufficiently similar—have learned from experience that a signature captured on paper provides more forensically useful information than the same signature captured only in image form. For example, the indentations on the paper registration form (indicating hand pressure with which a physical signature is made) can be compared to the paper signature captured at the polling place—such a comparison is impossible with current technology if the voter registration signature is available only in image form.

The signature requirement has one obvious drawback for voter registration—it makes impossible a voter registration process that operates entirely online. In those instances where voters may register entirely online, some other institution (generally the state's department of motor vehicles) has on file an original signature captured on a paper form. (In this case, the signature on file does not provide the voter's certification about the truth of the information provided—the electronic submission of such information provides the certification.)

If the registrant is not already in the state's VRD, the individual is considered to be a first-time applicant or someone whose previous voter registration was cancelled. HAVA requires certain procedures for verifying voter registration applications. With some exceptions specified in HAVA Section 303 (b), applicants are required to provide a current and valid driver's license number (or a state-issued nondriver's identification) or, lacking one, the last four digits of their Social Security number (SSN).[2] Those who register by mail are also required to present identifying information at the polls on Election Day (or with their mail-in ballots if they vote via mail) if their department of motor vehicles (DMV) or Social Security Administration (SSA) information cannot be verified. HAVA requires the state motor vehicle agencies and the SSA to enter into agreements with states to verify voter registration information. Currently, the state departments of motor vehicles and the Social Security Administration are using the first name, last name, month and year of birth, and last four digits of the SSN (SSN4) for the verification process.

Under these agreements, the applicant's information can be verified against the information on file with the DMV or the SSA. If an applicant has a drivers' license, he or she is required to provide the corresponding number, which is checked against the relevant state DMV database. State law determines

[2] If the applicant has neither a driver's license nor an SSN, the jurisdiction is required to provide the applicant with a unique voter identifying number.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

the necessary degree of agreement between the provided number and the DMV-recorded number,[3] and if the necessary degree of agreement is not present, the application cannot be accepted by the state for purposes of a federal election. In practice, insufficient agreement between the two numbers prompts many jurisdictions to contact the applicant and ask for a more accurate DMV number, but most do not perform any other searching for the applicant.

If the applicant does not have a driver's license, the applicant's information (name, date of birth, and SSN4) is checked against the SSA database. (In practice, many DMVs handle the entire verification request. The election officials submit the verification query to the DMV, which may involve a driver's license number or an SSN4. If the query involves a driver's license number, the DMV responds directly to the election officials. If the query involves SSN4, the DMV passes the request to the SSA using the AAMVAnet, a private network established by the American Association of Motor Vehicle Administrators, and the response from SSA is passed back to the DMV through the same network. The DMV then communicates the response to election officials.)

In the context of a nonmatch in verifying information of a new applicant (i.e., the applicant cannot be found in the DMV or SSA databases), election officials may submit name variants in follow-up inquiries or contact the applicant so that he or she can provide additional information. In addition, HAVA Section 303(b) requires that an applicant who cannot be matched at all be allowed to vote on Election Day provided he or she can present appropriate identification at the polling place.

2.2 LIST MAINTENANCE

A second important function of a VRD system is to maintain the list of eligible voters, that is, to keep voter registration information current and to remove the names of ineligible voters and duplicate registrations from the voter lists. Jurisdictions must perform periodic list maintenance in accordance with provisions of the NVRA.[4] Section 8 of the NVRA requires states to conduct a "general program that makes a reasonable effort to remove the names of ineligible voters" at voter request or as a result of a felony conviction (presuming that state law directs removal of felons from voter registration lists), mental incompetence (again presuming that state law directs such removal), death, or change of residence outside the jurisdiction that holds the voter's registration.

In addition, individual states use a variety of other databases and lists to indicate possible cancellations of voter registrations or changes of address. For example, election officials in Louisiana used a list of names, provided by the Federal Emergency Management Agency, of individuals claiming general assistance in the aftermath of Hurricane Katrina; this list indicated the state in which these individuals had filed. Election officials in Alaska sometimes use the Alaska Permanent Fund Division distribution list (which includes all applicants who want to receive a payment from the state's oil savings account earnings) to obtain current address information. Election officials in Massachusetts use the state's annual census list to verify addresses. State tax records are also sometimes used to indicate who has entered and left the state.

In some cases, the state is responsible for performing list maintenance. In others, counties are responsible for list maintenance using data supplied by the state (from various state agencies).

The NVRA requires that any program of systematic removal of names of ineligible voters must be completed not less than 90 days prior to a federal election. This time limit does not apply to registration cancellations due to death, felony conviction, or judgment of mental incompetence, which may occur within 90 days of an election. In the case of address changes or a failure to vote in federal elections (but not in the case of death, felony conviction, or judgment of mental incompetence), the NVRA requires election officials to notify voters that their registration may be cancelled prior to such cancellation. The

[3] For example, state law may allow a driver's license number provided by the voter that has all of the correct digits but possibly with some of the digits transposed.

[4] See 42 U.S.C. 1973gg et seq., including Subsections (a)(4), (c)(2), (d), and (e) of Section 8 of that act (42 U.S.C. 1973gg–6).

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

NVRA also requires states to maintain and provide access to all records concerning list maintenance for at least 2 years after the maintenance was performed—thus, names of individuals removed from an official VRD list must be made available to the public.

Duplicate Registrations

Duplicate registrations in a VRD often cause confusion, suspicion, and inefficiencies. For example, since voter turnout percentages are calculated on the basis of the actual number of voters on Election Day divided by the number of registered voters, a VRD with a large number of duplicate registrations can lead to underestimates of voter turnout. The same phenomenon has legal and operational significance in states where referendum propositions require a certain percentage of registered voters to approve the placement of any given proposition on the ballot. Election officials estimate the need for polling locations, allocate voting equipment, order ballots, and request budgets for mailings to voters based in large part on voter registration lists that are presumed to be accurate. And duplicate registrations raise suspicions or fears of voter fraud because they may open the door to persons voting more than once.

It is important to distinguish between two types of "duplicate" registrations. One type of duplicate (for purposes of this report, call this kind of duplicate "Type A") is a record in a database that is identical in all particulars to another record—this may occur, for example, if an individual has submitted more than one registration application, as he or she may do entirely by accident if a previous registration is forgotten. Removing Type A duplicates from voter registration lists is conceptually straightforward and technically easy.

A second type of duplicate (for purposes of this report, call this kind of duplicate "Type B") is present when two records in the VRD with nonidentical information correspond to the same individual. Type B duplicate registrations arise in many ways. Perhaps the most common source is a voter's change of address (for example, as the result of a move); a second common source is a change of name (for example, as the result of marriage).

The NVRA establishes procedures that must be followed before a Type B duplicate registration is removed due to a change of address (though not for other reasons), and HAVA establishes a requirement that states provide a unique identifier for every registered voter that is intended to facilitate handling of Type B duplicates. The EAC's *Voluntary Guidance on Implementation of Statewide Voter Registration Lists* further states that "if a state has identified a name on the voter list that it believes is either a duplicate name (or an ineligible voter), election officials should contact the individual."[5] Nevertheless, states establish the technical criteria for deciding when a Type B duplicate exists and process cancellations according to their own state-specific rules and guidelines.

Note also that the use of a commonly used unique identifier (e.g., full SSN or DLN) significantly reduces the technical complexity of managing Type B duplicates. Nevertheless, some matching issues arise even if unique identifiers are present (for example, what to do in the event that the unique identifier is recorded incorrectly).

Type B duplicates may exist within the confines of one VRD or may exist in two different VRDs.

Finding Duplicates Within a Single Voter Registration Database

As noted above, the first task in processing a voter registration form is determining whether the applicant is already on the list. The person may already be on the list but with a different address, or the person may have changed his or her name due to a marriage, divorce, or legal action. Or the person may have registered previously and simply forgotten that he or she had done so.

Checking for duplicates during data entry is one way to handle this problem. When the data from the voter registration form are entered, the system searches for possible duplicate registrations, perhaps

[5] See http://www.eac.gov/election/docs/statewide_registration_guidelines_072605.pdf/attachment_download/file.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

assigning a percentage score reflecting the likelihood that another record is a duplicate of the one being entered. (Ranking can be performed based on a number of fields, such as driver's license number (or SSN4 if no driver's license number (DLN) is available) and date of birth in addition to name fields.) When data entry is complete, the person entering the data can check the highest ranking records manually, for example by comparing signatures corresponding to those records to the signature on the form. (Not all state VRDs support making signatures available across jurisdictions, though many are moving in this direction.)

When a DLN is used, duplicate checking is quite straightforward, except for the possibility of an incorrectly entered DLN. In this instance, one approach is to check the DLN against the database, while using the date of birth and full name for secondary validation. Checking for duplicates can also occur during the data checking stage. So, if John Doe provides an incorrect DLN (causing the system to process the entry as a new record when the entry should be reflected as a change or transfer), the person performing data checking could use a date-of-birth search in the VRD to see if a duplicate is on file. In a more sophisticated implementation, the system would detect these additional duplicates and present them to the user for review without requiring the user to perform the date-of-birth search in each of these cases. The same scenarios apply for states that use the full SSN.

Although jurisdictions differ in their implementations, most jurisdictions undertake some kind of checking for duplicates within their VRDs.

Finding Duplicates Across Two Separate Voter Registration Databases

Finding duplicates across separate VRDs is of particular interest to election officials in states that see substantial residential mobility of their residents to and from one or more other states, particularly neighboring states.[6] Finding and, when appropriate, removing out-of-date voter registration records should result in more accurate voter registration databases that perform such list maintenance.[7] Such comparisons can serve two purposes:

- *Administrative cleanup*. Jurisdictions benefit from clean lists in election management, especially in keeping postage costs and ballot printing costs to a minimum. Although it is usually true that an individual breaks no law if he or she is registered to vote in two jurisdictions, states (or their constituent counties) may wish to delete the registrations of individuals who no longer reside in their jurisdictions. For such purposes, complete VRDs for each state must be used. Comparisons of complete VRDs are likely to result in a significantly larger set of possible matches than in the case of comparisons for fraud detection.

As an example of such an application, note that a number of states have performed data-matching exercises for their VRDs across state lines. For example, the Iowa Secretary of State performed a cross-check in 2006 of the Iowa VRD against the VRDs of Kansas, Nebraska, and Missouri using a criterion of exactly matching first name, middle name, last name, and date of birth, resulting in the identification of possible duplicate registrations numbering 1,464 for the Iowa-Kansas cross-check, 1,792 for the Iowa-Nebraska cross-check, and 3,350 for the Iowa-Missouri cross-check.[8]

---

[6] For example, election officials in some states in the Midwest (Kansas, Iowa, Missouri, Nebraska, Minnesota, and South Dakota) and in the south central United States (Kansas, Colorado, Arizona, Arkansas, Oklahoma, Louisiana, and Kentucky, with Mississippi and Tennessee interested in joining) have joined forces to share voter registration data in an attempt to identify individuals who may have voter registrations in multiple states. Personal communication to the committee from Brad Bryant, email of September 11, 2009.

[7] There are a number of regional consortiums that are intended to facilitate the comparisons of their statewide voter registration lists. See, for example, *Midwest voter registration data-sharing project moves forward: Advocates voice concern*. Available at http://www.mapj.org/?q=node/118.

[8] Iowa Secretary of State, undated document. Passed to the committee by Brad Bryant, Election Director for the State of Kansas.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Box 2.2 A Pilot Project to Compare Voter Registration Databases in Oregon and Washington

A pilot interstate voter registration database matching project between Oregon and Washington state explored the feasibility of using database matching to identify voters registered simultaneously in both states and sought to develop procedures for resolving those potential duplicates.[1] The database of just over 3.4 million Washington active voters was matched to the database of the just over 2 million Oregon active voters. Drawing from selected counties in each state, some of the voters found in both state databases were contacted in an attempt to determine whether the matches in fact reflected simultaneous registration (that is, whether the matches reflected duplicate records). Through such a process, participants in the pilot project hoped that each state would in the end have a more accurate voter registration database.

A simple data-export function was used to extract the information in textual form from the relevant databases, thus creating two files.[2] The structures of these files were highly consistent; the biggest data inconsistency was a minor difference in date formatting that was easily made consistent for purposes of comparison. Two matching algorithms were used. The first was based on a character-by-character match on the complete first name, complete middle name, complete last name, and date of birth. A second algorithm was based on a character-by-character match on the complete first name, the first character from the middle-name field, the complete last name, and the date-of-birth. All of the required information processing was completed in a few hours of computer time on a personal computer.

When the first algorithm (with complete middle name) was used, 3,482 pairs (possible duplicate registrations) were found. When the second algorithm (with middle initial only) was used, 8,292 pairs were found—this larger set of matches of course included the 3,482 matches found by using the first algorithm, as well as an additional 4,810 pairs.

Only a subset of the individuals in the second set living in selected counties were contacted directly. The reasons for this limitation were the pilot project nature of the experiment and the large-scale confusion and public misunderstanding that might have resulted from using the entire data set. A total of 1,312 pairs were identified. Of these, 686 pairs had a most-recent registration date in Washington, and a contact letter was sent to the indicated Oregon address. Similarly, 626 pairs indicated a most-recent registration date in Oregon, and a contact letter was sent to the indicated Washington address. The contact letter sent to Washington addresses asked the recipient if he or she wished to cancel his or her registration in Oregon (i.e., the state with the less recent registration) and vice versa. When, and only when, a contacted individual returned the form indicating that he or she wished to be removed from the registration list in the particular state, that information was recorded in the respective state election office and then forwarded to the appropriate county election official for resolution. No action was taken when the form was not returned.

Of the 686 letters mailed to Oregon addresses, 650 (95 percent) appear to have been delivered. Responses were received from 391 individuals (a response rate of 60% of delivered mailings). Of the 391 responses, 379 responses were forwarded to the appropriate county election official and resulted in cancelled voter registration records. The remaining 12 did not provide enough information for the removal process to proceed. The data for letters mailed to Washington addresses are quite similar.

---

[1] R. Michael Alvarez, Jeff Jonas, William E. Winkler, and Rebecca N. Wright, "Interstate Voter Registration Database Matching: The Oregon-Washington 2008 Pilot Project," available at www.usenix.org/events/evtwote09/tech/full_papers/alvarez.pdf.

[2] The method of data extraction is significant, because Oregon and Washington do not have statewide voter registration systems developed by the same vendor, suggesting that it is not necessary or important that states share similar voter registration systems in order to carry out efficient interstate data matching.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Another cross-check was performed in 2007 by Kansas against the VRDs of Iowa, Missouri, Nebraska, Minnesota, and South Dakota using a criterion of exactly matching first name, middle name, last name, and date of birth; this cross-check resulted in the identification of 11,205 possible matches.[9] These matches were sent to election officials in the relevant Kansas counties, who investigated further and canceled a voter's registration only if "the county election official [was] certain the records represent the same person and the Kansas record is the older record, meaning the record in the other state [had] a newer registration date" (quoting the words of a September 2007 state election newsletter). No information is available on how many Kansas registrations were cancelled. A similar 2007 exercise conducted by Iowa resulted in 5,753 possible matches in neighboring states, and by South Dakota in 2007 resulted in 2,800 possible matches. Note, however, that as discussed in Appendix B, the likelihood of individuals with common names (e.g., "John Smith") sharing birthdates is not negligible.

- *Fraud detection*. Because it is illegal for an individual to vote twice in a single election, two states might wish to compare their VRDs to identify individuals who have voted in both states. Such identification would be needed for criminal prosecution. However, only voters within a given VRD who have actually voted in a given election need to be compared for the purposes of detecting fraud. If the amount of fraud (voting in multiple jurisdictions) is small, then the number of matches will be very small—that is, very few individuals will actually be found to have voted in the same election. It is an empirical question (with one exception,[10] not known to have been examined in any given pair of states to the best of the committee's knowledge), but if it is true that the actual amount of voter fraud of this nature is small, then the resources required to further investigate through human examination the matches found under these conditions should be readily available at the state level. Clean lists can also enhance public confidence regarding the integrity of the overall electoral system, even if the incidence of provable fraud is very low.

Comparing VRDs to each other is a relatively straightforward computational task, because the data definitions of interest are likely to be highly consistent between databases. For example, formats of the date field vary (e.g., mm-dd-yyyy, dd-mm-yyyy, yyyy-mm-dd, yyyy-dd-mm), but these are easy to standardize for comparison purposes. Different spellings of a name (to include the presence or absence of name prefixes or suffixes and punctuation) present some difficulty but can usually be handled through the use of string comparators that account for typographical differences. Running the comparison may take a long time if the VRDs are large.[11]

Box 2.2 describes the results of a pilot experiment to compare two state VRDs. Methods to improve matching of a VRD against that of another jurisdiction are described in Appendix B, and include the use of name rooting (the generation of "equivalent names" from a given name, such as Bob and Rob from Robert) and third-party data. Address standardization is also helpful if address information is to be used for resolving proposed matches. And all such methods are further enhanced by the use of methods to account for typographical error.

It is important to note that significant progress can be made in identifying duplicate registrations that cross jurisdictional lines even if not all jurisdictions are interconnected. Although 100 percent coverage would indeed require universal interconnection, it is highly likely that the majority of duplicate registrations are contained in specific groups of jurisdictions, such as adjoining jurisdictions (e.g., Oregon and Washington), jurisdictions that serve as "bedroom" communities for another (e.g., Maryland, Virginia,

[9] See Sean Greene, "Midwest voter registration data-sharing project moves forward," Electionline Weekly, December 13, 2007, available at http://www.pewcenteronthestates.org/report_detail.aspx?id=33612.

[10] The exception is the study of Oregon and Washington described in Box 2.2.

[11] Running on a Dell workstation, the comparison of two state voter registration databases (each containing records for 2-3 million voters) took approximately 2½ hours to complete the generation of possible pairs of records. (See Alvarez et al., *Interstate Voter Registration Database Matching: The Oregon-Washington 2008 Pilot Project*.) Comparing databases ten times as large would increase the run time by approximately a factor of 100, or about 250 hours.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

and the District of Columbia), and jurisdictions that experience seasonal migration (e.g., New York and Florida).

Felony Convictions and Mental Incompetence

HAVA calls for coordination of state VRDs with state felony databases in accordance with state law. The Election Assistance Commission (EAC) recommends that states also coordinate with relevant federal databases and criminal conviction records from U.S. attorneys and federal courts. The use of multiple databases is helpful to overcome gaps in or omissions of data from external state files. However, HAVA does not specify how the coordination with other state agencies' databases is to take place and lacks specific guidance on standards or methods for removal of ineligible voters from the databases for the above reasons.

Note also that state law governs state policy regarding the relationship between voting eligibility and status as a felon. In some states, convicted felons are never permitted to vote after their conviction; in other states, the right to vote is reinstated automatically upon the end of the individual's sentence; in other states, the individual must apply for reinstatement after the end of his or her sentence or at a state-specified time afterward; in still other states, individuals may vote if they are granted probation or parole; and indeed, in a few states, some felons are allowed to vote even while imprisoned.

Matching a full VRD against an electronic list of felons or individuals deemed mentally incompetent is a computational task that is easier than that of matching the VRD of one state against that of another, simply because the number of such individuals is likely to be two or three orders of magnitude smaller than that of a full VRD. On the other hand, the data provided may not be complete enough to perform effective matching, or it may be provided in a form that does not facilitate effective computer-based matching (e.g., a "database" may in fact constitute a large number of fax or PDF images of the relevant documents; if so, the relevant data must be extracted from the images and rendered in a useful format).

Death

HAVA calls for coordination of state VRDs with state death databases in accordance with state law, and the EAC recommends that states also coordinate with Social Security death index databases.[12] In the case of removing deceased voters from a VRD, jurisdictions use a variety of sources to identify such individuals. Some jurisdictions rely on local obituaries and communications from local departments of vital statistics. But voters from one state may also die in another state, and thus states must obtain information regarding those deaths.

Today, there are at least two possible authoritative sources of information for out-of-state deaths. First, most jurisdictions (e.g., states) exchange data on deaths under an inter-jurisdictional exchange agreement.[13] Other formal agreements govern the exchange of data between jurisdictions and certain other internal and external organizations, which in principle can include election offices in various states. Such data are currently exchanged in a manual, paper-based manner (e.g., paper copies of death certificates).

STEVE (State and Territorial Exchange of Vital Events) is a secure messaging system currently under development by the National Association for Public Health Statistics and Information Systems (NAPHSIS) that will allow subscribers to electronically exchange the vital-event data they currently share. These data will be configured in a standardized format. STEVE is being deployed, but does not currently include all U.S. jurisdictions. When STEVE is fully deployed, state election offices should be able to receive comprehensive death information from all subscribers (which are expected to include all

[12] U.S. Election Assistance Commission, *Voluntary Guidance on Implementation of Statewide Voter Registration Lists,* July 2005. Available at http://www.eac.gov/election/docs/statewide_registration_guidelines_072605.pdf/attachment_download/file.

[13] See http://www.naphsis.org/index.asp?bid=1045.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

U.S. jurisdictions). Jurisdictions will post their data to STEVE in accordance with NAPHSIS contractual requirements for timeliness, which suggests that data recipients will receive data from jurisdictions in an uncoordinated manner—that is, from multiple sources.

A second source of authoritative data is the Social Security Administration's Death Master File (DMF). This file contains a list of individuals with Social Security numbers and whose deaths were reported to the Social Security Administration from 1962 to the present and on individuals who died before 1962, but whose Social Security accounts were still active in 1962. The file is updated on a monthly and weekly basis, and these updates are made available for a fee by the National Technical Information Service of the U.S. Department of Commerce.[14] A record in the file consists of

- Given name and surname
- Middle name
- Full date of birth
- Full date of death
- Social Security number
- State and Zip code of last residence
- Zip code of the address designated by the individual to be the address of record for SSA purposes (such as sending benefits)

A death certificate usually includes the full address of record, as well as a full name, full SSN, and a full date of birth. So, the DMF contains essentially all of the information that death certificates contain (less full address), and because a death certificate is generally regarded as authoritative evidence that a given individual has died, it is plausible to use the DMF in a similar fashion. (Whether the DMF should be the *only* source used depends in part on latency issues. The DMF may well lag the issuance of death certificates by a few months; if more current data on deaths are available on deaths, election officials must make a tradeoff involving currency of data versus inconvenience of access.) (It is for this reason, among others, that departments of vital statistics have sought to develop STEVE.)

Matching a full VRD against the full SSA Death Master File (necessary because anyone in the VRD may have died in the past year) is a computational task that is comparable to that of matching the VRD of one state against that of another, although this task only needs to be done once—subsequent comparisons should use the smaller incremental files from the Death Master File against the full VRD. The data are approximately comparable in format and structure, with the possibility that the Zip codes on file for the DMF do not necessarily correspond to the decedent's Zip codes of record for voting purposes.

Changes of Residence

The NVRA requires states to establish a program to use information supplied by the U.S. Postal Service (USPS) to identify registrants whose address may have changed; today, about 14 percent of the population changes an address every year.[15] (In addition, jurisdictions with colleges or universities face challenges resulting from the transient nature of much of the eligible voting population.) Identifying voters who have moved is often based on periodic mailings that election officials send to all voters in the jurisdiction by U.S. mail, indicating on the envelope "do not forward" but rather return to sender. Notices that are returned to the election official are an indication that the voter may have moved.

The USPS does not automatically notify election officials of an individual's change of address. Election officials must initiate address checks with USPS on their own. They have a choice of comparing selected records in the VRD or the entire VRD to the USPS National Change of Address (NCOA)

[14] NTIS Products: Social Security Administration's Death Master File, National Technical Information Service. See http://www.ntis.gov/products/ssa-dmf.aspx.

[15] See U.S. Census Bureau, *Geographical Mobility*: 2006. Highlights from this series are available at http://www.census.gov/Press-Release/www/releases/archives/mobility_of_the_population/010755.html, and detailed tables are available at http://www.census.gov/population/www/socdemo/migrate.html.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

database.[16] The change-of-address database can be accessed through a third-party provider or implemented local to the election office. (If the forwarding request has expired, a query will indicate that the new address is unavailable.) The utility of such a check varies, as some election offices have noted that NCOA information is not sufficiently timely for their purposes.

The NVRA requires election officials to notify the voter if they receive an indication that the voter has moved. In particular, when a change of address is received from the USPS process, the election official must send by forwardable mail a postage pre-paid and pre-addressed return card on which the voter may state his or her current address. If the voter remains within the jurisdiction of the election official and the voter responds to the notice, then the address can be updated based on the new information provided. If the new address is outside the jurisdiction of the election official, the voter is asked to return the card, and the voter registration record is handled accordingly. If the confirmation card is not returned and the voter does not vote in or by the second general federal election that occurs after the date of the notice, he or she may be removed from the VRD.

In addition, a state's department of motor vehicles and other mandated NVRA agencies (e.g., social service agencies) can be an important source of information regarding changes of address. These other agencies may well have more current address information than do the state DMVs, some of which have lengthened their drivers' license renewal times to save on costs. One state—Michigan—provides for the automatic updating of voter registration information based on changes received at its DMV, and has found that many changes of address are thus much more easily managed. (The reverse is also true—a change to a Michigan voter's registration address also automatically changes the voter's residence address at the DMV.) In general, federal law (NVRA) requires these agencies to provide change-of-address information to election officials unless the voter has specifically indicated that the change of address is not for voting purposes (as might be true for a voter in New York who relocates to Florida only for the winter); however, implementation of this requirement is uneven across the United States.

Voters who submit change-of-address notifications to election officials (e.g., by moving into a new community and checking the appropriate box on a voter registration form indicating a new address for the same voter) pose little or no computational problem for election officials. However, an automated check of a VRD against a USPS change-of-address database is not a straightforward task, because these databases do not include important identifiers such as date of birth or any part of the SSN. Note that when performing computer-based matching using name and address fields, address standardization is virtually mandatory. Address standardization can be accomplished using commercially available software or services that automatically process raw voter-provided addresses. Even so, a comparison based on standardized address fields is likely to be less accurate than one based on identification fields. (For example, a given change-of-address notification from the post office may or may not be associated with all voters in a household.)

An additional complication is that a change of mailing address (the primary purpose of the USPS NCOA database) does not necessarily reflect a change in residence or domicile (which determines eligibility to vote in a given electoral jurisdiction). Military personnel, who constitute the majority of UOCAVA voters,[17] in particular often prefer to keep their voter registrations at a main base or state of origin, and upon assignment to another location will file a change of address with postal authorities, even if it is not in fact a change of residence or domicile for voting purposes. Thus, reliance on USPS change-of-address databases as a proxy for changes of domicile for voting purposes may contribute to the disenfranchisement of mobile individuals who do not change domicile.

[16] For more information on the NCOA database and address change services provided by the U.S. Postal Service, see http://www.usps.com/ncsc/addressservices/moveupdate/changeaddress.htm. Commercial software costing in the range of $50,000 is available that checks addresses and formats them so that they can be checked against the NCOA (this cost does not include the cost of comparison to the NCOA database). A less expensive option available to states is to contract with a vendor licensed by the USPS, which can cost several thousand dollars per year to check the entire state database.

[17] UOCAVA voters are uniformed and overseas citizens who cast absentee ballots (UOCAVA is the acronym for the Uniformed and Overseas Citizens Absentee Voting Act).

Copyright National Academy of Sciences. All rights reserved.

3

Technical Considerations for Voter Registration Databases

3.1 DATA CAPTURE AND QUALITY

The data contained in a VRD can be characterized with respect to two different attributes—accuracy and completeness. For purposes of this report, accuracy refers to the factual correctness of the data that exist in the database, whereas completeness refers to the presence in the database of all individuals who *should* be in the database. If the database is perfect, it is both 100 percent accurate and 100 percent complete—that is, all of the data in the database are correct (and thus the database contains no individual who should not be in the database), *and* the database includes all of the individuals who should be in the database. Notice that in this formulation, accuracy does *not* subsume completeness, so that a database must be characterized with respect to *both* attributes.

This usage of the term "accurate" appears to be consistent with the meaning of the word in common discourse. However, the reader is cautioned that some other commentators and analysts use the term "accurate" to mean both "factually correct" and "complete."

As is the case with all other databases, the utility of a VRD depends strongly on the quality of the data it contains (the accuracy and completeness of the data), although a variety of processes can be applied to the data in order to improve their quality.

One common source of error in the data is data entry. Applicants typically submit handwritten voter registration forms that are sent to the election official. The applicant can make a mistake, forget to answer a question, or not write legibly. The form or its information could be altered in transmission (a field could get smudged or torn or otherwise damaged in postal handling, for example). Keying errors result in mistranscriptions.

Another source of error is the quality of other lists that are compared with VRDs. The quality of other lists similarly depends on the procedures for data collection and entry; methods employed to minimize errors in the data, such as removing duplicates and other anomalies from these secondary databases; and staff training and audits, among other aspects.

Moreover, the different purposes for which secondary data are collected can limit their use for other purposes and may not fully address what is needed for the purposes of voter registration databases. For instance, the USPS compiles change-of-address data when customers request mail forwarding through the USPS NCOA system. However, the USPS has defined its information services so as to serve its primary business function, that is, without particular concern for the needs of election officials—and

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

in particular, it does not collect date-of-birth information because such information is not related to the primary business purpose of the USPS. Thus, the NCOA system cannot be queried with name and date of birth to learn an individual's new address. Furthermore, because of privacy considerations, the USPS limits the disclosure of change-of-address information, and thus a name and old address must be presented before a new address can be provided. This limitation is significant because it means that an election official cannot simply query the NCOA database for the new address of an individual known to have moved.

A more detailed discussion of data capture and quality can be found in Appendix C.

3.2 DATABASE INTEROPERABILITY

Database interoperability arises as a requirement because election officials must perform a variety of tasks that involve other databases, ranging from other state VRDs to lists of deceased persons as described above. From a technical standpoint, database interoperability refers to the capability of two databases to exchange data (perhaps with a third-party application) and to use the exchanged data.[1] Data exchange involves transmitting and receiving data between two systems, by whatever means, in a way that maintains the usability (preserves the structure and formatting) of the data. Data use depends on the corresponding data fields having the same meaning in each database.

Transmitting and receiving data involve moving the electronic bits that represent the data in question through some channel. In practice, this involves either a communications network connecting the two database systems or use of a physical medium such as a CD-ROM to carry the data. Using a direct linkage (e.g., an Internet connection) provides for real-time communications—the data that are transferred to the receiving system can be kept current with changes. Use of a physical medium generally "batches" the data to be transferred, and thus changes to the sending system's database will arrive to the recipient with some delay and may not reflect the most recent changes.

As for the data that are passed through either approach, they must be formatted in a manner so that one system can write and the other can read. A common approach to achieve formatting compatibility is to use the sending system's ability to "export" its data into a known file format (e.g., a comma-delimited file) and for the resulting file to be transmitted or carried to the receiving system.

Data usability is guaranteed if all databases use the same data definitions.[2] However, in the situations faced by election officials, data definitions of the comparison databases (the databases containing the data with which VRD data must be compared) may well be different. Ensuring the similarity of data definitions goes beyond classic definitions such as "integer" or "character string"—it also includes issues such as formatting and data semantics.

For example, System A may define dates in a mm-dd-yyyy format, and System B in a dd-mm-yyyy format. The semantics of the two systems may differ: System A may use standardized addresses and strip all punctuation from name fields, whereas System B may not use standardized addresses and may retain punctuation in name fields. Or, System A may include name suffixes in the last name field, and System B may provide a separate field for name suffixes.

Such definitional differences may increase the difficulties of comparing fields unless the definitions of these fields can be reconciled. A variety of technical approaches have been developed for dealing with differing standards or incompatible definitions; see Box 3.1. In any event, data definitions must either match or be transformed in a way that preserves the semantics of the data.

---

[1] In colloquial usage, database interoperability sometimes has a broader meaning that entails data access, of which data exchange is a subset. Database interoperability without data exchange, for example, can refer to the ability of election officials in State A to view records and perform searches in the VRD of State B. Although such a capability can be helpful in individual instances, the inability to perform data exchange prevents any large-scale operation involving either database.

[2] On the other hand, a release of a given database system may have data definitions that are somewhat different from those of an earlier release. System developers know that such changes create operational chaos, and thus avoid such changes whenever possible.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Box 3.1 Approaches for Achieving Data Compatibility

There are a number of approaches for reconciling data definitions:

- *The data translator approach* requires two systems that need to interoperate to have a translator that converts one set of data definitions into the other. Data translators are probably the simplest and most straightforward approach to achieving data interoperability, although the data translator approach does not scale upward if interoperability among all databases is required.
- *The common format approach* calls for each system to use its own data definitions internally. However, exchanges of data with other systems are conducted by using a common data standard into which data must be translated before being transmitted to another system. Any system using these data then downloads them in the common format and retranslates the data into locally meaningful terms before the data are used.
- *The data server approach (an extension of the common format approach)* is based on the separation of data from the applications that use the data. When a system requires data, it connects to a data server that provides the data. Thus, enforcement of definitions can be limited to just a few servers rather than a myriad of applications. By moving the data into a system separate from the individual applications, this approach facilitates reuse of data in new, unanticipated ways.

Achieving interoperability between different systems is potentially complicated by the fact that these systems are built or acquired by a variety of agencies (election officials, departments of motor vehicles, departments of vital statistics, departments of correction) that are not generally subject to the same overall chain of command and thus may not implement compatible data definitions. These agencies are concerned primarily with developing systems optimized to serve their own mission needs. Thus, they generally have little interest (or funding or incentive) to focus very much attention on the needs of the other agencies with which they may some day need interoperability, and are likely to pay minimum attention even to mandated tasks that are outside their primary mission needs.

As an illustration, consider that voter registration databases must provide for recording the physical residential address of record for the voter as well as a mailing address; the former is essential for the determination of voting eligibility, precinct boundaries, and ballot style assignment (making sure that a given voter registered for a specific address receives the correct ballot for that address). The database systems of other agencies may only support fields for mailing address.

Lastly, any discussion of database interoperability is incomplete without mentioning its organizational dimensions. Specifically, interoperability between a state VRD and another database operated by federal, state, or county authorities depends on cooperation between the election official and the relevant federal, state, or county authority to exchange and/or process the relevant data. No technical solution can force agencies to cooperate with each other, and if such cooperation is not forthcoming, data exchanges may well be more infrequent or the data could be prepared more poorly than would otherwise be the case.

3.3 MATCHING

Database interoperability can be regarded as the process through which the data from one database can be made available in a useful and understandable format for a meaningful comparison to the data from another database. The matching process is the essence of the comparison in the context of VRDs.

Adding new voters to the VRD and maintaining the VRD both require a procedure by which attri-

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

butes of one data registration record are compared to attributes of another record (for example, a new voter registration application, a DMV driver's license, an SSA record, a record in a database of felons, and so on). This procedure, variously known as record linkage, identity matching, identity resolution, or simply "record matching," is "good" when it results in low rates of false positives (matches indicated when no match in fact exists) and false negatives (nonmatches indicated when a match does in fact exist).

In adding individuals to a VRD, poor procedures could have either or both of two undesirable consequences. First, they might result in improper indications of a nonmatch when a match should be indicated, a result that could be used (1) to disenfranchise voters (in the event that an applicant's information cannot be verified when it should be verifiable), or (2) to inflate the size of the VRD list mistakenly (in the event that an earlier registration for an applicant cannot be found and a new record is improperly added as though the individual were a new registrant). Second, they might result in improper indications of a match when a nonmatch should be indicated, a result that could be used to add ineligible names to the VRD list.

In maintaining the VRD, procedures of poor quality will result in improper indications of a match between the voter registration list and one of the databases of ineligible-to-vote individuals when a nonmatch should be indicated (a result that tends to remove voters from the voter registration list improperly) or improper indications of a nonmatch when a match should be indicated (a result that would keep felons, mentally incompetent individuals, and deceased people in the VRD).

The consequences of false positives and false negatives may vary depending on the purpose of the matching (and thus depending on the other databases against which VRD records are being matched). By law, the information on new voter registration applications must be checked against DMV or SSA records, and the consequences of a false negative (that is, no matches found when an individual is in fact represented in the DMV or SSA database) may be to wrongly keep the individual off the rolls—false negatives in this context may lead to a less *complete* VRD. List maintenance often calls for existing VRD records to be matched against felon or death records. The consequences of a false negative in the context of list maintenance are precisely the opposite: individuals may erroneously be kept on the rolls—false negatives in this context may lead to a less *accurate* VRD.

False positives (that is, a match improperly found when in fact the individual is not represented in the database being checked) have a different impact. In the context of adding individuals and checking against DMV or SSA records, false positives result in a less accurate VRD, because individuals may be improperly added to the list. In the context of list maintenance and checking against felon or death records, false positives result in a less complete VRD, because individuals may be improperly removed from the list.

Box 3.2 summarizes these conclusions regarding false positives and false negatives.

Because of data quality issues and the lack of a universally used unique identifier, record matching cannot be done perfectly in this context, that is, with zero false positives and zero false negatives.[3] The consequence is that achieving the goal of a simultaneously 100 percent accurate and 100 percent complete voter registration list is virtually impossible. At the same time, what counts as an acceptable rate of false positives or false negatives, or an acceptable tradeoff between accuracy and completeness, depends on the particular policy goals that are desired.

For example, given that a choice is necessary, a state could prefer to emphasize completeness over accuracy in its VRD. With this goal in mind, it may choose to minimize the rate of false positives in matching the VRD against a list of felons, a policy choice that almost certainly will increase the number of ineligible individuals on the list. Alternatively, a state could prefer to emphasize accuracy over com-

[3] If a unique identifier for every person were available, and if that identifier were used in all databases that were to be compared, and if those identifiers were always recorded correctly in the databases, perfect matching would be possible. But these conditions are essentially never realized in practice. When a DMV number or a full SSN number are unavailable, the matching process becomes one of probabilistic inference rather than logical deduction.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Box 3.2 Consequences of False Positives and False Negatives

	Adding new voters to the voter registration list	List maintenance
Consequence of false positive	Less accurate VRD (ineligible persons may be added to the rolls)	Less complete VRD (eligible voters may be improperly removed from the rolls)
Consequence of false negative	Less complete VRD (eligible voters may be kept off the rolls)	Less accurate VRD (ineligible persons may be kept on the rolls)

pleteness in its VRD. With this goal in mind, it may choose to minimize the rate of false negatives in matching the VRD against a list of felons, a policy choice that almost certainly will increase the number of legitimately eligible individuals removed from the list.[4] Inevitably, a number of voters in a given state will be disenfranchised given one policy choice that would not have been disenfranchised under the other. Also, if State A makes the first policy choice and State B the second, some similarly situated voters in these states will not be treated identically. (The committee does not make any normative judgment regarding either of these policy choices, and observes that the federal government appears to be more concerned that voters within a single state are treated alike than the possibility that voters in different states may be treated differently.)

From a technical standpoint, the hard problem in matching usually lies not in identifying potential matches (e.g., pairs of records that may have some but not all elements in common) but rather in how to handle the potential matches that are identified. (It is for this reason that the use of common unique identifiers greatly enhances matching outcomes—such use materially and significantly reduces the challenges of possible matches.) Determining whether two records refer to the same individual is usually the problematic step.

Record-matching procedures can, in principle, be executed by computer, by a human being, or both. Computer-based procedures for verification or maintenance have the advantages that they can perform matches very rapidly and can operate consistently (because they depend only on the specific data involved and the prescriptive rules as implemented). But computers using naïve matching rules (e.g., processing Liz and Elizabeth as different names) can also be "fooled" by data problems that suitably trained humans can often handle.

Human-based matching has the advantage of bringing to bear training and personal experience, which can be used to determine with confidence a match or nonmatch in more cases (Box 3.3). In some cases, humans can obtain additional information by contacting the individual(s) who may be involved, and use the information obtained to help resolve a match. A human can also compare signatures associated with each member of a proposed match and make a judgment about whether the signatures are

[4] Arguments might sometimes be put forth to make only a particular subset of the database maximally accurate or maximally complete. (Hypothetically, a particular subset might be "all female voters" or "all voters in precincts X, Y, and Z" which happen to have the highest fraction of registered Democrats or Republicans.) While legitimate policy reasons for doing so in some cases cannot be ruled out, such actions are inherently suspect and deserve the highest scrutiny before being implemented. For example, an election official might be motivated to maximize the number of voters in a particular socioeconomic class or other group in order to give his or her party of preference an advantage at the polls. Although the political motivation for wishing to take such action is clear, such an action would do serious injustice to the democratic process, and such a motivation would never be acknowledged publicly.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Box 3.3 An Illustrative Example of Human Exception Processing

- *Example 1—Users entering new voter registrations must check existing rolls for matches.*

New registration card	Existing voter
Mary Sinclair	Mary Sinclair
4131 Bayberry Street	731 Ascot Drive
4/28/63	4/28/63
SSN XXXXX3434 (4-digit SSN)	DL 00767234633

To address this ambiguity, if the user could confirm that the driver's license number is already known to be associated with an SSN with the same last four digits (3434), then this user could associate these records with high confidence. Another alternative would be to determine if Mary Sinclair on Bayberry Street used to live on Ascot Drive.

- *Example 2—System must match new voter registrations against records in SSA databases.*

New registration card	Closest record in SSA
Tom T Bowden	Taylor T Bowden
3121 Escondido Way	
11/04/77	11/04/77
SSN XXXXX1087 (4-digit SSN)	SSN XXXXX1087 (4-digit SSN)

With the match algorithm currently used by the SSA for matching inquiries from election officials, the SSA would return a "no-match" result.

If the algorithm were changed to include the closest matches to the submitted inquiry, the "Taylor T Bowden" record would be displayed. To address the potential ambiguity, the election official could seek to confirm that either Tom has a middle name of Taylor or Taylor has a middle name of Tom or Thomas; if so, the election official could associate these records with some degree of confidence if he or she concludes that the first and middle names have been transposed.

The human review process involves review and best judgment based on the attributes at hand. Because having more attributes improves match accuracy, having more attributes reduces the number of voters inappropriately categorized as ineligible.

sufficiently similar to indicate a genuine match. On the other hand, human-based matching is slow and thus impractical when large numbers of records are involved. Human-based matching is generally less consistent than computer-based matching but may be better (though still somewhat subjective) in other areas, such as comparing signatures. Human-based matching may also be biased—for example, a human matcher may have prejudices against Hispanics, and may be less likely to resolve in a favorable manner apparent matches in the database involving people with common Hispanic surnames compared to others.

These procedures can be used in tandem, so that any possible match or nonmatch (which depends on context) found by a computer-based procedure is directed to a human being before any action is taken.[5] For example, if the submission of a given name to the DMV and SSA results in a nonmatch, a

[5] These comments should not be taken to imply that the combination of computer plus human review is necessarily better than the computer alone in all circumstances. Indeed, the literature indicates that for human review to add to the quality of the

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

human being may inspect the original voter registration form to compare the handwritten data on the form with the data as transcribed into the database, correcting the database record if necessary and resubmitting it with the correct spelling as indicated on the handwritten form. A human being may also resubmit the query with a different but equivalent name. This different-but-equivalent name may be a common nickname (e.g., Bill, Will, Willie, Willy for William) or a different spelling of a name (Jazmine for Jasmine).[6] Helpful though such manual procedures are, they can break down under the stress of large numbers of applications, as may happen when applications are submitted near the deadline for submission of registrations. In addition, it is probably unrealistic even under normal conditions to expect a human to resubmit a large number of name variations—at most, trying a few alternatives is likely the best that can be expected.

In addition, match algorithms based on exact matches between corresponding data fields cannot account for typographical error. Blocking techniques and string comparators are helpful for dealing with this problem; when used, most query results would logically take the form of a list of records, sorted by a score indicating the likelihood of a match (that is, a fuzzy match) rather than a simple binary result (match or no match).[7]

A more detailed discussion of matching can be found in Appendix B. Some privacy issues that arise with matching are addressed in Appendix D.

3.4 SYSTEM AVAILABILITY

Availability is the property of a system related to a user's ability to use the system when necessary. Many factors influence the accessibility of a system, including how many users are trying to use the system at the same time, what kinds of tasks the system is handling at any given time, and whether or not an adversary is trying to reduce system availability.

Systems that are subject to large variations in the user load they support pose technical challenges. As compared to other times of year, VRDs in particular must typically support intense usage from many users in the period before registration deadlines occur, on Election Day or during other periods of voting (e.g., early voting), just before primaries, and so on. Furthermore, the demands on the system are different during these different periods—data entry tasks are likely to be most plentiful just before registration deadlines expire, whereas user queries to the database are likely to be most plentiful when voting is occurring.

In addition, VRDs also depend on other systems being available. For example, election officials make heavy use of DMV and SSA databases for verifying applicant-provided registration information, as required by HAVA. If these systems are unavailable during peak demand times, election officials may be unable to verify such information in a timely manner and thus may not be able to register a voter in time for a primary or an election.

For example, the Social Security Administration often performs system maintenance and upgrades over the Columbus Day weekend (mid-October). Although such actions are understandable given the SSA's primary mission, they also have major negative effects on election officials trying to process the enormous influx of voter registration applications that arrive before Election Day (in November). Some

---

outcome, human reviewers must be well trained (see, for example, H.B. Newcombe et al., "Reliability of Computerized Versus Manual Death Searches in a Study of the Health of Eldorado Uranium Workers," *Computers in Biology and Medicine* 13(3):157-69, 1983). Nonetheless, it tends to be true that the combination of good computer matching procedures and well-trained human reviewers is often superior in performance to the use of those procedures alone.

[6] Managing known name equivalents can also be performed in an automated fashion, but if automated assistance is not available, humans must undertake this task.

[7] Match algorithms are based on comparisons made at the level of individual fields or at the record level. String comparators compare text strings within individual fields and generate a score that reflects the amount of difference between the two strings. Blocking techniques bring together pairs via characteristics that are believed to contain less typographical error, and the remaining (or all) information in pairs is used in computing a matching score.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

voter registration databases do not have the capability to enter data from voter registration forms without verifying those data (that is, if verification cannot be attempted, data entry must stop). In short, the unavailability of SSA databases over the Columbus Day weekend means that election officials must halt all processing of applications if their VRDs do not support forms in a "verification pending" state.

Another issue, often classified as an issue of security, relates to deliberate denial-of-service (DOS) attacks against voter registration systems. A DOS attack attempts to flood a voter registration system with false requests for service, leaving no capability for processing legitimate requests. One DOS attack may target the servers hosting a VRD, thus preventing local election officials from accessing it. Another kind of DOS attack may target election officials by flooding them with fake voter registration forms. Although these forms will ultimately be rejected as being fake, it takes time to process each form, and processing fake forms prevents election officials from processing real forms.

3.5 SECURITY AND PRIVACY

Security

Security issues in VRDs arise for two reasons. First, state VRDs contain personal information associated with registered voters, and such information must be protected against disclosures not permitted by law. Second, the overall integrity of the VRD must be protected against unauthorized alterations (e.g., individual records being improperly added, deleted, or changed).

Insecure VRDs pose a number of dangers. Individual voters may be disenfranchised if records of their registration are improperly deleted from the VRD. Voter fraud may be possible if registration records are improperly added. A voter might fall victim to identify theft if sensitive personal information such as a Social Security number is compromised. And improper changes to a voter's record might also effectively disenfranchise him or her (e.g., an altered address might cause the voter to go to the wrong polling place) and at the very least have the potential for creating confusion and difficulty for a voter.

Security measures address the issues of both who is authorized to view or change information in the VRD and of what information within any record in the VRD may be viewed or changed. In the security context, viewing information includes seeing individual records and sending or transferring records en masse; changing information includes adding entirely new records, altering one or more fields within one or more records, and deleting records.

Appendix D describes some important best practices in security. However, these practices only work for data that are under the control of the relevant election official. In the event that the election official shares information with another party (e.g., on demand to a requestor as required by policy or applicable law), there are few if any practical technical measures that the election official can take to ensure the subsequent security of the released data (though some actions can be taken to increase the accountability of the party to whom data are released). Perhaps the only action that the election official can take is to ensure that the data released consist only of those data that are required to be released and no other data. Once the data leave the control of the election official, it is up to the recipient to abide by the terms of use and enforce any relevant security measures. Accordingly, the election official should find a way to bind the recipient—legally—to take the necessary precautions.

Appendix D addresses security issues in greater detail.

Privacy

Privacy is not the same as security, even though they are often discussed together. Privacy issues relate to policy regarding what information may be disclosed to which parties under what circumstances. Thus, a hypothetical law requiring that any registered voter's name and address (but not party affiliation or Social Security number) must be available without restriction to the public reflects a policy choice

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

rather than a security issue. A security issue arises if an unauthorized party is able to gain access through the VRD to the voter's Social Security number, which is supposed to be kept confidential.

Some of the information in VRDs is, by law, public information, although the specifics of which data items can be regarded as public information vary from state to state. In addition, states often limit the purposes for which such information may be used. Nevertheless, the electronic availability of such information raises concerns about the privacy of that information, because electronic access greatly increases the ease with which the information can be made available to anyone, including those who might abuse it.

Some transparency measures are required by law—for example, the NVRA requires public access to the outcomes of most list maintenance activities (excluding declinations, source of registration). Access to such information has been a critical enabler for the efforts of public watchdog groups in discovering problems with state list maintenance activities. Election officials sometimes advocate transparency measures—and most importantly a philosophy of open access to registration-related data—as an approach that helps to ensure the maximum possible accuracy of their files.

Many analysts of privacy issues point to fair information practices (FIPs) as a gold standard for privacy protection that balances privacy rights against user needs for personal information, and in the context of voter registration, the 2006 USACM report on statewide databases recommends the adoption of such practices as the basis for privacy policy regarding voter registration activities.[8] FIPs generally include notifying individuals with personal information that such information is being collected; providing individuals with choices about how their personal information may be used; enabling individuals to review the data collected about them in a timely and inexpensive way and to contest those data's accuracy and completeness, taking steps to ensure that the personal information of individuals is accurate and secure, and providing individuals with mechanisms for redress if these principles are violated.

From an operational standpoint, a full implementation of FIPs for VRDs is likely to prove problematic or undesirable for many jurisdictions. Perhaps the most salient issue is the tension between privacy of personal information and openness and transparency for public records. In its starkest terms, maintaining privacy involves withholding from public view certain information associated with individuals, while transparency involves the maximum disclosure of information, even if such information is associated with individuals.

Although a number of states, such as California, Hawaii, Idaho, Kentucky, Massachusetts, Minnesota, New York, Ohio, and Virginia, have enacted state privacy acts based largely on the provisions of the federal Privacy Act, these acts are typically formulated in such a way that they bar the disclosure of personal information unless disclosure is required by the relevant state's public records act, which may or may not allow the protection of all personal information associated with voter registration records. Such protection may be undesirable for policy reasons as well.

Appendix D addresses privacy issues in greater detail.

3.6 BACKUP

Backed-up files provide users with the capability to restore the VRD in the event of hardware failure (e.g., a fire or flood in the machine room), database corruption as the result of hardware or software problems, operator error, or a successful malicious attack (e.g., a cyber attack) against the database or the hardware.

There are two basic ways (not mutually exclusive) to backing up files. First, copies of the database can be stored and retrieved in the event of disaster. Second, mirrored or replicated facilities allow a system to continue operating even if the primary database is unavailable. How best to back up data

[8] U.S. Public Policy Committee of the Association for Computing Machinery, *Statewide Databases of Registered Voters: Study of Accuracy, Privacy, Usability, Security, and Reliability Issues*, 2006, available at http://usacm.acm.org/usacm/PDF/VRD_report.pdf.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

should reflect an assessment of threats and vulnerabilities (both accidental and deliberate), acceptable parameters for data loss and time-to-restore capability, and available financial resources.

Copying the database has the primary virtues of simplicity and low cost. For databases of modest size, backing up files in this manner is a task that could be accomplished in just a few hours using techniques available to any home PC user—one would simply copy the database file to some backup media late at night. The database could be locked at night for routine maintenance, and the entire file could be copied and stored away. A variety of automated tools are also available to simplify this process.

On the other hand, this simple approach to backup works only when the database in question is sufficiently small. A reasonable upper-bound estimate on the size of a voter's record is 200 bytes, assuming only textual information is stored.[9] The largest state voter database is that of California, with approximately 18 million registered voters, corresponding to a total database size of at most 3.6 gigabytes—files of this size can be copied easily in an hour or two.

But for many voter registration databases, textual information is not the only thing stored. VRD systems are increasingly incorporating capabilities for imaging the paper forms on which voters submit information. If only the voter's signature is stored, a high-quality image may require 100 kilobytes. If the entire filled-in form is imaged, 2 megabytes may be needed. Thus, the incorporation of image-handling capabilities into a VRD changes the storage requirements completely. A California-scale VRD that imaged the entire form for each registration might be 40 terabytes.

Although databases of terabyte scale do not come anywhere near stressing the current state of the art in file management and backup, they call for the use of database technology and hardware platforms that are considerably more sophisticated—and costly—than that of PC technology. These more sophisticated approaches—available in commercial database systems—provide for mirroring data in real time (that is, as it is written) onto redundant media and differential or incremental backup.[10] Such systems sometimes allow selective field backups, so that rarely used information (e.g., the large images of voter registration forms) would be backed up at a much lower frequency than fields that are used regularly (e.g., the much smaller text representations of the information contained in the voter registration forms).

3.7 THE IMPACT OF ELECTION DAY REGISTRATION AND PORTABLE REGISTRATION ON VOTER REGISTRATION DATABASES

Election Day Registration

A traditional VRD operates within a structure that requires a multi-week period between the deadline for new voters to submit voter registration forms and Election Day. Election officials use this period to enter the data from these forms into the VRD and to verify some of the data on these forms if required by HAVA (as in the case of mailed-in registration forms).

Election Day registration (EDR) eliminates this period, allowing voters to register on the same day on which they cast their ballot. On Election Day (or during a period of early voting), a person shows up at an appropriate location (which may be a polling place or a central election office) and presents the necessary identification to an election official. The official consults the registration list and if he or she is not already registered, the election official registers the voter immediately.

[9] In August 2008, the full VRD for Oregon consisted of 2,053,444 records, corresponding to approximately 280 megabytes of data, while the full VRD for Washington state consisted of 3,407,596 records, corresponding to approximately 465 megabytes of data. These totals point to an average record size of 136 bytes per voter. However, the records included only the minimum data needed to perform matching and pointers to the original records; thus, other information such as address, phone numbers, driver's license number, voting histories, and the scanned image of the voter's registration card, including the signature, were omitted.

[10] Differential backup saves all records that have been changed since the last full backup. Thus, a complete data restore involves only two operations—restoring the last full backup and then applying the differential backup. Incremental backup saves all records that have been changed since the last incremental backup. Thus, a complete data restore may involve many operations—restoring the last full backup and then applying the complete sequence of incremental backups in order. On the other hand, incremental backups are much less storage-intensive than differential backups.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

A number of states allow EDR today. Although this report takes no stand on the desirability of EDR, EDR appears to be a trend in the evolution of voter registration, and represents a middle ground between those who would relax or eliminate voter registration requirements and those who would tighten voter registration requirements.

Depending on how EDR is implemented, it may have no implications at all for the design and deployment of a statewide VRD, or it may have many deep and significant implications. The description below is not intended to be a complete discussion of the relationship between EDR, but rather a sketch of some of the important considerations that must be taken into account should any given state adopt EDR.

A VRD must perform two essential tasks for the registration of new voters. It must be able to take in information from a voter registration form (data entry), and it must be able to attempt to verify the necessary information with the DMV or SSA (data verification) for registration forms submitted by mail (a HAVA requirement).

If data entry is to take place on Election Day, sufficient data entry facilities must be available to handle the demand for EDR. These facilities may be located at some central location(s) or at polling places. Data entry at polling places has major disadvantages, such as a noisy or a sometimes-confused or chaotic environment that may make data entry more prone to error. It also requires a data entry station for each polling place, and additional training for poll workers. Data entry at a central location is likely to enable data entry facilities to be used in a more efficient and less error-prone manner, and voters should be able to cast their ballots at central locations in any event.

If EDR is implemented in such a way that data entry and data verification can take place *after* Election Day, there are few implications, if any, for the design of a VRD, simply because this operating scenario is no different from the traditional one involving a multi-week lag between submission of registration forms and Election Day. Assurances that a voter is legitimate would have to be provided by the first-time voter's presentation of the necessary identification. And because HAVA requires a match to either SSA or DMV data only in the case of mailed-in applications, a person who registers in person is not subject to data verification.

From a HAVA standpoint, it is not necessary to perform data verification for individuals submitting voter registration forms to election officials on Election Day if these individuals provide appropriate identification at the same time. However, states may have their own verification requirements for nonfederal elections, and in this case, the VRD must have access to the relevant databases on Election Day. As in the case of data entry, verification is likely to be performed in a more cost-effective and more secure manner from one or a few central locations rather than from polling places.

Portable Registration

Portable voter registration (PVR), defined in this report as the ability of a previously registered voter to vote even if his or her address has changed, has several variations. (In the majority of cases where PVR is implemented, the voter shows up at his or her new polling place or at a central location, submits a change of address form, and is immediately allowed to vote based on the new address.) PVR is required by NVRA for voters who move within a county (more precisely, whose new address falls within the jurisdiction of the same election officials and also is not included within a new congressional district). PVR is allowed but not required by federal law for changes of address within the same state, and several states allow in-state PVR as of this writing.[11] PVR that crosses state lines has not been implemented by any state.

[11] These states include Delaware, Florida, Oregon, Maryland, Ohio, Colorado, South Dakota, and Washington. These states variously allow the voter to use a regular ballot corresponding to the new address, a provisional ballot for the new address, and a regular ballot from the old address. In addition, eight other states have implemented EDR, which provides an in-place process for Election-Day address updates. See Adam Skaggs and Jonathan Blitzer, *Permanent Voter Registration*, New York University, 2009.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

PVR can, in principle, help to mitigate problems arising from a major source of duplicate registrations in a statewide VRD—registered voters who change address.

PVR does not necessarily have implications for the design of a VRD. There is no reason that a voter's change-of-address form must be entered into the VRD on Election Day—only that the voter be allowed to vote (preferably based on the new address). It does mean that poll workers must have access to the statewide VRD (or a suitable local copy of it, such as one in paper form or more likely as a DVD or CD-ROM that can be loaded on a personal computer at the polling place) in order to confirm that a voter was indeed previously and properly registered.

3.8 THOUGHTS ON A NATIONAL VOTER REGISTRATION DATABASE

Proposals are sometimes made to establish a national voter registration database. In principle, such a database could serve one of two purposes. First, it could be used to coordinate statewide VRDs to eliminate duplicate voter registrations across state lines and to facilitate interstate portability of voter registration. Second, it could be used in support of universal or automatic voter registration—an approach to voter registration in which the need for individuals to take affirmative action to register to vote is eliminated by shifting the burden of voter registration to the states in which these individuals reside.[12]

Conceptually, the first purpose is an extension of intrastate portability of voter registration. As noted in Section 3.7, statewide VRDs can facilitate intrastate portability and help to address problems arising from duplicate voter registrations within a state. A national VRD for this purpose could easily be constructed by amalgamating the statewide VRDs of all states and other voting districts and using the statewide VRD data export functions to move the data to the national VRD. Such a database would have to contain some 150 million to 200 million entries (the number of registered voters in the United States), and thus would be approximately 10 times as large as the largest statewide VRD in existence today.

Despite its larger size, however, performing list management (specifically—eliminating duplicates) on such a database is a relatively straightforward computational task. This task could not be managed on a single personal computer commercially available today in a reasonable time, but a mid-size departmental computer using commercially available software and a few dozen terabytes of disk storage would be able to do so with ease. Alternatively, generally available cloud computing services (an example of which is the Amazon Elastic Compute Cloud[13]) could be employed to perform the computational task. Cloud computing has the advantage of eliminating the need for capital investment in hardware. On the other hand, cloud computing is not a technology with which either the public or election administrators have much experience, and thus the use of cloud computing may suffer from a lack of transparency.

The substantially larger size of a national VRD would result in a large number of pairs of entries flagged as possible duplicate registrations. Even with a match rule based on an exact character-by-character matching on first name, last name, and date of birth, it can be expected that around 480 coincidental matches (that is, different individuals who share the same first name, last name, and date of birth) would be identified in comparing VRD lists from Oregon and Washington alone.[14]

The use of a universal national identifier would significantly increase the accuracy of any process

[12] See, for example, Wendy Weiser and Margaret Chen, "America's National Embarrassment: Why Is the Rest of the World So Much Better at Signing up the Vote?," *Foreign Policy*, July 29, 2009, available online at http://www.foreignpolicy.com/articles/2009/07/29/americas_backward_voter_registration_system. Note, however, that advocates of universal or automatic voter registration do not necessarily support a national VRD, and a national VRD is not a necessary component of universal voter registration. For example, Weiser also argues that a national VRD could prove costly and unwieldy, and errors in such a database might improperly disenfranchise voters. Eliza Newlin Carney, "Looking Abroad for Answers on Voter Registration," *National Journal*, July 20, 2009.

[13] See http://aws.amazon.com/ec2/.

[14] This estimate is based on the fact that individuals with a common name such as "Sharon Smith" may coincidentally agree on date of birth.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

intended to identify possible duplicate registrations. The notion of a universal national identifier is itself politically controversial, and the committee takes no stand on the desirability of adopting such an identifier. Accuracy in the resolution of the identified possible duplicates could be enhanced through the use of tertiary data, as discussed in Appendix C under the discussion of third-party data.

Which states would wish to participate in a national VRD? The most likely participants are jurisdictions likely to contain the majority of duplicate registrations, that is, adjoining jurisdictions, jurisdictions that serve as "bedroom" communities for another, and jurisdictions that experience seasonal migration. However, the committee notes that connection to a national VRD eliminates the need for multiple bilateral jurisdictional data exchanges. Thus, if most states will eventually participate in one bilateral data exchange, that exchange may as well be with a national VRD—and subsequent bilateral exchanges will not be necessary.

As for the second purpose, the committee recognizes the political controversy in universal voter registration, and is explicitly silent on the desirability of universal voter registration as a policy choice. Furthermore, a full examination of the technical dimensions of universal voter registration would require more time and resources than are available to this committee. It suffices here to make several observations:

- Universal or automatic voter registration generally calls for government authorities (especially state election officials) to use all available data sources (including those from state departments of motor vehicles (driver's license records), tax rolls, social services agencies, and so on) to assemble lists of eligible voters. Obtaining cooperation from all of these government data sources is likely to require significant effort on the part of political leaders.
- Significant coordination with federal immigration authorities and their databases may be needed to minimize the number of noncitizens added to the voter registration rolls. Such coordination is largely unnecessary today. In addition, noncitizens added inadvertently must be protected from legal harm so long as they do not try to vote.
- The use of tertiary data to identify eligible voters is likely to enhance the accuracy and completeness of automatically compiled voter registration rolls.
- Using any of these data sources is likely to be controversial from a privacy standpoint. As a general rule, privacy advocates are concerned when government authorities, whatever their mission, aggregate data from multiple sources. Some sources of data raise particular concerns—tax rolls, social services agencies, and private-sector sources might be included in this category.
- Standards for data quality assurance would have to be developed and adopted as a part of any attempt to implement universal voter registration.
- The overall cost of universal voter registration may be lower than today's state-centric system, especially if the effort expended by individual voters in registration is taken into account. Resource-strapped counties particularly may benefit from universal voter registration.
- Sustained funding for the voter registration enterprise will be even more necessary that it is today, given the larger role of government authorities in the process.
- To the extent that a national VRD is used to support election officials for checking voter registrations in real time, security (e.g., against denial-of-service attacks) and system reliability and availability will be issues of concern.

Copyright National Academy of Sciences. All rights reserved.

4

Sustainability and Long-Term Funding

Because the underlying information technologies will certainly improve significantly over the lifetime of a system's development and deployment, it is desirable to plan for the eventual incorporation of these improvements. Systems designers must thus pay particular attention to three areas:

- *Sustainability of the technological environment selected.* Technology selection and migration strategies have significant implications for sustainability. Initial choices of technologies from which systems will be built have long-lasting consequences, because they in effect freeze an enterprise's infrastructure. It is therefore important to select mainstream, broadly used platforms. But because the information technology industry is so dynamic, even broadly accepted technologies may later be abandoned by the marketplace. As a result, for example, a company that in 1985 had selected CP/M as its basic operating system would have had to convert long ago in order to remain current. Because maintaining applications that run on platforms based on an abandoned technological substrate is in the long run a very expensive task, applications developers must have migration strategies to port their applications to new technological environments when the old ones become too expensive to use. These comments are not intended to suggest that developers should abandon current systems in favor of the very latest technologies—it is prudent to select relatively stable technologies that are achieving widespread adoption and are likely to enjoy longer-term support. Indeed some old but previously mainstream technologies continue to be supported by vendors because their widespread use keeps them a profitable business.
- *Backward compatibility.* To at least minimally protect investments in design, applications, and training, and provide at least a limited measure of interoperability across versions, commercial information technologies usually incorporate considerable backward compatibility from generation N to generation N + 1, and usually provide tools to facilitate user transition to the newer generation. However, support for backward compatibility is not unlimited, and at some point, support for the earliest generations is usually abandoned. (Thus, generation N – 3 may no longer be fully supported.) Indeed, given the rate of evolution of the processing and storage capabilities of the underlying commercial technologies—and the advances in applications that these improvements enable—it is unrealistic to maintain backward compatibility forever. Election officials will have to provide guidance to system designers for how long backward compatibility for both platforms (e.g., underlying relational database systems) and applications (e.g., VRD compatibility with nonelection systems, such as DMV or SSA databases)

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

are to be maintained, if at all, and indicate a strategy for defining, batching, and sequencing system upgrades. In general, configuration control is required to provide operationally required interoperability and minimize deployment and training costs. The problem is made more difficult when the rate at which enterprise-wide upgrades take place is much slower than the rate of progress in the underlying technology.

- *Uneven rates of modernization*. Technology renewal and refresh for a widely deployed system rarely take place at a uniform rate at all installations. For example, County A in a given state will obtain funding for such purposes 3 years before County B, and thus will upgrade its configuration much sooner than County B. Depending on the nature and extent of the county configuration upgrade, there may be no impact on the county-state interface—this, of course, is ideal. If the upgrade does affect the county-state interface, there emerges a deeper technical issue than backward compatibility—managing and handling a new county-state interface with County A and the old county-state interface with County B means that everything from help desks to software interfaces to operating procedures may need to account for the differences in these interfaces, thus entailing greater support costs than if these differences did not exist.

In this context, configuration refers to the combination of the VRD application and the underlying platform on which it runs. Simultaneous technology renewal/refresh avoids these operational problems and the expense of supporting multiple configurations in the field. But it is very difficult to ensure the simultaneous deployment of a new configuration across a large organization. In states that provide counties with VRD applications (call these states Type 1 states), coordination of the deployment of new VRD applications is likely to be possible. But in states in which counties develop VRDs on their own and feed data to a statewide VRD (call these states Type 2 states), statewide coordination is likely to be virtually impossible. Furthermore, all counties face the question of when to upgrade the base platforms that run applications—and states rarely provide support for such upgrades.

As a practical matter, these observations suggest that all statewide VRD designers will have to support multiple underlying platforms. Also, in Type 2 states, statewide VRD designers must publish clear and complete standards for data exchange that provide guidance to county VRD designers so that new versions of county VRDs are compatible with the statewide VRD. Type 1 states would be well advised to roll out VRD upgrades simultaneously to the extent possible. (These comments are not intended to suggest that it is desirable to deploy for immediate use a new version of a VRD all at once in a "big bang." Indeed, rolling out a new version of a system on a small scale (e.g., in a few jurisdictions) in order to shake out operational problems is often a sensible step to take before large-scale deployment. But at the very least, operational testing and shakedown in such an environment should be part of a deliberate plan to introduce a new version widely. Such a plan might require simultaneous support for two versions for a short period of time, but at least simultaneous support would not need to be continued indefinitely.)

The Help America Vote Act provided a substantial one-time infusion of money for states to acquire modern information technology for supporting election administration, including the statewide voter registration systems that have been deployed. However, all experience with information technology suggests that the initial acquisition cost of information technology is a relatively small fraction of its life-cycle costs. Indeed, after initial deployment, funding streams will be required to support:

- *Maintenance and system upgrade* (keeping the systems running, installing operating system and applications patches to fix bugs);
- *Applications upgrade* (introducing new functionality and capabilities into existing applications in the VRD system, such as those needed to improve functionality or comply with federal or state rule changes);
- *Training* (both for end users of these systems and for new generations of system maintainers trained in the internal operations of the system); and

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

• *Technology renewal and refresh*. Given the rapidity with which information technology evolves, voter registration systems will inevitably have to migrate to new platforms, because the cost of maintaining an existing platform will eventually exceed the cost of migrating to a more modern one.

As for the magnitude of the funding streams required, one study places the total cost of ownership of personal computers in a work environment at more than five times the acquisition cost,[1] suggesting that as much as 80 percent of the initial system procurement cost must be budgeted every year to support nonprocurement expenses not related to data cleanup. Even if the use of more powerful computers and platforms (e.g., virtualized computers) could reduce the total cost of ownership to only twice the acquisition cost (unlikely), it only reduces the annual nonprocurement expenses to 50 percent of the initial procurement cost.

These costs do not account for data cleaning, which would add significantly to the total cost of operating the VRD system over time. (Cleaning data to facilitate comparisons with other databases is an essential component of database management. For example, addresses may need to be standardized if jurisdictions are to qualify for certain lower postage rates in their communications with voters—address standardization services are available but are not currently being offered for free. Other kinds of data cleanup may require communicating with individual voters, and so U.S. mail postage is likely to be a significant component of data cleanup expenses.)

In short, funding for VRD support will require—every year—a significant fraction of the sums spent for the acquisition of VRDs.

The above comments refer primarily to long-term sustainability of existing VRD systems. To implement many of the improvements described in the section on longer-term actions, additional funding may well be required. In some cases, such improvements will require a time-delimited investment associated with initial acquisition and deployment and a smaller stream of funding afterward; in other cases, they will require additional funding on a continuing basis as operating expenses. It is also possible that different kinds of spending will entail different political processes, as budget accounts for computer and system acquisition may well be separate from budget accounts for cleaning election-related data.

[1] See John Taylor Bailey and Stephen R. Heidt, "Why Is Total Cost of Ownership (TCO) Important?" *Darwin Magazine Online*, November 2003, available at http://www.darwinmag.com/read/110103/question74.html.

Copyright National Academy of Sciences. All rights reserved.

5

Actions Possible in a Relatively Short Time Frame

Presentations by representatives of various states to the committee since the November 2008 election indicated that the states' databases generally performed well in both the primary and the general elections of that year. Nevertheless, the committee believes that a number of meaningful nontechnical actions to further improve voter registration databases could be implemented in a relatively short time frame. (All but one of these actions were discussed in the committee's interim report of April 2008;[1] for the final report, the committee chose to add one additional short-term action, presented below as Recommendation S-10 (on sharing information regarding best practices). None of the short-term actions contained in the interim report were deleted, though Recommendation S-9 (on encouraging third-party groups to turn in forms promptly) was slightly modified in the interest of clarity.)

These actions focus on two areas: (1) education and dissemination of information and (2) administrative processes and procedures. Because these actions remain relevant to future elections, they are repeated below in full (with certain editorial changes to clarify the original intent of the committee).

These short-term actions are directed primarily at election officials at the state and local/county level, and the legislatures and county commissions that make policy regarding the conduct of elections at the state and local level. In some cases, the Election Assistance Commission has a useful role to play as well in facilitating and promoting their implementation.

5.1 PUBLIC EDUCATION AND DISSEMINATION OF INFORMATION

Recommendation S-1: Raise public awareness about the legibility and the completeness of voter registration card information.

Accurate and complete data are a basic element of a high-quality VRD. But as noted in Appendix C, the quality of the data in a VRD is no better than the data that are entered into the system. For example, illegible information impairs the ability of election officials to check registrations as required by HAVA

[1] National Research Council, interim report on *State Voter Registration Databases: Immediate Actions and Future Improvements*, National Academies Press, Washington, D.C., 2008.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

and/or state law, possibly placing additional downstream burdens on the voter (such as having to verify information by mail or having to provide an ID when voting the first time).

Efforts to raise public awareness about the importance of legibility and fully completing voter registration forms would help to reduce the amount of illegible or missing information on these forms when they are submitted for data entry. Properly undertaken, these efforts to raise public awareness of this particular issue could be integrated with ongoing efforts to encourage people to register to vote. Jurisdictions could take some or all of the following specific steps:

- Emphasize in the instructions for filling out voter registration forms the importance of legibility and completeness (for example, "Please print all responses; if your answers are illegible, your application may be mis-entered, rejected, or returned to you.").[2]
- Conduct media campaigns (perhaps undertaken by the Ad Council) emphasizing the importance of legibility and completeness in the information provided on voter registration forms.
- Coordinate with third-party voter registration groups and public service agencies, emphasizing the need for their field volunteers to attend to legibility and completeness as they distribute and/or collect registration materials.

5.2 ADMINISTRATIVE PROCESSES AND PROCEDURES

A variety of recommended administrative processes and procedures will also help to ensure higher-quality matching and increase voter confidence in VRDs. Note, however, that large volumes of registration forms usually need to be processed as registration deadlines approach, a workload that jurisdictions commonly rely on temporary staff to handle. Unless other arrangements are made to adjust workflow (such as ensuring that actions that require human judgment are routed to permanent staff), these temporary staff will, in many cases, have to carry out these recommended processes and procedures, suggesting that training them to do so will be necessary.

Recommendation S-2: Resubmit alternate match queries if the response returned from the Social Security Administration or department of motor vehicles is a nonmatch.

An election official can use any additional information available to generate match variations for a given name. For example, a match might be sought on standard name variations (for example, Bill versus William), or transposed fields (for example, last name and first name), or compound names separated, or on a maiden name if available.

In practice, humans tend to submit only a few variations. Moreover, manual name rooting is likely to be inconsistent across different officials or even across the same official in different states of fatigue. For this reason, the committee has made a recommendation regarding the use of automated name rooting (Recommendation L-8), although implementing such functionality is difficult to do in a short time frame.

Finally, it may be possible to resolve a nonmatch result by direct contact with the voter, either by phone, in writing, or via e-mail.

Recommendation S-3: Provide human review of all computer-indicated removal decisions.

Because inaccuracies in data may lead to false matching by automated processes, the committee urges jurisdictions to provide a human review of each and every decision to remove a registered voter from a VRD. One step in human review is for a trained election official to examine every computer-indi-

[2] Even the National Mail Voter Registration Form does not address this point.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

cated decision to see if it makes sense, subject to the availability of such personnel. Such an examination may well reduce the need for the second step, described below.

A second step in human review is advance notification of cancellation—contacting the voter prior to implementing a cancellation. Such contact could be effected through the use of computer-generated letter in the mail. (Such a procedure would not conflict with the requirements of the NVRA, which call for election officials to send to the voter a "confirmation" card that requests verification of his or her voter's address. After it is returned undeliverable, the voter stays on the rolls for two more federal elections before the registration can be cancelled. State law may or may not require a final notice of cancellation after all NVRA requirements have been met.)

For example, letters could be sent to individuals who are at risk for being removed from the voter registration list; these letters would have a "respond by date X or be deleted" notice. If a notice comes back as "undeliverable as addressed," the name of the individual would be deleted after date X. If duplicate records are of concern (that is, if two records appear for the same individual), the incorrect record can be deleted. To determine which record is correct, election officials could check available data sources (for example, tax records, real estate records, online search engines optimized for finding people such as www.zabasearch.com, and the telephone book) and/or contact the voter.

Recommendation S-4: Improve the transparency of procedures for adding voters and for list maintenance.

There is not enough transparency in the procedures used to add and remove voters from VRDs. To improve transparency, the states and local jurisdictions should:

- Make specific written procedures for the verification of new voters and the handling of removals publicly available. These procedures should address explicitly the specific field-level and record-level matching criteria used for each of these processes. These procedures both inform the public of what election officials intend to do and provide a standard by which audits, oversight, and accountability can be measured.

As an example of the need for making even small details of match algorithms public, Alvarez et al. found that using a match rule involving full first and last names, date of birth, and *middle initial* resulted in 2.4 times more matches than using a *full middle name*.[3]

Extrapolating from this example (not at all atypical), it is clear that "small technical details" in algorithms for matching records can have large policy consequences, such as systematic disenfranchisement of qualified voters (if such matching caused voters to be removed incorrectly)—a point that underscores the importance of transparency and openness regarding these algorithms.

- Publish these procedures widely, for example, on the election office Website.
- Collect relevant data, such as data on the outcomes of initial applications for registration:[4] How many applications were received? Of these, how many were approved and how many rejected? Of those rejected, what were the reasons for rejection—illegibility, incompleteness, person ineligible (cite reason for ineligibility), and so on. Data on how the state handles removals from the registry are also relevant: How many removals were made? Of these, how many were due to intrastate movement, death, and so on?

[3] R. Michael Alvarez, Jeff Jonas, William E. Winkler, and Rebecca N. Wright, "Interstate Voter Registration Database Matching: The Oregon-Washington 2008 Pilot Project," available at www.usenix.org/events/evtwote09/tech/full_papers/alvarez.pdf.

[4] Many jurisdictions already collect such data, and aggregations of some of these data are published in the EAC Election Day Survey. For more information on the Election Day Survey, see http://www.eac.gov/schedule/2008-election-day-survey/.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Box 5.1 Examples of Auditing Applied to Voter Registration Database Processes

Auditing Removals from Voter Registration Rolls

Election officials need the date of receipt of registration applications, the date on which a registration-related notice was sent to the voter, the date, if any, of any response from the voter, and the date on which the corrected or completed information was received; indexes of all of these dates must be kept if correspondence and documentation are to be located. In one state, a removal letter is kept with the original application, and these are sorted by year of first receipt and then alphabetically by name. In this state's experience, the individuals claiming that they had registered but not been found on the voter registration list had often been sent a copy of the letter, as could be demonstrated by referring to the voter's file.

It is also possible to keep copies of voter registrations that are cancelled and removed from a voter registration database (VRD). Keeping such records (possibly in a different file) together with the reason for the cancellation would provide data that might be used to improve matching algorithms and/or cancellation procedures.

Auditing Changes in Voter Registrations Records (1)

The main text of this report suggests that voter access to a VRD should be implemented through buffered access to a synchronized copy of the VRD, not to the VRD itself. One kind of audit procedure checks expected behavior against actual behavior. For example, an audit procedure could keep a log of which records were changed in the primary source (the VRD) since the last synchronization between the VRD and the copy. This log could be used to identify the records in the copy that are supposed to be changed—changes in the copy that do not match this list would indicate a problem that election officials could and should investigate further.

Auditing Changes in Voter Registrations Records (2)

As a general rule, corrections or updates to an individual voter registration record should be recorded in such a way that full reconstruction of previous versions of the record is possible as well as an accounting of who made each change. Without the complete revision history available, it is impossible to determine the state of a record at any given point in time and at the same time determine who made what change when. Thus, it may be necessary either to retain the old data in their original form (changes being appended rather than overwriting old data) or to have reliable backup and restore mechanisms so that the old data can be easily retrieved.

- Publish these data widely. For example, the Data.Gov Web site would appear to be an appropriate venue for publication.[5] Publication of these data would enable public scrutiny of the aggregate outcomes of list maintenance operations and thus prevent politically motivated purges from being performed in secret.
- Audit the processes to ensure that procedures are being followed (see Box 5.1 for examples).

Note that collecting and publishing the data suggested above can provide a basis for assessing how big a problem illegibility actually is, how many persons apply who are actually ineligible (for various

[5] See http://www.data.gov/. The Obama Administration is strongly calling for government to publish their transparency data in this single location making it easier for consumers of this data to locate it.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

reasons), and so on. The more of such data there are, the easier it will be for election officials to identify problems and to improve list maintenance procedures.

Recommendation S-5: Use printable fill-in online registration forms.

Typewritten or printed information is almost always more legible than handwritten information. Assuming they already have Web sites from which voters may obtain voter registration forms and other election-related materials, jurisdictions could encourage the use of fill-in online registration forms, such as fill-in PDF or Web forms that accept keyboard input (that can be printed, input and all); a number of states provide this service today. Although the form must still be printed, signed, and then mailed or delivered to the election officials, the information on the form will be much more legible. (Note that although the deployment of a new encoding of an old form—such as the National Mail Voter Registration Form—should be possible in a relatively short time frame (the EAC is a logical focal point for any such effort), it should not be regarded as a trivial effort that can be accomplished without some care and testing.)

Recommendation S-6: Perform empirical testing on the adequacy of processes for adding to and maintaining lists.

The only way to know how well a system is working is to test it. One way to test the adequacy of VRD adding and maintenance processes is to corrupt a copy of the most recent VRD by seeding it with artificial records with names and other identifying information from lists of felons, deaths, and mentally incompetent people and with duplicate records of individuals already in the database but with realistic types of error in them. Once corrupted in this way, the VRD can be matched against all of the usual databases (DMV, felons, and so on) to see what fractions of the corruption in each category were detected, thus providing estimates of rates of false positives and false negatives. Because "ground truth" is known in the form of the original seedings, the fractions of detected corruption are likely to be reasonable estimates of the effectiveness of the process overall.[6]

A corollary of such testing is that those who receive the data resulting from such testing (ultimately, the public at large) must be educated to interpret the data in context—and specifically to understand that no procedure for adding or removing voters can be perfect. At the same time, there is nothing to suggest that individual voters who are wrongly eliminated from the VRD cannot complain or seek correction of the problem through existing channels that are available for resolving such problems.

Another possible approach to testing is to audit actual acceptance, rejection, and removal decisions, not just to verify that procedures have been followed but also to estimate error rates. For example, it is often helpful to test the algorithm and matching procedures involved using a sampling approach in which some practical number of computer-indicated removals are randomly selected from the complete list. Election officials would then perform human review on those randomly selected names, and if the number of improper matches exceeded a certain threshold (which could be zero), the list-generation algorithm and parameters should be further investigated for possible flaws.

The committee emphasizes that sampling is a method for testing algorithms and procedures, not for performing list maintenance, and despite recommending the adoption of sampling for such testing, it reiterates the importance of human review of all indicated removals.

Recommendation S-7: Take steps to find and minimize errors during data entry.

[6] However, note that even the best state-of-the-art "error generators" are not capable of generating the full range of errors encountered in real databases. Thus, these estimates are likely not to account for certain kinds of errors; as a result, actual performance in realistic settings could be expected to be different and probably somewhat worse.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

A number of steps can be taken to minimize the effect of data entry errors.

- Sample audits can be undertaken to assess the degree of the problem and to identify the source—some data entry personnel, for example, may be much less accurate than others. Some systems produce daily data entry reports that can be compared against the original card for errors; such systems are used in a number of jurisdictions.
- The registrant can be provided with a copy of the data that were actually entered (for example, when a voter receives his or her registration card, which should in most cases reflect all of the data entered on behalf of the voter), reminded to check the data, and given information on how to contact the election jurisdiction if there are errors on the card. (To maintain security of personal information, the card should be mailed in a sealed envelope.)
- During the input process, the entered values can be tested against domains (for example, common names,[7] valid addresses including street name and postal code, valid phone numbers, valid dates of birth).
- Data can be entered twice by different people and compared for discrepancies (an expensive way to check, but effective in most instances).
- Discrepancies can be found when matching new inputs to previously known values (an ideal way to detect transposition keying errors in dates of birth, for example).

When errors or inconsistencies in the entered data are found, they should be immediately corrected. In some cases, an examination of the records themselves will indicate how corrections should be made; in other cases, it may be necessary to consult additional data sources or even the voter to make the necessary corrections. For example, election officials might provide a special telephone number for voters to call to make corrections.

The first two of these steps (sample audits and the voter being provided a copy) can be taken in a relatively short time frame. The other three require a nontrivial amount of new technology deployment.

Lastly, voter registration forms should ask the applicant to provide (voluntarily) e-mail addresses and/or SMS (text message)-enabled cell numbers in order to facilitate contact with the applicant—such contact may well be necessary in order to clarify other information that is unclear on the registration form.

Recommendation S-8: Allow selected individuals to suppress address information on public disclosures of voter registration status.

Although voter registration information is nominally public in most states, certain individuals (e.g., domestic violence victims, undercover police officers, witness protection program participants, and so on) have legitimate and understandable reasons for wanting to make address information inaccessible to the public, and an administrative process should be available to protect such information on request. Indeed, some states (such as New Mexico, California, Oregon, Missouri, and Kansas) already provide for such suppression under certain circumstances.

Defining who can and cannot suppress address information involves a balancing of privacy and other interests. Advocates of more open records might argue that the public interest is best served by a relatively narrow scope of individuals who should be granted such privileges, whereas some privacy advocates might argue for the broadest possible scope. The committee is silent on this particular point.

[7] Comparing the name entered as data against common names is a useful process for suggesting the possible occurrence of data-entry error. But *correcting* a name based only on such a comparison might well introduce error. "Jazmine" as the real first name of an individual might be compared against the more common name "Jasmine," but "correcting" the spelling of "Jazmine" to "Jasmine" would be an error. In such situations, the indication of a *possible* error suggests the need to re-check the transcription against the original voter forms in order to prevent the introduction of error at the error-correction stage.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

The committee notes that enacting this recommendation may require legislation in many jurisdictions.

Recommendation S-9: Encourage entities sponsoring voter registration drives to submit voter registration forms in a timely manner to reduce massive influxes at the registration deadline.

Election offices can be overwhelmed by the mechanics of data entry if large numbers of voter registration applications must be processed in a very short time. Such a volume reduces the time for error checking or multiple attempts to verify voter information, and often forces election officials to hire inexperienced temporary workers for data entry. These conditions in turn are likely to increase the error rate of data entry and may invalidate more registration applications that would be the case if more time were available to handle the applications.

In addition, forms that are not processed (not entered into the relevant VRD) in a timely manner can cause confusion for the applicant. Election officials often have to deal with a large number of inquiries from applicants who filled out an application and gave it to someone weeks or months before. Seeking information about whether their applications have been received and processed, these applicants are often told that their applications are not in the system, and are encouraged to submit another application because the election officials do not know if the party collecting applications will actually turn in the form. This chain of events is frustrating for the applicant, and it causes more work and confusion (more duplicates) if and when the original form is eventually submitted.

Some states (e.g., Oregon) require third-party groups to submit applications to election officials within a certain number of days of collecting them. Nevertheless, imposing such requirements more broadly may entail politically controversial state legislation.

Recommendation S-10: Improve information sharing regarding best practices and lessons learned regarding VRD acquisition, operation, and maintenance.

Election officials in various states operate a variety of voter registration databases and face many different problems in acquiring, operating, and maintaining their databases. Election officials, and the technologists who support them, would be likely to find value in reports of best practices and lessons learned in the ongoing database enterprise. The content of such venues would logically include both process and substance knowledge. In the former category might be lists of state database administrators and their contact information, provided so that others might call to inquire about how various things are done within their jurisdictions. In the latter category might be data related to database performance or published procedures for list maintenance.

The committee thus recommends the establishment of continuing venues that could act as repositories of such wisdom. Such venues should be regular and continuing, either through face-to-face meetings and/or a Web site. In the committee's view, the most logical focal point of action would be the National Association of State Election Directors.

Copyright National Academy of Sciences. All rights reserved.

6

Possible Future Improvements That Will Require Longer-Term Action

The material below describes actions that are almost certainly guaranteed to take more than several months to implement successfully. Indeed, given the time frame needed to implement changes that require the modification of computer systems (which involve at a minimum time to design, code, test, and document changes, and may require new procurements, procedures, and/or training), the committee cautions election officials against best-case planning scenarios in trying to implement any of them. In other words, none of the actions below should be placed on the critical path for an election that is coming up shortly.

As before, these longer-term changes are directed primarily at election officials at the state and local/county level, and the legislatures and county commissions that make policy regarding the conduct of elections at the state and local level. In some cases, the Election Assistance Commission has a useful role to play as well in facilitating and promoting their implementation. In addition, a number of the recommendations below are directed to the U.S. Congress, to the Social Security Administration, and to various nonelection agencies in the states and counties, because the effectiveness of statewide VRDs will depend on actions that these entities do or do not take in the future.

6.1 PROVIDE FUNDING TO SUPPORT VRD OPERATIONS, MAINTENANCE, AND UPGRADES

Recommendation L-1: Provide long-term funding for sustaining voter registration database operations.

The one-time infusion of federal funding provided by HAVA will not—and was never intended to—support VRD operations in the long run. A statewide VRD is a major investment in information technology, and its effective operation over time will require funding for operations, maintenance, and upgrades. The committee is silent on the appropriate source(s) for such funding, which might be some combination of federal, state, and local sources, but makes three critical points:

- Funding for operations, maintenance, and upgrades must be sustained over time—whatever amounts are allocated for such purposes must be continued year after year.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

- The amount required annually to support these activities is likely to be a significant fraction of the sums spent for the initial procurement of a full VRD system—40-50 percent would not be surprising.
- Giving short shrift to funding for operations, maintenance, and upgrades is likely to result in poorer performance and the occurrence of avoidable mishaps in the operation of VRD systems.

6.2 Improve Data Collection and Entry

Recommendation L-2: Develop and promote public access portals for online checking of voter registration status.

In anticipation of being able to vote on Election Day, prospective voters may wish to check their voter registration status so that any irregularities can be corrected in time. Web-based portals for checking the state VRD increase the ability of individuals to+ do so. For example, such a portal may ask the user to provide a name, birth date, and Zip code, and return either the user's current registration status or an indication that there is no record on file that matches the information provided. A number of jurisdictions across the country, including Kentucky, Washington, Oregon, Nebraska, and Nevada, provide this service today to voters today.

When protected against security and privacy violations, such portals serve the public interest in increasing transparency of the VRD and create another opportunity for the verification of voter information. They benefit individual voters who want to verify their information, and may provide an opportunity (if it is legal to do so, and if potential privacy concerns over retention of the data can be addressed) for third-party voter registration groups to confirm that the applications they have collected have been received, processed, and accurately entered in the voter registration database.

States that have developed such portals (for example, Nevada[1] and Nebraska[2]) have generally integrated them into their voter registration Web sites. These portals must access information stored in a state's VRD, which means that their development requires some sensitivity to and technical capacity for dealing with security issues. For example, data compromises have been reported in other instances when live queries have been allowed access to the primary database, suggesting that it may be safer to implement some sort of buffered arrangement whereby the portal provides access only to a synchronized copy containing only the minimum amount of information.

Another point to be considered is the prevention of automated exploitation that might circumvent existing legal restrictions on making the voter registration database available to commercial users; automated tests ("captchas") that distinguish between human and automated responses (for example requiring the user to type the letters displayed in a distorted image[3]) may be relevant in this regard, although this is an ongoing battle. Special steps must also be taken to prevent the display of voter registration information for individuals who need protection, such as victims of domestic abuse or individuals in witness protection, and in any event, the information to be displayed at all should be the minimum information needed for the voter to know that he or she is registered to vote and to inform the voter of the proper polling place (for example, driver's license numbers or SSNs (even SSN4) do not need to be displayed). Some states collect more information (for example, phone numbers, occupation, or e-mail addresses) on their application forms than is necessary for voter registration per se; such information poses increased privacy risks to the individual if needlessly disclosed.

Finally, for all states that provide online verification of voter registration information, it is important to inform voters that they can and should check their voter registration status well in advance of Election Day.

[1] See https://nvsos.gov/VoterSearch/.

[2] See https://www.votercheck.necvr.ne.gov/.

[3] A CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) is a program that differentiates between humans and computers by generating and scoring tests that humans can pass but current computer programs cannot. For more information, see http://www.captcha.net/.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Recommendation L-3: Allow voters to register and to update missing or incorrect registration information online.

As noted in Appendix C, typographical errors could be reduced significantly by eliminating the data transcription process and importing most or all of the relevant data from another system and/or allowing the voter to enter data himself or herself when necessary. However, the voter will always have to provide at registration some means of authenticating himself or herself at the polls, such as a signature. A mail-in registration form can contain a box for the voter's signature, but online registration requires the applicant to appear (or to have appeared) somewhere in person at some official government agency to provide a signature. If this signature is digitized, it can be made available to the election official along with the information needed to register to vote. A number of states today take advantage of the fact that applicants for driver's licenses must provide a signature; these states have developed online registration portals that enable citizens with driver's license application signatures on file to register to vote online without having to appear in person anywhere.

Registration portals can also leverage the fact that basic information about the individual, such as name, address, birth date, and so on, are often also stored along with the signature—suggesting that importing the relevant data from the original state agency with the signature into the voter registration database is feasible in principle. When the voter registration application required information not already on file, the user would enter the information himself or herself and then be given a chance to verify and correct the information.

In addition, individuals whose registration forms contain illegible or missing information could be notified of that fact and at the same time be given a special code or password that would grant entry to a secure Web page, whereupon the individual could correct or provide the missing data. In the longer term, it might be possible to realize real-time verification of an online application for voter registration, so that an applicant whose information did not match DMV or SSA information on file could be informed of that fact immediately, so that corrections could be made at the moment.

If the individual's signature is not already online with some other government agency, the individual will have to provide an original signature on a physical registration form. But such a form can be provided to the individual online, filled in online and the data captured, and then printed (and signed) for submission.

Such a procedure has several advantages for this committee's recommended short-term action regarding online fill-in forms, which would still require that the data be manually captured upon receipt at the election official's office. With online data capture, the individual's data can be stored temporarily and then entered officially into the VRD (i.e., made permanent) when the signed form is received. This procedure eliminates the need for further processing of the typed information on the form (i.e., no data reentry or Optical Character Recognition (OCR) scanning), reducing costs and increasing accuracy. In addition, during the period between online data capture and receipt of the form, election officials can "pre-verify" the data entered and contact the individual if the necessary match cannot be made. With contact information on file (such as e-mail addresses), election officials can also remind the individual to submit the form and can provide information regarding drop-off locations for the form at colleges, schools, and other locations. And online acknowledgment of receipt of the signed form can be provided as well.

Online registration would also help UOCAVA voters to register.[4] Today, the registration process for military and civilian voters overseas is cumbersome, requiring transmittal of completed registration forms by physical mail. Transmitting the information on voter registration forms would eliminate this sometimes-unreliable step.

[4] UOCAVA refers to the Uniformed and Overseas Citizens Absentee Voting Act of 1986, which enables states and territories to allow certain groups of citizens to register and vote absentee in elections for federal offices. Most states and territories also have their own laws allowing citizens covered by the UOCAVA to register and vote absentee in state and local elections as well.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Recommendation L-4: Encourage/require departments of motor vehicles as well as public assistance and disability service agencies to provide voter registration information electronically.

The NVRA requires state DMVs, public assistance agencies, and disability service agencies to facilitate the voter registration process. Today, this facilitation is mostly paper-based. Automatically providing information on new applications or changes of address to election officials would significantly reduce the burden of maintaining VRDs by reducing requirements for manual data entry and updating registrations with new addresses.[5]

As part of promoting cooperation and coordination between election officials and these other public service agencies, states may wish to develop and maintain performance metrics on the percentage of voter registration additions, modifications, and deletions that arrive electronically and on the number of electronic files that arrive from NVRA agencies that contain errors requiring correction. Making such figures public (e.g., through publication at www.data.gov) would provide a way of holding these agencies more accountable for their NVRA responsibilities.

The committee recognizes that election officials have no control over the budgets or operations of these agencies, a fact that often leads to a certain amount of bureaucratic politics as Agency A seeks to persuade Agency B to help carry out the mission of Agency A.

Recommendation L-5: Improve the design of voter registration forms.

The design of forms has a significant impact on their usability and their ability to capture the data that the form filler intends to record. For example, providing a specific separate space for each letter/number of the name/address often improves the legibility of forms completed, and may improve the suitability of the filled-out form for processing by optical character recognition software. In addition, the design of voter registration forms and data entry screens for VRD systems should be coordinated in order to minimize the data entry clerk's effort necessary to find information on the form.

Form design is often challenging and generally requires a significant degree of empirical testing to assess the usability of any given design. The committee finds considerable value in the work of the Design for Democracy project in designing election-related forms that are highly usable by lay people.[6]

Recommendation L-6: Encourage and if possible require departments of motor vehicles, public assistance and disability service agencies, tax assessors, and other public service agencies of state and local government in their communications with the public to remind voters to check and update their information.

Agencies of state and local government communicate with the public regularly, and each such communication is an opportunity to remind voters to check and update their information. Such reminders

[5] Recommendation L-4 is consistent with the EAC's *Voluntary Guidance on Implementation of Statewide Voter Registration Lists*, III-D.2-d. This particular guidance notes that states should "ensure that the coordination of information in the verification process is accurate and efficient. Verification of voter registration information shall be accomplished through electronic transmission. Further, to the greatest extent allowed by State law and available technologies, this electronic transfer between statewide voter registration lists and coordinating, verification databases should be accomplished through direct, secure, interactive and integrated connections." See www.eac.gov/election/docs/statewide_registration_guidelines_072605.pdf/attachment_download/file.

[6] See, for example, the informative reference on the design of forms for use by election officials by Marcia Lausen, *Design for Democracy: Ballot and Election Design*, University of Chicago Press and American Institute of Graphic Arts, 2007. (The Design for Democracy project recommends for voter registration forms using capital and lowercase letters rather than all capital letters; prioritizing information for registrants over information for administrators; keeping type font, size, weight, and width variations to a minimum; not center-aligning text or headings; using contrast and graphics to support hierarchy and to aid legibility; and not using decorative art or illustration.) More information on the Design for Democracy project can be found at http://www.aiga.org/content.cfm/design-for-democracy.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

could be helpful in increasing the accuracy and completeness of the data contained in VRDs. Further, the online environment for state and local agencies provides opportunities for less passive forms of reminder—for example, individuals who use online government services to indicate a change of address (for example, on tax or property assessment records) can be offered reminders to update their registration information, or can even be routed automatically to online voter registration services to effect a similar change of address.

Because these additions would generally entail only small changes to existing applications in these other service agencies, they would be significantly less expensive than implementing the previous recommendation on developing and promoting portals for online checking of registration status and thus might well be a first long-term step that states could take.

Recommendation L-7: Consider providing tracking tags for voter registration forms to improve administrative processes.[7]

If a jurisdiction were to provide tracking tags, voter registration forms would have a tear-off tracking tag, and online registrants would be told to make a copy of the online form. First-time voters would be instructed to keep the tag or the copy of the online form and to bring it with them when they try to vote. States would keep data (and report such data to the EAC) on how many individuals attempted to vote and were not registered but had their tags or presented copies of their form. States would then be encouraged to lower the number of individuals in this category. In order to discourage attempts to improperly discredit or disrupt the voter registration process (e.g., through the use of fake tags and false claims that an individual was not registered), it might be necessary to provide for statutory penalties for the inappropriate use of these tags.

In addition, the tag might also include a tracking number or bar code to match it with the registration form itself, facilitating the association of specific individuals with specific forms. Blocks of numbers could also be allocated to different organizations to use. On the other hand, because including such a number or code would almost certainly have to be a government function, requiring such numbers or codes might run afoul of the NVRA, which specifically allows private duplication of voter registration forms in order to facilitate their widest possible distribution. In addition, numbered forms would entail additional costs for printing. Some states (e.g., Missouri and New Mexico) provide numbered registration forms today. The committee, however, takes no position on the general desirability of tracking numbers or codes at this time.

Although the use of these tags is not intended to substitute for a proper voter registration or for provisional voting, such tags would provide a factual basis for investigating, at least partially, claims from one political party that supporters of the other party have "pocketed" voter registration forms—that is, when conducting voter registration drives, receiving registrations for people of the opposite party and never turning them in. This activity is against the law, but there can be no proof as to whether it has occurred unless there is some form of receipt given to the person registering. If there were tags, then people who possessed them but were not in the VRD would be proof of some problem, including the possibility that registration forms had gone missing.

Election officials would also note in the language on the form explaining the tag that the tag is intended for administrative purposes, and is in no way a substitute for a valid and properly processed voter registration form. That is, in the absence of clear explanations to the contrary, citizens may believe

[7] The corresponding recommendation in the interim report was more emphatic than the version offered in this, the final report. The reason for the change was that in the committee's information gathering after the release of the interim report, additional information came to light suggesting that implementing such a recommendation might be problematic for certain jurisdictions. Although the committee continues to support the original rationale provided for the recommendation in the interim report, it now recognizes a more complex cost-benefit tradeoff than was reflected in that report. The new language of Recommendation L-7 is intended to reflect this realization.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

that they will be allowed to vote, even if not properly registered, if they can present a tag or a copy of an online registration form to poll workers.

The committee recognizes that the NVRA (Section 8(a)(2)) already requires that election officials provide notice to applicants on the disposition of all voter registration applications. But this requirement can only be met when the applications indeed make it into the hands of these officials—if they never arrive, notice cannot be given, and individuals who never receive a notice cannot prove that they should have received notice.

Another important benefit of such tags is that they can facilitate reminders to third-party voter registration groups to turn in forms that they have been holding for an excessive period of time. Election officials can keep track of numbered registration forms as they are distributed to third-party groups, and if the applicant has the tag when he or she calls the election office, tardy groups can be identified and reminded to turn in the forms they are holding.

6.3 IMPROVE MATCHING PROCEDURES

Recommendation L-8: Upgrade the match algorithms and procedures used by election officials, the Social Security Administration, and departments of motor vehicles.

To the best of the committee's knowledge, many (if not most) of the matching procedures used by the states have been developed on the basis of intuitive reasoning without further systematic validation or mathematically rigorous analysis, do not reflect the state of the art in matching techniques, and have not been validated scientifically, in the market, or otherwise. The best computer matching procedures that have been developed and compared by both researchers and industry do not appear to be widely used by the states for voter registration purposes. State-of-the-art matching techniques have been successfully used in a variety of commercial and government applications. The committee believes that there are several areas in which matching involving VRDs can be improved, and thus recommends that election officials engage the relevant technical community when considering improvements in matching techniques as described in the section "Improving Record-Level Matching" in Appendix B.

The enhanced methods should improve (1) the capability for locating of duplicates in a state's VRD, (2) the matching of voters against the state DMV file and the SSA files, and (3) the matching of registered voters against any secondary federal or state list (for example, of deaths, felons, and so on). The effectiveness of these enhanced methods could readily be demonstrated by applying them to a particular state's VRD file and showing (especially through confirmed communication with the voters) how rates of false positives can be quite low even while significantly lowering rates of false negatives.

The committee believes that matching procedures can be substantially improved by implementing four changes that are described below and in greater detail in Appendix B.

- *Automated name rooting*. Matching processes should handle equivalent common names (e.g., Bill, William, Will, Willie) and different spellings of those names (e.g., Jazmine and Jasmine, Mohamed and Muhammad) in a more automated fashion in order to avoid the problems associated with manual processing of equivalent names. This process should be implemented either at the election office or, more ideally at the highest point of integration (e.g., at the DMV or the SSA which provides lookup services for many users). One option is for the system to generate all the name variants. A second option is to assign to each name a most-rooted form (e.g., Bob = Robert and Rob = Robert), and when rooted forms match, putatively different names can be regarded as members of the same name family. The false positive rate will be low if other attributes such as date of birth and SSN4 or driver's license number are taken into account.
- *Automated name ordering*. Different cultural conventions may affect how names are represented in a database. For example, the Hispanic name Lucia Vega Garcia may be recorded in a database as Lucia Vega, Lucia Garcia, Lucia Vega-Garcia, or Lucia VegaGarcia, depending on how the data entry clerk chose

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

to represent the fact that she uses Vega Garcia as her "last name." Thus, matching processes should be able to handle in a more automated fashion different representations arising from ordering and spacing variations. As with name rooting, this process should be implemented either at the election office or, more ideally at the highest point of integration (e.g., at the DMV or the SSA, which provide lookup services for many users). (More discussion of the issues associated with name ordering is contained in Appendix B and Appendix C.)

- *Wildcard matching capabilities*. These capabilities may be useful for searching and matching in the presence of incomplete information.[8] Wildcard matching, especially for "*" on name fields located at the beginning of the string, may be impractically slow on large databases because the match may require examining every record in the database. This would be especially true in searching SSA and DMV databases, given the need for relatively quick response. However, if the universe of relevant search can be narrowed (e.g., by using the first few characters of the name, or by using other fields such as a date of birth), a wildcard match can be performed in a much shorter amount of time.
- *Blocking and string comparators*. Used in matching, these techniques—described in Appendix B—return a score indicating the degree of similarity of two fields, rather than the simple "match or no match" outcome of naïve matching algorithms.

The above changes should be implemented in applications of the Social Security Administration and state departments of motor vehicles for processing verification queries from election officials. As a technical matter, it is easier to implement such changes in the query processing application rather than in the query generation application (if done in the generation application, an inordinately large number of queries would be generated). In addition, implementation of blocking and string comparators is likely to be a nontrivial programming task, and such a task may be beyond the resources and technical capabilities of many jurisdictions. Lastly, from the point of view of reducing duplication of effort, implementing it once at the SSA or the DMV makes much more sense than implementing it in multiple jurisdictions. Individual jurisdictions may also wish to adopt these changes to improve intrastate matching (such as in the case of lists of state felons) and when they compare their own VRDs with those of other states.

Although it is not likely that software for implementing these functions will be free for the taking from the Internet, a number of sites provide good points of departure for technical personnel interested in improving matching capabilities.[9]

Finally, any new matching procedure used in a VRD or to support a VRD should be rigorously evaluated and benchmarked in a public (and preferably peer-reviewed) study against the procedure currently in use in the existing VRD. Although an exact character-by-character match on the first name, middle name, last name, and date-of-birth fields is easily implemented, it must be regarded as a very weak default baseline (and calling such an algorithm a default baseline is not a recommendation that it should be used—it is only a recognition that it is often used). As discussed in Appendix B, it is somewhat common for two different individuals who have a common name such as "John Smith" to also agree on the full date of birth.

Recommendation L-9: Use commonly used unique identifiers for voter identification when available and when necessary privacy safeguards are in place.

[8] Traditionally, the "*" character refers to a string of arbitrary length and arbitrary content, while the "?" refers to a string of length one (1) and arbitrary content. Thus, the string "SMITH*" matches SMITH, SMITHSON, and SMITHSONIAN, or any other string starting with SMITH. The string "R?B" matches RIB, ROB, RUB, and RCB. Conceptually, using wildcard searches is a generalization of automated resubmission for different name variants?—OB matches both ROB and BOB, as well as COB, DOB, and so on.

[9] See, for example, http://datamining.anu.edu.au/projects/linkage.html; http://www.cdc.gov/cancer/npcr/tools/registryplus/lp_tech_info.htm; http://www.mathcs.emory.edu/Research/Area/datainfo/FRIL/; http://www.cs.umd.edu/projects/linqs/ddupe/; http://www.the-link-king.com/; and http://members.shaw.ca/andre.wajda/linkpro.html.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

From a technical standpoint, the use of a commonly used unique identifier generally enhances the accuracy of matching. Today, the full SSN is the only commonly used unique identifier in the United States. Thus, if technical considerations were the only relevant considerations, the committee would recommend its use, even though today's SSN has a number of technical flaws even as a unique identifier. (These flaws include the lack of a check digit and the scarcity of 9-digit SSNs relative to the population of the United States.[10])

But technical considerations are not the only relevant ones. Because it is linked to so many other kinds of personal records, the use of the SSN for voter ID purposes inevitably raises significant privacy issues, especially when it must be disclosed under public records acts in the name of openness.[11] Similar issues would arise in the United States with any effort to assign a new and unique voter ID identifier to every voter, because of concerns that its use could not be limited to the voter ID application. Thus, the use of a unique identifier for voter identification, which today could only be the SSN, is necessarily conditional on resolving these privacy issues (a task not within the committee's charge and one that has been examined in many other contexts without definitive policy resolution).

It is also relevant that under today's law, only six states are allowed to collect full SSNs for purposes of voter registration; these states were "grandfathered" at the time NVRA was passed because they were already using full SSNs as identifiers. The use of the full SSN nationally for voter registration purposes would require legislative change at the national level, and would quite likely be highly controversial.

Recommendation L-10: Establish standards or best practices for matching algorithms.

Standards or best practices for matching algorithms would have three components.

- *Pre-packaged software implementations of tested and debugged matching algorithms*. A repository of such implementations to which states have free or low-cost access could significantly reduce the financial and logistical burden on individual states to implement such procedures and promote the adoption of these procedures. Broad adoption of such packages would provide greater uniformity in how similarly situated voters in different states are treated.
- *Specifications of acceptable levels of false positives and false negatives and the necessary thresholds for defining matches and nonmatches*. When comparison algorithms return numerical scores rather than a binary result, it is necessary to define threshold values for those scores that determine matches and nonmatches. Best practice usually calls for establishing two thresholds, X and Y (X greater than Y), such that for scores greater than X, a match is indicated; for scores less than Y, a nonmatch is indicated; for scores between X and Y, manual review is indicated.
- *A standardized voter registration data set with known characteristics that can be used to evaluate the performance of specific algorithms and thresholds with respect to rates of false positives and false negatives*. (In concept, a similar data set is the data set associated with the USPS CASS system.[12]) Vendors would then be able to demonstrate in a consistent manner how well their implementations of matching algorithms perform—results of tests involving these implementations could then be compared to the acceptable levels of false positives and false negatives described above.

A number of entities (such as the National Association of State Elections Directors, the EAC, or the National Institute of Standards and Technology) could establish a repository for algorithms, threshold

[10] William E. Winkler, "Should Social Security Numbers Be Replaced by Modern, More Secure Identifiers?," *Proceedings of the National Academy of Sciences USA* 106(27):10877-10878, July 7, 2009, available at http://www.pnas.org/content/106/27/10975.full.

[11] Indeed, in *Greidinger v. Davis* (92-1571), the 4th Circuit Court of Appeals held that a voter has a legitimate privacy interest in preventing disclosure of an SSN to the public when that number is provided for matching and other internal election-related purposes.

[12] See http://www.usps.com/ncsc/addressservices/certprograms/cass.htm.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

values, and standardized data sets—such a repository would support the adoption of best practices and standards for improved matching algorithms.

Recommendation L-11: Use the Social Security Death Master File and STEVE (when deployed) for list maintenance.

In order to purge VRDs of deceased voters, many jurisdictions rely on sources such as newspaper obituaries and information provided by their state departments of vital statistics. Relying only on such data means that these jurisdictions are likely to have a difficult time indentifying voters on their voter registration rolls that die in other jurisdictions (e.g., a New Mexico voter who dies in Texas).

The SSA Death Master File (DMF) is widely regarded as a high-quality database. Use of this database—a national database—would enable jurisdictions to address the dying-out-of-state problem (some jurisdictions, such as Kentucky and occasionally Missouri, already do). Such use is consistent with HAVA, and the committee believes that when a very close match between the VRD record and the DMF record can be accomplished, such a match can be considered sufficient evidence to cancel a voter registration without further investigative action.[13] (Similar points and conclusions apply to a high-quality database from a state department of vital statistics.)

However, nongrandfathered states are not allowed to capture a full SSN in a voter registration record, but rather simply SSN4. Matching is more difficult without using a full SSN, and in such cases, even a full match on the remaining fields (as well as the SSN4) should be taken only as an indicator of a possible death that warrants further investigation to see if the person is really dead. Furthermore, because the DMF is a relatively high-quality database, it is likely that with the use of high-quality matching algorithms, the number of false positives (death wrongly indicated) would be exceptionally small, and thus election officials would not be wasting significant resources on further investigation.

At the time of this writing, the STEVE system for exchanging death information is not widely deployed, and thus does not yet provide such information comprehensively. But when it is widely deployed, it is likely to provide information to election officials with death information in a more timely fashion than does the SSA DM, and election officials should either subscribe to STEVE on their own or work their own state departments of vital statistics to obtain STEVE data. VRD systems will need to be configured to accept data from STEVE, perhaps accumulating them as they arrive and performing list maintenance on a "batch" basis.

Recommendation L-12: Use third-party data when available to resolve possible matches.

As discussed in Appendix C, third-party data such as telephone books or multiple previous addresses where an individual has resided can be used effectively to resolve pairs of records identified as possible matches. For example, two records may have the same name and similar dates of birth (12/01/80 and 01/12/80). Third-party data could be used to determine if these two records refer to the same individual; if, for example, these data indicated that both records shared a number of common addresses for the last 20 years, a higher likelihood of this possible match being a true match would be indicated.

Third-party data are likely most useful in applications where a false positive has high consequences—where individuals would be wrongly disenfranchised. In addition, today most uses of third-party data involve manual processing and review by humans. Automated processes to use third-party

[13] For purposes of this discussion regarding the DMF, a very close match is defined as (1) an exact character-by-character match of all of the key fields—full name, full SSN, full date of birth, and full 5-digit Zip code—both in the VRD and in the DMF; (2) an exact character-by-character match of the full SSN, full date of birth, and full Zip code—both in the VRD and in the DMF, and a close name match. A close name match is one in which the first names match as common equivalents (e.g., Bob versus Robert) and a "text match." A text match requires an exact match for a first name field if the length is 3 or fewer, an n–1 character match for n between 4 and 7 inclusive, and an n–2 character match for n larger than 7; or (3) a close name match, exact SSN match, and an exchanged date of birth (m/d versus d/m) match.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

data would reduce the number of cases necessitating human review and judgment and would improve the overall accuracy, quality, and repeatability of matching.

Recommendation L-13: Develop procedures for handling potential disenfranchisement caused by mistaken removals from voter registration lists.

Any given removal of a name from a voter registration list may have been performed in error. Indeed, a great deal of experience with information technology suggests that even a combination of automated and human matching can sometimes result in inappropriate action because of data errors, inherent ambiguity in the data, algorithm deficiencies, human error, and so on. For example, a felony may have been reduced to a misdemeanor by the court without that fact being made known to election officials. Other sources of error exist as well, and there is an inherent unfairness in changing a voter's status and potentially disenfranchising him or her without providing an opportunity for contesting the removal.

Procedures for addressing disenfranchisement could be handled in a number of different ways. For example, one approach is to provide the person removed from a voter registration list with the opportunity to contest that decision before the removal is made final, though understaffed and/or underfunded election offices might find this approach onerous in light of small staffs, high mailing costs, and other pertinent issues. In addition, notification of voters removed from the list may be upsetting to the families of those individuals suffering from the pain of a relative's death or the person's being declared mentally incompetent. Another approach might be to allow a voter who was inadvertently removed to vote provisionally. Such an approach is mandated by HAVA for federal elections, but it could be adopted for state and local elections as well.[14]

Developing such procedures might well require new legislation and administrative processes.

6.4 IMPROVE PRIVACY, SECURITY, AND BACKUP

Recommendation L-14: Implement basic practices for backing up important data.[15]

Basic backup practices include:

- *Backing up regularly*. Backup of data every night (or at least every night after data are entered into the VRD) is a sensible practice.
- *Keeping backup media for as long as necessary*, based on an explicit risk assessment for determining appropriate data retention periods.
- *Practicing restoration of backups*. File backups are useless if they cannot be restored. Although in principle file backups should be easily usable, experience shows that such is not necessarily the case. Most installations learn a lot from the first time they try to restore a backup, and subsequent restores go much more smoothly. Of course, precisely because problems may occur, attempts to restore a backup should occur only at times when such problems would cause minimal disruption.
- *Storing backups offsite*. Backup media should be stored in a physical location that is some distance away from the main site where the database is used—such a precaution protects against a single catastrophe destroying the database at the main site and the backup media. Offsite storage requires both backing up the data and arranging for an alternative facility. Many commercial facilities and services exist for this purpose.

[14] Of course, provisional ballots will be counted only if, in fact, the caster of the provisional ballot is indeed eligible to vote. Since eligibility is determined by a proper and accurate registration of the voter, the caster of the provisional ballot must be able to challenge what he or she believes to be an improper removal from the VRD. Thus, such individuals must also receive the information they need to understand why they were removed and how they might correct the error(s).

[15] See, for example, ca.com/files/whitepapers/backup_recov_wp.pdf.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

- *Maintaining backup logs*. Operators should know what is backed up, when it was backed up, and where the backups are located.
- *Encrypting backups*. Backup media are a treasure trove of information for miscreants to steal. They are especially vulnerable to loss or theft while in transit and remain vulnerable while in storage. Thus, backups should be encrypted. It is true that encrypted files are often difficult to restore (passwords can be lost, files corrupted, and so on), but a combination of good backup logs and at least occasional practice of file restore procedures generally suffices to make encryption a reasonable safety precaution to take. Decryption keys should also be stored securely, preferably in locations separate from the backup media.
- *Performing full backups when possible, and incremental or differential backups when necessary*. Because full backups are much easier to restore than incremental and/or differential backups, full backups are recommended if it is possible to perform them within the necessary time constraints (usually, an 8-hour night shift). Incremental or differential backups may be necessary if a full backup would take more than 8 hours, and in such cases, making full backups may have to be done on a weekly or a monthly basis. Incremental backups are also more likely to fit onto a single storage unit (e.g., one DVD), and the ability to use a single unit for backup rather than multiple units may make it feasible to fully automate the backup.
- *Cycling backups*. A robust example of scheduling backups might be a schedule in which data are backed up every day separately (e.g., Monday through Sunday). At the end of the week, the Sunday backup is kept as the backup for the week, and the backup media from the other days are recycled or reused. Sunday backups are kept for the month, and then an end-of-month backup is stored. End-of-month backups are kept for the entire year, and the end-of-year backup is kept in perpetuity or until data destruction practices come into play. This type of cycling backup reduces the risk that by the time some data corruption is found (e.g., 3 months after it occurred), a backup prior to the data corruption cannot be located.
- *When possible, investing in real-time backup in the form of data replication or mirroring*.

In addition, an Election Day full backup should be performed in order to have a permanent record of those who were deemed eligible to vote on Election Day. Such a record provides statistics that would not otherwise be available—and such data could be helpful both to election officials and to researchers.

All parties—state or county—that store data for any length of time should take some responsibility for backing up the data in their possession. However, the most essential backup points are located with the systems of record (that is, where the data are posted for all to use). Secondary aggregations such as state-level databases in bottom-up configurations have backup obligations as well, though they may need backup less frequently than local offices making daily changes that need backup every day, since the state-level database can in principle be recreated from the data contained in the local systems.

Recommendation L-15: Implement basic security measures.

Good security policies and procedures start with a commitment to security being an integral part of an organization's operating practice. All too often, organizations give lip service to security but in practice are never willing to pay any price (in either operational or fiscal terms) to improve security. The reality is that cybersecurity expenses must be regarded in the same way as expenses for disaster insurance and door locks. Such purchases entail some degree of expense and inconvenience for organizations and individuals and they are ideally never needed, but they are intended as a hedge against the presence of security threats.

Best practices (described further in Appendix D) for security include:

- Establishing and enforcing access control policies that group people by established roles and assign to these roles the minimal level of access needed to carry out their job functions.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

- Limiting the number of people with administrative privileges that afford the ability to grant access to others.
- Training authorized users of the system in security practices, such as choosing and protecting passwords and resisting "social engineering" attacks. (A "social engineering" attack is one based on duping an authorized VRD user into taking some action that compromises the security of the system.)
- Securing all communications channels used by the system via end-to-end cryptography to protect both the confidentiality and the integrity of the data.
- Limiting connectivity between internal and external networks through the use of mechanisms such as firewalls.
- Deploying mechanisms such as commercially available intrusion detection and antivirus systems to reduce the risk of cyberattacks or insider misuse.
- Minimizing the use of VRD systems for other purposes, and minimizing the amount of non-VRD-related software installed on it.
- Limiting the number of access points to the VRD with access to particularly sensitive information such as complete or last-four digits of Social Security numbers.
- Obtaining independent security review of the VRD system before deployment and periodically thereafter through penetration testing.
- Tracking and logging all changes to VRD data and systems.

Recommendation L-16: Take measures to help ensure system accessibility during critical times.

In some cases, technical fixes can be implemented to enhance system accessibility. For example, a VRD can be designed in such a way that applicant-provided data that cannot be immediately verified are accepted, stored, and flagged as "verification-pending." Such a feature would enable election officials to continue with data entry if the nonelection databases on which they depend are unavailable during periods when the volume of voter registration forms is high.

On the other hand, DOS attacks against Internet-based VRDs are difficult to mitigate—the only known solution with broad applicability is the acquisition of additional bandwidth to "soak up" falsified requests for service. Such a solution is expensive and is likely not to be cost-effective, given relatively few DOS attacks in the elections environment.

Absent such measures, election officials can only make contingency plans for a DOS (e.g., ensuring that copies of a statewide VRD are widely distributed on a computer-readable DVD to polling places on Election Day, saving paper forms for later entry when automated entry is not available). As a general rule, the best contingency plan for electronic outages is the ability to use (temporarily) whatever paper-based procedures were in place before the VRD system was introduced. Such a measure requires clear documentation and a modicum of training for Election Day poll workers and election officials.

In the election environment, a specific measure recommended by the committee is to make the entire VRD available to poll workers. (In some cases, the entire VRD may refer just to a county VRD, and in other cases, to the full state VRD.) The easiest and most secure method for doing so is probably to write the relevant fields of each record in the VRD to a file and then to distribute the file to every precinct.

Distribution could take place over the Internet (but would most likely require a broadband connection, which is not available to every county) or via CD-ROMs or even paper. The former has the advantage of currency—using broadband transmission, file creation could reflect the most recent updates. The latter have the disadvantage of latency—physical media take time to mail, and by the time they arrive at the election offices, they will be a few days out of date. On the other hand, they do not require any special technology to use.

Lastly, other agencies—such as state DMVs and the Social Security Administration—should take steps to ensure the availability of critical election-related databases during times of peak electoral business. The committee calls special attention to the Columbus Day weekend, which occurs in near prox-

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

imity to Election Day. Although the holiday is recognized by the federal government and many state agencies, it is also a period in which election officials process enormous numbers of voter registration applications. Accessibility to DMV and SSA databases during this period is extraordinarily important from an elections management standpoint.

Recommendation L-17: Consider fair information practices as a point of departure for protecting privacy in voter registration databases.

Although fair information practices are often regarded as a reasonable framework for balancing privacy of personal information against the needs of users, judgments about protecting privacy have to be subject to a balancing test against other interests to be served by public policy. A full implementation of FIPs for voter registration databases is likely to conflict with other legal requirements for openness and to interfere with administrative efficiency. For this reason, the committee believes that FIPs should be only a starting point for election officials thinking through their privacy policies. That is, it is the spirit and philosophy underlying the FIPs rather than a literal reading that should guide the efforts of election officials in designing VRD systems.

For example, FIPs afford the individual a high degree of control over the disclosure of his or her personal information. But openness of individual voter registration information also serves valid public purposes (e.g., as a tool for helping to prevent or reduce voter fraud) and facilitating communications between election candidates and voters. One possible way of balancing these interests would be to provide selected individuals—but only those individuals—with opportunities for limited disclosure of information (such as addresses).

Recommendation L-18: Take steps to protect voter privacy when voter registration data are released on a large scale.

Although voter privacy is important even when just one voter's information is at stake, large-scale compromises of personal information can be particularly damaging. Obviously, election officials must do what they can to protect information while it is within their control. But once the information has been released (putatively in accordance with the applicable law), election officials have no effective technical control over how that information will be actually used. To the extent that they can do so, election officials should find a way to bind the recipient—legally—to take the necessary precautions.

Election officials can take some steps to trace how data are used. As discussed in Appendix D, they can seed the data before they are transferred with one or more fake record(s) that can be used to indicate subsequent misuse.

Recommendation L-19: Review appropriate nonelection uses of voter registration data.

States use voter registration data for a number of purposes other than election administration. One of the more common uses is for juror selection—voter registration lists are often one of the sources used to compile lists from which potential jurors are selected. Many states also make such lists available to political parties to facilitate communications with voters in their parties. A number of states regard voter registration lists as public information, and disseminate them to any party willing to pay a nominal fee, though they may place restrictions on the use of such data (e.g., not for commercial purposes).

Nonelection uses of and/or restrictions on voter registration data are sometimes contested by parties wishing to use the data for their own purposes, which may include commercial purposes, nonpartisan educational purposes, and so on. The committee notes that most state and local policy regarding how voter registration data can be used and by whom was developed in a technological environment that did not make it easy to aggregate personal information on a widespread basis, in which commercial

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

use of all forms of individual data was not commonplace, and in which privacy concerns were not as salient in the public eye as they are today.

For this reason, the committee believes it is appropriate for state policy makers to review their policies regarding nonelection use of voter registration data (who should have access, under what circumstances, and for what purposes) with particular attention to whether the users, uses, and restrictions entailed are consistent with the purpose for which voter registration information is requested and collected.

6.5 IMPROVE DATABASE INTEROPERABILITY

Recommendation L-20: Encourage and if possible require state, local, and federal agencies to cooperate with election officials in providing data to support voter registration.

The starting point for achieving database interoperability between a VRD and the databases of other agencies is a willingness and a desire of those other agencies to share data with election officials. But because sharing data with election officials does not generally further the primary mission of those agencies, it is not difficult to imagine that devoting resources to this task would be low on their priority lists.

Broadly speaking, there are two ways to seek the necessary cooperation in the face of agency reluctance—invocation of higher authority, and incentives to cooperate. When the other agencies in question are under state control, the governor's office may have an important role to play in persuading them to provide data to election officials in a timely fashion. When the other agencies in question are under local control, state incentives and/or directives may be necessary to secure cooperation.

The NVRA requires various state agencies to make voter registration forms available to people seeking the services that these agencies provide. The data that service-seeking individuals must provide to the relevant agency in many cases has all of the information needed to perform voter registration for these individuals to vote (and these individuals must usually provide signatures to obtain services). These data are electronically captured and thus could, in principle, be easily available to VRDs.

But the reality is that the NVRA requirement is met by merely delivering completed forms to election officials. Data already captured in electronic form are generally not transmitted to election officials, leaving them with the same data entry task as always. Nothing in the NVRA forbids these agencies to transmit their electronic data to election offices, but doing so would require these agencies to identify individuals who would like to register to vote and then to make their data available to election officials in electronic form. At the very least, these agencies would have to invest in some redesign and reimplementation of some parts of their information technology systems. Raising the priority of these agencies for implementing changes that primarily benefit election officials is likely to require direction from higher authority, such as governors or state legislatures, and the NVRA may itself need to be clarified to allow electronic transfer of registration information or modified to require such transfer.

A second kind of support involves access to useful federal databases, such as the USPS National Change of Address database and the Social Security Death Master File. Although such databases are in principle available to state election officials, access suitable for the needs of these officials is often expensive relative to the financial resources. In some cases, election officials cannot approach the relevant agency directly, but must instead go through a qualified commercial provider.

In general, these barriers to access arise from the fact that state/local election registrars are regarded as customers on a par with other private sector or commercial entities. Providing privileged low-cost access to these databases for election officials to help maintain accurate and complete voter registration rolls would seem to serve a worthy public purpose, and the committee would support providing such access even if it would require legislation to do so.

Recommendation L-21: Use inexpensive data export functions to facilitate data exchange.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

It is commonly believed that direct linkages (real-time electronic interfaces) between systems are essential for effective data exchange. Although direct linkages generally provide the most current and recent data (because they have direct unmediated access to the database in question), they are often expensive to deploy and complex in operation. By contrast, data can also be exchanged using the sending system's ability to "export" its data into a known file format (e.g., an Excel spreadsheet, or a comma-delimited file). Such a file can be written onto a physical medium or sent electronically. This approach may not capture the most recent data, but since most systems support a data-export function, it entails a very low cost of operation.

In addition, file comparisons can usually be performed offline, that is, using applications separate from the core VRD application that read the exported files. Offline applications have the major benefit that they can be developed without rewriting the VRD system itself, and they thus pose little danger to its functionality.

Recommendation L-22: Develop national standards for data-exchange formats for voter registration databases.

Standardized field definitions for database records greatly facilitate the common processing of records derived from different databases, as might be entailed, for example in a search for voter registration records for the same person in two different databases. For example, one database may record dates in a yyyy-mm-dd format and the other in a dd-mm-yyyy format. Or, one database may include a suffix such as Jr. or Sr. as part of the last-name field, and another might include a separate field for suffixes.

One approach to implementing standardized field definitions is for every database to adopt the same conventions for recording data. Thus, any export of that data outside the system's boundaries would automatically be rendered consistent with any other database. On the other hand, requiring all systems of record to convert to the same standard field definition necessarily entails operational costs (e.g., disruption to current ongoing operations, costs of reprogramming internal logic of the applications using the databases, and so on) and is thus impractical.

A second approach, and one recommended by the committee, focuses on standards only for data interchange—that is, standardized field definitions only for data that are intended to be used by another database system. With such an approach, the internal logic of applications can remain unchanged (because the data are stored in the same format as before). But the data are converted into this standard form when data are prepared for export.

Standards for name, date, and SSN representation are, in principle, not difficult to implement. The last-name field should include name suffixes, such as Jr. or III.

The implementation of this recommendation would do much to facilitate the matching of records within different VRDs. But the committee also notes that the data of interest for matching VRD records—name, SSN, date of birth—are sufficiently standardized in their definitions that with the use of string comparators, these data fields can be used for matching purposes.

Copyright National Academy of Sciences. All rights reserved.

7

Conclusion

In the years since HAVA mandated the nationwide adoption of statewide VRDs, the states have been largely successful in deploying their initial VRD implementations. Nevertheless, there are a number of immediate opportunities for states to improve the operation of their VRD systems. In addition, if the promise of statewide VRDs for improving voter registration is to be realized, the states will have to address some longer-term issues. These issues can be successfully addressed only with coordinated, concerted, and sustained support for continuing improvement on the part of many parties, including state election officials, nonelection state and local agencies, state legislatures, voter advocacy groups, and the federal government.

Copyright National Academy of Sciences. All rights reserved.

Copyright National Academy of Sciences. All rights reserved.

Appendixes

Copyright National Academy of Sciences. All rights reserved.

Copyright National Academy of Sciences. All rights reserved.

A

Background and Context

THE ROLE AND STRUCTURE OF VOTER REGISTRATION

Voter registration (described briefly in Box A.1) plays a central role in elections in most states. Today, in all states except one (North Dakota), [1] a voter must be registered for his or her vote to count in an election; some states allow same-day voter registration on Election Day.

During reforms of the Progressive era, voter registration procedures spread throughout the states, beginning in urban areas, launched at least in part in an attempt to reform how elections were carried out. These reforms aimed to restore fairness in the conduct of elections by, for example, minimizing the influence of urban political machines over elections. However, many believe that these procedures also caused voter turnout to decline sharply. The use of strict registration rules to verify the eligibility of a voter, such as requiring in-person registration during limited weekday hours, effectively limited the participation of many eligible voters who could not afford to take time off work to register to vote.[2] These rules were eventually eased by a series of federal mandates.

The U.S. Constitution (Article I, Section 4 and Article II, Section I) gives states the power to make rules governing federal elections, subject to the authority of Congress to make or alter such rules.[3] As a result, elections management in the United States is largely a mosaic involving many individual election decision makers and administrators in multiple jurisdictions in different states. Thus, the procedures

[1] North Dakota does not formally require voter registration as a condition of voting and was exempted from certain provisions of HAVA. For more background information, see www.nd.gov/sos/forms/pdf/votereg.pdf. On the other hand, North Dakota maintains a "central voter file," which contains most of the information that the VRD systems of other states contain, including the voter's complete legal name, complete residential address, complete mailing address, a unique identifier for the individual generated and assigned by the state, and the voting history for the last 4 years. North Dakota's central voter file is used for purposes of "preventing and determining voter fraud, making changes and updates, and generating information, including pollbooks, reports, inquiries, forms, and voter lists." (Chapter 16.1-02, North Dakota Code, available at http://www.legis.nd.gov/cencode/t161c02.pdf.) Thus, many of the issues described in this report regarding VRDs are also likely to be found in North Dakota.

[2] Alexander Keyssar, *The Right to Vote: The Contested History of Democracy in the United States*, Basic Books, New York, 2000.

[3] More precisely, Article I, section 4 of the Constitution gives the Congress plenary power over congressional elections (which the states may exercise in the absence of congressional legislation) while Article II gives state Legislatures greater power in the rules for presidential elections. In practice, most of Congress's authority exercised in legislation such as the Help America Vote Act and the National Voter Registration Act comes from its Article I, section 4 power over congressional elections.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Box A.1
A Thumbnail Description of Voter Registration

States generally require that a voter be a U.S. citizen, at least 18 years of age, and a resident (in some cases, a resident for some minimum period of time, such as 30 days). Some states also limit voter eligibility on the basis of criminal status (for example, incarcerated felons may not be permitted to vote), and some on the basis of mental competency, although the specifics of these limitations vary.[1]

As a general rule, a voter registers to vote in a specific geographic jurisdiction that is determined from the residential address that he or she provides for the purpose of voting. Citizens can register to vote at election offices. Depending on the state, citizens can also obtain voter registration materials in many places, including military facilities, assisted living facilities, high schools, vocational schools, social service agencies, nursing homes, and libraries, or through voter registration drives, or by downloading materials from the Internet. In addition, the National Voter Registration Act requires all states to provide such materials at their departments of motor vehicles, departments of human services, and public assistance agencies. By filling out the required forms and providing the necessary identification, citizens in all states can also register to vote by mail. In at least three states (Washington, Kansas, and Arizona), a citizen can register to vote through the Internet if he or she already has a driver's license or a state-issued ID from that state.

The voter completes the registration form and it is returned to the election office. The returned materials are accompanied by an original signature that serves as an authentication mechanism when voter registration must be checked in the future. If the voter registers at a department of motor vehicles, the relevant information may be extracted from the information on file or provided at the department of motor vehicles (DMV) and transmitted electronically to the election office, along with the signature on file with the DMV as an authentication device for the voter at the polls. Overseas voters, and members of the U.S. armed forces and their dependents, can sometimes register to vote by fax.

The voting address of record determines the precinct from which the voter may cast his or her ballot, whether at the polling place, or by absentee or mail ballot, or by an early vote. A precinct is a subdivision of a local election jurisdiction, and all voters in a given precinct vote at one polling place. (Sometimes, a number of small precincts are consolidated at one polling place, and sometimes election officials can require that all voters from certain precincts vote by mail.) A local election jurisdiction is an administrative entity responsible for the conduct and administration of elections within it, and may be a county or a municipality (a city or town).

SOURCE: Adapted largely from National Research Council, *Asking the Right Questions About Electronic Voting*, Richard Celeste, Dick Thornburgh, and Herbert Lin (eds.), The National Academies Press, Washington, D.C., 2005.

[1] A description of the legal restrictions on felons and voting rights in a large number of states can be found in American Civil Liberties Union, *Purged! How Flawed and Inconsistent Voting Systems Could Deprive Millions of Americans of the Right to Vote*, October 2004, available at http://www.aclu.org/VotingRights/VotingRights.cfm?ID=16845&c=167.

and regulations governing the electoral process for voters are virtually guaranteed to be different from state to state.

Nonetheless, federal supremacy does put some constraints on elections as the states administer them. For example, amendments to the Constitution prohibit racial or gender discrimination in the right to vote, prohibit poll taxes for federal elections,[4] and grant individuals the right to vote at age 18.

[4] Poll taxes for state elections are unconstitutional under the equal protection clause of the 14th Amendment pursuant to the Supreme Court's 1966 decision in *Harper v. Va. Bd. of Elections.*

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

The one-person, one-vote principle emerges mainly from Supreme Court interpretations of the equal protection clause of the 14th Amendment, and subsequent legislation.

In addition, starting with the 1960s civil rights legislation, Congress gradually expanded federal oversight of election administration and registration provisions, although states continue to have considerable discretion in how to implement federal requirements. The Voting Rights Act of 1965 aims to broadly protect voter rights by prohibiting discriminatory voting practices and by preventing an individual from being denied the right to vote "because of an error or omission on any record or paper relating to any application, registration, or other act requisite to voting, if such error or omission is not material in determining whether such individual is qualified under the State law to vote in such election." Subsequent legislation aimed at facilitating voter registration and increasing the accessibility of absentee ballots for particular classes of voters includes the Voting Accessibility for the Elderly and Handicapped Act of 1984 and the Uniformed and Overseas Citizen Absentee Voting Act of 1986.

The National Voter Registration Act of 1993 (NVRA) added two requirements to voter registration. The first was to increase voter registrations by requiring applications to be made available at a number of physical locations—motor vehicle agencies, all offices that provide public assistance or services to persons with disabilities, other places that states could designate (for example, public libraries), and nongovernmental offices that agree to serve as voter registration sites—and by mail. The second focused on the maintenance of voter lists by establishing rules under which names could be removed from the voter registration list. It also mandated that states monitor and report on their implementation of the NVRA. Figure A.1 illustrates the various list maintenance options under the NVRA.

Following the passage of the NVRA, a variety of proposals were made to further enhance voter registration by the creation of centralized statewide voter registration databases. Following the Florida recount in the 2000 presidential election, the Help America Vote Act (HAVA) was passed in 2002 to undertake a number of electoral reforms.

HAVA aimed to improve election administration by allocating funds to upgrade and certify voting systems and by creating the U.S. Election Assistance Commission (EAC) to provide voluntary guidance to states. Another goal of HAVA was to establish more uniformity within individual states and to empower the states to take a stronger role vis-à-vis local election officials. Finally, HAVA included several provisions related to voter registration databases. It required states to shift to centralized voter registration lists at the state level and away from the estimated 3,000, mostly locally administered, voter registration lists. It requires that each state's database contain the name and registration information of each legally registered voter in the state and that each legally registered voter be assigned a unique identifier. HAVA specifies that the state list is the official voter registration list for federal elections. It also requires election officials to perform regular maintenance regarding the accuracy and completeness of the registration lists.[5]

THE POLITICAL LANDSCAPE OF VOTER REGISTRATION

The tensions that gave rise to laws related to voter registration persist today. In an ideal world, voter registration lists would include all those individuals eligible to vote and none of the individuals not eligible to vote. In addition, all of the data in the database would be factually correct. For purposes of this report, the term "accuracy" refers to the factual correctness of the data that exist in the database and also the notion that the database contains none of the individuals not eligible to vote. Completeness refers to the presence in the database of all individuals who *should* be in the database. If the database is perfect, it is both 100 percent accurate and 100 percent complete—that is, all of the data in the database are correct (and thus the database contains no individual who should not be in the database), *and* the database includes all of the individuals who should be in the database. Notice that in this formulation,

[5] HAVA uses the term "accuracy" to mean a list both from which ineligible individuals have been eliminated and for which safeguards have been established to ensure that eligible individuals have not been improperly eliminated.

Copyright National Academy of Sciences. All rights reserved.

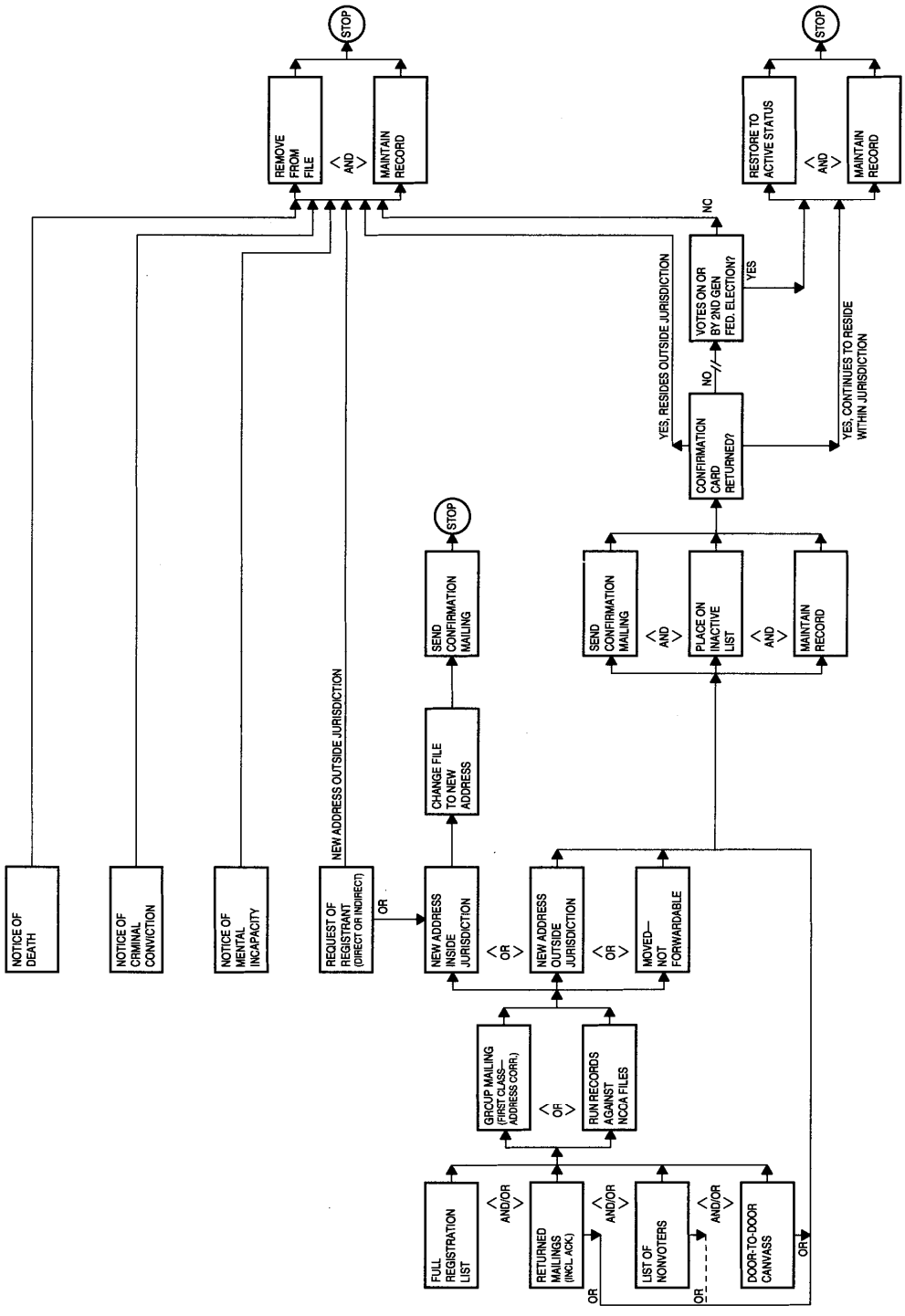


FIGURE A.1 Voter registration list maintenance options under the National Voter Registration Act. SOURCE: National Clearinghouse on Election Administration, "Implementing the National Voter Registration Act of 1993: Requirements, Issues, Approaches, and Examples," Federal Election Commission, Washington, D.C., January 1, 1994, p. 5-1.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

accuracy does *not* subsume completeness, so that a database must be characterized with respect to *both* attributes.

It is often true in practice that efforts to maximize in the voter registration database (VRD) the number of individuals eligible to vote conflict with efforts to minimize the number of individuals in the VRD who are *not* eligible to vote. One view of this tension emphasizes the risks of voter fraud and highlights the need to maintain the integrity of the voting list by placing the greatest effort on minimizing the number of individuals in the VRD who are *not* eligible to vote. This side argues that if election fraud were to occur, it could undermine public confidence in an election.

A different view of these tensions emphasizes the importance of inclusivity in a representative democracy. Individuals with these concerns believe that the number of eligible but unregistered voters could be decreased through better access to and easier voter registration procedures. This side contends that confidence in the election process could be lost if methods and procedures used to improve the accuracy of voter registration lists cause eligible voters to be removed erroneously, and that overly strict or onerous procedures could suppress registering and/or voting. Additionally, there is concern that the barrier of registration might skew a representative government toward certain interests because the political views and values of those who do not vote as a result of registration issues may differ from those of individuals who do vote. Completeness serves the end of inclusivity by ensuring that all eligible individuals who have sought to register to vote are not erroneously deleted from the VRD.

These two views are commonly identified with specific political parties. Another set of concerns about voter registration, generally not associated with one party or another, stems from the fact that exercising the right to vote in the United States requires the active participation of the voter to register—and some individuals in policy-making or operational positions have been known to be dismissive of efforts to ease the voter registration process or to reduce voter effort in maintaining registration by saying, in effect, "If the person isn't willing to do X, then he or she shouldn't be voting anyway."

Ultimately, voter registration lists cannot be perfect with respect to either completeness or accuracy, in part because the voting population changes by the day and even by the hour. But today's political environment raises the stakes significantly for even small deviations from perfection in either direction. Today's political campaigns and debates are rancorous and bitter. In addition, many elections today are close—a reflection of an electorate that has been about evenly divided—and close elections are breeding grounds for postelection suspicion, on the theory that even a small amount of fraud or accident or mishap or improperly followed procedure might have tipped the election the other way. While the presidential election of 2000 is perhaps the most salient example, outcomes in other close races have been very closely scrutinized by supporters of the losing side for irregularities in all aspects of the voting process, including voter registration.

These tensions and political sensitivities point to the need for voter registration procedures and practices that are transparent, consistent, and robust, and for the use of approaches that balance the inherent tensions. This report does not aim to resolve these tensions, but they must be kept in mind as technical, policy, and procedural challenges of implementing and maintaining statewide voter registration databases are considered.

OTHER USES OF VOTER REGISTRATION LISTS

Voter registration lists are used for a number of purposes other than establishing the eligibility of an individual to vote in an election. For example, voter registration lists are used by candidates and political parties to reach out to potential voters by phone and by mail. At the local level, they are used to estimate the financial, personnel, and logistical requirements for elections. They are used to track absentee ballots and voter histories. They are used in some jurisdictions to establish signature and vote thresholds for referenda and petitions. They are used, at least in part, to establish jury pools. All of these uses require voter registration information to be as accurate and complete as possible.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Some of these applications have led to privacy concerns, and although most voter registration data are generally public information, there are sometimes restrictions on making such information broadly available. For example, some states restrict the sale or use of voter registration lists for commercial solicitation purposes. Concerns have also been raised about the safety of battered men or women if the contact information contained in their voter registration were to be disclosed publicly, and some jurisdictions have enacted special protections in this instance.

THE BASIC REQUIREMENT FOR STATEWIDE VOTER REGISTRATION DATABASES

HAVA Section 303 requires each state to establish and maintain a "single, uniform, official, centralized, interactive computerized statewide voter registration list" that contains the voter registration information for all eligible voters in the state and requires that the VRD be electronically accessible by any election official in the state. But although HAVA provides some criteria for developing and maintaining this database, and the Election Assistance Commission has issued its 2005 *Voluntary Guidance on Implementation of Statewide Voter Registration Lists*,[6] the states still maintain a degree of discretion in how to conform to HAVA. Such discretion, exercised in different ways by different states, inevitably leads to various problems and inconsistencies within and between statewide voter registration databases.

States have taken different architectural approaches to building systems to meet the centralized voter registration list requirement. Under the so-called top-down approach followed by many states, state election officials maintain a single, unified database and local election officials provide the state with the information needed to update the database. Some states instead opted for a bottom-up approach, in which local jurisdictions continue to maintain their own registration lists but also provide periodic updates to a separate statewide system. Other states have adopted a hybrid architecture that combines elements of both the top-down and the bottom-up approach. Kentucky and Michigan had already implemented statewide voter registration databases before the enactment of HAVA, but most states have had to implement new systems to comply with HAVA.

Does HAVA mandate a particular architectural approach to the implementation of VRDs? This issue has been argued both in the affirmative and in the negative at length, and the committee takes no position on this question. HAVA does require that the control of the VRD be maintained at the statewide level. However, the nature of the decision making used within a state to determine eligibility for inclusion in the voter registration list is at least as important as the particular technical architecture used. As a result, any assessment of whether a system conforms to the requirements and expectations of HAVA should consider the locus of decision making regarding an individual's eligibility to vote.

It should also be noted that although guidance regarding database structures or system attributes has been promulgated through the EAC's *Voluntary Guidance on Implementation of Statewide Voter Registration Lists*, the guidance remains voluntary, and the agency charged with enforcing HAVA—the Department of Justice—has not issued guidelines or regulations of its own. Thus, state election officials may proceed at their own risk that some design decision might be challenged later as not being HAVA-compliant.

[6] Available at www.eac.gov/election/docs/statewide_registration_guidelines_072605.pdf/attachment_download/file.

Copyright National Academy of Sciences. All rights reserved.

B

Matching Records Across Databases

As noted in Appendix A, HAVA and the NVRA direct the states to implement a variety of procedures that require the "coordination" of voter registration databases (VRDs) with other databases. The central technical issue in such coordination (known in this appendix as "matching" or, more precisely, record-level matching) is finding individuals who are represented in both the VRD and another database (or the reverse—finding an individual who is represented in only one of these databases). (In the case of removing duplicate registrations, the "coordination" occurs within the same database.)

THE BASIC PROCESS OF MATCHING RECORDS ACROSS DATABASES WITHOUT UNIQUE IDENTIFIERS[1]

The basic element of a VRD is a record with data contained within specific fields associated with an individual—first name, last name, street address, date of birth, and so on. Databases may differ in the number of fields that a given record contains (for example, one database may include a field for telephone number and another might not) or in definitions of the fields (for example, one database may have one field for street name and number together (123 Main Street), and another may have separate fields for street name (Main Street) and street number (123)).

Matching records across databases (that is, record-level matching) involves the comparison of corresponding fields between databases. HAVA requires states to check the information provided on a new voter registration application against the databases of the state's motor vehicle agency if the applicant provides a driver's license number. An applicant must provide a driver's license number if one is available, and the election officials must verify the applicant's information with the state department of motor vehicles. If the applicant does not have a driver's license, he or she must provide the last four digits of his or her Social Security number (SSN4), in which case the applicant's information is verified with the Social Security Administration (SSA). (In practice, many DMVs handle the request. The election officials submit the verification query to the DMV, which may involve a driver's license number or an SSN4. If the query involves SSN4, the DMV passes the request to the SSA using the AAMVAnet, a private network

[1] For an overall background document that covers many elementary aspects of matching records (that is, record linkage), see William E. Winkler, "Matching and Record Linkage," pp. 355-384 in *Business Survey Methods*, Brenda G. Cox et al. (eds.), Wiley, New York, 1995.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

established by the American Association of Motor Vehicle Administrators, and the response from SSA is passed back to the DMV through the same network.) Individual states also have the authority to—and often do—use additional databases and criteria to verify voter registration information.[2]

The matching process is greatly simplified if each individual has used the same unique identifier (such as the driver's license number or the full Social Security number) in each database.[3] In this case, matching records across databases is simplified. However, in the absence of a unique identifier, it is necessary to use combinations of fields in order to match records. Matches based on the comparison of corresponding fields such as first name, last name, address, and date of birth are inherently inferential, and thus subject to higher rates of error. (Some combinations, such as first name, last name, date of birth, and last four digits of the Social Security number, have a high likelihood of uniquely identifying an individual.[4])

Errors in record-level matching may be false positives (a match is indicated when in fact the two records refer to different individuals) or false negatives (a nonmatch is indicated when the two records refer to the same individual). What is an acceptable upper limit on a given type of error depends on the application in question. For example, if the voter registration database is being checked against a database of felons or dead people, a low rate of false positives is needed to reduce the likelihood that eligible voters are removed from the VRD. Just how low a rate is acceptable is a policy choice.

In this report, the term "field-level match" denotes the process of comparing individual fields, so that the "first name" field of a record in Database 1 is compared to the "first name" field of a record in Database 2. In addition, a field-level match can be indicated on the basis of different match rules, which might include:

- Exact match—the fields are exactly equal, character by character for every character.
- Fuzzy or approximate match, which is intended to deal with typographical variation. At its simplest level, it allows comparison of fields with very simple errors ("Smith" versus "Smoth"). Fuzzy matching methods can be developed intuitively as seems to be the case in many VRD applications or based on principles that computer scientists have shown to work consistently well in practice.
- Content equivalence—"Road" and "Rd," or "Bill" and "William" are treated as equal.

The need for such rules arises for many reasons, not the least of which is that when asked for information, people often provide inconsistent information unintentionally. They use nicknames, include or omit middle initials, use abbreviations or not, and so on—and forget what they have done on previous occasions. An area code for a phone number may have changed. A street address might be recorded with digits transposed in the house number, or a street name spelled incorrectly, or with the wrong Zip code.

A record-level match occurs when several field-level matches are indicated. The decision about how many field-level matches are needed to define a record-level match is an important influence on the accuracy of the match. For example, a record-level match rule that required only field-level matches on first name and last name would lead to many more false positives than a rule requiring field-level matches on first name, last name, and date of birth. If the former rule were used instead of the latter to remove

[2] See Election Assistance Commission, *Impact of the National Voter Registration Act on Federal Elections 2005-2006*, Table 12, "Verification of Applications," p. 72, available at http://www.eac.gov/clearinghouse/docs/the-impact-of-the-national-voter-registration-act-on-federal-elections-2005-2006/attachment_download/file.

[3] In fact, even the full SSN is flawed as a unique identifier, as the SSA has been known from time to time to issue the same SSN to different individuals. Identity theft in which an individual appropriates someone else's SSN has also happened. Lastly, because the SSN lacks a check digit and is most often entered manually (rather than swiped as credit cards are), typographical errors often occur with no way of catching them at the point of entry.

[4] One way to estimate how many combinations exist is to consider that the population of the United States is currently approximately 300 million. The number of possible four-digit SSNs is 10,000. A plausible estimate of the number of distinct birth dates (month, day, year) is perhaps 365 x 70 = ~25,000. Thus, there are around 250 million possible combinations of birth date and four-digit SSN, which corresponds approximately to about one such combination for every American.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

voters from registration lists (for example, if the voter registration list were compared against a list of state felons), many more eligible voters would be improperly removed.[5] (In principle and sometimes in practice, matching algorithms can also consider differences as well as similarities. For example, if the name and date of birth are the same but the Social Security number and gender values are inconsistent between the records, a nonmatch might be indicated under some circumstances.)

States have considerable discretion to decide for themselves the criteria to be used for matching, although these criteria cannot be used to disenfranchise legitimate voters.[6] Some states will use fuzzy matching and others exact matching for checking any given data field. States also vary in the fields that they check—for example, some will compare addresses and others will not. In general, some election offices may be using match criteria without sufficient consideration of possible false-positive and false-negative error rates associated with different variants of the methods.

The details of matching algorithms and the parameters used to drive them may have a substantial impact on the output of any matching process. For this reason, what appears to be a technical decision can have enormous policy significance. Box B.1 illustrates the levels of detail with which match criteria must be specified.

Finally, a manual review of matches is sometimes performed. That is, under some circumstances, a voter registrar will review a match (or a nonmatch) indicated by automated processes.

COMPLICATIONS IN MATCHING

Apart from the issues involved in the matching criteria, a variety of data issues also complicate matching. Data quality (addressed in more detail in Appendix C) is impaired by many different sources of error, including illegible handwriting, incomplete or lost forms, and keypunching errors.

Another problem occurs because certain names are quite common. For example, it is known that the name "John Smith" occurs between 30,000 and 60,000 times in national lists. This means that there are between 1.5 and 3.0 John Smith's for each date of birth. Assuming there are 500 individuals named John Smith in a given state, then a certain (low) proportion of them will have the same date of birth. With certain other commonly occurring names, some chance agreements on dates of birth would be expected as well.[7]

This point suggests that more accurate record-level matching will take into account the possibility of chance agreement on date of birth for certain commonly occurring combinations of first and last name, which will in turn require knowledge of the most common names in any given state. Such information can easily be computed from either state-held databases (such as the department of motor vehicles (DMV) or voter registration databases, whichever is of higher quality as indicated by fewer

[5] An example of such a problem was a case with a record-level match conducted to identify felons in the voter registration database in Florida before the 2000 election. In matching the Florida VRD to a national list of felons, the applicable rule used exact field-level matches on the first four letters of the first name, middle initial, gender, and last four digits of the Social Security number (when available) and used approximate matches for last name (matching on 80 percent of the letters in the last name) and date of birth. Certain name variations were also explicitly taken into account (Willie could match William; John Richard could match Richard John). The result of this match was that approximately 15 percent of the names removed from the VRD were improperly removed. See Gregory Palast, "The Wrong Way to Fix the Vote," *The Washington Post*, Sunday, June 10, 2001, Outlook section, p. 1, available at http://www.legitgov.org/palast_wrong_way_fix_vote.html. To remediate the issues raised in this case, Choicepoint—the firm responsible for conducting the match—agreed to a very detailed set of criteria described in Box B.1.

[6] A description of the various practices employed by the various states in late 2005 can be found in Justin Levitt, Wendy R. Weiser, and Ana Muñoz, *Making the List: Database Matching and Verification Processes for Voter Registration*, Brennan Center, New York University, 2006, available at http://www.brennancenter.org/dynamic/subpages/download_file_49479.pdf.

[7] See, for example, Michael P. McDonald, "The True Electorate: A Cross-Validation of Voter File and Election Poll Demographics," *Public Opinion Quarterly* 71(4):588-602, 2007; Michael P. McDonald and Justin Levitt, "Seeing Double Voting: An Extension of the Birthday Problem," *Election Law Journal* 7(2):111-122, 2008.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Box B.1
The Detailed Nature of Match Criteria—An Illustration

As an illustration of the detail with which match criteria must be specified, consider the following criteria taken from the consent decree in *National Association for the Advancement of Colored People v. Katherine Harris, Secretary of State of Florida et al.* (Case No. 01-120-CIV-Gold/Simonton, United States District Court, for the Southern District of Florida).

> Notice of Filing Fully Executed Copy of June 28, 2002, Choicepoint Settlement Agreement . . .
>
> 9. The matching criteria described in Paragraph A.8 . . . [are] as follows:
>
> ChoicePoint will identify all matches on the comprehensive list resulting from the processing described in Paragraphs A.2-A.7 that do not match based on all of the following data fields:
>
> - Validated 9 digit Social Security Number
> - Non-normalized (i.e., as name appears in original source data) Last Name
> - Non-normalized (i.e., as name appears in original source data) First Name
> - Non-normalized (i.e., as name appears in original source data) Middle Name
> - Suffix
> - Race
> - Gender
> - Date of Birth
>
> ChoicePoint will perform Social Security Number validation in accordance with guidelines established by the Social Security Administration.
>
> Records will be deemed to match under the criteria listed above if a middle name in one record begins with the same letter as a middle initial shown in the match record assuming all other fields listed above match.
>
> Records will be deemed not to match under the criteria listed above if they share common blank data fields among the fields listed above, except for cases in which the middle name field or suffix field is blank in both records.
>
> Records will be deemed not to match under the criteria listed above if one of the fields being compared contains data and the same field in the match record contains no data.

typographical errors, more current entries, and so on) or commercially available databases (such as credit header records[8]).

Matches involving common names may require additional processing (perhaps manual) and involve the use of additional information not contained in databases. For instance, a prior address may confirm a match on a name when date of birth is missing. An e-mail address, phone number, or other corroborating information may confirm a match when there is a typographical error in any of the first name, last name, or date of birth.

At the same time, using other fields may entail other complications. For example, addresses may be represented differently in different databases; for example, in one database, "123 Main Street" represents an address, whereas in another database, addresses are represented in three fields (house number ("123"), street name ("Main"), and suffix ("Street")). Address standardization is often required to fix this problem.

[8] Credit headers refer to information in the credit report such as name, address, and phone number, not the credit history portion of the report.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Finally, the above technically oriented comments *presume* that the databases to be matched against the VRD are in fact available. But in the real world of state voter registration databases, fragmented state control over state social service agencies and departments of motor vehicles, and state/county tensions regarding authority over voter registration, the politics of database availability are at least as challenging as the technology for matching. Achieving integration or interoperability of the information systems of election officials and of other state and/or local agencies may be deeply problematic if strong political leadership is not available to demand cooperation. Database-providing agencies not under the authority of state election officials (whether state or county) may well give low priority to meeting the election needs of the state, resulting in difficulties for state election officials in gaining access without undue delay or difficulty. For example, a database-providing agency may demand that election officials provide a voter registration list in a particular format that is hard or time-consuming to generate before the agency is willing to perform a match between the two databases. A more serious problem occurs when the database-providing agency is made responsible for matching the voter registration data against its own data—the agency may be unable to devote serious resources to doing so, or lack the inclination or skills to do the matching properly. An agency may be unmotivated to resolve or address possible interoperability problems.

THE POSSIBLE IMPACT OF INADEQUATE RECORD-LEVEL MATCHING

According to the EAC report *Impact of the National Voter Registration Act on Federal Elections 2005-2006*,[9] there were 36,277,749 voter applications received by 45 reporting states. Among those received, there were 10,938,385 changes of address or party; 2,196,608 duplicate applications; and 1,138,955 invalid or rejected applications—resulting in a total of 17,281,234 new registrants.[10] The percentage of applications not entered into the database because they were "invalid or rejected" or "duplicate applications" was about 9 percent, a total of 3,335,563 in the 45 reporting states. For comparison purposes, Table 4b from page 50 of the EAC report indicates that 333,663 people from 34 reporting states were removed from voter registration lists due to presumed felony convictions.

Once it is known that an application is not a duplicate, and not just a change of address or party, the application needs to be checked against the relevant databases. Table 12, "Verification of Applications," on page 72 in the EAC report[11] shows that each state has its own unique set of criteria for verifying the applications, ranging from states like Pennsylvania, which verifies only through the DMV and the SSA, to Montana, which verifies against the DMV, the SSA, Vital Records, "Match Against Voter Registration Databases," "Tracking Returned Voter ID Cards," "Tracking Returned Disposition Notices," and "Verify Through Other Agency." According to Table 13, "Data Fields for Comparison to Identify Duplications," in the EAC report, 15 states verify using the address; 48 verify the date of birth; 38 verify the driver's license number; 46 verify the names provided by the registrant; 40 verify "Social Security number" (although surely that is just the last four digits in most cases, since according to Table 11, pages 68-69, in the EAC report, only 7 states use the full SSN); and 10 verify "other" data.

Consider two points. First, the state with the highest rate of "invalid or rejected" applications (Table 3, p. 38, in the EAC report) also reported in this survey that it verifies application information only through the DMV and the SSA (Table 12). Second, the state reporting in this survey the highest percentage of applications rejected because they were duplicates also reports in this survey that it uses

[9] Available at http://www.eac.gov/clearinghouse/docs/the-impact-of-the-national-voter-registration-act-on-federal-elections-2005-2006/attachment_download/file.

[10] The EAC report also notes that it "may also have under-reported various voter registration activities because several States were in the middle of converting their local voter registration files into a statewide system in 2005. As a result, some States indicated that their local jurisdictions stopped keeping track of various registration functions and activities because they understood the State would be compiling this information" (p. 10).

[11] In this and the next paragraph, the tables (and page numbers) referred to are in the EAC report *Impact of the National Voter Registration Act on Federal Elections 2005-2006*, available at http://www.eac.gov/clearinghouse/docs/the-impact-of-the-national-voter-registration-act-on-federal-elections-2005-2006/attachment_download/file.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

only date of birth and names provided by the applicant to identify duplications (Table 13 in the EAC report). These points do not prove a causal relationship between use of a small number of non-VRD databases or a small number of fields in verification and a high percentage of rejected applications, but presuming that the data reported are valid and accurately reported, these points raise the question of how a broader set of criteria would have changed the percentage of applications rejected.[12]

AN IMPORTANT EXAMPLE OF MATCHING IN PRACTICE

To illustrate the issues described above, consider a record-level match based on exact matches for an individual's first and last name, the month and year of birth, and the last four digits of an SSN. This example is significant because HAVA requires the Social Security Administration to check the name, date of birth, and the last four digits of the SSN ("the applicable information") in support of the federal voting process (usually to verify information for first-time voter applicants who do not provide a driver's license number to be checked against state DMV records), and to notify the voter registrar if the person so identified is deceased. (This requirement does not mean that the SSA mechanism is the only means through which voter information can be verified—states with other mechanisms available to them can select another method. According to the Brennan Center, 24 states in late 2005 planned to use the process described above.[13])

The requirement of using only the last four digits of the SSN increases the number of false positives, even though the absolute number of false positives is still quite low. The limitation to the use of the last four digits of the SSN reflects a balancing between a more effective matching of records and concerns about privacy.

Upon receipt of the applicable information, the SSA queries its database and returns one of five responses: no match found; one unique match, death indicator absent; one unique match, death indicator present; multiple matches found with at least one lacking a death indicator; or multiple matches found but all with death indicator. As noted above, the query is based on searching for exact matches on the applicable information. At its November 2007 workshop, the committee heard testimony that this particular strategy for matching was developed by the SSA through the efforts of a working group involving the National Association of Secretaries of State, National Association of State Election Directors, American Association of Motor Vehicle Administrators, and five states. However, to the best of the committee's knowledge, no testing of match criteria was conducted in advance of deployment, and the error rates that such a strategy would entail were unknown at the time of deployment.

This strategy has a number of limitations that would prevent records from being matched when they should be matched. For example, the search query does not account for content equivalence of names (so that Bill and William are regarded as completely different names). Using only the first and last name causes difficulty, because the number of multiple and compound names is increasing rapidly in the population. In addition, a full legal name was not originally required to obtain an SSN, and thus many SSA records do not contain the full legal names of individuals. Changes in last name (for example, of women who change their last names through marriage) are also problematic, as someone may not report a change of last name to the SSA until it is needed to determine Social Security benefits. In addition, individuals were not required until 1972 to provide SSA proof of identity when applying for an SSN. Finally, individuals may still have been assigned SSNs even if their applications did not contain birth date information.

[12] The committee recognizes that the issue of data validity is an important one. For example, states may have reported their figures using definitions or criteria that were not uniform across all reporting jurisdictions. Issues with terminology are also known to cause difficulties for survey design. Until such matters are resolved, these data can only be regarded as providing tentative indications of possible relationships.

[13] Justin Levitt, Wendy R. Weiser, and Ana Muñoz, *Making the List: Database Matching and Verification Processes for Voter Registration*, Brennan Center, New York University, 2006, available at http://www.brennancenter.org/dynamic/subpages/download_file_49479.pdf.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Data provided by the SSA to the committee's second workshop in November 2007 indicate that 55 percent of queries result in at least one match being indicated; queries using the full SSN result in a match rate of about 88 percent. The cost per query is at less than one cent ($0.0062), which is low enough to allow election officials to vary the queries themselves in the event that a nonmatch response is received (for example, querying on "Bill" if "William" did not return a match).

As an example of a matching procedure in action, consider the elements of a new voter registration application card as shown on the left below and the SSA record on the right (presume these records are, in fact, supposed to refer to the same person):

New Registration Card	*SSA Record*
Tom T Bowden	Taylor T Bowden
3121 Escondido Way	
11/04/77	11/04/77
SSN 000001087	SSN 000001087

In this case, the SSA would return a response of "no match found." However, if the voter registrar could determine that either Tom has a middle name of Taylor or Taylor has a middle name of Tom or Thomas, then this registrar could associate these records with some degree of confidence if he or she concluded that the first and middle names have been transposed. But in the absence of other information, the registrar has no way to make such a determination.

States vary in their treatment of what happens in the event that an applicant's information cannot be matched against the SSA or DMV databases. In some cases, a state may grant the applicant a conditional registration that requires the voter to present an ID at the polls before voting (indeed, in some states, all first-time voters are required to present an ID at the polls, regardless of whether a match is found); others may provide a provisional ballot to the voter on election day. As of June 2009, a Florida law that requires a nonmatch to result in an applicant not being registered was being challenged.[14] In still other cases, states register the voter without a provisional status (though they may flag first-time voters who have registered by mail).

IMPROVING RECORD-LEVEL MATCHING

In general, three approaches can be used to improve record-level matching: allowing more data (that is, using more data fields or more complete data fields in performing the match), improving the quality of the data contained in the relevant databases (including the use of tertiary/external data), and introducing systematic field-level matching algorithms to augment certain locally developed matching techniques.

The first approach often runs afoul of privacy concerns, and it requires policy makers to be willing to make a tradeoff between less privacy and better record-level matching. In this case, experiments with using more data fields or more complete data fields are necessary to determine the incremental benefit in record-level matching (for example, using an additional field in the match or using the last six digits of the SSN for matching instead of only the last four). The second approach, improving data quality, is addressed in more detail in Appendix C.

For purposes of this report, "ad hoc matching" is used to mean matching developed on the basis of intuitive reasoning that is not further validated systematically or analyzed with mathematical rigor. By contrast, systematic matching is based on a formal mathematical approach that develops metrics to measure match efficacy. With metrics in hand, policy makers can set scales for three relevant areas—what

[14] See *Florida State Conference of the NAACP v. Browning*, available at http://moritzlaw.osu.edu/electionlaw/litigation/Florida-NAACPv.Browning.php.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

determines a match, what determines a nonmatch, and what is indeterminate. In addition to good techniques for dealing with typographical error (discussed in next section), implementation of systematic techniques for matching can use some or all of the following elements:

- *Use of modern matching techniques* (also known in the statistical literature as techniques for record linkage). For example, a model introduced by Fellegi and Sunter[15] formalizes ideas of Howard Newcombe based on likelihood ratios in which it becomes somewhat easier to estimate record linkage parameters (even without training data). Training data is a large representative "truth" set of truly matching and nonmatching pairs of records. In the Fellegi-Sunter model each pair is given a score (or weight). The higher the score, the more likely a pair is to be a match.
- *Use of preprocessing to standardize data elements*. Preprocessing involves breaking fields into components and standardizing components, and a common preprocessing application is the use of address standardization software in which a house-number-and-street-name type of address may be broken into house number, street name, direction words (such as East, Southwest, and so on), and street type (Drive, Avenue) that are given standard spellings or abbreviations. Other methods can facilitate use of name information.[16] Although some of the methods described in this appendix are a good starting point, individual states may need to have specific methods for the types of idiosyncrasies and errors relevant to their individual needs.
- *Accounting for the relative frequency of occurrence of values of strings such as first and last names*. A relatively rare name such as "Zabrinsky" has more distinguishing power than a common name such as "Smith." The primary purpose of the frequency-based (or value-specific) matching is to downweight pairs having the more commonly occurring values of strings. If one has a large file representing an entire state, then one can compute the frequency-based scores associated with different strings by comparing the entire file against itself. The entire file becomes the surrogate training data. These ideas were introduced by Newcombe and extended by Fellegi and Sunter[17] and by Winkler[18] (Box B.2) in demonstrating how to implement frequency-based matching. In production matching software for the Decennial Censuses (1990 and beyond), Winkler had methods that automatically created the frequency-based weights. The distinguishing power of a particular name may vary considerably by geography. In Minnesota, for example, names such as "Garcia" and "Martinez" were relatively rarer and given more distinguishing power; in California the names are much more common and given less distinguishing power.
- *Estimation of optimal matching parameters (probabilities in the Fellegi-Sunter model) for classifying pairs as matches or nonmatches*. The probabilities can be computed by comparing an entire state file against itself, using a simple unsupervised learning method such as a properly applied expectation-maximization algorithm,[19] or an alternative method.[20] The optimal parameters have the effect of better separating matches from nonmatches. Although this improves matching, it does not yield estimates of error rates.
- *Providing methods for estimating false match rates*. Estimates of matching rates vary according to the matching scores (or weights). A certain false match rate will be associated with the designation of all

[15] Ivan P. Fellegi and Alan B. Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association* 64(328):1183-1210, December 1969.

[16] See William E. Winkler, "Business Name Parsing and Standardization Software," unpublished report, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., 1993; and William E. Winkler, "Advanced Methods for Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 467-472, 1994.

[17] Ivan P. Fellegi and Alan B. Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association* 64(328):1183-1210, December 1969.

[18] William E. Winkler, "Frequency-based Matching in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 778-783, 1989.

[19] William E. Winkler, "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 667-671, 1988.

[20] William E. Winkler, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 354-359, 1990.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

pairs above a value U1 as matches. If all pairs above a value U2 are designated as matches where U2 > U1, then the typical result is a lower false match rate and fewer pairs designated as matches. Belin and Rubin[21] and Winkler[22] have given unsupervised learning methods for estimating false match rates in situations for which there are no training data.

- *Providing methods for estimating false nonmatch rates*. Estimates of false nonmatches may partially be accomplished via methods of Winkler,[23] although these techniques may need to be modified if they are to be used on state DMV and VRD files.
- *Use of indexes and keyed search strategies to speed up the matching process when necessary*. Although most changes to VRDs are incremental, an operation involving entire database-to-database comparisons may sometimes be necessary. If two databases each have 5 million records, the number of possible pairs that must be compared is 25×10^{12}, a number that is much too large to search with most computer systems available to states. Optimized candidate selection strategies may be needed to reduce significantly the number of pairs that must be compared if the databases involved are large.
- *Use of automated name-matching logic that is guided and enhanced by culturally sensitive syntactic and semantic knowledge that accounts for different naming conventions*. As discussed in Appendix C ("Data Capture and Quality"), different cultures have different conventions for how names are formed. For example,

—Common American naming conventions regard certain names as equivalent (for example, Bill, Billy, and Will for William). The use of *automated name rooting and name equivalency tables* could be used to automatically generate common variants of a given name. Such tables would greatly reduce the need for multiple manual queries using name variants.

—Hispanic and Asian naming conventions for what parts of a name should be considered a surname do not fit easily into the conventional American convention of "first-name, middle name, last name." The use of *automated name ordering* could be used to automatically generate permutations of all types of ethnic surnames from the text string that makes up the complete name.

Implementation of name rooting and name ordering at the SSA would benefit all states that verify voter registration information using the SSA. Notably, name rooting and name ordering could be used as a component of any intrastate query mechanism as well.

MATCHING IN THE PRESENCE OF TYPOGRAPHICAL ERROR

One of the most difficult problems in matching is finding appropriate matches in the presence of typographical errors in the data. If the amount of typographical error in the files to be compared is small, then it is relatively easy to find pairs that agree on name and date-of-birth characteristics (for example).[24] However, if there is significant typographical error, then it is not possible to bring together pairs using straightforward character-by-character matching on name and date of birth. For instance if first name,

[21] Thomas R. Belin and Donald B. Rubin, "A Method for Calibrating False-Match Rates in Record Linkage," *Journal of the American Statistical Association* 90(430):694-707, 1995.

[22] William E. Winkler, "Automatic Estimation Record Linkage False Match Rates," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM. Also available at http://www.census.gov/srd/papers/pdf/rrs2007-05.pdf.

[23] William E. Winkler, "Matching and Record Linkage," pp. 355-384 in *Business Survey Methods*, Brenda G. Cox et al. (eds.), Wiley, New York, 1995; William E. Winkler, "Approximate String Comparator Search Strategies for Very Large Administrative Lists," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 2004.

[24] Whether these pairs in fact refer to the same person is an entirely separate question, because name and date of birth do not uniquely identify an individual. For instance, a given large state may have 1,000 individuals with the name "John Smith" and it is likely that some of the "John Smith" pairs will agree on date of birth. It may well be necessary to conduct other follow-up (such as manual examination of other data fields such as street address and Zip code) or to use data from third parties to help delineate the true match status of the pair.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Box B.2
Accounting for Commonly Occurring Names

The earliest computerized record linkage methods[1] do effectively account for the commonly occurring name plus "chance" date-of-birth phenomenon.

Newcombe's matching classification rule was to use the fields in pairs of records to compute a *matching score*. The idea was that agreement on individual fields was more likely to occur among "truly matching" pairs. Pairs above a certain upper bound were designated as matches; pairs below a certain lower bound were designated as nonmatches; and pairs with in-between scores were held for clerical review (when auxiliary information might be used to fill in missing information or "correct" contradictory information). If the upper bound is raised, then the false positive (false match) rate decreases. If the lower bound is decreased, then the false negative (false nonmatch) rate decreases.

The frequencies (probabilities) used in computing the scores can be estimated a priori using the frequencies in the large administrative lists, recognizing that matters such as "the list of most common names" will change slowly over time (which requires periodic adjustment of that set and the probabilities that those names will occur). Efficiently computed frequencies (conditional probabilities) are optimal in the sense that they can minimize the size of the clerical review region. Further, in many situations such as with voter registration databases or department of motor vehicles files, it is possible to estimate or give reasonable approximations of the error rates even without training data.[2] The earliest matching parameter and error-rate estimation procedures are the easiest to implement and most likely appropriate for VRD

[1] Howard B. Newcombe et al., "Automatic Linkage of Vital Records," *Science* 130(3381):954-959, October 1959; Howard B. Newcombe and James M. Kennedy, "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information," *Communications of the Association for Computing Machinery* 5(11):563-566, November 1962.

[2] William E. Winkler, "Comparative Analysis of Record Linkage Decision Rules," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 829-834, 1992; William E. Winkler, "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 274-279, 1993; William E. Winkler, "Automatic Estimation Record Linkage False Match Rates," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM, 2006, also at http://www.census.gov/srd/papers/pdf/rrs2007-05.pdf; Thomas R. Belin and Donald B. Rubin, "A Method for Calibrating False-Match Rates in Record Linkage," *Journal of the American Statistical Association* 90(430):694-707, 1995.

last name, and year of birth have 3 percent typographical error, then 9 percent (3 fields times 3 percent error in each field) of truly matching pairs may be missed with exact character-by-character matching.

An example of typographical error is provided in Box B.3. To overcome some of the difficulties caused by typographical error, modern techniques for matching are based on the computation of a score that indicates the degree of match rather than the generation of a yes-no result for any given comparison.

Comparisons can be made at the level of individual fields or at the record level.

String comparators compare text strings within individual fields; the Jaro-Winkler (JW) and edit-distance string comparators have been described elsewhere,[25] and code (C, C++, JAVA) is readily available on the Internet. The text strings to be compared are arbitrary, and in particular can represent names (or parts of names) or dates of birth (in some standardized format). These techniques provide

[25] William E. Winkler, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 354-359, 1990; William E. Winkler, "Overview of Record Linkage and Current Research Directions," Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., 2006, available at http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

files. The most general version of the parameter estimation procedures[3] generalize the iterative scaling procedures of Della Pietra et al.[4]

The frequency-based methods[5] automatically adjust match scores downward for the most frequently occurring first and last names. The effect of the downward adjustment is that pairs of records that are associated with commonly occurring names such as "James Smith" fall into an indeterminate region in which additional information (possibly via clerical review and contacting the voter) is required to determine matching status. In many situations, it is straightforward to obtain the extra matching information for the indeterminate pairs. Most other (much less commonly occurring names) can be matched effectively because the false positive rate is much less than 0.004 percent when using the combination of name, date of birth, and last four digits of the SSN (that is, typically they uniquely identify).

If the state VRD files can be examined a priori, then for each common first-name-last-name combination, we can find the most frequent dates of birth and lower the matching score of the associated pairs of records. We first lower the matching score for the common name combination and then again for the common dates of birth. To match the pairs with the lowered matching scores, we would need additional corroborating information such as telephone number or middle initial. If driver's license number or the last four digits of the SSN are available, then string comparators can be used to check whether the pairs of corresponding numbers are almost the same. The corroborating information might vary somewhat in differing states. In particular, some states request telephone numbers and/or e-mail addresses.

In this situation, it is possible to repeat analogous procedures to raise the worst-case false positive probabilities for certain specific name-date-of-birth combinations while significantly reducing the false match probabilities associated with the same name but different dates-of-birth combinations. This approach has the effect of significantly increasing the number of pairs of records for which match status can effectively be computed.

[3] William E. Winkler, "On Dykstra's Iterative Fitting Procedure," *The Annals of Probability* 18(1):1410-1415, July 1990; William E. Winkler, "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 274-279, 1993.

[4] Stephen Della Pietra et al., "Inducing Features of Random Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(4):380-393, April 1997.

[5] Howard B. Newcombe et al., "Automatic Linkage of Vital Records," *Science* 130(3381):954-959, October 1959; Howard B. Newcombe and James M. Kennedy, "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information," *Communications of the Association for Computing Machinery* 5(11):563-566, November 1962.

an automated mechanism for reducing the overall matching score from the score associated with exact character-by-character agreements on individual fields to account for partial agreement, thus accounting for very minor typographical error between two strings that are nearly the same. For instance, a comparison of "John" with "John" might yield a value of 1.0; a comparison of "Johm" with "John" might yield 0.90; and a comparison of "Smith" with "Smeth" might yield 0.94. These techniques often outperform ad hoc methods of "fuzzy matching."

The Jaro-Winkler comparator is a fast alternative to "edit distance" (as much as 10 times faster) that measures the minimum number of insertions, deletions, and substitutions to get from one string to another and returns equally high-quality results with administrative lists of the types that are similar to voter registration databases or department of motor vehicles files.[26]

[26] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg, "A Comparison of String Metrics for Matching Names and Addresses," *Proceedings of the Workshop on Information Integration on the Web*, International Joint Conference on Artificial Intelligence, Acapulco, Mexico, pp. 73-78, August 2003; William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg, "A Comparison of String Distance Metrics for Name-Matching Tasks," *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification*, Washington D.C., August 2003.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Box B.3
Example of Typographical Error

	First name	Last name	Date of birth
1a.	Robert	Smith	04211964
1b.	Rovert	Snith	04221963
2a.	Susan	Janes	bbbb1977
2b.	Sue	Jones	06171976

NOTE: Date format is mmddyyyy; "b" represents missing.

Comparisons at the record level are often based on a multiple pass strategy (sometimes called *blocking* or *binning*) in which pairs are brought together via characteristics that are believed to contain less typographical error and the remaining (or all) information in pairs is used in computing a matching score.[27] For instance, a search might be performed on first initials "J" and "S" and year of birth to retrieve records for which all remaining information is considered to compute a matching score against a record in another database for John Smith.

Blocking increases the number of possible pairs to be considered over what would be obtained if perfect agreement between fields were required. For example, a given blocking pass may bring together pairs that agree exactly on the date-of-birth field and also on the first character of the surname field. Because the first character of the surname typically is less likely to be in error (or is assumed to be so), this criterion is insensitive to some basic kinds of typographical error, e.g., "Smith" versus "Smoth." For each of these pairs, a matching score is computed using the rest of the information in the available data fields. For example, the first-name field and the entire surname field are compared using string comparators, and the match score between the pair may be defined as the sum of the two field-level scores. When record-level match scores are available for individual pairs, a threshold can be established (on the basis of experience) to the minimum score necessary for a pair to be considered a match.

It is common to use multiple passes through the data using different criteria. For example, a set of blocking criteria might be as follows:

- *Pass 1: date of birth and first character of surname.* As indicated above, this pass accounts for typographical errors in any part of the first name and in any part of the surname except the first character. Thus, it captures Bob and Rubert for Robert, Smoth for Smith).
- *Pass 2: day of birth, month of birth, and first three characters of surname.* This pass accounts for errors in the year of birth, which are known to be less accurate in many computer files than the day of birth and month of birth. However, using only day of birth and month of birth would usually result in too many pairs for efficient computation, and so a part of the surname is used to reduce that number. Thus, this pass accounts for first names and last names with typographical error in the last portions of these fields and for reporting/transcribing variations in year of birth.

[27] Howard B. Newcombe et al., "Automatic Linkage of Vital Records," *Science* 130(3381):954-959, October 1959; Howard B. Newcombe and James M. Kennedy, "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information," *Communications of the Association for Computing Machinery* 5(11):563-566, November 1962.; Ivan P. Fellegi and Alan B. Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association* 64(328):1183-1210, December 1969.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

• *Pass 3: first three characters of surname and first three characters of first name.* This pass accounts for any errors at all in the birth date (such as mistranscribed year of birth and mistakenly exchanged day of birth and month of birth).

In practice, the ordering of these passes matters. After Pass 1 is completed, all of the matching pairs above the relevant threshold score (indicating a match) are removed from the dataset and Pass 2 is performed on the remaining data. (More precisely, if a given pair exceeding threshold is composed of "name-a" and "name-b," name-a is removed from File A and name-b is removed from File B. This process is repeated for all matching pairs, and then Pass 2 is performed on the reduced File A and File B.) Pass 3 operates on a similarly reduced File A and File B, except that these reduced files do not include names found in pairs that matched on Pass 2 and Pass 1.

At the end of these multiple passes, all of the pairs exceeding the relevant threshold for one of the blocking criteria are considered matches. In general, the number of such pairs will be larger—sometimes substantially larger—than the number of pairs that would result if the matching criterion simply specified an exact match on first name, last name, middle initial, and date of birth.

Other technical approaches to blocking and string comparators can be found in Fienberg et al.[28]

MATCHING RECORDS USING THIRD-PARTY DATA

The use of blocking and string comparators is likely to generate a number of possible matches that may well be too large to investigate comprehensively through human review. In such cases, it may be possible to use third-party data (such as telephone books, credit header records, records of property ownership, and so on, discussed further in Appendix C) to resolve many of these ambiguities without human intervention, thus improving match accuracy.

For example, consider the two records R-1 and R-2 in Box B.4. If a human judge were faced with such a possible match, he might make a manual request from the neighboring county to compare signatures, or contact the voter, or prepare a letter to send to both addresses. However, if a search of a tertiary data source such as credit header data turned up record R-3, it would provide fairly strong evidence that records R-1 and R-2 in fact refer to the same individual. Alternatively, if the search turned up record R-4, it would provide some confidence that records R-1 and R-2 did not refer to the same person.

Note that the use of tertiary data in such a manner does not depend on a pairwise comparison between two data sources. Many list comparison systems are designed to compare one input file to another. If there is a third input file to process, the first output file is then compared to the third file (i.e., again a pairwise comparison). The approach illustrated above—a simple case of entity resolution—considers the data from all sources as their union (in the logical, set-theoretical sense).[29]

MATCHING RECORDS WITH UNIQUE IDENTIFIERS

Many of the difficulties described above can be reduced or eliminated through the use of a unique identifier (UID) for every voter, such as a driver's license number. If every voter has a single UID, records for a voter can be matched more simply.

In practice, even UIDs are sometimes improperly keyed in transcribing from a handwritten application or improperly recorded on the application (for example, because digits were transposed or one digit

[28] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg, "A Comparison of String Metrics for Matching Names and Addresses," pp. 73-78 in *Proceedings of the Workshop on Information Integration on the Web*, International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 2003; William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg, "A Comparison of String Distance Metrics for Name-Matching Tasks," *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification*, Washington D.C., August 2003.

[29] Jeff Jonas blog entry, Entity Resolution Systems vs. Match Merge/Merge Purge/List De-Duplication Systems (http://jeffjonas.typepad.com/jeff_jonas/2007/09/entity-resoluti.html).

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Box B.4
Illustrative Records

Record R-1: As written on registration form

County A
Daniel R Smith
123 Post Street
My City
DLN 0873457345
DOB 6/1944

Record R-2: As captured by the Social Security Administration

County B
Dan Randal Smith
456 Adele Lane
Your City
SSN4 5657
DOB 6/1944

Record R-3: As provided by credit header data (version 1 of Record R)

Daniel Randal Smith
DOB 6/1944
Current address: 123 Post Street, My City
Previous address: 456 Adele Lane, Your City
SSN4 5657

Record R-4: As recorded by credit header data (version 2 of Record R)

Daniel Richard Smith
DOB 6/1944
Current address: 123 Post Street, My City
Previous address: 789 Temple Hills, Some Other City
SSN4 1212

is illegible). If there is an error in the UID, a search could be performed using the name and the date of birth to find all possible UIDs associated with those names and dates to find the UID that is most similar to the one recorded in error—that UID would likely be the "correct" UID for the person in question.

A more general strategy would be needed when there is a possibility of typographical error in every field. The matching strategy is to search the entire file and apply suitable proximity metrics that indicate that the UID, first name, last name, and date of birth are sufficiently close to the query record. The feasibility of this strategy depends on the frequency with which invalid UIDs are encountered, because it is not practical to sequentially read every record in the database and perform substantial computation on every record in the file for every query.

The most general strategy involves substantial restructuring of the database to facilitate fast searches. Keys such as first character of first name plus last name plus date of birth, telephone number, or house number plus street name are defined and added to the database to allow fast searches. Using all appropriate fields, only records with proximity scores sufficiently close to the query record are retrieved for review. Definition of the keys and the order in which they are applied requires certain experience and skill.

Copyright National Academy of Sciences. All rights reserved.

C

Data Issues

As noted in Appendix B, the quality of data with which matching procedures must work has a significant impact on the rate of false positives and false negatives that result from such procedures.

SOURCES OF VOTER REGISTRATION INFORMATION

The NVRA requires state departments of motor vehicles to incorporate the voter registration application into the application for driver's licenses in a way that does not require the applicant to duplicate any information (except for a second signature). Thus, the DMV is responsible for passing to voter registrars the information needed to register a voter. In most states, the forms are simply sent from DMV offices to the local elections office, where a second manual data entry into the VRD takes place. In a few states, the data from the form are entered into DMV records, and then the proper information is extracted and sent to the registrar electronically (eliminating the need for a second data entry). State DMVs are also required to transmit changes of address received for driver's licenses to the appropriate voter registrar for a change of registration address unless the individual involved indicates otherwise.

The NVRA also requires public assistance and disability service agencies to provide voters with voter registration forms that voters complete manually and then return to the agency or department for delivery to the voter registrar, or to certify in writing that the individual applying for assistance or service has declined the opportunity to register to vote.[1] (However, the committee also recognizes that election officials are not generally in the chain of command for these agencies, a fact that often leads to a certain amount of bureaucratic politics as Agency A seeks to persuade Agency B to help carry out the mission of Agency A.) The availability of registration forms in these many locations increases the opportunities for eligible voters to register, but can also result in duplicate registrations that are sent to election agencies, and if voters themselves fill out the form manually, they can and do make mistakes.

[1] The committee received testimony during its second workshop that many state assistance and service agencies are not following through with this obligation.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

DATA CAPTURE AND QUALITY

Under all procedures used for voter registration in the United States today, the prospective voter must take action to register to vote.[2] Through such action, the voter provides certain pieces of information that eventually wind up in a voter registration database. If this process could be guaranteed to be error-free, many fewer problems of data quality would exist. But unfortunately, this is not the case.

It is useful to distinguish between three categories of error that may be introduced in the journey of these pieces of information from the voter's head to the database. Usually, the voter provides handwritten information on a form. The form is transmitted or carried to the voter registrar, where the data are transcribed from the form into machine-readable form, usually by a data-entry clerk who performs this task manually. Once in machine-readable form, the data may then be processed in some minimal fashion before it is stored permanently in the database. All of these steps can result in some kind of error.

A variety of problems complicate the data capture process. For example, data capture efforts are often compromised by:

- *Illegibility*. The information on most voter registration forms is handwritten, and in many cases, the handwriting is difficult to read, entirely illegible, or misunderstood. This makes the act of entering this information more challenging and increases the potential for errors in voter registration records to be entered in the database.
- *Inaccurate or incomplete voter registration information*. Applicants may fill out the forms inaccurately or incompletely if they misunderstand what information is required. Although applicants make such errors in all venues in which they fill out applications, they are more likely to make errors when the venue is crowded, noisy, and chaotic and when those available to help applicants do not have time or are not knowledgeable enough to answer questions about the applications. These conditions are often met during voter registration drives that take place in locations other than election offices—shopping centers, university campuses, and other locations that attract large crowds. In addition, voter registration drives are frequently staffed by volunteers, some of whom may not have sufficient knowledge of process and procedures in collecting voter information; this may be especially true when volunteers are brought in from out of town.
- *Missing voter registrations*. For example, Jim Dickson of the American Association of People with Disabilities testified to the committee that the volume of voter registration applications received from state social service and disability agencies (a service to potential voters that the NVRA directs these agencies to provide) has dropped significantly since the initial implementation of the law in 1995, although the committee notes that the causality of this drop remains unclear—that is, it is unknown whether this drop reflects failures in the social service agencies to meet their legal obligations; a change in the demographics and/or preferences of those applying for social services; problems in conveying completed applications to voter registrars; or some other reason(s).
- *Repeated (duplicate) registration applications*. An individual may submit multiple voter registration applications "just to be sure," or because s/he may have forgotten that s/he is already registered to vote. Although voter registrars are supposed to have mechanisms in place to screen duplicate registrations, the screening process does not always work smoothly, and sometimes the same individual may be registered more than once.
- *Inconsistencies in submitted information*. In filling out forms, individuals are often unintentionally inconsistent in the information they provide, especially if a period of time has elapsed between multiple form-fillings (either across registrations or between registrations and other activities such as applying for a driver's license or an SSN). An individual may use a nickname in one case and the full legal name in another, or include a middle initial in one and omit it in another. Such inconsistencies may arise because of a lack of clarity in the instructions given to the individual about what specific information to

[2] Exceptions arise from the fact that some states allow same-day registration and that North Dakota does not require voter registration per se.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

provide or a lack of recall about what s/he entered on a previous occasion. In other cases, the information requested may have changed (names sometimes change upon marriage, for example).

- *Data entry errors*. Typographical errors are made by hitting one key when another was intended. Transposition errors transpose two letters in a field, or even two fields. Even with carefully handwritten registration forms, it is possible that transcription/keying error may approach 5 percent or more in fields such as first name, last name, and date of birth if the data entry clerks lack adequate training and monitoring.[3]
- *Systematic errors stemming from different data representation conventions*. Among the most important are those associated with dates and names.

  —In many countries (including most of Europe), 01/03/2007 means March 1, 2007, whereas in the United States it means January 3, 2007. A naturalized U.S. citizen is perhaps more likely to make such a mistake than an individual raised in the United States.
  —In many Asian nations, the family name is always stated first. Kim Jong-il is a Korean name; the family name is Kim, and the given name is Jong-il. However, it would be easy for an American to recognize Kim as a first name, perhaps as an abbreviation for Kimberly, and Jong-il as a last name.
  —Names normally rendered in an alphabet other than a Roman alphabet may well be spelled inconsistently when transcribed into a Roman alphabet. This problem is of particular concern to those of Russian, Asian, Israeli, and Arabic descent.
  —Hispanic naming conventions are complex and very difficult to fit into a conventional "first name, middle name, last name" structure. The complexities include:[4]

    o Marriage-related name changes for females (Appelido de Casada) and/or widowhood (viuda de, v. de);
    o Incomplete collection of all surnames, due to bearer preference or to data collection constraints;
    o Inconsistent white-space placement, causing merger of phrasal prefixes (DE LA, DELA) and/or merger of prefix and surname stem (DE LA FUENTE, DELAFUENTE);
    o Use of initials in surnames, especially for high-frequency matronymic elements (RODRIGUEZ DE G.);
    o Use of familiar/nickname forms of given names (FRANCISCO-PACO);
    o Use of orthographic shortened forms of given names (FRANCISCO-FCO, MARIA-MA);
    o Presence of a surname from a non-Hispanic culture in an otherwise Hispanic name which continues to follow Hispanic nomenclature patterns.

These factors generate a wide range of errors. Table C.1 summarizes a variety of error types that may also exist in name fields; Table C.2 describes some possible errors in date-of-birth fields. Voter registrars are left with the problem of managing an environment in which such errors are common.

Problems with data capture and errors in the voter registration database can have an important effect on the individuals whose data are involved. The voter believes that he or she is properly registered, but the registration may have been rejected as a result of the inaccurate, incomplete, or illegible information on the form, or the voter may not know to bring to the polls on Election Day the additional identification required because of a problem with his or her form. In some cases, the voter may be entirely absent from the voter registration rolls.

[3] See Joseph J. Pollock and Antonio Zamora, "Automatic Spelling Correction in Scientific and Scholarly Text," *Communications of the ACM* 27(4):358-368, April 1984. In a highly controlled situation, keying error rates were in excess of 2 percent (in keystrokes). A 1-2 percent error rate in keystrokes could easily yield a 5 percent error rate in fields.

[4] Leonard Shaefer, Chief Scientist, IBM Global Name Recognition, personal communication to the committee, e-mail to Jeff Jonas of August 31, 2009.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

TABLE C.1 Illustrative Sources of Error in Names

Source of Error	Name on Voter Registration Form[a]	Name in Database
Typos	Pierce	Peirce or Pearce or Perce or Pierrce
Transliteration	Mohammad	Muhammed
Marriage	Mary Pierce (maiden name Owens)	Mary Owens or Mrs. Martin Pierce
Nickname	Sam Pierce	Samuel Pierce
Transposed field	Bao Lu	Lu Bao
Double names	"Mary Ann" (first) "Pierce" (last)	"Mary" (first) "Ann" (middle) "Pierce" (last)
Hyphenated name	"Mary" (first) "Owens-Pierce" (last)	"Mary" (first") "Owens" (middle) "Pierce" (last)
Punctuation	al-Amin	al Amin
Omitted middle name or initial	John Philip Pierce	John Pierce
White space	Mario De La Fuente	Mario Delafuente

[a]Handwriting assumed to be readable.

SOURCE: For all rows but the last two: Justin Levitt, Wendy R. Weiser, and Ana Muñoz, *Making the List: Database Matching and Verification Processes for Voter Registration,* Brennan Center, New York University, 2006. Reprinted with permission.

TABLE C.2 Illustrative Sources of Error in Dates of Birth

Source of Error	On Voter Registration Form	In Database (Voter, DMV, and/or SSA)
Typos	01/03/05	02/03/05 or 1/00/05 or 1/03/05 or 11/03/05
Transposed field	01/03/05	03/01/05 or 05/01/03
Invented default	01/03/05	01/01/05 (submitted only as January 2005)

SOURCE: Justin Levitt, Wendy R. Weiser, and Ana Muñoz, *Making the List: Database Matching and Verification Processes for Voter Registration,* Brennan Center, New York University, 2006. Reprinted with permission.

Errors in databases will accumulate if action is not taken to correct them promptly. For example, assume that 16 percent of all records in a database reflect at least one change in a field per year. After 3 years, 40 percent of the records will be different. This means that if the database is not updated yearly, 40 percent of the records in the database will be in error.

In addition, it may become more difficult over time to correct errors that occurred at previous time periods in the absence of mechanisms to keep track of individuals uniquely (for example, through driver's license numbers or through secondary systems that keep history)—that is, errors can compound as multiple matches and corrections take place. For instance, if a state VRD file has dates of birth corrected using a semiautomatic procedure that utilizes matching with a state DMV file, then incorrect matching or an erroneous date of birth in the DMV file will induce error in the state VRD file. Subsequent matching against state social services files or SSA files to determine whether an individual is deceased will either fail or possibly induce additional error.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

IMPROVING DATA CAPTURE AND QUALITY

A number of approaches are available for improving the quality of data within a VRD. However, all such approaches require certain skills and resources *on a continuing basis*. This last point is important—because of ongoing changes in the population eligible to vote, a continuous effort to maintain data quality in a voter registration database is needed if the database is not to fall into an error-filled state. Inadequate resources for database maintenance will result in greater amounts of error.

The remainder of this section addresses a variety of ways for improving data quality. However, one often-used method for improving data quality is not an option for voter registrars—starting over from scratch. In many cases, databases with errors that accumulate over time eventually become so filled with erroneous data that it is more cost-effective to rebuild the databases from scratch than to try to clean them up. Voter registrars in Kentucky did so in 1973, requiring all voters to re-register. However, "starting from scratch" for a VRD would mean purging everyone from the VRD, and since the NVRA establishes specific criteria for removing voters from registration lists, such an act would be contrary to existing law.

Human-Assisted Data Cleaning

Many traditional systems for managing administrative lists incorporate procedures that improve data capture and remove some typographical variations. The data-capture procedures are intended to improve the quality (legibility and completeness) of the information on written forms and the subsequent keying of the data-derived information into computer files. In traditional systems, list cleanup is often performed by skilled specialists who can determine name variations or possible missing information in the main administrative files. Using experience and auxiliary information, the specialists might determine that the date of birth (in the form mmddyyyy) "06139182" might really have been meant to be "06131982."

The intent of the corrections by the specialists was to remove typographical errors in the main administrative list. A cleaned-up list allows more effective searching of large files and effective comparison of pairs of records. For a new record "John Smith" with date of birth "06131982," it is much easier to search for "John Smith" in the corrected administrative list and compare dates of birth or search for "06131982."

Note that some types of typographical error simply cannot be identified using such a technique. Although automated accounting for the presence of typographical errors in a database is often possible, certain "errors" may not in fact be errors. "Bill" is only one character away from "Bull"—and indeed the "i" in Bill may be a mistyped "u," but "Bull" is used as a first name from time to time as well. There are no known ways to handle such "errors" automatically without the availability of tertiary reference data. (In some instances, it may be possible to check a possible spelling error against the gold standard of the original data form that first captured the data.)

In some instances, such as UK national health files or U.S. SSA files, a full-time staff locates, follows up, and corrects for certain types of errors. This effort can significantly reduce the number of individuals who are represented in the lists two or more times. If these cleaned-up lists are used in verifying information associated with other lists, then these other lists are much less likely to induce additional error than are lists that have not undergone intense cleanup.

Voter-Assisted Error Correction

New registrants can sometimes be given the opportunity to correct erroneous information. For example, the name and address provided on a registration card may be legible, but the date of birth illegible. If enough legible information is provided, voter registrars can contact the voter to inform him/her of the problem and ask them to resubmit correct information.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

In many polling places today, voters can correct registration information—a poll worker notes an error on the registry or on another log, and the election officials can update their registry as part of the postelection canvass. In addition, voters in many states now receive confirmation cards that confirm their registrations; these cards provide the voter with an opportunity to review the information that is part of their registration.

To help minimize keying errors, registrars might ask individuals with access to the relevant facilities to correct their information online through a Web site; security would be provided by a special code or password returned to the individual with the data correction request to ensure that only the proper individual could view or correct the information.

Electronic Transmission of Voter Registration Applications

Important sources of voter registration applications include departments of motor vehicles and social service agencies. Today's processes usually require individuals to register using handwriting on paper forms, a process that is highly subject to error upon data entry. But there is no reason in principle that the information collected by the DMVs and social service agencies (which is almost surely being captured in electronic form for use in DMV or social service agency systems) that is relevant to voter registration could not be transmitted electronically to voter registrars, thereby eliminating errors associated with repeated keying (once for the agency in question and a second time for the VRD). Some states also require that the voter provide a signature for the voter registration record, which is used for verification against pollbooks or ballot return envelopes in the mail-in voting process. An electronic transfer of voter registration forms must therefore accommodate in some way the need for the signature.

Though recommended by the Election Assistance Commission in its *Voluntary Guidance on Implementation of Statewide Voter Registration Lists*,[5] electronic transmission is not required by any present regulation and would entail some nontrivial work to implement on a large scale, such as agreement on the format for transmission and the construction of additional software to permit the exchange of information.

Use of Other Databases (Including Third-Party Data)

Yet another way to correct errors in an existing database is to match as many of its records as possible with those in another complete, (nearly) error-free database (or several such databases) and to use these other databases as "truth" for error correction. If there are no such complete high-quality databases available, then the use of other databases can still be useful to triangulate on the correct information, but the error correction process will take a lot more work under these circumstances.

At the same time, the fact that other databases may contain data with fewer errors does not mean that the information they provide should automatically be used to update the voter's registration. Discrepancies between the voter's registration information as represented in the VRD and data in these other databases are indicators of possible errors in the VRD, but in most cases voter registrars are required by law or policy to follow up on such discrepancies by contacting the voter to inquire as to which information is accurate—the voter database or the other database used in the match.

Third-party data, or secondary data, of high quality can be used to reduce ambiguity in record-level matches because they can be used to associate the same identity with a different record using data values based on a different time period or on differences in the values recorded. Sources of such data include telephone books and credit header data (credit records), which can be used to determine or validate middle names, addresses, dates of birth, and so on. Other generally available sources of data sometimes worth consideration include databases of property ownership, magazine subscriptions, and so on. Data aggregators, such as Lexis-Nexis, Choicepoint, and Acxiom, collect data from a variety of

[5] Available at http://www.eac.gov/election/docs/statewide_registration_guidelines_072605.pdf/attachment_download/file.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

disparate sources and sell data on a record-by-record request basis over an Internet connection, although the expense of access to such data may be a significant barrier to their use.

Third-party data vary in quality, with some sources worse than others. In addition, data collected to serve one purpose are sometimes less well suited for another purpose. These issues with quality may affect judgments about the suitability of available third-party data for correcting errors in a VRD.

Some 94 percent of the parties responding to an unpublished 2007 National Association of State Election Directors survey on voter registration practices indicated that they did not use secondary data sources such as phone directories or real-property records to reconstruct a voter's information if information supplied by the voter on a voter registration card was missing or incomplete.

A special source of third-party data for a given state is the VRDs of other states. That is, under most circumstances, an individual can vote in only one jurisdiction. Generally, it violates no law for an individual to be registered to vote in more than one jurisdiction, but the presence of the same person in the VRDs of two states suggests that one of those registrations does not accurately reflect the status of that individual. A number of states have agreed to exchange voter registration data in a couple of ongoing collaborations. Only preliminary data from these collaborations are available at this point, and the committee looks forward to analyzing more detailed data from these projects in the future, including information on the fields they are matching, the number of potential duplicates on the lists, and the number of actual duplicates they remove from their lists. A start at tracking some efforts at interstate checking of duplicate registrations can be found in the EAC report *Impact of the National Voter Registration Act on Federal Elections 2005-2006*.[6] On page 76 of that report can be found the fact that at least three groups of states have checked for such duplicates at least once: District of Columbia, Virginia, and Maryland; Minnesota, Missouri, Nebraska, Kansas, and Iowa; and Kentucky, South Carolina, and Tennessee.

Still another method to improve data quality in names is to check names against standard name inventories. For example, the Census Bureau compiles a list of first names and surnames found in the census.[7] Using such lists as spell-check dictionaries, a data entry clerk could be notified in real time that a given first name or surname was not represented in these lists, possibly signaling a keypunching error and/or the need to recheck the name entered against the information on the voter registration form.

Use of ethnically specific lists may be especially useful for entering name data associated with those ethnic groups. For example, the Passel-Word list of Hispanic surnames includes a large majority of surnames used by Hispanic residents.[8] To prevent non-Hispanic names from being flagged as possible errors, the name "spell-checker" could generate a string comparator score for the match between the entered name and the closest match on the Hispanic surnames list. If that score were higher than some threshold X, it would indicate a possible Hispanic name, and a flag would be issued. If the score were lower than X, no flag would appear. To the extent that inventories are available for other ethnic groups (Asian, Arabic, Russian, and so on), they might also be used for name checking.

Of course, because a given individual may really have a surname or a first name that is not represented in the inventory, nonmatches should never be used to *correct* data in an automated fashion.

Improving match accuracy can contribute to improved completeness of a VRD. Match accuracy, whether performed by automated processes or manual review, can be improved by tertiary, third-party, data. When such external data are carefully harnessed for improved match accuracy, systems can more often resolve ambiguities without human involvement. Reducing the number of exceptions necessitating human review and judgment increases the repeatability of list maintenance.

Such data can be used in two ways. First, such data can be acquired across the entire population and

[6] Available at http://www.eac.gov/clearinghouse/docs/the-impact-of-the-national-voter-registration-act-on-federal-elections-2005-2006/attachment_download/file. See also Thad Hall and Michael Alvarez, *The Next Big Election Challenge: Developing Electronic Data Transaction Standards for Election Administration,* IBM Center for the Business of Goverment, 2005, available at http://www.vote.caltech.edu/media/documents/AlvarezReport.pdf.

[7] See, for example, http://www.census.gov/genealogy/names/nam_meth.txt.

[8] See, for example, R. Colby Perkins, *Evaluating the Passel-Word Spanish Surname List: 1990 Decennial Census Post Enumeration Survey Results*, U.S. Dept. of Commerce, Economics and Statistics Administration, Bureau of the Census, Washington, D.C., 1993.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

made available for error-correction processes. Second, data can be selectively made available only when they are needed to resolve ambiguities in any putative record-level match—an approach that minimizes privacy concerns because it obtains additional data on individuals only when they are needed.[9]

When using third-party data to enhance matching accuracy, additional logging and accountability requirements must be introduced. Each third-party record requested and received must be retained and retained in its original form until it is no longer needed (for example, until the point that the voter has confirmed any changes that may have resulted from the use of such data). Furthermore, any third-party record used to improve a match should be logged and accounted for similarly. In addition, government matching with third-party datasets raises privacy concerns (such as concerns if credit header data are merged with voter history data, for example).

COLLATERAL ISSUES IN IMPROVING DATA QUALITY

Application of the techniques discussed above is intended to improve the quality of the data in a VRD by making the data more accurate—that is, these techniques allow erroneous data to be changed into correct data. But their success in doing so is not guaranteed—use of the techniques may introduce additional error, or the original data may in fact have been correct. Thus, it may well be advisable to keep the old data as well as the new, but with a flag that indicates that the old data have been corrected. In addition, a policy must be established regarding notification of the voter if a field is changed. The cost of such notification must be weighed against the value of ensuring with high confidence that the updated data are correct.

[9] This technique is explained in detail in Paul Rosenzweig and Jeff Jonas, "Correcting False Positives: Redress and the Watch List Conundrum," Legal Memorandum 17, The Heritage Foundation, June 17, 2005, available at http://www.heritage.org/Research/HomelandSecurity/lm17.cfm.

Copyright National Academy of Sciences. All rights reserved.

D

Security and Privacy

Voter registration systems are known to be points of risk in election administration systems. The ostensible purpose of voter registration is to make the election system more secure against fraud in the first place. When a voter registration system is computer-based, security thus becomes an issue.

Security is the property of a computer system whereby the system does what is required and expected in the face of deliberate attack.[1] For purposes of this report, privacy refers to policies that protect the information contained within the voter registration database (VRD) against inappropriate access.

As the comments in this appendix indicate, privacy and security issues related to VRDs are not merely technical issues. Indeed, a mix of policy and technology is relevant to their consideration, and these issues are nothing else if not hard to resolve.

SECURITY[2]

The security of the VRD is necessary to ensure that the VRD properly performs its function as an accurate and complete list of registered voters. Although the security of VRD systems has not been subject to the levels of scrutiny directed at electronic voting systems, it is nonetheless important. Security issues in VRDs arise for three reasons. First, state VRDs contain personal information associated with registered voters, and such information must be protected against disclosures not permitted by law. Second, the overall integrity of the VRD must be protected against unauthorized alterations (e.g., individual records being improperly added, deleted, or changed). Third, the VRD system must be avail-

[1] Reliability in the face of human, machine, or network failure is also an important dimension of system trustworthiness, but this appendix focuses on security against deliberate attack.

[2] There is an extensive body of National Research Council work on computer security issues, beginning with *Computers at Risk: Safe Computing in the Information Age*, 1990, and continuing with *Cryptography's Role in Securing the Information Society*, 1996; *Trust in Cyberspace*, 1999; *Realizing the Potential of C4I: Fundamental Challenges*, 1999; *Making IT Better: Expanding IT Research to Meet Society's Needs*, 2000; *Cybersecurity Today and Tomorrow: Pay Now or Pay Later*, 2002; *Software for Dependable Systems: Sufficient Evidence?*, 2007; and *Toward a Safer and More Secure Cyberspace*, 2007, all published by the National Academy [Academies] Press, Washington, D.C. In addition, an extensive discussion of security and privacy issues specifically with reference to voter registration databases is contained in U.S. Public Policy Committee of the Association for Computing Machinery, *Statewide Databases of Registered Voters: Study of Accuracy, Privacy, Usability, Security, and Reliability Issues,* 2006, available at http://usacm.acm.org/usacm/PDF/VRD_report.pdf. Excerpts from the executive summary of this report relevant to privacy and security are provided in Box D.1.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Box D.1
Excerpts from a 2006 Study of Voter Registration Databases Relevant to Privacy and Security

The following material is reprinted from the executive summary and the main text of *Statewide Databases of Registered Voters: Study of Accuracy, Privacy, Usability, Security, and Reliability Issues,* a 2006 report by the U.S. Public Policy Committee of the Association for Computing Machinery.

2. Accountability should be apparent throughout each VRD.

It should be clear who is proposing, making, or approving changes to the data, the system, or its policies. Security policies are an important tool for ensuring accountability. For example, access control policies can be structured to restrict actions of certain groups or individual users of the system. Further, users' actions can be logged using audit trails (discussed below). Accountability also should extend to external uses of VRD data. For example, state and local officials should require recipients of data from VRDs to sign use agreements consistent with the government's official policies and procedures.

3. Audit trails should be employed throughout the VRD.

VRDs that can be independently verified, checked, and proven to be fair will increase voter confidence and help avoid litigation. Audit trails are important for independent verification, which, in turn, makes the system more transparent and provides a mechanism for accountability. They should include records of data changes, configuration changes, security policy changes, and database design changes. The trails may be independent records for each part of the VRD, but they should include both who made the change and who approved the change.

4. Privacy values should be a fundamental part of the VRD, not an afterthought.

Privacy policies for voter registration activities should be based on Fair Information Practices (FIPs), which are a set of principles for addressing concerns about information privacy. FIPs typically address collection limitation, data quality, purpose specification, use limitation, security safeguards, openness, individual participation, and accountability. There are many ways to implement good privacy policies. For example, we recommend that government both limit collection to only the data required for proper registration and explain why each piece of

able and functional when needed, both to perform the "real-time" updates required by HAVA and, most critically, on or before Election Day to enable real-time queries or to create poll books.

Security measures address the issue of both who is authorized to view or change information in the VRD and of what information within any record in the VRD may be viewed or changed. In the security context, viewing information includes seeing individual records and sending or transferring records en masse; changing information includes adding entirely new records, altering one or more fields within one or more records, and deleting records.

The security of systems is usually conceptualized in terms of confidentiality, integrity, and availability.[3] These apply in the context of VRD systems (where "system" is intended to include the human and organizational aspects of a system as well as the technology):

[3] See for example, NRC, *Toward a Safer and More Secure Cyberspace*, Seymour E. Goodman and Herbert S. Lin (eds.),The National Academies Press, Washington, D.C., 2007.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

personal information is necessary. Further, privacy policies should be published and widely distributed, and the public should be given an opportunity to comment on any changes. . . .

6. Election officials should rigorously test the usability, security and reliability of VRDs while they are being designed and while they are in use.

Testing is a critical tool that can reveal that "real-world" poll workers find interfaces confusing and unusable, expose security flaws in the system, or that the system is likely to fail under the stress of Election Day. All of these issues, if caught before they are problems through testing will reduce voter fraud and the disenfranchisement of legitimate voters. . . .

Security Against Technical Attacks

. . . [M]echanisms should be deployed to detect any penetration of system defenses, as well as any insider misuse. For example, application-specific intrusion detection systems could be used to monitor the number of updates to the VRD. Any large spike in activity, whether by an authorized user or in the aggregate, might warrant human attention. In addition, officials could consider contracting with a third-party network security monitoring service to detect network intrusions and attempted attacks on the system. . . .

. . . Officials should consider including an independent security review and publication of the software as part of the acceptance testing for the system. Claims that the security of the system will be endangered by such a review should be treated with extreme skepticism or rejected outright. . . .

SOURCE: U.S. Public Policy Committee of the Association for Computing Machinery, *Statewide Databases of Registered Voters: Study of Accuracy, Privacy, Usability, Security, and Reliability Issues,* 2006, available at http://usacm.acm.org/usacm/PDF/VRD_report.pdf. (c) 2006 ACM. Excerpted with permission. ISBN: 1-59593-344-1. Permission to make digital or hard copies of portions of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permission from permissions@acm.org.

- *Confidentiality*. A secure system keeps protected information away from those who should not have access to it. Examples of failures that affect the confidentiality of a VRD include an unauthorized party obtaining voter information on a large scale or a spouse abuser obtaining the address of his/her spouse from a VRD even if such information is supposed to be protected from disclosure.
- *Integrity*. A secure system produces its intended results or information, regardless of whether or not the system has been attacked. When integrity is violated, the system may continue to operate, but under some circumstances of operation it does not provide accurate results or information that one would normally expect. Failures of integrity of a VRD include both inclusion of noneligible individuals and unauthorized exclusion of eligible registered voters, as well as unauthorized modifications to data fields such as addresses, birth dates, or voting histories.
- *Availability*. A secure system is available for normal use even in the face of high load or an attack. An example of a failure in availability might be a system that is clogged with so much bad data that the system no longer operates reliably (typically this refers to electronic attempts to overwhelm a system but also could occur in the nonelectronic domain; for example, a flood of bogus paper voter registration applications might attempt to overwhelm the data-entry staff in a particularly critical jurisdiction).

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

A number of security breaches of VRDs have been reported.[4] For example, on October 23, 2006, an official from the not-for-profit Illinois Ballot Integrity Project reported that his organization demonstrated that it was possible to use the Chicago voter database remotely to compromise the confidentiality of names, SSNs, and dates of birth of 1.35 million residents. According to a spokesman for the Chicago Election Board, the problem arose because the city's database allowing voters to locate their voting precinct once asked voters for detailed information such as Social Security numbers, and even though the Web site was updated to require only names and addresses to make a query, the links to the Social Security numbers and the dates of birth were never eliminated.[5]

Security threats can arise even in systems that are not connected to the Internet. Although Internet connections are often an important source of vulnerability, they are most assuredly not the only source. The recent history of computer security is replete with examples of security compromises that had nothing to do with the Internet, such as data on stolen laptops, attacks from insiders abusing their privileges, and "social engineering" attacks involving humans posing as other humans, often over the telephone, in order to learn credentials such as passwords that could enable them to access systems and files they should not be able to access.[6]

Developing secure systems is a challenging task, and much has been written about such matters.[7] Below, some best practices for security measures are highlighted.

- Access control policies should be established and enforced that group people by established roles (based on function, jurisdiction, etc.) and assign to those roles the appropriate (minimal) level of access needed to carry out their job functions. In doing so, the "principle of least privilege" should be followed: access should be kept to the minimal necessary levels. This reduces the possibility of both intentional misbehavior and accidental mistakes.
- The number of people with administrative privileges should be limited. Very few users should have the ability to grant access to others. However, there should also be rules and procedures that allow trusted election officials to temporarily increase privileges available to others during emergencies or time-critical situations (such as on Election Day) in a controlled and fully audited manner. The specific number chosen here should balance the competing concerns of minimizing administrative privileges to minimize abuse and increasing them to ensure availability. This balance will vary depending on the size and other specifics of a jurisdiction, but a reasonable number might be expected to be at least 3 and no more than 10.
- Authorized users of the system should receive security training, including how to choose and protect passwords and how to resist "social engineering" attacks (attempts to deceive someone into performing certain actions).
- Encryption should be used to protect the confidentiality of data. For example, all communica-

[4] See http://www.privacyrights.org/ar/ChronDataBreaches.htm. This site contains descriptions of a number of data breaches involving actual VRDs, and a number of others of potential relevance to VRDs.

[5] See http://abcnews.go.com/Politics/story?id=2601085; http://www.electiondefensealliance.org/chicago_voter_registration_database_wide_open.

[6] For example, video surveillance cameras caught two intruders in Mississippi on June 23, 2006, stealing hard drives from 18 computers. Data files contained names, addresses, and SSNs of current and former city employees and registered voters as well as bank account information for employees paid through direct deposit and water system customers who paid bills electronically. See http://www.privacyrights.org/ar/ChronDataBreaches.htm.

[7] There is an extensive body of National Research Council work on computer security issues, beginning with *Computers at Risk: Safe Computing in the Information Age*, 1990, and continuing with *Cryptography's Role in Securing the Information Society*, 1996; *Trust in Cyberspace*, 1999; *Realizing the Potential of C4I: Fundamental Challenges*, 1999; *Making IT Better: Expanding IT Research to Meet Society's Needs*, 2000; *Cybersecurity Today and Tomorrow: Pay Now or Pay Later*, 2002; *Software for Dependable Systems: Sufficient Evidence?*, 2007; and *Toward a Safer and More Secure Cyberspace*, 2007, all published by the National Academy [Academies] Press, Washington, D.C. In addition, an extensive discussion of security and privacy issues specifically with reference to voter registration databases is contained in U.S. Public Policy Committee of the Association for Computing Machinery, *Statewide Databases of Registered Voters: Study of Accuracy, Privacy, Usability, Security, and Reliability Issues,* 2006, available at http://usacm.acm.org/usacm/PDF/VRD_report.pdf.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

tions channels used by the system should be secured via end-to-end cryptography to protect both the confidentiality and the integrity of the data. In many cases, this will be handled by the network or application layer (e.g., via the use of https on Web interfaces) rather than in the database system itself. Stored data—or at least sensitive data fields, such as SSN—should be encrypted as well, and under some circumstances, the data need not be decrypted for it to be used.[8]

- Firewalls should be used to severely limit connectivity between internal and external networks.
- Mechanisms (such as commercially available intrusion detection and anti-virus systems) should be deployed to detect and prevent any penetration of system defenses or insider misuse.
- It is easier to secure a computer if less software is installed on it. To the extent feasible, computers used for administering VRD systems should be dedicated for this purpose. (Election offices with limited resources may find it difficult to refrain from using computers for multiple purposes.) Furthermore, the number of computers that have the complete VRD system and/or the complete VRD database (particularly sensitive information such as complete or last four digits of Social Security numbers) should be limited.
- Election officials should obtain independent security review of the VRD system before deployment and thereafter whenever significant changes are made to the VRD system. Periodic security review is also helpful, though state regulations may make such review more difficult.
- All changes to the VRD contents and system must be tracked (e.g., via immutable audit logs and associated policies for monitoring them) for accountability purposes.[9] These include changes on individual VRD records, large-scale or batch updates, source code, database schemas, system configuration, and access control policies. Such logs also guard against individuals with authorized access viewing those records for unauthorized purposes; such unauthorized purposes may include satisfying curiosity (e.g., viewing details about a famous person) and making illicit money (e.g., selling an SSN). Immutable audit logs serve as a deterrent (because the use of such a log has been made known), a forensic tool when a breach is believed to have occurred, and a useful tool when conducting sample audits.
- Any realistic assessment of a system's security involves actual testing of the system's security by an adversary that is motivated to compromise it (such as a "red team" commissioned to find vulnerabilities). Although testing cannot necessarily reveal all security problems (and does nothing by itself to eliminate such problems), testing can often identify some remaining failures.
- Recovery from security failures and/or accidental mishap must be possible. This topic is discussed in more detail in Section 3.6 (Backup) in the main body of the report.

These measures address security issues for data under the control of the relevant election registrar. In the event that the election registrar releases data to another party (e.g., on demand to a requestor as required by policy or applicable law), there are few if any practical technical measures that the election registrar can take to ensure the subsequent security of the released data. Perhaps the only action that the election registrar can take is to ensure that the data released consist only of those data that are required to be released and no other data. Once the data leave the control of the election registrar, it is up to the recipient to enforce any relevant security measures.

PRIVACY

Distinct from security issues, privacy issues relate to *policy* regarding what information may be disclosed to which parties under what circumstances. Thus, a hypothetical law requiring that any registered voter's name and address (but not party affiliation or Social Security number) must be available

[8] For example, consider the use of a full SSN to facilitate matching of records. An individual's full SSN is usually regarded as sensitive information, but the use of a full SSN can greatly enhance the accuracy of performing matches. But encrypting the SSN creates a new but still unique identifier, which can itself be used as the match key without revealing the true SSN.

[9] Immutable audit logs are further described in http://www.markle.org/downloadable_assets/nstf_IAL_020906.pdf.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Box D.2
Codes of Fair Information Practice

Fair information practices are standards of practice required to ensure that entities that collect and use personal information provide adequate privacy protection for that information. As enunciated by the U.S. Federal Trade Commission (other formulations of fair information practices exist),[1] the five principles of fair information practice include:

- *Notice and awareness.* Secret record systems should not exist. Individuals whose personal information is collected should be given notice of a collector's information practices before any personal information is collected and should be told that personal information is being collected about them. Without notice, an individual cannot make an informed decision as to whether and to what extent to disclose personal information. Notice should be given about the identity of the party collecting the data, how the data will be used and the potential recipients of the data, the nature of the data collected and the means by which it is collected, whether the individual may decline to provide the requested data and the consequences of a refusal to provide the requested information, and the steps taken by the collector to ensure the confidentiality, integrity, and quality of the data.
- *Choice and consent.* Individuals should be able to choose how personal information collected from them may be used, and in particular how it can be used in ways that go beyond those necessary to complete a transaction at hand. Such secondary uses can be internal to the collector's organization, or can result in the transfer of the information to third parties. Note that genuinely informed consent is a sine qua non for observation of this principle. Individuals who provide personal information under duress or threat of penalty have not provided informed consent—and individuals who provide personal information as a requirement for receiving necessary or desirable services from monopoly providers of services have not, either.

[1] See http://www.ftc.gov/reports/privacy3/fairinfo.htm.

without restriction to the public reflects a policy choice rather than a security issue. A security issue arises if an unauthorized party is able to gain access through the VRD to the voter's Social Security number, which is supposed to be kept confidential. That said, technical measures to enhance security sometimes protect privacy as well.

Some of the information in VRDs is, by law, public information, although the specifics of which data items can be regarded as public information vary from state to state. In addition, states often limit the purposes for which such information may be used. Nevertheless, the electronic availability of such information raises concerns about the privacy of that information, because electronic access greatly increases the ease with which it can be made available to anyone, including those who might abuse it.

Many analysts of privacy issues point to fair information practices as a reasonable framework for privacy protection that balances privacy rights against user needs for personal information, and in the context of voter registration, the 2006 USACM report on statewide databases recommends the adoption of such practices as the basis for privacy policy regarding voter registration activities.[10] Fair information practices (FIPs) generally include notice to and awareness of individuals with personal information that such information is being collected, providing individuals with choices about how their personal information may be used, enabling individuals to review the data collected about them in a timely

[10] U.S. Public Policy Committee of the Association for Computing Machinery, *Statewide Databases of Registered Voters: Study of Accuracy, Privacy, Usability, Security, and Reliability Issues*, 2006, available at http://usacm.acm.org/usacm/PDF/VRD_report.pdf.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

- *Access and participation.* Individuals should be able to review in a timely and inexpensive way the data collected about them, and to similarly contest that data's accuracy and completeness. Thus, means should be available to correct errors, or at the very least, to append notes of explanation or challenges that would accompany subsequent distributions of this information.
- *Integrity and security*. The personal information of individuals must be accurate and secure. To assure data integrity, collectors must take reasonable steps, such as using only reputable sources of data and cross-referencing data against multiple sources, providing consumer access to data, and destroying untimely data or converting it to anonymous form. To provide security, collectors must take both procedural and technical measures to protect against loss and the unauthorized access, destruction, use, or disclosure of the data.
- *Enforcement and redress.* Enforcement mechanisms must exist to ensure that the fair information principles are observed in practice, and individuals must have redress mechanisms available to them if these principles are violated.

For reference purposes, the original "Code of Fair Information Practices" promulgated in 1972 by the Health, Education, and Welfare Advisory Committee on Automated Data Systems is provided below:[2]

> The Code of Fair Information Practices is based on five principles:
>
> 1. There must be no personal data record-keeping systems whose very existence is secret.
> 2. There must be a way for a person to find out what information about the person is in a record and how it is used.
> 3. There must be a way for a person to prevent information about the person that was obtained for one purpose from being used or made available for other purposes without the person's consent.
> 4. There must be a way for a person to correct or amend a record of identifiable information about the person.
> 5. Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take precautions to prevent misuses of the data.

[2] U.S. Department of Health, Education and Welfare, Secretary's Advisory Committee on Automated Personal Data, *Systems, Records, Computers, and the Rights of Citizens,* 1973.

and inexpensive way and to contest the data's accuracy and completeness, taking steps to ensure that the personal information about individuals is accurate and secure, and providing individuals with mechanisms for redress if these principles are violated. Box D.2 describes two versions of a code of fair information practices.

In the context of government-operated voter registration systems, many tensions arise between these principles and the application of existing policy and law. For example, one of the thorniest issues regarding privacy is the tension it sometimes poses with transparency. In its starkest terms, maintaining privacy involves withholding from public view certain information associated with individuals, while transparency involves the maximum disclosure of information, even if such information is associated with individuals.

As an illustration of how these tensions play out, consider a proposition regarding the public disclosure of the reason(s) for removing specific individuals from voter registration lists. On one hand, the removal of a voter from a VRD is often associated with a stigmatizing condition, such as being a felon or being declared mentally incompetent. Those mistakenly removed from a VRD may experience adverse consequences from such association, and even if the removal is correctly performed, those individuals are still arguably entitled to some measure of privacy. Thus, a person balancing the scales in favor of privacy would argue that the reasons for removing individuals from the VRD should be kept confidential, as they are in some states already.

On the other hand, advocates of greater transparency argue that removals from a VRD should be

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

subject to public oversight in the same way that additions are. They point out that convictions and even arrest records are generally public, and thus argue that not disclosing reasons for removal from a VRD does not really protect the privacy of these individuals anyway. At the same time, they argue that associating reasons for removal with specific individuals is critical to determining the qualification of voters—and that statistical tabulations alone would not provide the detail needed to investigate individual errors that might indicate systemic problems.

The committee noted significant value without much negative impact on privacy in statistical tabulations of the reasons for voters being dropped from a VRD and publication of such tabulations, as well as in personal and private notification of individual voters of the reason(s) for being dropped. But the different points of view described above were reflected in the committee, and thus the committee takes no position on the desirability or undesirability of the above proposition.

Other privacy advocates have raised concerns about the widespread availability of complete voter registration information in the context of the physical security of battered men or women. Such individuals have good reason to keep their addresses private, and might be apprehensive with good reason about the availability of their addresses to their batterers. A second concern relates to abuse of lists of validated addresses for commercial marketing purposes—many citizens would be upset to know that the information they provide to exercise their right to vote in a democracy is also being used for commercial purposes. Addressing such issues properly belongs to state policy makers, who can develop (and sometimes have developed) regulation and law to protect citizen interests—for example, some states allow only political parties to obtain voter registration lists.

Another tension arises because some state laws also allow election officials to change voters addresses of record without their explicit consent (e.g., when the officials receive a notice of a forwarding address).

Finally, FIPs require that the personal information provided by individuals be used only for specified purposes. But election officials rely on third parties to collect voter registration information, and they have no effective control over how those parties actually use the information they collect. And in some cases, election officials must release voter registration lists to political parties. The committee has no specific knowledge of whether third parties do in fact use voter registration information for their own purposes, but it recognizes the possibility of doing so as a potential conflict with implementing FIPs in a voter registration context.

One way to guard against large-scale misuse of voter registration data (e.g., using voter data for commercial purposes after agreeing contractually to only use the data for political purposes) involves seeding the database before it is transferred with one or more record(s) that can be used to detect later misuse. For example, a seeded record may indicate that John Cue Smith is a registered voter, at the registered address of 123 Special Street in a town within the relevant jurisdiction. If a piece of mail later arrives for John Cue Smith at this address promoting the sale of tennis shoes, the possible misuse of the database may be worth investigating.

A second set of privacy issues arises from matching and linking records. For example, voter registration lists may be matched against a list of convicted felons. If a list of voters removed from the VRD is made public, those removed from the list improperly or removed for other reasons (that is, all nonfelons removed from the list) may be tainted by association in the public eye. Similarly, if a voter registration list is made public that indicates the source of an individual application, those who registered to vote at public assistance agencies might regard their privacy rights as having been violated. Although overt public disclosure would violate the NVRA, accidental disclosure through a security breach might have a similar result. This could in turn reduce the likelihood that people will seek out public assistance if seeking it will automatically place that information in a voter registration record that is publicly accessible. Alternatively, where registration is not automatic, it may reduce the number of individuals who take advantage of the ease of registering at the public assistance agency and thereby undercut the goal of the program.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

A third set of privacy issues arises from insider access to the VRD. Insiders such as election officials could be expected to have access to the full set of information associated with any individual record, and possibly to some of the information in matched records existing in other databases. Although most election officials are trustworthy in this regard, a few might seek to use this access—improperly—for personal benefit or gain, and measures (such as immutable audit logs) are needed to deter and/or investigate such inappropriate insider access.

A fourth set of issues arises in the context of transferring a VRD to another party en masse. Such a bulk transfer may occur, for example, when two VRDs must be compared to each other (e.g., for the purpose of identifying duplicate registrations between them), to judicial authorities for jury selection, to political parties, or to any other party in accordance with applicable law. Because bulk transfers—by definition—involve personal information on a very large scale, potential threats to privacy are magnified in such circumstances.

For example, voters may well provide personal information for voter registration without knowing that such information may be used for other purposes. Even if such uses are entirely legal, it is still desirable to protect voter privacy to the maximum extent consistent with law. Thus, voter registration records transferred for comparing VRDs should only include the records that need to be used or matched, i.e., active records, and the fields contained on each record should be limited to the fields necessary to perform matching (such as name and date of birth but not party affiliation) and the voter's state-assigned voter ID. (The latter is necessary because without such a pointer, a record cannot be recalled or updated and reconciliation audits become problematic.)

Bulk transfers of data are also likely to persist in the absence of specific actions taken to decommission (remove from service) the data involved. Persistence after the data have served the original purpose of the transfer increases the likelihood of unintended disclosure and/or repurposing inconsistent with the original reasons for bulk transfer.

Lastly, bulk transfers of data—by definition—involve large quantities of data. Without specific knowledge of precisely what data have been transferred (i.e., a complete copy of what was transferred), it can be very difficult to determine who needs to be notified in the event that a problem arises (e.g., a data compromise). All too often, the only information kept regarding the bulk transfer are the selection criteria used to generate the data to be transferred and the number of records sent—given changes to the database in the intervening period, this information is almost certainly insufficient to reproduce the transferred dataset.

Copyright National Academy of Sciences. All rights reserved.

E

Workshop Agendas

WORKSHOP 1—AUGUST 6, 2007 (WASHINGTON, D.C.)

10:30 a.m. **Welcome to the Workshop**
Sharon Priest and Olene Walker, Committee and Workshop Co-chairs

10:45 a.m. **Panel 1: Overview of the Issues**

What are key voter registration issues and how do they affect the establishment of statewide voter registration databases as mandated by HAVA?
Moderator: Sharon Priest/Olene Walker

Panelists:
Gracia M. Hillman, Commissioner, U.S. Election Assistance Commission
Caroline C. Hunter, Commissioner, U.S. Election Assistance Commission
Leonard M. Shambon, formerly with Wilmer Cutler Pickering Hale and Dorr
Robert A. Pastor, Executive Director, Carter-Baker Commission and Director of the Center for Democracy and Election Management, American University

Q&A with presenters

12:00 p.m. Lunch Available
Continue discussion from first panel session and prepare for afternoon sessions

12:45 p.m. **Panel 2: Status of Voter Registration Database Efforts**

What are the different types of and approaches to voter registration systems? What are the benefits and tradeoffs? Do you build it on your own or do you contract it out? What are some upcoming challenges that will need to be addressed in the near term (1-2 years) and in the longer term (5+ years)?
Moderator: Bruce McPherson

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Panelists:
Deborah Markowitz, Secretary of State, Vermont, and Immediate Past President of the National Association of Secretaries of State
Brad Bryant, President, National Association of State Election Directors and Deputy Assistant for Elections, Kansas
Linda Lindberg, General Registrar, Arlington County, Virginia

Q&A with presenters

2:15 p.m. Break

2:30 p.m. **Welcome and Brief Overview for Web Cast Audience**
Sharon Priest and Olene Walker, Committee and Workshop Co-chairs

2:35 p.m. **Panel 3: Record Matching: Technical/Operational Issues and Problems**

What types of technical problems can occur in record linking? What is the impact on data quality? What type of data cleaning is required? What are potential solutions to these problems?
Moderator: William Winkler

Panelists:
Gio Wiederhold, Professor (Emeritus), Computer Science, Medicine, and Electrical Engineering, Stanford University
William Cohen, Associate Research Professor, Machine Learning Department, Carnegie Mellon University
Michael Franklin, Professor, Electrical Engineering and Computer Sciences, University of California, Berkeley

Respondents:
James Willis, Principal, Banyan Social Technology, and Former Director, eGovernment for Rhode Island
Frank Olken, Computer Scientist, Lawrence Berkeley National Laboratory

Q&A with presenters

4:00 p.m. **Panel 4: Interoperability and Database Operations in Other Domains**

What kinds of problems or issues exist in nonelection domains (i.e., government and nongovernmental settings), including technical and organization dimensions? What is the range of possible solutions?
Moderator: Paula Hawthorn

Panelists:
John Glaser, Vice President and Chief Information Officer, Partners Healthcare System
Dan Schutzer, Executive Director, Financial Services Technology Consortium
Vivek Narasayya, Senior Researcher, Data Management, Exploration and Mining Group, Microsoft Research
Ken Orr, Founder, Ken Orr Institute (participating by phone and Web conference)

Q&A with presenters

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

5:30 p.m. Adjourn

5:30 p.m. Open reception

WORKSHOP 2—NOVEMBER 29-30, 2007 (WASHINGTON, D.C.)

November 29, 2007

8:30 a.m. **Welcome to the Workshop**
Olene Walker, Committee and Workshop Co-chair

8:40 a.m. **Panel 1: Data Providers Issues and Challenges**
Moderator: William Winkler

Panelists:
Peter Monaghan, Director, Information Exchange and Computer Matching, Social Security Administration
William L. Farrell, Director, Office of Systems Security Operations Management, Social Security Administration
Walter A. Jackson III, Senior Systems Analyst, Systems Analysis Division, American Association of Motor Vehicle Administrators
James Wilson, Program Manager, Address Technology, U.S. Postal Service
Garland Land, Executive Director, National Association for Public Health Statistics and Information Systems

Q&A with presenters

10:15 a.m. Break

10:45 a.m. **Panel 2: Data Providers Issues and Challenges—continued**
Moderator: Paula Hawthorn

Panelists:
Kimball Brace, President, Election Data Services (remote participation)
Clark Bensen, Principal Consultant, Polidata
Keith Cunningham, Director of the Board of Elections for Allen County, Ohio

Q&A with presenters

11:45 a.m. Lunch available
Continue discussion from morning sessions and prepare for afternoon panels

1:00 p.m. **Panel 3: IT Operations—State and Local**
Moderator: John Lindback

Panelists:
Ray Palmer, Information Technology Manager, Office of the Governor, Utah
Mike Stewart, Chief Information Officer, Office of the Secretary of State, Kansas
Paul Miller, Technical Services Manager, Elections Division, Office of Secretary of State, Washington

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Shane Hamlin, Assistant Director of Elections, Office of Secretary of State, Washington

Q&A with presenters

3:00 p.m. Break

3:15 p.m. **Panel 4: Impact of Technical Implementation on Policy**
Moderator: Olene Walker

Panelists:
Wendy R. Weiser, Deputy Director, Brennan Center for Justice at New York University School of Law
James C. Dickson, Vice President of Government Affairs, American Association of People with Disabilities
Melanie L. Campbell, Executive Director, National Coalition on Black Civic Participation
Lloyd Leonard, Senior Director for Advocacy, League of Women Voters

Q&A with presenters

4:45 p.m. **Panel 4: Impact of Technical Implementation on Policy—continued**
Moderator: Michael Alvarez

Panelists:
Vincent Keenan, Executive Director, Publius
Michael P. McDonald, Associate Professor, George Mason University and Non-Resident Senior Fellow, Brookings Institution
Chris Thomas, Director, Bureau of Elections, Michigan Department of State
Ernie Hawkins, CERA, Chair of Election Center Board of Directors, California

Q&A with presenters

5:45 p.m. Reception

November 30, 2007

8:30 a.m. **Welcome and Overview**
Olene Walker, Committee and Workshop Co-chair

8:35 a.m. **Panel 5: Security and Privacy Issues**
Moderator: Jeff Jonas

Panelists:
Peter G. Neumann, SRI International Computer Science Laboratory
Glenn Newkirk, President, InfoSENTRY Services Inc. (remote participation)
James J. Horning, Chief Scientist, Information Systems Security Operation, SPARTA Inc.
Bradley A. Malin, Assistant Professor, Department of Biomedical Informatics, Vanderbilt University

Q&A with presenters

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

10:00 a.m. Break

10:30 a.m. **Panel 6: IT Operations—Vendors**
Moderator: John Lindback

Panelists:
Thomas H. Ferguson, Director, Saber Corporation
Neil McClure, Chief Technology Officer, Hart InterCivic

11:15 a.m. Workshop adjourns

WORKSHOP 3—MAY 7-8, 2008 (PORTLAND, OREGON)

May 7, 2008

Closed Session

8:00 a.m. **Welcome and Overview**
Olene Walker and Fran Ulmer, Committee and Workshop Co-chairs

Review workshop agenda and questions to raise during workshop

Open Session

8:30 a.m. **Welcome and Overview**
Olene Walker and Fran Ulmer, Committee and Workshop Co-chairs

8:40 a.m. **Overview and Demo—Oregon State Voter Registration Database System**
John Lindback, Dave Franks, and Ericka Hass, Oregon Secretary of State's Office

9:40 a.m. Break

10:00 a.m. **Roundtable Session 1—State Voter Registration Databases: Lessons Learned**
Discussion Facilitator: Olene Walker

- Reactions to study committee's interim report, *State Voter Registration Databases: Immediate Actions and Future Improvement*
- How has your state addressed implementation or operational challenges at either the state or local level?
- What is your state's experience with matching against DMV, SSA, or other state agency data sources? What is the status of any inter-state matching efforts, if any?
- What has been working well and what requires more attention?

12:15 p.m. Lunch available

Continue discussion from morning sessions and prepare for afternoon panels

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

1:15 p.m. **Roundtable Session 2—State Voter Registration Databases: Medium-Term Improvements (1-5 years)**
Discussion Facilitator: Bruce McPherson

- What are possible areas or potential solutions to improve voter registration and/or state voter registration databases in the medium term?
- How useful would any of the following proposed solutions be, including
  —The use of the tear-off receipts for voter registration?
  —The creation of a software repository to do better matching so states do not have to implement it themselves?
- How can we accommodate the need for a signature during electronic transfers?
- How can we get a better handle on the estimated costs for implementing the medium- and long-term recommendations?

Format:
Introduction and brief overview—Bruce McPherson
General discussion
Recap highlights—Jeff Jonas

3:15 p.m. Break

3:30 p.m. **Roundtable Session 3—State Voter Registration Databases: Long-Term Improvements (5+ years)**
Discussion Facilitator: Fran Ulmer

- What "big picture" system changes could be considered by regions or the federal government to provide easier exchange and sharing of databases?
- What methods would make it easier for a highly mobile voting population to vote (without having multiple jurisdiction registration processes limit that right)?
- What options and possibilities are available for automatic voter registration?
- What are the benefits/drawbacks of moving to same-day voter registration?
- Are there possible legislative changes that should be considered?

Roundtable Participants:

Gail Fenumiai, Director, Division of Elections, Alaska
Barbara Gruenstein, Municipal Clerk, Municipality of Anchorage, Alaska
Lee Kercher, IT Division Chief, Secretary of State, California
Dean Logan, Acting Registrar-Recorder/County Clerk, California
Bruce McDannold, Elections Division Lead, Secretary of State, California
Nancy Blankenship, County Clerk, Deschutes County, Oregon
Annette Newingham, Supervisor of Elections, Lane County, Oregon
Thad Duvall, Douglas County Auditor and President of Washington State Association of County Auditors
Peggy Nighswonger, Director of Elections, Wyoming
Lynne Fox, Uinta County Clerk, Wyoming

5:30 p.m. Adjourn public workshop

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Closed Session

6:00 p.m. **Working Dinner—Committee Members Only**

8:00 p.m. Adjourn working dinner

May 8, 2008

Closed Session

8:30 a.m. **Welcome and Overview**
Olene Walker and Fran Ulmer, Committee and Workshop Co-chairs

Review workshop agenda and questions to raise during workshop

Open Session

9:00 a.m. **Welcome and Overview**
Olene Walker and Fran Ulmer, Committee and Workshop Co-chairs

9:05 a.m. **Security Issues**
Discussion Facilitator: Rebecca Wright

- Describe the process of determining the risk assessment and the security model for a state and discuss how the security model was implemented.
- For centralized state voter registration systems, including the voter registration database and any additional election management functions, (1) what different types of functionality are states adding and deploying with their voter registration systems (e.g. online registration)? (2) what minimum levels of security would you expect a state to have for the different functionalities?
- What are some time-tested and valid ways to test your security without causing too much disruption?
- How much redundancy/back-up should be built into a system of this size and importance in order to accommodate security and reliability concerns?
- What kinds of security advice can we give that will apply to a broad range of potential statewide voter registration systems and will continue to hold for at least the next 5 years?
- Is there anything about security that is unique or specific to voter registration databases, as opposed to other domains? Or, alternatively, are there other well-studied domains that have the same or very similar security issues as statewide voter registration systems?

Participant: Randy Cobena, Vice President, Government Solutions, Saber

Q&A

9:30 a.m. **Interactive Discussion on Security for Voter Registration Databases**
Panelists: Paula Hawthorn, Rebecca Wright, and John Lindback

10:30 a.m. Adjourn public workshop

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Closed Session

10:30 a.m. **Discussion and Reflection**
Reactions to workshop and identifying issues to address in final report

12:00 p.m. Lunch

12:30 p.m. **Report Development and Workshop Planning**

- Review project schedule
- Workshop planning, possible topics/issues for the next three workshops
- Planning for the final report (target date for first draft is February/March 2009) including scope, outline, and assigning report sections
- Current set of issues/topics (deferred in the interim report) for possible inclusion in the final report is provided below. This list is not fixed and may continue to change based on committee deliberations.
  —Develop/promote public access portals for checking voter registration status.
  —Encourage/require DMV, public assistance, and disability agencies to provide voter registration information electronically.
  —Encourage DMV, public assistance/service agencies, tax assessors, etc. to remind voters in their communication to the public to check and update their information.
  —Improve matching procedures.
  —Establish software repository of tested matching algorithms.
  —Provide voter registration receipts to improve administrative process.
  —Allow voters to register and update missing or incorrect registration information online if a signature is already on file with a state agency.
  —Develop procedures for handling disenfranchisement caused by mistaken removals from voter registration lists.
  —Improve the design of voter registration forms.

3:00 p.m. Adjourn

WORKSHOP 4—JULY 30-31, 2008 (KANSAS CITY, MISSOURI)

July 30, 2008

Closed Session

8:30 a.m. **Welcome and Overview**
Olene Walker and Fran Ulmer, Committee and Workshop Co-chairs

Open Session

9:00 a.m. **Welcome and Overview**
Olene Walker and Fran Ulmer, Committee and Workshop Co-chairs

9:10 a.m. **Welcome and Overview of Interstate Matching Efforts**
Hon. Ron Thornburgh, Secretary of State, Kansas
Hon. Mark Ritchie, Secretary of State, Minnesota

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

9:40 a.m. **Past, Present, and Future of Interstate Data Sharing and Data Matching**
Brad Bryant, Election Director, Kansas

10:00 a.m. **Roundtable Session 1—Interstate Data Sharing and Data Matching: Initial Attempts and Areas for Improvement**
Discussion Facilitator: Fran Ulmer

- What is your state's experience with and status of any interstate matching efforts?
- What has been working well and what requires more attention?

11:45 a.m. Lunch available

12:45 p.m. **Roundtable Session 2—Interstate Data Sharing and Data Matching: Looking Toward the Future**
Discussion Facilitator: Michael Alvarez

- Why conduct interstate data matching? What are the goals?
- Should any broader frameworks be created to facilitate interstate data sharing?
- Should the efforts remain regional, expand to affinity states, move toward a national approach?
- What broader improvements could be made across states, such as data field/format standardization?
- Should attempts at interstate data matching remain ad hoc and bottom up (e.g., grow slowly and build to develop a standard over time) or would a more formalized, top-down approach be more effective?
- How can best practices be shared among the states that are engaging in interstate data sharing efforts?
- As more states take part in interstate data matching efforts, what privacy policies and/or public relation strategies may need to be formulated?

Roundtable Participants:

Brad Bryant, Election Director, Kansas
Mike Stewart, CIO for the Secretary of State's Office, Kansas
L. Neal Erickson, Deputy Secretary of State for Elections, Nebraska
Josh Daws, CIO for Secretary of State's Office, Nebraska
Mark Ritchie, Secretary of State, Minnesota

2:15 p.m. Break

2:30 p.m. **Other Perspectives on Interstate Data Matching**
Robert Brandon, Co-founder and President, Fair Elections Legal Network

3:15 p.m. **Third-Party Voter Registration**
Discussion Facilitator: Wendy Noren

- Share your reactions to the committee's recommendations for third-party voter registration groups in its interim report, *State Voter Registration Databases: Immediate Actions and Future Improvements*

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

- Describe your organization's experience with voter registration, noting particular successes as well as challenges.
- What changes, if any, is your organization working on to improve your voter registration processes?

Panelists:
Mary Merritt, President, League of Women Voters, Missouri
Mary Potter, Member, Jackson County Republican Committee
Michael Slater, Deputy Director, Project Vote, Association of Community Organizations for Reform Now (ACORN)

4:30 p.m. **Why Have Voter Registration?**
Jim Silrum, Deputy Secretary of State, North Dakota

5:15 p.m. Adjourn public workshop

Closed Session

6:00 p.m. **Working Dinner—Committee Members Only**

8:00 p.m. Adjourn working dinner

July 31, 2008

Closed Session

9:00 a.m. **Welcome and Overview**
Olene Walker and Fran Ulmer, Committee and Workshop Co-chairs

9:10 a.m. **Discussion and Reflection**
Reactions to workshop and identifying issues to address in final report
Workshop planning, discuss topics/issues and possible speakers for the next workshops

10:30 a.m. **Top Ten Lists for Security and Privacy**
Rebecca Wright—security check list
Jeff Jonas—privacy concerns

11:30 a.m. Lunch

12:00 p.m. **Developing an Outline for the Final Report**
Each committee member will come up with three to four issues that the final report should address. Each issue should be recorded on a separate sheet of paper and should relate to one of the six broad areas identified in the interim report (see list below). After the committee has finished this exercise, the group will discuss each of the major areas and address other issues that did not fit into one of these categories but may be important to include in the final report.

- **Online Access** (e.g., develop/promote public access portals for checking voter registration status, to register/update registration information, etc.)

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

- **Electronic Transmission of Data** (encourage/require state agencies to share/transmit data electronically)
- **Voter Education** (e.g., encourage state agencies to remind voters in their communications to the public to check and update their information, etc.)
- **Matching** (e.g., improve matching procedures, establish software repositories of tested matching algorithms, etc.)
- **Process Improvements** (e.g., provide voter registration receipts, develop procedures for handling disenfranchisement caused by mistaken removals form voter registration lists, improve design of voter registration forms, etc.)
- **Privacy and Security**

1:00 p.m. **Small Group Discussions and Writing Sessions**
Continue thematic discussions and/or break into small groups to further refine focus and create outlines for the identified report sections.

3:00 p.m. Adjourn

WORKSHOP 5—DECEMBER 4-5, 2008 (ATLANTA, GEORGIA)

December 4, 2008

Closed Session

8:30 a.m. **Welcome and Overview**
Olene Walker, Committee and Workshop Co-chair

8:45 a.m. **Conflict and Bias Discussion—Ms. Denise Lamb**
Jon Eisenberg, CSTB Director

Open Session

9:00 a.m. **Welcome and Overview**
Olene Walker, Committee and Workshop Co-chair

9:10 a.m. **Remarks from the EAC Commissioners**

9:25 a.m. **Panel 1: VRD Experiences—Northeast and Midwest**
Discussion Facilitator: John Lindback

Panel Participants:

Joseph E. McLain, Help America Vote Act Administrator, Indiana
Barbara Hansen, Statewide Voter Registration System Director, Wisconsin
David Burgess, Deputy Secretary for Planning and Service Delivery, Pennsylvania
George Gilbert, Guildford County Elections Director, North Carolina

11:00 a.m. **Panel 2: VRD Experiences (cont.)—South**
Discussion Facilitator: Sarah Ball Johnson

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Panel Participants:

Donald Palmer, Elections Director, Florida
Marc Burris, IT Director, North Carolina
Wesley Tailor, Director of Elections, Georgia
Adam Thompson, HAVA Director, Alabama

12:25 p.m. Lunch available

1:00 p.m. **Panel 3: VRD Experiences (cont.)—Southwest**
Discussion Facilitator: Bruce McPherson

Panel Participants:

Ryan High, HAVA Administrator, Nevada
Kelli Fulgenzi, BOE Administrator, New Mexico
J. Wayne Munster, Director of Elections, Colorado

2:25 p.m. Break

2:40 p.m. **Panel 4: Technical Performance Assessments**
Discussion Facilitator: Rebecca Wright

Panel Participants:

Peggy Taff, Bureau Chief Voter Registration Services, Florida
Lani Smith, IT Manager, Nevada
Trevor Timmons, Chief Information Officer, Colorado
Patricia Lemus, Business Analyst/Technical Liaison, New Mexico
Adam Thompson, HAVA Director, Alabama
David Burgess, Deputy Secretary for Planning and Service Delivery, Pennsylvania

4:40 p.m. Break

4:50 p.m. **Panel 5: VRD Performance Successes and Challenges**
Discussion Facilitator: William Winkler

Panel Participants:

Adam Skaggs, Voting Rights Fellow, Brennan Center for Justice
Eric Fischer, Senior Specialist in Science and Technology, Congressional Research Service

5:45 p.m. Adjourn public workshop

Closed Session

6:15 p.m. **Working Dinner—Committee Members Only**

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

8:00 p.m. Adjourn working dinner

December 5, 2008

Closed Session

9:00 a.m. **Welcome and Overview**
Olene Walker, Committee and Workshop Co-chair

9:10 a.m. **Oregon & Washington Inter-state Interoperability Project**
John Lindback, Mike Alvarez, and Jeff Jonas

10:10 a.m. Break

11:20 a.m. **Discussion and Reflection**
Reactions to workshop and identifying issues to address in final report

12:15 p.m. Lunch

12:35 p.m. **Potential Revision to Project Statement of Task**
Herb Lin, NRC Staff

1:00 p.m. **Project Status Update**
Enita Williams, NRC Staff
Project timeline, workshop planning—dates, locations, themes, and writing schedule

1:15 p.m. **Developing an Outline for the Final Report**

3:30 p.m. Adjourn

WORKSHOP 6—MARCH 19-20, 2009 (BOSTON, MASSACHUSETTS)

March 19, 2009

Closed Session

8:00 a.m. **Welcome and Overview**
Olene Walker and Fran Ulmer, Committee and Workshop Co-chairs

Open Session

9:00 a.m. **Welcome and Overview**
Olene Walker and Fran Ulmer, Committee and Workshop Co-chairs

9:10 a.m. **Remarks from the EAC Commissioners**

9:30 a.m. **Panel 1: VRD Experiences**
Discussion Facilitator: John Lindback

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

- Describe the performance of your state's VRD systems in preparing for the November 2008 election.
- Were there notable successes? Challenges?
- How and to what extent did technical issues (including criteria for matching names) arise in adding new registrants to the approved list?
- How and to what extent did technical issues (including criteria for matching names) arise in performing list maintenance?
- How and to what extent did technical issues (including criteria for matching names) arise in election-related legal proceedings such as recounts and law suits?
- What are some of the looming census and redistricting issue your states face? Are there plans in place to manage redistricting changes through the state's VRD systems?

Panel Participants:

Michelle Tassinari, Elections Director, Massachusetts
Keryn Cadogan, Voter Registration Information System (VRIS) Applications Manager, Massachusetts
William Campbell, Clerk, Woburn Board of Registrars of Voters, Massachusetts
Robin Fields, Secretary of State's Office, Ohio
Eleanor Speelman, Legal Council, Secretary of State's Office, Ohio
Sarah Whitt, SVRS Functional Lead, Wisconsin
Herb Thompson, Department of Administration, Wisconsin

10:50 a.m. **Panel 2: Maintenance Challenges**
Discussion Facilitator: Denise Lamb

- What are the best practices/state of the art in VRD maintenance procedures?
- Have there been attempts at interstate collaboration? If so, describe these efforts and whether they were successful.
- What are some potential next steps to address technical challenges to interstate collaboration?

Panel Participants:

Deirdre Bishop, Assistant Chief, Census Redistricting Data Office, Bureau of the Census
Jay Varner, SABER/EDS
Adam Gigandet, Chief Information Officer, New York State DMV
Keryn Cadogan, Voter Registration Information System (VRIS) Applications Manager, Massachusetts
Sarah Whitt, SVRS Functional Lead, Wisconsin
William Campbell, Clerk, Woburn Board of Registrars of Voters, Massachusetts

12:00 p.m. Lunch available

12:45 p.m. **Sponsor Update—Related VRD Research Efforts**

- Results of a number of surveys EAC commissioned and mentioned at the last workshop

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

- Efforts in the pipeline, collaborations underway with relevance to VRD system maintenance and performance in order to avoid duplication of effort.

Karen Lynn-Dyson, Research Director, U.S. Election Assistance Commission

1:05 p.m. **Panel 3: Intra-State Social Services Collaboration**
Discussion Facilitator: Paula Hawthorn

- How are data shared/transmitted among the various intra-state organizations?
- What are some potential next steps to address technical challenges to intra-state collaboration?

Panel Participants:

Leesa Shem-Tov, NAPHSIS State and Territorial Exchange of Vital Events (STEVE) Project Manager
Adam Gigandet, Chief Information Officer, New York State DMV
Gregory Fulchino, Data Processing Manager/Jury Census Manager, Massachusetts Juror Commissioners Office
Brian Mellor, Senior Counsel, Project Vote

2:05 p.m. Break

2:15 p.m. **Panel 4: UOCAVA**
Discussion Facilitator: Sarah Ball Johnson

- What are some of the unique challenges in registering military and overseas voters?
- How are UOCAVA voters managed in various states VRD systems?
- What are some technological registration methods being explored for UOCAVA voters? What can state VRD managers learn from these experiments?
- Also learn more about specific activities such as the Alliance for Military and Overseas Voting Rights January 29th Meeting; Project SERVE; and efforts by the Pew Center for the States

Panel Participants:

Susan Dzieduszycka-Suinat, Overseas Voting Foundation (via conference call)
Carol Paquette, SERVE Project and Project BRAVO
Karl Cowart, Voting Information Officer, Hanscom Air Force Base
Bob Carey, Executive Director, National Defense Committee
Pat Hollarn, former Okaloosa County Supervisor of Elections, Florida (via conference call)
David Becker, Project Director-Making Voting Work, Pew Center on the States

4:15 p.m. Break

4:30 p.m. **Panel 5: New Administration, New Legislation Aims for Voter Registration**
Discussion Facilitator: Gary Cox

- Insights into what new legislation is in the pipeline in Congress for elections.
- What are some of the goals of the new administration with respect to VRD systems, HAVA funding, UOCAVA funding, etc.?

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Panel Participants:

Peter Schalestock, Legislative Counsel, Congressional Committee on House Administration—Minority side
Thomas Hicks, Legislative Counsel, Congressional Committee on House Administration—Majority side
Adam Ambrogi, Legislative Counsel, Senate Rules and Administration Committee —Majority side (invited)
Michael Merrell, Legislative Counsel, Senate Rules and Administration Committee —Minority side (invited)

5:45 p.m. Adjourn public workshop

Closed Session

6:15 p.m. **Working Dinner—Committee Members Only**

8:00 p.m. Adjourn working dinner

March 20, 2009

Closed Session

8:00 a.m. **Welcome and Overview**
Olene Walker and Fran Ulmer, Committee and Workshop Co-chair

8:45 a.m. **Discussion and Reflection**
Reactions to workshop and identifying issues to address in final report

10:15 a.m. Break

10:30 a.m. **Discussion and Reflection (continued)**
Brainstorming Exercise
Reactions to workshop and identifying issues to address in final report

11:30 a.m. **Break-Out Group or Individual Writing Session**

12:15 p.m. Lunch

12:45 p.m. **Developing an Outline for the Final Report**

2:00 p.m. Break

2:15 p.m. **Final Thoughts**

3:00 p.m. Adjourn

Copyright National Academy of Sciences. All rights reserved.

F

Biographical Information

COMMITTEE MEMBERS

Frances Ulmer, *Co-chair*, is the chancellor of the University of Alaska Anchorage, bringing to this position 30 years of experience in public policy in Alaska. Previously, she was a fellow at the Institute of Politics at Harvard University's Kennedy School of Government and a Distinguished Visiting Professor of Public Policy at the Institute of Social and Economic Research. In the early 1980s, she was the mayor of Juneau, then became a member of the Alaska House of Representatives (1986-1994), and in 1994 became the first female lieutenant governor of Alaska. In that year, she was appointed to the North Pacific Anadromous Fish Commission by President Bill Clinton and served on this international board for 11 years. She has participated in numerous panels, task forces, commissions, and forums as a speaker, moderator, and panelist to address the intersection of science, economics, politics, and policy. She currently serves on the Board of Trustees of the National Parks Conservation Association, the Advisory Board of the Union of Concerned Scientists, and the Alaska Nature Conservancy Board. At the national level, Ms. Ulmer has served as a member of the above-mentioned North Pacific Anadromous Fish Commission, the Federal Communications Commission's State and Local Advisory Committee, and the Federal Elections Commissions Committee. She has a B.A. in political science and economics and a Law Degree from the University of Wisconsin.

Olene Walker, *Co-chair*, was the first woman governor of the state of Utah. Before being appointed as governor, she served as the first woman lieutenant governor of Utah. During her time in office, Dr. Walker spearheaded many important initiatives, including education programs, budget security measures, health care reform, and workforce development. She also worked to implement the federal "motor voter" legislation in Utah and oversaw the plan to bring Utah into compliance with the Help America Vote Act (HAVA). She has chaired the National Conference of Lieutenant Governors and is a past president of the National Association of Secretaries of State. She was the first lieutenant governor ever to serve as the president of that organization. Dr. Walker received her bachelor's, master's, and doctoral degrees from Brigham Young University, Stanford University, and the University of Utah, respectively.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Rakesh Agrawal, NAE, is a Microsoft Technical Fellow at the newly founded Search Labs. His areas of expertise are in developing fundamental data mining concepts and technologies and pioneering key concepts in data privacy, including Hippocratic Database, Sovereign Information Sharing, and Privacy-Preserving Data Mining. He is the recipient of the ACM-SIGKDD First Innovation Award, ACM-SIGMOD Edgar F. Codd Innovations Award, ACM-SIGMOD Test of Time Award, VLDB 10-Year Most Influential Paper Award, and the Computerworld First Horizon Award. He is a member of the National Academy of Engineering, a fellow of the Association for Computing Machinery, and a fellow of IEEE. *Scientific American* named him to the list of 50 top scientists and technologists in 2003. Prior to joining Microsoft in March 2006, Dr. Agrawal was an IBM fellow and led the Quest group at the IBM Almaden Research Center. Earlier, he was with the Bell Laboratories, Murray Hill, from 1983 to 1989. He also worked for 3 years at India's premier company, the Bharat Heavy Electricals Ltd. He received M.S. and Ph.D. degrees in computer science from the University of Wisconsin-Madison in 1983. He also holds a B.E. degree in electronics and communication engineering from IIT-Roorkee, as well as a 2-year postgraduate diploma in industrial engineering from the National Institute of Industrial Engineering (NITIE), Bombay.

R. Michael Alvarez is a professor of political science at the California Institute of Technology (CalTech). His research interests have been in the areas of elections and electoral behavior, survey methodology, statistics and political methodology, and more recently, election administration. Professor Alvarez is currently the co-director of the Caltech/MIT Voting Technology Project and recently co-authored a book published by the Brookings Institution Press, *Point, Click and Vote: The Future of Internet Voting*. Professor Alvarez received his Ph.D. and M.A. degrees in political science from Duke University and his B.A., magna cum laude, in political science from Carleton College.

Gary W. Cox, NAS, is a professor of political science at the University of California, San Diego. In addition to numerous articles in the areas of legislative and electoral politics, Professor Cox is author of *The Efficient Secret* (winner of the Samuel H. Beer dissertation prize in 1983 and of the 2003 George H. Hallett Award), coauthor of *Legislative Leviathan* (winner of the Richard F. Fenno Prize in 1993), author of *Making Votes Count* (winner of the Woodrow Wilson Foundation Award, the Luebbert Prize, and the Best Book in Political Economy Award in 1998), and coauthor of *Elbridge Gerry's Salamander: The Electoral Consequences of the Reapportionment Revolution*. His latest book, *Setting the Agenda*, was published in 2005. A former Guggenheim Fellow, Professor Cox was elected to the American Academy of Arts and Sciences in 1996 and to the National Academy of Sciences in 2005. He received a Ph.D. from the California Institute of Technology in 1983.

Paula Hawthorn, retired, serves as a consultant and continues her involvement with the University of California, Berkeley. She received her Ph.D. in electrical engineering and computer science from the University of California in 1979. Her thesis topic was on the performance of database systems. She has spent much of her career as a manager of database development, including vice-president of Software Development for start-ups such as Britton Lee and Illustra, and both management and individual contributor positions at Hewlett-Packard (working on database performance issues) and Lawrence Berkeley National Laboratory.

Sarah Ball Johnson currently serves as the executive director of the Commonwealth of Kentucky's State Board of Elections. She has 15 years of experience in election administration on the state level. She has a bachelor of arts degree in business administration from Transylvania University and a master of public administration degree, specializing in state and local government, from the University of Kentucky. She participated in four international election observation trips, to Slovakia, Kosovo, Macedonia, and Nigeria. She is a member of the National Association of State Election Directors and serves as the secretary of the association's executive board. She serves on the Election Assistance Commission Board

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

of Advisors and serves on the Election Assistance Commission Standards Board. She is a member of the Election Center.

Jeff Jonas is a distinguished engineer and chief scientist of Entity Analytic Solutions at IBM. He is responsible for shaping the overall technical strategy of next-generation identity analytics and the use of this new capability in the overall IBM technology strategy. The IBM Entity Analytic Solutions group was formed based on technologies he developed as the founder and chief scientist of Systems Research & Development (SRD). SRD was acquired by IBM in January 2005. He applies his real-world experience in software design and development to drive technology innovations while delivering higher levels of privacy and civil liberties protections. He is a member of the Markle Foundation Task Force on National Security in the Information Age and actively contributes on issues of privacy, technology, and homeland security to leading national think tanks, privacy advocacy groups, and policy research organizations, including the Center for Democracy and Technology, Heritage Foundation, Center for Strategic and International Studies, and the Office of the Secretary of Defense Highlands Forum.

Denise Lamb is the chief deputy clerk for Elections in Santa Fe County, New Mexico, and has held that position for four years. She began her work in election administration in 1991 as a legislative analyst for the New Mexico Secretary of State and in 1993 was responsible for the implementation of the National Voter Registration Act in the state. Ms. Lamb was named as State Election Director in 1994 and held that position until 1997, returning in 1999-2004. Denise Lamb is a past-president of the National Association of State Election Directors and was co-chair of that group's Voting Systems/Independent Test Authority Accreditation Board. Ms. Lamb was New Mexico's Election Director during the state's project to transition to a statewide voter file, beginning in 1999 and finishing after the passage of the Help America Vote Act. Ms. Lamb's responsibilities in Santa Fe County include supervision of all local, state, and federal elections held within the jurisdiction which has 88,500 registered voters. She also works on Native American voting rights issues, poll worker training, and with the legislature and county clerks on election legislation.

John Lindback is a senior officer for voter registration modernization at the Pew Charitable Trusts' Pew Center on the States. From March 2001 to July 2009, he served as director of elections for the state of Oregon. His duties for the state included enforcing laws governing the conduct of elections in Oregon, enforcing Oregon's campaign finance laws, administering the state's initiative and referendum process, and publishing state voters' pamphlets. Previously, he worked for 6 years as chief of staff for the lieutenant governor of Alaska, a job that included administrative oversight of Alaska's statewide election system. In 2008 he served as president of the National Association of State Elections Directors (NASED). He also served as an Oregon representative to the U.S. Election Commission's Standards Advisory Board and he serves as an advisor to Design for Democracy, an organization promoting better designs of ballots and other elections materials. His other experience in the public sector includes work as budget analyst, legislative finance aide, and public information officer. Mr. Lindback's first career was newspaper reporting. After earning a journalism degree from the University of Arizona in 1976, he reported on politics and government for newspapers for 12 years.

Bruce McPherson was the 30th California secretary of state. The first 26 years of his career he worked in the newsroom of the family-owned *Santa Cruz Sentinel*, serving as sports editor, news reporter, city editor, and editor. During this time he served on, and was president of, numerous community organizations. In his 11 years in the California legislature, he focused his attention on education, environmental protection, and public safety. In the aftermath of the resignation in early 2005 of California's secretary of state, he was nominated by Governor Arnold Schwarzenegger to be secretary of state. Mr. McPherson was confirmed unanimously in both the Senate and the Assembly. While in office, he updated the information technology required to meet election laws, and he oversaw three statewide elections and

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

two special elections. Mr. McPherson graduated from Cal Poly–San Luis Obispo with a B.S. degree in journalism in 1965. He subsequently was given an honorary degree in humane letters from Cal Poly–San Luis Obispo in 2005.

Wendy Noren is county clerk of Boone County, Missouri, a position she has held since 1982, and she managed the election division of the office for 4 years prior to that. Ms. Noren is responsible for keeping records of the orders, rules, and proceedings of the County Commission. In addition, she is responsible for inspecting and reviewing all voter precinct boundaries within the county and conducting elections. Throughout this period, she has served as a programmer for all of the voter registration functions. Over the past 25 years, she has been one of the first to implement emerging technology for the county's voter registration system—often years before most jurisdictions. As both the programmer and user, she has a unique perspective on the critical components of a voter registration system. Other administrative responsibilities of the clerk include maintaining payroll files, administering employee benefits, administering the records management budget, and procuring adequate insurance and bonding for the county's assets and elected officials.

William Winkler is a principal researcher with the U.S. Census Bureau. He is a fellow of the American Statistical Association. He has published more than 130 papers and has developed eight (and counting) generalized computer systems for record linkage, edit/imputation, multipurpose and multiway sampling, text classification, and masking for public-use microdata. Dr. Winkler holds a Ph.D. in probability theory from Ohio State University.

Rebecca N. Wright is an associate professor of computer science at Rutgers University. She is also deputy director of the DIMACS Center for Discrete Mathematics and Theoretical Computer Science. Prior to that, she was a professor of computer science at Stevens Institute of Technology and a researcher in the Secure Systems Research Department at AT&T Labs and AT&T Bell Labs. Her research spans the area of information security, including cryptography, privacy, foundations of computer security, and fault-tolerant distributed computing. She was a co-author of a study, "Statewide Databases of Registered Voters: Study of Accuracy, Privacy, Usability, Security, and Reliability Issues," commissioned by USACM. She was an invited speaker in the National Academy of Engineering's 2007 U.S. Frontiers of Engineering Symposium. She received a Ph.D. in Computer Science from Yale University in 1994 and a B.A. from Columbia University in 1988.

CSTB STAFF

Herbert S. Lin is chief scientist, CSTB at the Computer Science and Telecommunications Board, National Research Council of the National Academies, where he has been the study director for major projects on public policy and information technology. These studies include a 1996 study on national cryptography policy (*Cryptography's Role in Securing the Information Society*), a 1991 study on the future of computer science (*Computing the Future*), a 1999 study of Defense Department systems for command, control, communications, computing, and intelligence (*Realizing the Potential of C4I: Fundamental Challenges*), a 2000 study on workforce issues in high technology (*Building a Workforce for the Information Economy*), a 2002 study on protecting kids from Internet pornography and sexual exploitation (*Youth, Pornography, and the Internet*), a 2004 study on aspects of the FBI's information technology modernization program (*A Review of the FBI's Trilogy IT Modernization Program*), a 2005 study on electronic voting (*Asking the Right Questions About Electronic Voting*), and a 2005 study on computational biology (*Catalyzing Inquiry at the Interface of Computing and Biology*). Prior to his NRC service, he was a professional staff member and staff scientist for the House Armed Services Committee (1986-1990), where his portfolio included defense policy and arms control issues. He received his doctorate in physics from MIT.

Copyright National Academy of Sciences. All rights reserved.

Improving State Voter Registration Databases: Final Report

Kristen R. Batch was an associate program officer for the Computer Science and Telecommunications Board of the National Academies through August 2008. Since joining CSTB in 2002, she worked on studies that produced *Toward a Safer and More Secure Cyberspace*, *Engaging Privacy and Information Technology in a Digital Age*, *Asking the Right Questions About Electronic Voting*, *Signposts in Cyberspace: The Domain Name System and Internet Navigation*, *A Review of the FBI's Trilogy Information Technology Modernization Program*, and *The Internet Under Crisis Conditions: Learning from September 11*. While pursuing an M.A. in international communications from American University, she interned at the National Telecommunications and Information Administration, in the Office of International Affairs, and at the Center for Strategic and International Studies, in the Technology and Public Policy Program. She also received a B.A. from Carnegie Mellon University in literary and cultural studies and Spanish, and she received two travel grants to conduct independent research in Spain.

Morgan Motto, program associate, was with CSTB from December 2007 until April 2009 supporting several projects. Previously, she worked with the Board on Environmental Studies and Toxicology (BEST). Prior to coming to the NRC, Ms. Motto worked as a project manager for international affairs and technology at the U.S. Pan Asian American Chamber of Commerce. She earned a B.A. in international affairs and East Asian studies from the Elliott School of International Affairs at the George Washington University.

Enita Williams is an associate program officer with the Computer Science and Telecommunications Board of the National Academies. She formerly served as a research associate for the Air Force Studies Board of the National Academies where she supported a number of projects including a standing committee for the Special Operations Command (SOCOM) and a standing committee for the intelligence community (TIGER). Prior to her work at the National Academies, she served as a program assistant with the Scientific Freedom, Responsibility and Law Program of AAAS, where she drafted the human enhancement workshop report. Ms. Williams graduated from Stanford University with a B.A. in public policy with a focus on science and technology policy, and an M.A. in communications.

Eric Whitaker is a senior program assistant at the Computer Science and Telecommunications Board of the National Academies. Prior to joining the CSTB, he was a realtor with Long and Foster Real Estate, Inc., in the Washington, D.C., metropolitan area. Before that, he spent several years with the Public Broadcasting Service in Alexandria, Virginia, as an associate in the Corporate Support department. He has a B.A. in communication and theater arts from Hampton University.

Copyright National Academy of Sciences. All rights reserved.