UIC John Marshall Journal of Information Technology & Privacy Law

Volume 13
Issue 4 *Journal of Computer & Information Law - Summer 1995*

Article 8

Summer 1995

A Normative Regulatory Framework for Computer Matching, 13 J. Marshall J. Computer & Info. L. 585 (1995)

Roger A. Clarke

A NORMATIVE REGULATORY FRAMEWORK FOR COMPUTER MATCHING

by ROGER A. CLARKE†

Computer matching is any computer-supported process in which personal data records relating to many people are compared in order to identify cases of interest.[1]

ABSTRACT

Computer matching is a powerful data surveillance tool widely used by government agencies since its emergence in 1976. Computer matching involves the merger of data from multiple sources: data gathered for different purposes, subject to different definitions, and of variable quality. It is a mass *dataveillance* technique[2], for its purpose is to generate suspicions that errors, misdemeanors or fraud have occurred. For many years, computer matching activities were carried on in semi-secrecy. The purpose of this paper is to propose a framework within which effective regulation can be imposed on this dangerous technique.

This article commences by providing background to computer matching's origins and nature. Its impacts are then discussed, in order to establish that there is a need for controls. Intrinsic controls are assessed, and found wanting. A set of features for a satisfactory external control regime is then presented. This article provides a basis for evalu-

† Roger A. Clarke is the Reader in Information Systems in the Department of Commerce at the Australian National University. He has been involved in information privacy and data surveillance matters since 1972, holding the positions of Director of the Australian Computer Society's Community Affairs board, and Vice-Chairman of the Australian Privacy Foundation.

The research underlying this paper was undertaken during a protracted period. The author's literature search and analysis are supplemented by considerable field work, undertaken primarily in the United States and Australia.

1. Roger A. Clarke, *Computer Matching and Digital Identity*, Address at *Proc. Conf. Computers, Freedom & Privacy, Ass. Comp. Machinery* (San Francisco, March 1993).

2. Symposium, *Surveillance, Dataveillance, and Human Freedom*, 4 COL. HUM. RTS. L. REV. 1 (1972).

ating the protective measures which are in force in at least four jurisdictions, and guidance for legislators in others.

I. INTRODUCTION

Computer matching is the comparison of machine-readable records containing personal data relating to many people, in order to detect cases of interest. The technique is called *computer matching* in the United States, and *data matching* in Canada, Australia and New Zealand.[3]

Computer matching became economically feasible in the early 1970s, as a result of developments in informational technology. The technique has been progressively developed since then, and is now widely used, particularly in government administration and particularly in the four countries mentioned above. It has the capacity to assist in the detection of error, abuse and fraud in large-scale systems, but in so doing may jeopardize the information privacy of those whose data is involved, and even significantly alter the balance of power between consumers and corporations, and between citizens and State.

This article's scope is restricted to computer matching in the public sector. This is because, first, the technique's primary applications are in these areas; second, access to both primary and secondary sources is even more difficult in the private sector than it is in relation to government; and third, the issues which arise in the private sector are somewhat different, and justify separate treatment elsewhere. Throughout this article, reference is made to circumstances in the United States and Australia. The comments made are at least generally applicable in Canada and New Zealand, which are also advanced in their use of the technique. On the basis of research undertaken in Europe, computer matching appears to have been less actively applied there. Because of differences in legal contexts (including civil rights), and the patterns of information technology use in personal data systems, the application of this paper outside the four countries mentioned requires care.

Background is provided concerning the origins of computer matching, and the nature of this technique. Consideration is then given to its impacts and implications, and the natural or intrinsic controls which act to limit its use and to ensure that where it is used, it is used fairly. The conclusion is that intrinsic controls are largely ineffectual, and that external controls need to be imposed. The final section of the article proposes a framework to support assessment of the adequacy of existing control regimes and to design of regulatory regimes in countries which do not yet have them.

---

3. J.B. RULE, PRIVATE LIVES AND PUBLIC SURVEILLANCE: SOCIAL CONTROL IN THE COMPUTER AGE (Schocken Books 1974).

II. BACKGROUND TO COMPUTER MATCHING

Computer matching is an outgrowth of the increasing data intensity of public administration during the twentieth century. Conventional surveillance mechanisms, such as direct observation and electronic eavesdropping, have been increasingly supplemented and even supplanted by the monitoring of people through their data. Dataveillance is the "systematic use of personal data systems in the investigation or monitoring of the actions or communications of one or more persons."[4] Two kinds of dataveillance need to be distinguished: personal dataveillance, in which an identified person in monitored, generally for a specific reason; and mass dataveillance, in which groups of people are monitored, generally to identify individuals by interest. Notably, dataveillance differs from conventional surveillance in several respects; for example, it is less directly intrusive, and is much cheaper. A rich selection of dataveillance literature is available for an in depth review.[5]

Computer matching has been the subject of surprisingly limited literature. There are government reports of various kinds, but little of an academic nature. Neither the term nor its equivalent can be found in landmark documents on informational technology,[6] including the U.S.

---

4. Roger A. Clarke, *Information Technology and Dataveillance*, 31 COMMUNICATIONS OF THE ACM 5 (May 1988), republished in C. DUNLOP AND R. KLING, COMPUTERS AND CONTROVERSY (1991).

5. K.C. LAUDON, COMPUTERS AND BUREAUCRATIC REFORM (Wiley 1974); R.E. Smith (Ed.), PRIVACY JOURNAL, November 1974; Robert E. Smith and Robert D. Snyder, (Eds.) *Compilation of State and Federal Privacy Laws*, PRIVACY JOURNAL, (annual ed.); A.F. WESTIN, M. BAKER, AND PROJECT ON COMPUTER DATABANKS, DATABANKS IN A FREE SOCIETY: COMPUTERS, RECORD KEEPING, AND PRIVACY, (Quadrangle Books 1974); R. Kling, *Automated Welfare Client Tracking and Welfare Service Integration: The Political Economy of Computing*, 21 COMMUNICATIONS OF THE ACM 6, 484-93 (June 1978); JAMES RULE, DOUGLAS MCADAM, LINDA STEARNS AND DAVID UGLOW, THE POLITICS OF PRIVACY (New American Library 1980); DAVID BURNHAM, THE RISE OF THE COMPUTER STATE, (Random House 1983); J.B. Rule, *Documentary Identification and Mass Surveillance in the United States*, 31 SOCIAL PROBLEMS 222 (1983); G.T. Marx and N. Reichman, *Routining the Discovery of Secrets*, 27 AM. BEHAV. SCIENTIST 4, 423-52 (1984); G.T. Marx, *Surveillance: A Dangerous Game Played With Matches*, 4 ABACUS 1, 60-64 (1986); T. ROSZAK, THE CULT OF INFORMATION (Pantheon 1986); U.S. Congress, Office of Technology Assessments, *Federal Government Information Technology: Electronic Record Systems and Individual Privacy*, OTA-CIT-296 (U.S. Government Printing Office 1986); KENNETH C. LAUDON, DOSSIER SOCIETY: VALUE CHOICES IN THE DESIGN OF NATIONAL INFORMATION SYSTEMS, (Columbia University Press 1986); Roger A. Clarke, *Just Another Piece of Plastic for Your Wallet: The Australia Card*, 5 PROMETHEUS 1 (1987) (republished in 18 COMPUTERS & SOCIETY 1 (1988), addendum, 18 COMPUTERS & SOCIETY 3 (1988)); D. H. FLAHERTY, PROTECTING PRIVACY IN SURVEILLANCE SOCIETIES, (University of North Carolina Press 1989); COLIN J. BENNETT, REGULATING PRIVACY: DATA PROTECTION AND PUBLIC POLICY IN EUROPE AND THE UNITED STATES (Cornell University Press 1992).

6. *See, e.g.*, A. F. WESTIN, PRIVACY AND FREEDOM (Atheneum 1967); A. F. WESTIN, INFORMATION TECHNOLOGY IN A DEMOCRACY (Harvard University Press 1971); U.S. Depart-

Privacy Act of 1974,[7] and other national studies.[8] The term came into currency following publication of descriptions of "Project Match" conducted in 1977 by the Department of Health, Education & Welfare ("HEW" now the Department of Health and Human Services).[9] The development of computer matching techniques and operations can be traced through a series of important published works.[10]

---

ment of Health, Education and Welfare, Report of the Secretary's Advisory Committee on Automated Personal Data Systems, *Records, Computers and the Rights of Citizens* (MIT Press 1973).

7. Privacy Act of 1974, Pub. L. 100-503, Oct. 18, 1988, 5 U.S.C.A. §552a (West 1995).

8. *See, e.g.*, U.S. Federal Advisory Committee on False Identification, *The Criminal Use of False Identification* (U.S. Govt. Printing Office 1976); Privacy Protection Study Commission, *Personal Privacy in an Information Society* (U.S. Govt. Printing Office, 1977); New South Wales Privacy Committee, *Guidelines for the Operation of Personal Data Systems* (1977); Sir Norman Lindop, *Report of the Committee on Data Protection*, Cmnd. 7341, HMSO, London (December 1978); Organization for Economic Cooperation and Development, *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data* (O.E.C.D. Pub. and Info. Center 1980).

9. Office of Management and Budget, President's Commission for Integrity and Efficiency, *Guidelines to Agencies on Conducting Automated Matching Programs* (U.S. Govt. Printing Office 1979); Office Management and Budget, *Privacy Act of 1974: Supplemental Guidance for Matching Programs*, 44 Fed. Reg. 23, 138 (1979).

10. The U.S. Office of Management and Budget, Office of Technologies Assessments and Department of Health and Human Services documents have published extensively in this area. *See, e.g.*, Office of Management and Budget, *Computer Matching Guidelines* (U.S. Govt. Printing Office 1982); Office of Management and Budget, *Privacy Act of 1974: Revised Supplemental Guidance for Conducting Matching Programs*, 47 Fed. Reg. 21, 656 (1982); Office of Management and Budget and President's Commission for Integrity and Efficiency, *Model Control System for Conducting Computer Matching Projects Involving Individual Privacy Data* (U.S. Govt. Printing Office 1983); Department of Health and Human Services ("DHHS"), *Computer Matching in State Administered Benefit Programs: A Manager's Guide to Decision-Making* (U.S. Govt. Printing Office 1983); DHHS, *Inventory of State Computer Matching Technology* (U.S. Govt. Printing Office 1983); DHHS, *Nationwide Impact of Verifying Resources by Matching Recipients of Public Assistance with Bank Records* (U.S. Govt. Printing Office 1983); U.S. Department of Health and Human Services, *Computer Matching in State Administered Benefit Programs*, (U.S. Govt. Printing Office 1984).

Canadian and Australian studies provide additional insight. *See, e.g.*, PRIVACY, Australian Law Reform Comm., Rep. No. 22 (1983); Australian Public Service, *Review of Systems for Dealing with Fraud on the Commonwealth* (March 1987); Privacy Commissioner, *Data Matching Review: A Resource Document for Notification of the Privacy Commissioner of Proposed Data Matches* 112 PRIVACY COMMISSIONER (Kent, Ottowa, Canada 1989); Privacy Commissioner, Human Rights and Equal Opportunities Commission, *Data Matching in Commonwealth Administration: Discussion Paper and Draft Guidelines* 56 (available from: Privacy Commissioner, G.P.O. Box 5218, Sydney, Australia) (October 1990); K. O'Connor, *Paper for the Communications & Media Law Association* (available from: Privacy Commissioner, Human Rights & Equal Opportunities Commission, G.P.O. Box 5218, Sydney) (April 1990).

Other seminal works in the field of information technology include: R.P. Kusserow, *The Government Needs Computer Matching to Root out Waste and Fraud*, 27 COMMUNICATIONS

During the 1980s, many hundreds of matching programs have been conducted, particularly by federal government agencies. Areas of application include the administration of taxation, social welfare, housing and health insurance.[11] In the United States, fiscal pressure was brought to bear on the states to ensure that they conducted and participated in matching programs.[12]

There is no authoritative definition of the term *computer matching*. Following extensive examination and analysis of the meaning of the term, both in literature and practice, the author adopted the following definition:

> [c]omputer matching is any computer-supported process in which personal data records relating to many people are compared in order to identify cases of interest.[13]

The procedure involved varies, but the model in Exhibit 1 is sufficiently rich to enable the analysis of the vast majority of programs. The steps in computer matching are:

(a) from each database being used as a source for the program, *selection* of all (or a sub-set) of the available data-items from all (or a sub-set) of the available records;

(b) optionally, "*data scrubbing*" operations, to change the organization, format and/o content of one or more of more of the files into a form suitable for the matching step;

(c) "*matching*," whereby a "matching algorithm" is applied to the files of personal data records in order to find "raw hits" or "matches." Generally, these are matched pairs of records which are deemed to refer to the same data subject. Alternatively, the algorithm may involve a search for records on one file for which a match on the other cannot be found;

(d) "*inferencing*", which involves an "inferencing procedure" being applied to the outcome of the matching process (i.e., either to the contents of matched pairs of records, or to the existence or the non-existence of matches). The purpose of this step is to draw conclusions about the person to whom the data purports to relate, or to his or her behavior, actions or proclivities;

OF THE ACM 6, 542-45 (1984); Department of Health and Human Services, Office of the Inspector-General, *Computer Matching in State-Administered Benefit Programs* (U.S. Govt. Printing Office 1984); J. Shattuck, *Computer Matching is a Serious Threat to Individual Rights*, 27 COMMUNICATIONS OF THE ACM 6, 538-41 (1984).

11. *See* U.S. Department of Health and Human Services, *Computer Matching in State Administered Benefit Programs*, (U.S. Govt. Printing Office 1984).

12. *See* R.P. Kusserow, *supra* note 10 at 542-45.

13. Roger A. Clarke, *Computer Matching and Digital Identity*, Address at *Proc. Conf. Computers, Freedom & Privacy, Ass. Comp. Machinery* (San Francisco, March 1993).

(e) *"filtering"* of the "raw hits" to produce "solid hits," to ensure efficient use of investigative resources and to avoid unjustified administrative action;

(f) *analysis* of the resulting information, and *decisions and action* arising from it;

(g) optionally, *creation of new records or amendments or extensions to existing records*; and

(h) optionally, *quality analysis* activities may generate feedback from any and all stages, requiring a return to earlier steps.[14]

III. THE NEED FOR REGULATION

The contention of this paper is that external controls over the practice of computer matching are essential. This section presents an analysis of the impact of computer matching, first at an abstract and then a more detailed level, concluding that these impacts are very substantial. The section assesses factors which act as natural or intrinsic controls, and concludes that they are inadequate.

This article does not assert that computer matching is valueless or should never be applied. There is a limited literature on the benefits (financial and otherwise) of computer matching, and very little of what is available results from careful and independent assessment.[15] This pa-

---

14. *Id.* A detailed history and description of the technique is found in an earlier article by the author. *See* Roger A. Clarke, *Dataveillance by Governments: The Technique of Computer Matching*, 7 INFORMATION TECHNOLOGY & PEOPLE 2, 46-85 (June 1994).

15. Office of Management and Budget and President's Commission for Integrity & Efficiency, *supra* note 10 at 4-27; Department of Health and Human Services ("DHHS"), *Computer Matching in State Administered Benefit Programs: A Manager's Guide to Decision-Making* (U.S. Govt. Printing Office, 1983); DHHS, *Inventory of State Computer Matching Technology* (U.S. Govt. Printing Office 1983); DHHS, *Nationwide Impact of Verifying Resources by Matching Recipients of Public Assistance with Bank Records* (U.S. Govt. Printing Office 1983); Comptroller General of the United States, General Accounting Office, *Action Needed to Reduce, Account For, and Collect Overpayments to Federal Retirees* (U.S. Govt. Printing Office 1983); Comptroller General of the United States, General Accounting Office, *Computer Matches Identify Potential Unemployment Benefit Overpayments* (U.S. Govt. Printing Office 1983); Comptroller General of the United States, General Accounting Office, *GAO Observations on the Use of Tax Return Information for Verification in Entitlement Programs* (U.S. Govt. Printing Office 1984); General Accounting Office, *Better Wage-Matching Systems and Procedures Would Enhance Food Stamp Program Integrity* (U.S. Govt. Printing Office 1984); D. H. Greenberg and D. A. Wolf, *Is Wage Matching Worth All the Trouble?*, PUBLIC WELFARE, 13-30 (1985); D. H. GREENBERG, D. A. WOLF, AND J. PFIESTER, USING COMPUTERS TO COMBAT WELFARE FRAUD: THE OPERATION AND EFFECTIVENESS OF WAGE MATCHING (Greenwood Press 1986); General Accounting Office, *A Central Wage File for Use by Federal Agencies: Benefits and Concerns*, GAO/HRD-85-31 (1985); General Accounting Office, *Social Security: Pensions Data Useful for Detecting Supplemental Security Payment Errors*, GAO/HRD-86-32, 14 (1986); General Accounting Office, *Computer Matching: Assessing Its Costs and Benefits*, GAO/PEMD-87-2, 102 (1986); General Accounting

per focuses on the negative aspects of this process.

A. The Abstract Impacts of Computer Regulation

There is a large and highly varied literature expressing cautionary, critical and in some cases, hysterical comments about dataveillance in general and computer matching in particular.[16] This sub-section outlines the more general arguments regarding the power relationship between the citizen and the State, arbitrary interference, due process, the loss of data context and social context, social equity considerations and risks to public morality and law and order.

(1) *The Power Relationship Between Citizen and State*

Computer matching has been in many cases undertaken without knowledge or consent of any of the many people to whom the data relates. Data subjects have had no clear right to prevent this happening, because until 1988 no laws creating such a right existed in either the United States or Australia.[17]

Even where consent is sought, it is often not meaningful consent. In some cases, the wording of a consent form provides an individual with no appreciation of the import and consequences of giving consent. There are also instances in which "pseudo-consent" forms are used; for exam-

---

Office, *Computer Matching: Factors Influencing the Agency Decision-Making Process*, GAO/PEMD-87-3BR, 30 (1986); N. Holden, J. A. Burghardt, and J. C. Ohls, *Final Report for the Evaluation of the Illinois On-Line Cross-Match Demonstration*, Mathematica Policy Research (reporting for U.S. Dept. of Health and Human Services, Office of Family Assistance) (1987); General Accounting Office, *Welfare Eligibility: Deficit Reduction Act Income Verification Issues*, GAO/HRD-87-79FS, 93 (1987); General Accounting Office, *Veterans' Pensions: Verifying Income with Tax Data Can Identify Significant Payment Problems*, GAO/HRD-88-24, 100 (1988); D. Greenberg and R. Yudd, *Food Stamp Program Operations Study: Computer Matching: A Review of Exemplary State Practices in the FSP* (report for the U.S. Dept. of Agriculture, Food and Nutrition Service, available from: The Urban Institute, 2100 M. Street NW, Washington D. C. 20037) (July 1989); General Accounting Office, *Veterans' Benefits: VA Needs Death Information From Social Security to Avoid Erroneous Payments*, GAO/HRD-90-110, 14 (1990); General Accounting Office, *Federal Benefit Payments: Agencies Need Death Information From Social Security to Avoid Erroneous Payments*, GAO/HRD-91-3, 23 (1991); General Accounting Office, *Welfare Benefits: States Need Social Security's Death Data to Avoid Payment Error or Fraud*, 13, GAO/HRD-91-73 (1991).

16. K. J. Langan, *Computer Matching Programs: A Threat to Privacy?* 15 Colum. J. L. & Soc. Probs. 2 (1979); J. Kirchner, *Privacy: A History of Computer Matching in Federal Government Programs*, Computerworld (December 14, 1981); N. Reichman and G.T. Marx, *Generating Organizational Disputes: The Impact of Computerization*, Address before the Proc. Law & Society Ass. Conf., (transcript available from the author) San Diego, June 6-9, 1985; S. Davies, Big Brother: Australia's Growing Web of Surveillance (Simon & Schuster 1992).

17. Notably, the U.S. Computer Matching and Privacy Protection Act of 1988 laid a foundation for the process in the United States. Computer Matching and Privacy Protection Act of 1988, Pub. L. 100-503, Oct. 18, 1988, 5 USCA §552a (West 1995).

ple, where agencies seek an "authority" from benefit recipients (ostensibly to enable the agency to gather information from another agency) when they have legal authority to gather it. The absence of effective subject knowledge and consent mechanisms can be expected to result in individuals distrusting organizations, and in the provision of low-quality data.

In many cases the bargaining position is so disproportionately weighted in favor of the agency that the data subject has no real option but to comply; for example, the provision of the person's identification code issued by some other agency (such as the Social Security Number in the United States, the Social Identification Number in Canada, or the Tax File Number in Australia) may be made a condition of application for a benefit, specifically to enable front-end verification and matching.

The scope that computer matching offers to cross-system enforcement has also been discovered by important agencies:

> [u]se of the data-matching arrangements would enable each agency to identify cases where recovery of an outstanding debt owing to their agency was becoming difficult and that an income support payment was being made by one of the related agencies. Action could then be taken to recover the outstanding debt by withholding an amount from the person's current financial assistance payment.[18]

The legislative authorization which exists for such recovery was passed in an age before information technology had delivered efficient mass dataveillance capabilities.

Automated maintenance of databases using data arising from matching (for example, cross-notification among agencies of a data subject's change of address) is consistent with a paternalistic approach by the State to the lives of its citizens. Associated with this paternalism is a loss of initiative:

> [o]nce people start fearing the government, once they think they are under surveillance by government, whether they are or not, they are likely to refrain from exercising the great rights that are incorporated in the First Amendment to make their minds and spirits free.[19]

The "chilling effect" of mass data surveillance could cause a society to lose its morality, its creativity, and its resilience. With the recent increased understanding of the depths of despair to which residents of Eastern European Communist dictatorships plunged in recent years, this "chilling effect" ceases to be mere sociologist's jargon, and becomes a palpable factor. Put more colorfully,

---

18. Department of Social Security and the Data Matching Agency, *Data-Matching Program (Assistance and Tax): Report on Progress*, 41 (October 1991).

19. Remarks by Senator S. Ervin during the U.S. Senate's 1971 Hearings into Federal Data Banks, Computers and the Bill of Rights *quoted* in Laudon, *supra* note 5 at 357.

> [w]hat we confront in the burgeoning surveillance machinery of our society is not a value-neutral technological process . . . . It is, rather, the social vision of the Utilitarian philosophers at last fully realized in the computer. It yields a world without shadows, secrets or mysteries, where everything has become a naked quantity.[20]

(2) *Arbitrary Interference*

Matching involves trawling through data which refers to large populations, in order to identify relatively small numbers of individuals who may have committed an error, default, dishonesty or fraud. Many commentators have addressed this issue, making the following criticisms:

- data matching programs are old-fashioned fishing expeditions posing as high technology;[21]
- data matching has been criticized as simply a modern version of the general warrant once used to search homes without discrimination or reasonable cause;[22]
- computer matches are inherently mass or class investigations, as they are conducted on a category of people rather than on specific individuals;[23]
- computer matching is like investigators entering a home without any warrant or prior suspicion, taking away some or all of the contents, looking at them, and keeping copies of what is of interest, all without the knowledge of the occupier;[24] and
- data matching is the information society's equivalent of driftnet fishing.[25]

Such concerns have not only been voiced by privacy advocates. The U.S. Civil Service Commission's General Counsel opposed the original Project Match in 1977, on the grounds that the matching of disparate records violated the Privacy Act:

> [a]t the matching stage there is no indication whatsoever that a violation or potential violation of law has occurred . . . . It cannot fairly be said . . . that disclosure of information about a particular individual at this preliminary stage is justified by any degree of probability that a violation or potential violation of law has occurred.[26]

---

20. ROSZAK, *supra* note 5 at 186-87.

21. *Data Matching Review: A Resource Document for Notification of the Privacy Commissioner of Proposed Data Matches*, 4 PRIVACY COMMISSIONER 112 (1989).

22. *Privacy*, AUST. L. REFORM COMM., Rpt. No. 22 (1983).

23. *See* O.M.B., *Computer Matching Guidelines supra* note 10 at 40.

24. *Data Matching in Commonwealth Administration: Discussion Paper and Draft Guidelines*, PRIVACY COMMISSIONER vi, (available from: Human Rights & Equal Opportunities Commission, G.P.O. Box 5218, Sydney, Australia) (1990).

25. O'Connor, *supra* note 10 at 4.

26. Remarks *quoted in* J. Berman and J. Goldman, *A Federal Right of Information Privacy: The Need for Reform*, 15 (available from: Benton Foundation Project on Communications & Information Policy Options, 1776 K. Street NW, Washington, D. C. 20006) (1989).

The practice is therefore not merely an invasion of information privacy. In respect of each individual, computer matching is an arbitrary action, because no prior suspicion existed. The analogy may be drawn with the powers of policemen to interfere with individuals' quiet enjoyment. If a police officer has grounds for suspecting that a person has committed, intends, or is likely to commit an offense, the officer generally has the power to intercept and detain that person. Otherwise, with rare, carefully defined (and in a democratic state, well-justified) exceptions (such as national security emergencies and, in some jurisdictions, random testing of drivers for alcohol or drugs), even a police officer has no power to intercept or detain a person. Computer matching is therefore in conflict with democratic standards relating to arbitrary interference. There is no justification for the handling of any one person's data, other than that the person is a member of a population which is known to contain miscreants.

Yet it is generally the case that a relatively very small proportion of the people whose data is involved in matching actually prove to be of interest to the organization conducting the match. Research conducted by the author shows that typically between 1% and 9% of records generate raw hits, and 0.1-2.0% survive the filtering process and reach the analysis stage. In the case of the Australian Department of Social Security's parallel matching scheme, the proportion of raw matches which have resulted in downward variations in benefits has been only about 0.5%, with 0.2% leading to debt recovery action in relation to overpayments.[27] Computer matching therefore represents an arbitrary interference with personal data which, in relation to the vast majority of the people whose data is processed, is demonstrably unjustifiable.

(3) *Due Process*

Basic among the freedoms which residents of advanced Western nations value is the freedom from arbitrary imprisonment and oppression by agents of the State. This is entirely dependent on an independent judiciary applying the law in a fair manner.

> Legal due process requires fair and equal treatment, decision making in accordance with known, uniform laws, elimination as much as possible of arbitrary and capricious behavior, and the right of appeal to assure proper procedures were, in fact, followed. All of these factors require a presumption of innocence, full judicial hearings free from the slightest taint of coercion, threats, or consideration of advantage to either the accused or the judicial system .... Information systems which contain inaccurate, incomplete, or ambiguous information lead to violations of elemental notions of fairness in treating individuals and threaten the

27. Data Matching Program (Assistance and Tax): Report on Progress, Department of Social Security, at 88-90, 107-111 Canberra (October 1992).

> specific due process guarantees afforded by the Constitution and statutes. Information must be accurate, unambiguous and complete; it must be open to challenge and review by all parties; and the procedure for creating records must be uniform and apply equally to all cases.[28]

A related issue is the increasing tendency of parliaments and government agencies to provide for the reversal of the onus (burden) of proof, such that the agency's determination holds unless the person involved successfully prosecutes his or her innocence. For example, a taxation authority may extend an established practice in relation to taxation assessment to a newly acquired function, such as child maintenance administration.

Government agencies which face difficulties and delays in relation to prosecution and debt recovery may be expected to seek to have their administrative decisions treated as direct legal authority. Because the data on which those decisions are based is by its nature suspect, granting their requests threatens fundamental freedoms.

(4) *Acontextuality*

Many people regard their provision of data to an organization as being for a particular purpose only, and perceive its use for any other purpose without consent as being, in effect, a breach of contract or trust. Privacy laws and policies throughout the Western world, and encapsulated in international documents such as the OECD Guidelines [1980],[29] reflect that viewpoint. Yet computer matching generally violates this restraint.

The principle is not merely based on a democratic ideal. People, and matters relating to them, are complicated, and organizations generally have difficulty dealing with atypical, idiosyncratic cases or extenuating circumstances.[30] Achieving a full understanding of the circumstances generally requires not only additional data which would have seemed too trivial and/or too expensive to collect, but also the application of common sense, and contemporary received wisdom, public opinion and morality.[31]

When data is used in its original context, its quality may be sufficient to support effective and fair decision-making, but when data are used outside their original context the probability of misinterpreting them increases greatly. This is the reason why information privacy principles place such importance on relating data to the purpose for which it

28. Laudon, *supra* note 5 at 133, 255.

29. Organization for Economic Cooperation and Development, *Guidelines On The Protection of Privacy and Transborder Flows of Personal Data*, OECD Pub. and Info. Center (1980).

30. MARX ET. AL., *supra* note 5 at 436.

31. ROSZAK, *supra* note 5.

is collected or is to be used, and why sociologists express concern about the acontextual nature of many administrative decision processes.

(5) *Social Equity Concerns*

The easiest targets are those people about whom records exist, and whose records are accessible by government agencies. Hence some classes of people are subjected to frequent examination by several different agencies, whereas people who live relatively undocumented lives (for example, those who operate in the so-called "black economy") escape attention: "[i]n practice, welfare recipients and Federal employees are most often the targets."[32]

In one of a series of GAO Reports urging the commencement of additional matching schemes, four anecdotes were selected to illustrate the kinds of cases which had come to light during pilot projects, and which, by implication, would give rise to millions of dollars of program savings. One related to an income-support recipient identified by the GAO as "a 78-year-old housebound veteran who in 1984 received a pension of about $3,500.00."[33] Tax records suggested that he received over $4,000.00 in interest that year, and not the zero interest claimed on his Veterans' Affairs income questionnaire. This would have precluded him from receiving any pension. The tolerance level between declared and apparent income was set at $100.00, and earnings of $1,000.00 or more were treated as being "substantial." [34] Welfare recipients might reasonably complain that the precision applied was of a different order of magnitude from that used in pursuing white-collar criminals and in assessing the taxation payable by self-employed businessmen.

In addition to being readily investigated because of their dependent relationship with one or more of the matching agencies, many of these easy targets are ill-equipped to defend themselves, because they are generally little versed in the ways of information technology, government and the law. Reversal of the onus of proof is especially problematical for people with low incomes and for welfare recipients:

> [c]omputer-generated data is doubly dangerous. On the one hand, its supposed reliability and objectivity generates a degree of reverence that makes it most difficult to challenge. On the other hand, it is highly prone to error and misinterpretation. Proving these errors, however, is extremely difficult.[35]

---

32. Office of Technologies Assessment, *supra* note 10 at 40.

33. General Accounting Office, *supra* note 15 at 18, 23, 28.

34. *Id.*

35. *Testimony of the Food Research and Action Center in Computer Matching and Privacy Protection Amendments of 1990: Hearing Before the Government Information, Justice and Agriculture Subcommittee on H.R. 5450, 102d Cong., 1st Sess. 50* (1990).

The Department of Health and Human Services has acknowledged that social equity problems are not merely a moral issue:

> [c]omputer matching, if approached too aggressively, can have some unfortunate side-effects. It can threaten and compromise the service orientation of a department, and in the process engender considerable resistance by front-line eligibility assistance workers. And, not least of all, it can actually undermine public credibility in public assistance programs.[36]

(6) *Risks to Public Morality and to Law and Order*

Beyond these rational and describable concerns, computer matching attracts a significant amount of sensationalist treatment, particularly in the lower echelons of the media. The "Big Brother"[37] specter is invoked intemperately, inaccurately, and with monotonous regularity, and discussions of human identification numbering are inevitably punctured by allusions to "the mark of the devil" in Revelations.[38]

Such irrationality often underlies popular movements, and some of these movements are successful in achieving their objectives. It is arguable that a loss of public confidence in and support for organizations could arise, if they are perceived to focus on minor transgressions by "the little people," rather than addressing larger, but inevitably more difficult, issues. A prevailing climate of suspicion is likely to result in alienation of data subjects from their social institutions.

Danger exists that data surveillance techniques such as computer matching may fuel the disaffection of a sufficient number of people to encourage anarchic developments in social organization. In sympathy with the "black economy," the "black information society" may be stimulated - a proportion of society who mislead and lie as a matter of course, on the not illogical basis that government agencies, remote from the realities of everyday existence and highly impressed with their information-based processes, can be rendered impotent by manifold inconsistencies among their copious data.

B. Specific Areas of Concern About Computer Matching

This sub-section focuses on more specific matters relating to a number of aspects of the data, the matching step, and use of the results.

---

36. Kusserow, *supra* note 10.

37. George Orwell, 1984 (1948).

38. Revelations 13:17 (New International). "He also forced everyone, small and great, rich and poor, free and slave, to receive a mark on his right hand or on his forehead. . . which is the name of the Beast or the number of his name. Id. at 13:16-17.

(1) *The Sources of the Data*

As depicted by Exhibit 2, there are many ways in which data may come to be in an organization's possession:

- it may be created by the organization, through observation of the data subject or their behavior;
- it may be created by the organization as a by-product of a transaction between the organization and the data subject;
- it may be acquired by the organization from the data subject;
- it may be acquired by the organization from another organization; and
- it may be re-acquired from an organization to which it has been disclosed.

Data is collected with particular purposes in mind, and even within-agency matching may involve data with different purposes; for example, the Australian Taxation Office holds data relating to taxation, and data relating to maintenance payments to spouses and children. The use of data arising in respect of one function to support a distinctly different function represents a breach of the vital OECD Data Protection Guideline regarding Purpose of Collection.[39] Government agencies in some countries have circumvented this protection by relying upon vague statements of objectives for computer matching, such as "the administration of the Department's programs."

Information privacy concerns are much greater where the computer matching involves two or more sets of data acquired by different organizations, especially where the purposes of the two systems are not the same.

In the case of data transfer between agencies in the same tier of government (for example, between U.S. Federal agencies, or between the agencies of a particular state), government agencies sometimes adopt the pretense that they are members of a monolithic public or civil service. This enables them to claim that all data transfers between government agencies are internal rather than external transfers. It is vital to information privacy that such totalitarian tendencies be resisted, and all inter-agency data flows treated as being inter-organizational.[40]

(2) *The Meaning of the Data*

As a matter of administrative effectiveness and efficiency, clear definitions should exist of all data-items, groups of related data-items (such as the elements of an address), and records and files involved in computer matching. In practice, however, the meaning of a significant pro-

39. OECD, *supra* note 29.

40. Clarke, *supra* note 13.

portion of data is either never explicitly defined or is subject to change over time without the changes being formally recorded. Although it is arguable that meaning is a concept applicable only when data is used, it is incumbent on organizations to make clear the definitions intended to be associated with data existing within its systems.

Many seemingly common data-items are capable of a variety of interpretations. Is a de facto spouse a "spouse?" Is a non-dependent child, or a dependent adult son or daughter, a "child?" Is a child of separated parents a dependent of each or of both of them? Is "income" to be understood as gross, or net of expenses incurred in earning it, or net of tax as well? Does it include only earned income, or also unearned income? Over what period of time is it to be averaged - a single weekly or bi-weekly pay period, a year, or many years (as is applied in some countries to farmer's income)? When matching and inferencing processes are applied to such common data-items, there is significant risk of designer or investigator misunderstanding and/or machine misprocessing. The greatest dangers arise not from unusual data-items, but in respect of widely used but subtly different data.

Some clusters of organizations have responded to inconsistencies among their definitions of critical data-items by migrating toward a common standard; for example, common identifiers can be used precisely as they are defined by their originating bodies (for example, the DHHS in respect of the SSN and the Australian Taxation Office in respect of the Tax File Number). Where names and addresses are part of the matching or inferencing algorithms, many conventions need to be harmonized to account for the enormous richness involved: for all of the sophistication of the name standardization routines applied, the Australian Department of Social Security reported as late as 1992 problems with surnames which had "non-alphabetical characters such as an apostrophe (for example, O'Connor) or contained a blank character (for example Da Vinci)."[41] The first example was particularly apt, as it was the name of the Australian Privacy Commissioner!

In the late 1980s, there was a tendency among Australian Commonwealth Government agencies to define "income" in all benefits programs in the same manner as income was defined for the purposes of taxation administration. The problems involved were exemplified by a difficulty encountered in the large-scale parallel matching program. At least one element of income (lump sum payments) continued to be treated differently in the various statutes authorizing the programs, resulting in erroneous inferencing.[42] In any case, alignment of key data-items to assist

41. Department of Social Security and the Data Matching Agency, *Data-Matching Program (Assistance and Tax): Report on Progress* 73, (Canberra, Australia 1992).

42. *Id.* at 51, 73-74.

front-end verification and computer matching represents a compromise of the primary objectives of each of the programs in order to serve a secondary purpose. It is clear evidence of subordination of service to control objectives.

A further aspect of consequence in particular circumstances is the "temporal applicability" of a definition. Remedial actions by organizations are generally retrospective, and the definition of an item may be different at the time at which the action is taken, from the time to which the action relates (for example, the current definition of income for taxation purposes may well be different from that applicable to the particular taxation return under dispute).

(3) *The Relevance of Data*

Any decision making process is at risk of reaching inappropriate or unfair conclusions if the people involved are permitted to use data which has no logical bearing on the matter - every decision-maker has experiences, biases and foibles which may lead him or her to apply possibly unconscious, but decidedly extraneous criteria. For this reason, a key privacy protection is the requirement that personal data not be used or disclosed unless it is relevant to the decision being made. Data is relevant to a decision if it can be shown that, depending on whether the data is or is not available to the decision-maker, a different outcome could result.

To establish whether data is relevant to a decision, two tests need to be applied. The first is whether the data-item is in any circumstances capable of influencing the decision. In many jurisdictions, some items of information are specifically precluded from being used in some classes of decision and should therefore not be available to decision-makers; for example, gender, marital status, age, sexual preferences, religious persuasion and political affiliation may be illegal bases for discriminating between alternative applicants for a job. There are other data items whose use may not be illegal, but which are irrelevant, or only tangentially relevant to a decision; for example, whether a woman has children and if so, their age, is generally irrelevant to an employment decision (although her availability to work during the hours after her children come home from school might be relevant).

The second test of relevance is concerned with the particular value of the data-item in each particular case; for example, whether a person suffers from a disability or chronic disease such as color blindness or asthma is irrelevant to most employment decisions, and even information about deafness or the lack of a limb is only relevant to some.

(4) *The Quality of the Data*

The concept of data quality, sometimes referred to as "data integrity," is a complex of criteria which must be satisfied if decision-making quality is to be achieved.[43] This sub-section identifies key considerations.

The criterion of accuracy is important, but easily misunderstood. At least two tests must be applied. The first is correctness, by which is meant the extent to which the content of the data-item corresponds to the attribute of the real-world entity which it purports to measure. One common source of difficulties is lack of information about the measurement scale against which the data is to be understood; for example, an examination mark of 85 looks impressive until it is discovered that the possible mark was 250. Another example is the tendency for Grade-Point Averages awarded by educational institutions to creep upwards: even if the score of 85 was from a possible 100, the mean may have been 90 and the standard deviation 5.

Particular care is needed with textual data, particularly opinions and value-judgments, since the scale of measurement is especially unclear. For example, if a debtor is recorded as being in default on a payment when the debt is actually in dispute, the data is inaccurate. To protect the data's integrity, it is important for the identity of the person making the judgment and the date to be recorded with the data.

The other dimension of accuracy is the precision with which the data is recorded. For example, time of birth can be recorded as two digits representing the year (for example, 01), four digits representing year (for example, 1901), date of birth (for example, March 6, 1901) or date and time of birth (for example, 4:30 a.m., March 6, 1901). An extreme example is the need when casting a horoscope for full precision of birth details, down to at least the minute and location of birth, but more workaday examples arise in respect of the detail needed in an audit trail of electronic financial transactions. Precision difficulties can arise in relation to, for example, benefits entitlement and rights of citizenship, permanent residency or asylum. Date and time of birth may need to be supplemented by the time-zone in which the birth occurred (or some surrogate such as city of birth). A related contemporary issue is the precision with which DNA prints need to be measured and recorded in order to have evidentiary value.[44]

Timeliness may be assessed from the perspective of the recording of the data, or of its use. On the one hand, the data must be "up-to-date," in

43. New South Wales Privacy Committee, *supra* note 8; *see* Neumann [1976-] Risks Forum, in ACM Software Eng. Notes, since 1,1 (1976); *also* on UseNet, as Comp. Risks.

44. *See, e.g.*, William C. Thompson, *Evaluating the Admissibility of New Genetic Identification Tests: Lessons From the "DNA War"*, 84 J. CRIM. L. & CRIMINOLOGY 22 (1993).

the sense that it reflects the present state of the relevant entity, and not some previous state. A typical difficulty in this regard is taxpayer address held by a taxation authority, which generally does not receive notifications of change-of-address by its data subjects, but only an address current at the time of submission of an annual return. The other sense of timeliness is the promptness of delivery of the data to the decision-maker such that it is available at a time when it is useful to the decision; for example, a person's taxable income for a particular financial year is not knowable during that year, or indeed perhaps until long after the end of that year.

A further data quality criterion is completeness. By this is meant the requirement that data not be provided in such a way that, due to the absence of some associated item of data, a misinterpretation is invited. For example, a series of defaults on payments is incomplete if it is not accompanied by the information that the person concerned lives alone, and has been in a coma for the whole of the relevant period. A criminal record which shows a judgment and sentence is incomplete if it fails to also record the subsequent reversal on appeal, or pardon, or even the existence of an as yet unresolved appeal. Many less dramatic, but nonetheless significant situations occur.

A factor often overlooked is the erosion of data quality over time; for example the correspondence of a photograph to its subject decreases, in some circumstances very quickly. Some simple measures can be used to mitigate the effects of erosion, such as the recording of birth date rather than age, and the recording of the month during which an address was last confirmed. In some jurisdictions, criminal records data is expunged after the expiration of a pre-determined period after the date of the offense or the date of release from custody. The justification for this is that the value of the data in sentencing decreases to the point where it is better to invite the person to retain their now "clean" record.

There is no absolute measure of data quality, not even in respect of a specific data-item. Moreover there are many ways in which low data quality can arise. Some errors are intentional on the part of the data subject or some other party, but many are accidental, and some are a result of design deficiencies such as inadequate coding schemes and inadequate data capture or validation procedures. Data quality costs money, and in general higher data quality costs considerably more money. It is only natural for each organization to select an appropriate trade-off between cost and quality according to its perception of the needs of the function for which the data is used. For many organizations, it is cost-effective to ensure high levels of accuracy only of particular items (such as invoice amounts), with broad internal controls designed to ensure a reasonable chance of detecting errors in less vital data.

Data quality does not simply occur, but arises because particular features of the collection process or system ensure quality. The general term for these features is data quality assurance. One element is the application of validation or editing rules at the time of capture. Another is the application of controls at the time of processing or use. For example, before a decision is made adverse to the interests of a data subject, they may be informed of the reason for the proposed decision, and given the opportunity to contest the quality of the data on which it is based. A critical feature is the existence of an audit trail, such that the origins of data can be traced, in order to resolve disputes about quality.

The vast majority of data systems operators are quite casual about the quality of most of their data. For example, the Office of Technologies Assessments reported that few federal agencies have conducted internal audits of data quality.[45] Even in systems where a high level of integrity is important, empirical studies have raised serious doubts.[46] In a study of criminal records systems, 54% of Federal records, and 74% of Federal disseminations were found to have significant quality problems, and sample surveys of state systems suggested that quality problems were even more extreme.[47] Data quality appears to be generally not high, and while externally imposed controls remain very limited, it seems likely that the low standards will persist. The quality of data which is used to make important decisions should not only be subject to controls, but those controls should also be subject to external audits of data quality, preferably by an organization which has an appropriate degree of independence and which represents the interests of all stake-holders.

All of these criteria are important in normal administrative systems. They become even more important when computer matching is undertaken. The reason is that quality problems are compounded by computer matching, because the whole purpose of the matching activity is to detect differences. Quality problems in relation to data-items used in the matching activity can result in differences which are spurious, but which are interpreted by the data scrubbing, matching, inferencing and filtering algorithms as being significant.

The quality level appropriate to the original function may not be appropriate to the new purpose for which the computer matching is performed. HUD confirmed that there have been instances in which data quality was so low that a proposed matching program was aborted dur-

45. Office of Technologies Assessment, *supra* note 10 at 26.

46. *Id.*

47. Kenneth C. Laudon, *Data Quality and Due Process in Large Interorganisational Record Systems*, 29 COMMUNICATIONS OF THE ACM 1, 4-11 (1986); LAUDON, *supra* note 5 at 139-43.

ing the preliminary examination stage.[48] Moreover, data from multiple sources collected for multiple purposes has been the subject of different trade-offs, and this alone creates dangers of erroneous matching and inferencing. This issue is particularly important in interagency and intersystem matches, and when the matched data is used to make decisions seriously adverse to the interests of the data subject, as in matching programs relating to welfare, taxation and criminal justice administration.

The complexities of each system (particularly a country's major data systems such as taxation, social welfare and health insurance) are such that few specialists are able to comprehend any one of them fully. It challenges the bounds of human capability to appreciate and deal with the incompatibilities between data collected from different systems, of varying meaning and quality, and to handle the matched data with appropriate care. Computer matching should never be undertaken without the greatest caution and skepticism. This is especially so in the case of large-scale, repetitive and routinized programs.

(5) *Other Aspects of the Data*

Depending on the nature of the data, there may be additional characteristics which require consideration. One such question is sensitivity. This is usually regarded as referring to race, religious and political affiliations, and sexual preferences. In some cultures and under some circumstances, a wide variety of data may be regarded as sensitive, including financial matters, medical information (for example, the nature of treatments and drugs prescribed to an individual) and household structure (for example, the absence of a male may mark the household as a target for crime). For some, even their date of birth is sensitive date.

The question also arises as to whether any privileges exist in relation to the data. In some countries this is the case in confidential relationships, such as those between doctor and patient, and priest and parishioner.

(6) *The Context*

Related to the notion of "completeness" is the context within which the data is to be understood. It would, for example, be unworldly to expect the social mores of the time and culture to be captured into a database as a step in ensuring the completeness of data. On the other hand, when the data is about to be used, it is desirable that the decision-maker consider whether the context in which it arose is relevant. Otherwise, since the decision-maker is remote from the person affected by his

48. Roger A. Clarke, *Computer Matching in the Social Security Administration*, (unpublished working paper available from the author) (January 1992).

decision, some insensitivity in decision-making is an almost inevitable result of computer matching.

One example of a situation in which context may play an important role is information about the person's racial background; for example, the United States, Canada, Australia and New Zealand all make provisions for the cultures of their small remaining indigenous populations. They also recognize the transitional period required for new immigrants. Other examples of contextual matters are organizational norms and practices, both official and informal; the geographic location of relevant activity (for example, financial inducements to purchase, which would be illegal in the host country, may be legal or at least tolerated, and even commonplace, in a foreign country); the level of law and order in the region at the relevant time (for example, loan default might be treated more leniently when it occurs during a period of riots, civil war, invasion or natural catastrophe); and changing attitudes to, for example, the Committee on Un-American Activities, the Vietnam War, Yassir Arafat, Oliver North, and, in the coming years, perhaps even Communism, the Ayatollah Khomeini, and Colonel Mumar Qhadaffi.

These points have been made more graphically elsewhere:

> [i]nformation, is no more than it has ever been: discrete little bundles of fact, sometimes useful, sometimes trivial, and never the substance of thought [and knowledge] . . . . The data processing model of thought . . . coarsens subtle distinctions in the anatomy of [the] mind . . . . Experience . . . is more like a stew than a filing system . . . . Every piece of software has some repertory of basic assumptions, values, limitations embedded within it . . . . [For example], the vice of the spreadsheet is that its neat, mathematical facade, its rigorous logic, its profusion of numbers, may blind its user to the unexamined ideas and omissions that govern the calculations . . . [result:] garbage in - gospel out.[49]

(7) *The Matching Step*

In any matching program, the possibility exists that the data-items may have been incorrectly captured, and hence the question arises as to the extent to which validation has been undertaken. For example, one person may volunteer another's identification code or name in his or her dealings with an organization, and as a result computer matching will associate that data with the wrong person. The same problem may arise through intent, or as a result of error by the individual or by the organization that records the data. The risk of such errors arises at the time that an organization originally collects its standing or master-file data, and again on every occasion that a transaction takes place.

---

49. ROSZAK, *supra* note 5 at 87, 95, 98, 118, 120.

Matching may work fairly well in circumstances in which people have an apparent interest in being readily associated with the data stored about them. The procedures are less effective where the control functions of the system are perceived by the data subject to dominate the service objectives, as has always been the case with taxation authorities, and is often so with welfare agencies. Matching is much more difficult to apply when the data subject has a clear self-interest in frustrating the organization's purpose, for example, in criminal investigation and national security operations.

Matching algorithms may result in zero, one, several or many "putative hits," and some means of ambiguity resolution is needed. One approach is, for each record, to accept the first "hit" and suspend the search. Alternatively, for each record, the search can be run on the whole file, and, where more than one "putative hit" arises, a further criterion can be applied, such as:

- treat all putative hits as "hits";
- abandon that record (in order to focus on easier ones);
- present the full list of possibilities to a human decision-maker;
- order the set according to some a priori criteria for assessing the likelihood of each putative hit being the correct match;
- reduce the set by extending the matching algorithm to incorporate further procedures or rules.

Different alternatives may be appropriate to particular circumstances. In many cases, however, it is not possible to reach an unambiguous conclusion, and serious danger exists that any administrative judgment made or action taken will be based on justification that is inadequate, ineffective, inequitable or wrong in law.

In principle, a matching algorithm's effectiveness may be measured on the basis of the error-rate. This might be interpreted as the proportion of spurious matches or false hits, plus undetected matches or false misses. In practice, however, this is not easy, because it is seldom economically practicable to identify the errors. As a result, it appears that quality assurance and the exercise of control over the effectiveness of matching algorithms are very difficult and expensive, and the auditability of those controls very limited.

Because so few contemporary identification schemes use a physiological identifier, they are (at best) of moderate integrity. Rather than individuals themselves, the subject of surveillance is data which purports to relate to them. As a result, there is a significant risk of wrong identification, and the incorrect association of data with the wrong person.[50]

50. Roger A. Clarke, *Human Identification in Information Systems: Management Challenges and Public Policy Issues*, INFORMATION TECHNOLOGY & PEOPLE (forthcoming 1995); L. Henderson, *Case Study in Data Integrity: Australian Passports Office*, DEPART-

Identification quality problems are compounded by computer matching. The rules applied by each organization in deciding on its person-identifier are of necessity somewhat arbitrary, and the strong likelihood is that the choices made by the two organizations are different, at least in regard to the details. Characteristics as simple as the sequence of parts of a name, and the inclusion or exclusion of second initial in an identifying routine, can become significant technical challenges when two sets of records are merged.

To address such difficulties, there is a trend toward the use of common identifiers. This has been noticeable in the United States through the widespread use of the Social Security Number (SSN), in Canada where the Parliament has recently acted to curb the proliferation in use of the Social Insurance Number (SIN), and in Australia where successive attempts have been made to establish an Australia Card number (which failed)[51] and to enhance and extend the Tax File Number (which is progressively succeeding).[52] The movement in Australia has followed the conclusions reached by a report on fraud in government, which said:

> [a]gencies should record identification and locator information, as far as possible, in standard form to facilitate matching .... Naturally, clients are free to call themselves as they please .... The Commonwealth is equally entitled to call the client by a standard name.[53]

Matching techniques which are based on common data-items appear likely to more easily achieve a reasonable level of quality than those which are based on complex rules interrelating a set of data-items with imprecise and differing definitions. Even in those cases, however, allowance must be made for both false hits and false misses. For example, persons may have intentionally or unwittingly provided an inappropriate but valid identifier to one of the organizations involved, or the organization may have incorrectly processed the data by applying the right number to the wrong person's data. This tends to generate both a false miss and a false hit.

It is important that matching algorithms be subject to controls, and that the controls be subject to audit. This may imply for many organizations which conduct computer matching a significant increase in the degree of formality to be applied.

MENT OF COMMERCE, AUSTRALIAN NAT'L. UNIV. (unpublished working paper available from the author) (March 1992).

51. Clarke, *supra* note 13.

52. Roger A. Clarke, *The Resistible Rise of the National Personal Data System*, 5 SOFTWARE L.J. 1 (1992).

53. Aust. Govt. Publ. Service, *supra* note 10.

(8) *The Use of the Outcomes of Computer Matching*

When matching has been completed, and decisions are made and action is taken on the basis of the combined data, further concerns arise. At the heart of the freedoms which have been won for citizens of free countries during the last few centuries is the right to know one's accuser, the accusation, and the evidence on which the accusation is based. This implies that the data subject be informed that the origin of a decision lay in a particular computer matching run. Moreover, the data subject should have access to the data on which the decision was based, in order to provide both validation, and necessary information to enable defenses to be mounted.

It may seem unreasonable to call into question whether due process is being respected by organizations who are using the results of computer matching runs. Certainly there have been claims by various U.S. agencies in hearings before congressional committees that due process provisions pre-existed the 1988 Act regulating computer matching.[54] It is increasingly common, however, for guilt to be presumed. In taxation administration, for some time now the burden has been on taxpayers to prosecute their innocence. This inversion was knowingly authorized by a Parliament concerned to ensure effective administration. Much more disturbing is that the inversion of the onus of proof is tending to proliferate, without explicit legislative consideration.[55]

Further concerns relate to the retention and reuse of data arising from a matching run, and the subsequent redisclosure of such data. A case study demonstrating how strongly presumptive such uses can be involved a change to a record in the Australian Passports Office, on the basis of a matching run against the Electoral Register. As a result of inadequate procedures in the Electoral Office, the inference drawn by the Passports Office's matching run was wrong. By the time the mistake became evident, the audit trail no longer existed and careful negotiations by a person well aware of the processes within government agencies were necessary to ensure that the whole story came to light and the specific problem could be overcome.[56]

C. Intrinsic Controls Over Computer Matching

The previous section identified a large number of considerations which make it necessary for computer matching to be subject to controls. This section identifies factors which tend to restrain agencies from ap-

54. *See supra* note 4.

55. Laudon, *supra* note 5 at 376.

56. Clarke, *supra* note 50.

plying computer matching, or to cause them to do so carefully, and assesses their effectiveness.

The use of matching may be limited by the exercise of countervailing political power by the class of data subjects affected by the process, by their representatives, or by the general public. Given the imbalance of power between organizations and individuals, it is not realistic to expect this factor to be of any great significance except in particular circumstances; for example, a particular individual who is affected by the process may be influential, or may attract the active support of some influential person or group, or may attract the attention of the mass media to the matter. In practice, the most common classes of subject who are affected by computer matching are welfare recipients and government employees, neither of which have much political influence. Occasional negative publicity may arise in respect of a particular program or case, but is likely to last only a short time, and have negligible impact on the agency's ongoing activities.

Another potential control is that a computer matching activity may incur the displeasure of some organization, such as a competitor. In practice there is little evidence of such a mechanism having been operational.

There is the possibility of self-restraint being practiced by the agency itself. One basis for this could be that matching is inconsistent with the agency's objectives. In earlier years, many benefit paying agencies regarded their service-related aims as primary. During the past two decades, with the growth in aggregate benefits payments, in the visibility of overpayments and in the dominance of economic over social concerns, welfare-related agencies in the U.S. and Australia have placed increasing emphasis on their control objectives. Self-restraint on the basis of incompatibility with client-service objectives appears, on the basis of empirical research, to be very limited. There is, however, evidence of a more self-interested motivation on the part of the Internal Revenue Service to limit the use of its data in other agencies' matching programs. It has expressed opposition to such use on the grounds that this would compromise the voluntary tax system.[57]

Other potential bases for agency self-restraint exist. For example, key members of the agency's management team might be guided by professional norms, or by an appreciation of the delicacy of public confidence in its institutions and the resultant need to respect constitutional rights and moral concerns. Relevant decision-makers appear not to have been subject to any explicit professional requirements for care in the use of powerful informational technology tools, however.[58] The findings of a

57. General Accounting Office, *supra* note 15 at 5.

58. Shattuck, *supra* note 10.

General Accounting Office study were that "we seldom encountered any expression of concern about the potential invasion of privacy that went beyond a matter of compliance with existing legislation and regulations."[59]

A factor that should not be overlooked is the extent to which general blundering constrains excesses by agencies in their use of computer matching. The quality of both computer-based and manually maintained data, and of data processing, case analysis and prosecution must all be high if the potential benefits are to be attained; for example, two major agencies failed over a series of years to exchange benefit-recipient death information purely because of misunderstandings.[60] In addition, long delays and vague plans to refer cases to the FBI "for potential fraud review" have been detected,[61] and agencies are frequently unwilling to take further action where the recipient has died[62] or is no longer receiving benefits.[63]

The intrinsic factor that might be considered to offer the most significant degree of control over computer matching is economics: the theory being that surely government agencies will not apply the technique where it is not worthwhile.

However, economic constraints will only be effective if the relationship between a program's costs and benefits are visible, and this is only likely if formal cost/benefit analysis is applied. The use of cost/benefit analysis in relation to computer matching was evaluated as part of this lengthy study, and is reported in the author's *Computer Matching by Government Agencies: The failure of Cost/Benefit as a Control Mechanism*.[64] The conclusions reached were that cost/benefit analysis is seldom performed voluntarily, that there are many serious deficiencies in those few analyses which have become publicly available, and that programs are continued even after they have been clearly demonstrated to be financially unjustifiable.

IV. A NORMATIVE REGULATORY FRAMEWORK

The previous section argued that the negative impacts of computer matching are sufficiently serious to demand controls, and that natural or intrinsic controls are inadequate. The final section of this paper presents

59. General Accounting Office, *supra* note 15 at 3.

60. *Id.*

61. *Id.*

62. Social Security Administration, Office of Program Integrity and Reviews, *SSA Matching Operations Report: October 1990 - March 1991*, Pub. No. 31-004 (1991).

63. General Accounting Office, *supra* note 15 at 4.

64. Roger A. Clarke, *Computer Matching by Government Agencies: The Failure of Cost/Benefit Analysis as a Control Mechanism*, INFORMATION AND THE PUBLIC SECTOR (forthcoming 1995-96).

a framework for a regulatory regime for computer matching. It is normative because, in the absence of "good" models, it is impractical for it to be derived from experience. The general principles proposed here are derived from previous proposals[65] and the analysis presented in the earlier sections of this paper.

This framework is intended to be generally applicable. In the few jurisdictions where regulatory regimes exist, it provides a basis for evaluating the existing controls. In other jurisdictions, this framework may assist in the creation of codes appropriate to the particular social and economic culture, and legal structure and processes.

General principles are enunciated relating to the environment and infrastructure of regulation. These specify the need for effective information privacy laws, a privacy protection agency, effective computer matching laws, the denial of monolithic government, exceptional treatment rather than exemption, publicly visible justification and adaptive control. The detailed requirements of a legally enforceable code of practice are laid out in the Appendix.

1. *Effective Information Privacy Laws*

The establishment of extrinsic controls over computer matching are very unlikely to be even embarked upon until comprehensive information privacy laws are in place. Unfortunately, the primary consideration in the formulation of privacy laws throughout the world has been that the efficiency of business and government should not be hindered. What was provided was an "official response" which legitimated dataveillance measures in return for some limited procedural protections commonly referred to as "fair information practices."[66]

In the United States, the early discussions of protections (and in particular, those of the Department of Health, Education, and Welfare)[67] resulted in the enactment of the Privacy Act 1974.[68] President Ford threatened to veto that statute, and forced the proposed Privacy Commission to be reduced to a short term Study Commission.[69] The Commission's Report[70] implicitly accepted the need to make surveillance as publicly acceptable as possible, consistent with its expansion and efficiency.[71] Agencies had little difficulty subverting the Privacy Act. The

65. LAUDON, *supra* note 5 at 384-90; FLAHERTY, *supra* note 5 at 359-407.

66. RULE, *supra* note 2; Smith, *supra* note 5; FLAHERTY, *supra* note 5; W. MADSEN, HANDBOOK OF PERSONAL DATA PROTECTION (Macmillan, London, 1992); Bennett, *supra* note 5 at 96-110.

67. *See supra* note 10.

68. *See supra* note 7.

69. Laudon, *supra* note 47 at 6.

70. Privacy Protection Study Commission, *supra* note 8.

71. RULE, ET AL., *supra* note 5 at 75, 110.

President's Council for Integrity and Efficiency ("PCIE") and the Office of Management and Budget ("OMB") have worked not to limit computer matching, but to legitimize it.

The legitimization process has been evident in developments in other countries and in international organizations. The OECD's 1980 Guidelines for the Protection of Privacy were quite explicitly motivated by the economic need for freedom of transborder data flows. In the United Kingdom, the Government stated that its Data Protection Act of 1984 was enacted to ensure that U. K. companies were not disadvantaged with respect to their European competitors. The purpose of the "EC'92 Directive," which has been under discussion within the European Community for several years, is the application of the limited "fair information practices" tradition uniformly throughout Europe.[72] Because it would mandate European Community nations' prohibition of the export of personal data to countries that do not provide "an adequate level of protection," it would significantly increase the influence of international instruments such as the OECD's 1980 Guidelines.

There have been almost no personal data systems, or even uses of systems, which have been banned outright. Shattuck[73] reported that during the first five years, the unsympathetic interpretation of the U.S. Privacy Act of 1974 (by the Office of Management and Budget) resulted in not a single matching program being disapproved. Few sets of information privacy principles appear to even contemplate such an extreme action as disallowing some applications of informational technology because of their excessive privacy-invasive nature. Exceptions include those of the New South Wales Privacy Committee,[74] which are not legally enforceable and, with qualifications, Sweden. This contrasts starkly with the conclusions of observers:

> [a]t some point . . . the repressive *potential* of even the most humane systems must make them unacceptable[75] . . . [w]e need to recognize that the potential for harm from certain surveillance systems may be so great that the risks outweigh their benefits (emphasis in original).[76]

The first requirement of a control regime for computer matching is comprehensive and universally applicable data protection legislation which achieves a suitable balance between the various economic and social interests, rather than subordinating informational privacy concerns to matters of administrative efficiency.

---

72. *Amended Proposal for a Council Directive Concerning the Protection of Individuals in Relation to the Processing of Personal Data*, EUROPEAN COMMUNITY, O.J. No. L. 123 (1992).

73. Shattuck, *supra* note 10 at 540.

74. New South Wales Privacy Committee, *supra* note 8.

75. RULE, ET AL., *supra* note 5 at 120.

76. Marx, *supra* note 5 at 48.

2. *A Privacy Protection Agency*

Privacy protection regimes based on cases being brought by private citizens against the might of large information rich and worldly-wise agencies have not worked, and are highly unlikely to do so in the future. To achieve appropriate balance between informational privacy and administrative efficiency, it is necessary for an organization to exist which has sufficient powers and resources to effectively represent the information privacy interest.[77]

It is has been argued that, in the United States, "a small agency of perhaps sixty professionals drawn from legal, social science and information systems backgrounds" would provide sufficient resources to address the problem.[78] From the perspective of Australia and Canada, this would seem parsimonious for such a large country, but there can be little doubt that, given an appropriate set of powers, and sufficient independence from the Executive, it could make a significant difference to the balance of power.

It would be valuable to complement such a body with an effective option for individuals to prosecute and sue organizations which fail to comply with legal requirements. This can only come about if the requirements of organizations are made explicit, and this in turn is only likely to come about if detailed codes, prepared by a privacy protection agency on the basis of research and experience, are given statutory authority. In addition to valuable discussions in the literature, elements of these can be found in Australian and Canadian legislation and practice.[79]

There are two competing models. The conventional model requires the agency to balance information privacy against other interests (such as administrative efficiency), and is based on negotiation and conciliation rather than adversary relationships. This risks the capture of the privacy protection agency by the other much larger, more powerful, information-rich and diverse agencies, as is occurring in Australia. Laudon argues strongly for the alternative - an agency which is explicitly an advocate for information privacy, and which can represent that interest before the courts and Congress.[80]

3. *Effective Computer Matching Laws*

Legislation is not an effective extrinsic control over computer matching practices if its primary function is to legitimize existing activities or

77. Australian Public Service, *supra* note 10; Office of Technologies Assessment, *supra* note 15 at 57-59, 113-22; FLAHERTY, *supra* note 5 at 359-407.

78. LAUDON, *supra* note 5 at 383.

79. Privacy Protection Study Commission, *supra* note 8; Australian Law Reform Commission, *supra* note 22; LAUDON, *supra* note 5 at 382-84; FLAHERTY, *supra* note 5.

80. LAUDON, *supra* note 5 at 384.

proposed programs. The legislative packages brought forward by or for government agencies are based on those agencies' perceptions of their own needs, and do not reflect broader, public interests. If legislators support the balancing of economic against social objectives, they must be suspicious of agency initiated bills, and must instead specify the philosophy and objectives of bills brought before parliament or legislatures.

Computer matching is a specific technique, and cannot be appropriately enabled or controlled by general legislation.[81] Either Parliaments must expend the effort to become competent in dealing with the issues, or they must create a specialist agency and invest it with the responsibility of bringing specific statutory instruments before it, dealing with specific programs or specific agencies.

Laudon concluded that "a second generation of privacy legislation is required."[82] This second generation has since begun, with the United States somewhat improving control over computer matching with its 1988 Act;[83] with Canada rolling back the uses of the Social Insurance Number and regulating data matching in 1989; and with Australia issuing draft Guidelines and passing its first (and admittedly limited) second generation legislation in 1990 (after finally catching up with the first generation only at the beginning of 1989). The challenge is to regulate computer matching in such a way that it clears the path for worthwhile applications of the technique, while preventing unjustifiable information privacy intrusions.

4. *Denial of Monolithic Government*

When it suits their interests, agencies adopt the attitude that they form a collective governmental monolith, and therefore all data transfers are internal to government. This is inimical to informational privacy interests, and it is necessary for parliaments and legislatures to make clear that agencies are independent organizations for the purposes of data transfer; *all* data transfers are therefore subject to the rules regarding collection and dissemination.

In addition, there is a danger that privacy protections may be subverted by the concentration of functions and their associated data into large, multi-function agencies. Hence boundaries must be drawn not only between agencies but also between functions and systems.

"Virtual centralization" of data by way of network connection, and data linkage via common identifiers, also embody extremely high information privacy risks. The "national databank" agenda of the mid-1960s is being resuscitated by government agencies, and with it is coming pres-

81. Office of Technologies Assessment, *supra* note 15.

82. LAUDON, *supra* note 5 at 400.

83. *See supra* note 15.

sure for a general-purpose identification scheme. These must be strenuously resisted if the existing balance of power between individuals and the State is not to be dramatically altered.

5. *Exceptional Treatment Rather Than Exemption*

Care must be taken to ensure that exemptions do not rob privacy protection legislation of its effectiveness. The general principles of information privacy must be applied to all agencies and all systems, and the regulatory regime for computer matching to all programs. The widely practiced arrangement of exempting whole classes must therefore be rolled back.

It is entirely reasonable, on the other hand, for the specific nature of controls to reflect the particular features of an organization, system or program. This applies particularly to operations whose effectiveness would be nullified in the event of full or even partial public disclosure. In such instances, the privacy protection agency needs to be explicitly nominated as the proxy for the public, authorized in law to have access to sensitive operational data, but precluded in law from disclosing details to the public. Not only government agencies but also government business enterprises and private sector organizations have tenable claims for exceptional treatment along these lines.

Finally, the favored status traditionally granted to defense, national security, criminal intelligence and law enforcement agencies must be rolled back. Parliaments must make these agencies understand that they are subject to democratic processes and that their distinctive and challenging functions and operational environments dictate the need for careful formulation and implementation of privacy protections, not the need for exemption.

6. *Publicly Visible Justification*

After all financial costs have been considered, the financial worth of many computer matching programs is at best in doubt, and in some cases, negative. This is not necessarily to say that those programs should not have been undertaken, because there are many factors, especially deterrence on the one hand and informational privacy invasion on the other, which are not reducible to financial measures.

As Laudon noted,

> [a] pattern has emerged among executive agencies in which the identification of a social problem (such as tax evasion, welfare fraud, illegal immigration, or child maintenance default) provides a blanket rationale for a new system without serious consideration of the need for a system.[84]

84. LAUDON, *supra* note 5 at 385.

This "blanket rationale" must be discarded. Computer matching programs must be subjected to conventional cost/benefit analysis, including estimation of the full range of actual and opportunity costs and financial benefits, quantification of as many as possible of the non-financial costs and benefits, and description of the non-quantifiable factors. Guides exist as to the nature of costs and benefits which should be considered.[85]

It is insufficient, however, for a requirement to be imposed without a control mechanism to ensure its satisfactory implementation. The justification of each program needs to be reviewed by an organization whose interests are at least independent of those of the proponent organization, and perhaps even adversarial. While it may be unreasonable in some cases to ask agencies to make their full strategy publicly available (because of the harm this might do to the program's effectiveness), the arguments against publishing the justification for the program are far weaker. Prior cost/benefit analyses and post-program evaluations must be undertaken, must be subject to review by the privacy protection agency and must be available to the public.

7. *Adaptive Control*

Technological developments have rendered some of the early information privacy protections ineffective: "new applications of personal information have undermined the goal of the Privacy Act that individuals be able to control information about themselves."[86] If a proper balance between informational privacy and other interests is to be maintained, processes must be instituted whereby technological change is monitored, and appropriate modifications to law, policy and practice brought about. This needs to be specified as a function of the privacy protection agency, and the agency funded accordingly.

Despite the long-standing dominance of the arguments of Westin,[87] it is clear that the needs of administrative efficiency are at odds with individual freedoms. The power of computer matching techniques is far greater than it was two decades ago, and refinements and new approaches are emerging which will require further regulatory measures in the near future.

---

85. Kusserow, *supra* note 10 at 33-40; Privacy Commissioner of Canada, *supra* note 10; Clarke, *supra* note 64.

86. J. Thom and P. Thorne, *Privacy Legislation and the Right of Access*, 15 AUSTRAL. COMP. J. 44, 145-50 (1983); Roger A. Clarke and G. W. Greenleaf, *Database Retrieval Technology and Subject Access Principles*, 16 AUST. COMP. J. 1 (1984).

87. A. F. WESTIN, PRIVACY AND FREEDOM (Atheneum 1967); A. F. WESTIN (ED.) INFORMATION TECHNOLOGY IN A DEMOCRACY (Harvard Univ. Press 1971); A. F. WESTIN AND M. BAKER, DATABANKS IN A FREE SOCIETY (Quadrangle 1974).

8. *Detailed Requirements*

In addition to the general principles, detailed requirements must be imposed on organizations involved in computer matching. These are presented in the Appendix to this paper, and comprise pre-conditions to the commencement of a program, requirements regarding the conduct and aftermath of a program, general requirements, and definitions.

This framework has been applied to the protective regime which applies in the United States (based on the Computer Matching and Privacy Protection Act 1988),[88] and that proposed for Australia (the Privacy Commissioner's Data Matching Guidelines).[89] Both are seriously deficient in comparison with the normative framework.

V. CONCLUSIONS

Computer matching is a powerful dataveillance technique, capable of offering benefits in terms of the efficiency and effectiveness of government business greater than its financial costs. However, it is also highly error-prone and privacy-invasive. Unless a suitable balance is found, and controls imposed which are perceived by the public to be appropriate and fair, its use will result in inappropriate decisions and harm to people's lives. In a tightly controlled society, this is inequitable. In a looser, more democratic society, it risks a backlash by the public against the organizations which perform it, and perhaps also against the technology which supports it.

Brief background has been provided to the origins and technique of computer matching. Its impact has been analyzed, and the need for controls established. Intrinsic controls have been shown to be insufficient to ensure that computer matching programs are accurate, equitable and socio-economically justified. A framework has been presented, which provides a basis for the preparation of effective regulatory measures in jurisdictions where none exists, and a tool for assessing the control regimes in place in those few countries which do have them.

Because of the current fashion of highly information-intensive procedures, and the inadequacy of the controls, the current boom in identification-based computer matching activity may be expected to continue for some years. Further refinements may be confidently expected in the data scrubbing, matching, inferencing and filtering steps, including the application of such techniques as direct access, multiple-file matching, associative memory, expert systems, neural networks and fuzzy logic.

A longer-term scenario can be constructed on the basis of experiences in information privacy issues generally. Occasional public back-

88. *See supra* note 15.

89. *See supra* note 35.

lashes will follow the publication of information about proposed new schemes, unjustified dataveillance, blunders and unfair behavior on the part of one agency or another. These will generally be short-lived, and after the proposal has been temporarily withdrawn or the current Secretary or Minister has resigned, the grand momentum of government agency policies will resume. The level of public morality in relation to the provision of information to government agencies will fall lower, and the intrusiveness of government agency questioning of and about data subjects will increase, in order to provide the necessary additional data. The level of public confidence in government agencies will spiral lower still. Improvements to the integrity of identification, which had been withheld in the past, will be instituted. Faced with substantial failure, those schemes will be enforced using seriously repressive measures. The climate of public suspicion and animosity will be exacerbated, and the quality of data will fall lower still.

In due course, as the proportion of the public routinely indulging in multiple identities and noise-laden data increases, two other dataveillance tools will become more prominent. Identification-based matching will be supplemented and then steadily supplanted by content-based matching, as the techniques increase in sophistication while decreasing in price. Meanwhile, the capacity of distributed databases will be increasingly applied to the linkage of records about individuals, within and beyond the community of government agencies. Initially, this application will be significantly constrained by technical difficulties and, to a limited extent, by public opposition and the law. However, the public's short span of attention, delay of the government community with only limited interference from transitory congressmen or parliamentarians, the readily revoked economic imperative to use data efficiently, the unenforceability of most data protection laws, and the effective revocability of the remainder of those laws (through seemingly minor but debilitating amendments) will combine to enable government agencies to achieve a 21st century "virtual national databank" more powerful, more extensive, and far more practicable than that conceived in the 1960s. A wide variety of inventive techniques will be used by many individuals to sustain their private space, and flourishing black economies and black information societies will make a mockery of government statistics.

Like any other scenario, this somewhat apocalyptic vision is a projection rather than a prediction. The analysis in this paper suggests, however, that it is a very plausible trajectory. If it is to be resisted, and the values of democracy and individual freedom sustained, powerful countervailing regulatory regimes must be instituted. Applications of information technology are changing the landscape of the societies in which we live. This paper has proposed a basis for exercising control over one currently important dataveillance technique.

APPENDIX: DETAILED REQUIREMENTS

This Appendix contains the detailed set of requirements which need to be imposed within the general framework presented in section 4 of this paper. Definitions are first provided. The requirements are then expressed which are applicable respectively prior to the commencement of a matching program, while it is being conducted, and during its latter stages. A final section contains requirements applicable to all phases.

A. Definitions

The meanings of key terms used in these Detailed Requirements are as follows:

° COMPUTER MATCHING is any computer-supported process in which personal data records relating to many people are compared, in order to identify records of interest;

° A COMPUTER MATCHING RUN is an event in which one file is compared against one or more files;

° A COMPUTER MATCHING PROGRAM is a set of one or more computer matching runs, which are very similar in nature (in terms of the data which are accessed, the matching and inferencing criteria applied, etc.), and which are undertaken to assist a single organization or set of organizations in addressing a single set of objectives;

° A MATCHING ALGORITHM is a procedure or set of rules, whereby the person to whom a given record relates is inferred to be the same person to whom one or more other records relate;

° A HIT, RAW HIT OR MATCH is two or more records which appear, on the basis of a particular matching algorithm, to relate to the same person; or, in an inverse matching process, a record on one file which fails to find a match on the other;

° AN INFERENCING PROCEDURE is the means whereby the contents of the matched records, or the existence or non-existence of a match are analyzed, and conclusions are drawn about the person to whom the data purports to relate;

° RECORD means a collection of data items each of which represents an attribute of the same real-world entity;

° FILE means a collection of records, each of which has similar structure and identifiers, and each of which refers to a real-world entity of the same class;

° A PARTICIPATING ORGANIZATION is any government agency, corporation or other legal entity which takes part in computer matching, whether as a source organization, as the matching organization, as a client organization, or in any other manner. Which organization or organi-

zations is or are responsible for the satisfaction of each particular requirement is to be understood from the context;

° A SOURCE ORGANIZATION is a participating organization which makes data available to a computer matching step;

° A MATCHING ORGANIZATION is a participating organization which undertakes a matching step;

° A CLIENT ORGANIZATION is a participating organization which receives data resulting from a computer matching step;

° THE LEAD ORGANIZATION, in respect of any particular computing matching program, is that one of the participating organizations designated by them to take the responsibility for coordinating the compliance of all participating organizations with these requirements;

° THE PRIVACY PROTECTION AGENCY is the organization within the jurisdiction which is responsible for protecting information privacy and assisting individuals to protect their own privacy;

° COST/BENEFIT ANALYSIS (CBA) is a technique which facilitates evaluation of a proposed policy measure or a completed program. It involves first the identification and description of all significant advantages and disadvantages of a measure, taking into account the interests of all stakeholders; second the quantification of all costs and benefits which can practicably be quantified, expressed wherever possible in financial terms; and third the presentation of net benefits and non-financial costs and benefits in a manner designed to enable the application by decision-makers of explicit and implicit values and criteria. The key elements of CBA are expressed in Exhibit A1.

EXHIBIT A1: COST/BENEFIT ANALYSIS

SOURCES OF INFORMATION

- ° post-evaluation reports on prior programs
- ° prior experiences in the particular area of government administration
- ° experiences in other jurisdictions within Australia and overseas
- ° post-evaluation reports on specially-designed pilot matching programs

STEPS TO BE UNDERTAKEN AND DOCUMENTED

° identification and description of all benefits, costs and impacts, including those:

- ° arising in all participating organizations;
- ° arising in other organizations; and
- ° incurred by individuals;
- ° in relation to all of the above benefits, costs and impacts:

- ° description of the mechanisms whereby the anticipated benefits, costs and impacts are expected to arise;
- ° estimates of all financial benefits and costs;
- ° the basis upon which each estimate has been made, including the base data, estimated recovery rates, and the rationale for the use of each figure; and
- ° where financial estimates are not possible or appropriate, quantitative estimates of relevant factors;
- ° where the timings of financial inflows and outflows are significantly different, the application of discounted cash flow techniques;

° computation of net financial benefits;

° juxtaposition of net financial benefits against costs and benefits which cannot be expressed in financial terms;

ITEMS REQUIRING SPECIAL ATTENTION

° a broad perspective must be adopted, reflecting the agency's public responsibilities; the objective must be to maximize benefits or minimize costs to the general public;

° cash costs, the commitment of existing resources and the time and effort of existing staff are all to be included;

° costs of pilot programs, and costs attributable jointly to the program in question and other programs and operations are to be identified and treated according to the established principles of economics and management accounting. For the purposes of performance evaluation, this will generally require apportionment between programs on some more or less arbitrary basis; for the purposes of deciding whether or not to proceed with a proposed scheme, it will generally be appropriate to exclude sunk costs and apportion joint costs and benefits depending on whether or not the associated programs proceed;

° in post-evaluation reports, actual measures should be provided wherever possible;

° consideration should be given to factors which constrain the ability to collect debts, such as public policies, court rules and practices, practicalities and economics;

° consideration should be given to the scope for persons to casually or with criminal intent subvert the scheme, and possible countermeasures against those risks;

° the mechanism of deterrence and the basis for any financial or other quantitative estimates of its effect must be particularly carefully explained;

° opportunity costs are to be taken into account, i.e. the net benefits foregone because the resources are assigned to this program rather than to alternative programs;

° the effort invested in the cost/benefit analysis should reflect the scale and privacy-intrusiveness of the program;

WHEN COST/BENEFIT ANALYSIS IS TO BE APPLIED

° in all cases, as part of the justification process of a new program;

° in the case of particularly sensitive programs, progressively during the program;

° in the case of completed programs, as part of the post-evaluation report;

° in the case of long-running programs, annually.

B. REQUIREMENTS PRIOR TO COMMENCEMENT OF A PROGRAM

1. *Pre-Conditions*

A computer matching program is only to be undertaken if the following conditions apply:

(i) all uses and disclosures of personal data which are involved in the program must be PERMITTED (i.e. required or authorized, and not precluded) by a law which is expressed in reasonably precise and limited terms (i.e. computer matching is sufficiently privacy-invasive that its use should not be based on broad and general powers);

(ii) the OBJECTIVES, and the administrative and social JUSTIFICATION, of the computer matching program must be explicitly identified and documented;

(iii) explicit consideration must be given to ALTERNATIVE MEASURES less privacy-intrusive than computer matching, and those alternatives must be demonstrably less effective in achieving the objectives;

(iv) a PRIOR COST/BENEFIT ANALYSIS must be undertaken and documented, and must demonstrate that the program is very likely to yield financial and non-financial benefits to society greater than the financial and non-financial costs which are likely to be incurred by all parties, including the participating organizations, all other organizations involved and the public;

(v) the QUALITY OF DATA to be used in the program must be assessed, and be shown to be sufficiently high to enable the objectives to be achieved;

(vi) except where secrecy is essential to the operation, and is explicitly justified in the Terms of Reference, measures must be taken to ensure PUBLIC KNOWLEDGE of the program. Where a secrecy provision is invoked, the privacy protection agency must be provided with details of the program and the justification for the secrecy, subject to appropriate non-disclosure provisions;

(vii) consideration must be given to any PRIVILEGES AND SPECIAL SENSITIVITIES that may relate to any of the data; and
(viii) the program must have been NOT DISALLOWED by the privacy protection agency.

2. *Terms of Reference*

At the commencement of every computer matching program:
(i) the participating organizations must identify one of themselves as the lead organization;
(ii) the lead organization must prepare a clear statement of the program's Terms of Reference, dealing with at least the matters listed in Exhibit A2 concerning the program's objectives, rationale, procedure, legal authority, justification, public knowledge and safeguards;
(iii) all participating organizations must examine the Terms of Reference, satisfy themselves as to that document's completeness and adequacy, and document their agreement to the Terms; and
(iv) the Terms of Reference must be reviewed, and not disallowed, by the privacy protection agency.

EXHIBIT A2: REFERENCE TERMS FOR A COMPUTER MATCHING PROGRAM

1. OBJECTIVES

The objectives of the program, expressed in sufficient detail to enable the reader to gain a clear understanding of the purposes.

2. RATIONALE AND PROCEDURE

° the target class or classes of individual
° the sources of data which are to be used
° the nature of data which is to be used
° the quality of data which is to be used
° the extent to which the proposed use of data relates to the purpose for which it was collected
° the scope and scale of the program (for example, all or some regions or classes of client)
° the matching techniques to be used
° the basis on which cases will be identified, inferences drawn and filtering undertaken
° the action which will be taken in relation to the identified cases
° any further use to which the data will be put
° the nature of benefits which will accrue as a result of this action
° the nature of costs, both financial and non-financial, which will arise

° the frequency of the program (for example, once-off, occasional or regular)

° any other factors important to the justification of the program

3. Legal Authority and Constraints

The legal authorities and constraints under which each participating organization is to perform its various activities, including:

° the disclosure of data

° the receipt of data

° data selection, data scrubbing, matching, inferencing and filtering

° the disclosure of resulting data

° the receipt of resulting data

° the use of resulting data

4. Justification

° the alternative means of satisfying the objectives which have been considered

° why the alternatives are inadequate in comparison with the computer matching program

° reference to post-evaluation reports of previous computer matching programs

° prior cost/benefit analysis of the program

° reference to the results of pilot tests undertaken, or if none has been undertaken, then justification as to why not.

5. Public Knowledge

° the measures to be taken to ensure public knowledge of the program

° if secrecy is essential to the operation, then explicit justification for it, and the surrogate measures being undertaken.

6. Safeguards

What measures are being taken to deal with:

° any special sensitivities that may relate to any of the data (such as racial origins, religious and political beliefs, opinions and affiliations, medical and financial data)

° any privileges that may relate to any of the data (for example, doctor-patient confidentiality)

7. Data Quality

Each source organization must, for every data item which it makes available to a computer matching run, clearly document the matters listed in Exhibit A3. The matching organization must ensure that these matters have been documented, and refer to that document during the design and operation of the computer matching run. If data is to be acquired from a source organization which may not have fully complied with these requirements (for example, an agency of a government, or a corporation, outside the jurisdiction), special care is needed by the matching organization. The lead organization must, in consultation with other participating organizations, design a quality assurance mechanism to ensure that outcomes are measured, analyzed, documented and fed back to the appropriate organization(s).

Exhibit A3: Data Quality Factors

The following must be clearly documented:

1. Definition

For every data item supplied to a matching organization, whether or not it is actually used in any part of the program:

° the precise definition of the data item in the originating system (for example, the particular meaning of "spouse," "child," "dependent," "income," or "asset" which applies in that system)

° if a data item's meaning has changed over time, how the relevant meaning is determinable (for example, by comparing the date on which the data was collected with the dates of changes in relevant definitions)

2. Relevance

For every data item supplied to a matching organization, whether or not it is actually used in any part of the program:

° the purpose(s) for which the data item was collected

° the purpose(s) for which the data item is to be used

° the relevance of the data item to the purpose(s) of use

3. Accuracy

° the precision with which the data was collected (for example, age in years at the time of a particular transaction, birth year, or full date of birth)

° where the data is textual or an opinion rather than factual in nature, where the source and the date it was expressed are to be found

° the extent to which the data item is subjected to validation procedures (for example, evidence may be required before date-of-birth is cap-

tured; date may be validated using a standard date-routine; date may be captured as supplied by the data subject; or date may be optional, non-validated data)

4. Timeliness

° if the data item's temporal applicability is limited, how the period of its applicability is determinable (for example, a benefit entitlement indicator may have associated commencement and termination dates; or an income item may have another item associated with it, showing the year to which it was applicable)

° the time-lag between the occurrence of a relevant event and its reflection in the data (i.e. the data item's "up-to-dateness")

° if the data item is volatile, the frequency with which the data item's correctness is checked (for example, income may be captured at time of application for benefit, and on such subsequent occasions as the person concerned advises change of circumstance or is interviewed as part of an audit program)

5. Completeness

If the data item is liable to misinterpretation unless used in conjunction with additional data, where that additional data is to be found.

6. Controls and Audit

For every data item supplied to a matching organization, whether or not it is actually used in any part of the program:

° whether documentary evidence and an audit trail exist to support the data and, if so, where they are to be found

° any relevant control measures used to ensure data quality

° any audit programs to which the data has been, or is regularly, subjected

7. Conduct of the Program

7.1 *The Data Scrubbing Step*

The organization which undertakes data scrubbing must:

(i) when designing the process, acquire and give consideration to the Terms of Reference and to the information available concerning data quality;

(ii) clearly document the algorithms applied to data to massage its format or content;

(iii) identify the impacts the processing is expected to have on the proportions of false hits and false misses;

(iv) gather statistics on the effects of the algorithms; and

(v) analyze the quality of the data scrubbing step.

7.2 *The Matching Step*

The matching organization must:

(i) acquire and give consideration to the Terms of Reference and to the information available concerning data quality and the data scrubbing step;

(ii) clearly document the matters listed in Exhibit A4, regarding the matching algorithm, the data definitions and the procedure for recognizing hits; and

(iii) gather statistics and analyze the quality of the matching step.

7.3 *The Inferencing Step*

The organization which undertakes inferencing must:

(i) when designing the process, acquire and give consideration to the Terms of Reference and to the information available concerning data quality, and the data scrubbing and matching steps;

(ii) clearly document the algorithms applied to matched data in order to draw conclusions about the person to whom the data purports to relate;

(iii) gather statistics on the effects of the algorithms; and

(iv) analyze the quality of the inferencing step.

EXHIBIT A4: MATCHING TECHNIQUES

The following matters must be clearly documented:

1. ALGORITHM

The precise matching algorithm used (for example, phonex equivalence of family name, phonex equivalence of first given name *and* birth date within two years).

2. RECORD DEFINITION

For every source which is used in the computer matching process, the precise meaning of a record appearing on that file (for example, "a person using that name and supplying those details has submitted at least one return to the organization during the last fifteen years")

3. DATA ITEM FORMAT, DEFINITION AND QUALITY

For every data item used in the matching process:

° the precise meaning within the source system(s). If the definition(s) have varied during the life of the system, the various definitions and their dates of applicability must be documented

° the precise specification of the content rules; for example, in respect of Chinese names, different systems may adopt different conventions for identifying a "surname";

° the quality factors affecting the data-item (for example, name variants, misspellings, use of initials, omitted second given names, address updated only annually, non-validated income or date of birth)

4. Procedure for Recognizing "Hits"

° the sources of error inherent in the matching technique, in particular:

- ° data items which, on the basis of their labels, appear to be the same, but whose precise definitions are different in a manner which may in some circumstances be significant (for example, name, address, spouse, income, number of dependents)
- ° the use of different identifiers by the same person (for example, professional and social names, and original and anglicized forms)
- ° low data quality
- ° differing dates of applicability of the different sources of data

° the measures considered, any trials undertaken, and the measures adopted, to address each such source of error, and improve the accuracy of the matching algorithm

° the strategy adopted regarding trade-off between false hits and false misses

° estimates of the proportions and numbers of false hits and false misses which will result from these errors

° the measures considered, and the measures adopted, to ensure that false hits and false misses do not result in misjudgments and erroneous decisions and actions harmful to the interests of individuals

4.1 The Filtering Step

The organization which undertakes filtering must:

(i) when designing the process, acquire and give consideration to the Terms of Reference and to the information available concerning data quality, and the data scrubbing, matching and inferencing steps;

(ii) clearly document the algorithms applied to data in order to draw conclusions about the person to whom the data purports to relate;

(iii) gather statistics on the effects of the algorithms; and

(iv) analyze the quality of the filtering step.

5. Requirements During the Latter Stages of a Program

5.1 *Use of the Results*

An organization must only use for the purposes identified in the program's Terms of Reference data which is disclosed for the purposes of a computer matching program or arises from such a program.

Before using data arising from a matching program, an organization must:

(i) review all information prepared by participating organizations, including the Terms of Reference and the information available concerning data quality, and the data scrubbing, matching, inferencing and filtering steps;

(ii) consider whether special account needs to be taken of the wider context of the matched data (such as the social values or political conditions prevailing at the appropriate times);

(iii) where any organization involved in the computer matching program may not have fully complied with these Guidelines (for example, an agency of a government, or a corporation, outside the jurisdiction), consider what special care is required; and

(iv) based on these findings, design and document procedures appropriate to the circumstances.

No organization may use the results of a computer matching program until and unless the conditions listed in Exhibit A5, regarding the data handling processes, data quality controls, due process, opportunity to appeal and the consequences of the decision or action, have been satisfied.

These requirements apply to all uses of such data, including:

- the making of an administrative decision about an identified individual;
- the taking of administrative action about an identified individual;
- use in or in relation to a court process; and
- the creation, amendment and extension of any record.

Where an organization uses in a court process data that has been input to or generated from a computer matching program, the data, its source and relevant information concerning its quality must be disclosed as part of the pre-trial discovery procedures.

Exhibit A5: Using Data from a Computer Matching Program

1. The Data Handling Process

The means whereby meaning is extracted from the matched data, including the data scrubbing, matching, inferencing and filtering steps, must be explicitly and clearly defined.

2. Data Quality Controls

° any reasonable cross-checking or verification of the relevant data must be undertaken before the decision is made or the action taken. This may include:
  ° checks against the organization's own records
  ° where appropriate, and subject to information privacy and other relevant laws, checks against the records of other organizations
° except where secrecy is essential to the operation, and has been explicitly justified, the individual concerned must be, at an appropriate point in time, advised of the information concerned, and given the opportunity to provide information which may counter or mitigate the information that has been derived from the computer matching program, before the decision is made or the action taken.

3. Due Process

° except where secrecy is essential to the operation, and has been explicitly justified, the individual concerned must be made aware, before the decision is made or the action taken, that:
  ° such a decision is to be made or action is to be taken
  ° information relevant to the decision or action has arisen from a computer matching run
° the information must be of a nature that could reasonably be expected to be accepted by a court of law as evidence

4. Opportunity to Appeal

° an avenue to contest or appeal against the decision or action must be effectively available to the individual concerned
° the existence of that appeals avenue must be communicated to the individual concerned, before the decision is made or the action taken

5. Consequences

Any secondary effects of the decision or action must be considered before the decision is made or the action is taken (for example, the temporary withdrawal of a benefits card as a disciplinary measure may result in the person no longer being able to take advantage of additional benefits normally available from other sources, such as pensioner discounts)

5.1 *Retention and Destruction*

An organization must not retain personal data used in or arising from a computer matching program any longer than is reasonably necessary for

the fulfillment of the objectives specified in the Terms of Reference. In general:

° in the case of data which is *not* the subject of a "hit", the data should be retained no longer than 2 months after the date on which the computer matching process was undertaken;

° in the case of data which *is* the subject of a"hit", the data should be retained no longer than 2 months after the date on which a decision is taken not to proceed further with the matter. A decision on whether or not to proceed further with the matter should be made no later than 2 months after the date on which the computer matching process was undertaken. If, 2 months after the date on which the computer matching process was undertaken, no decision has been taken to proceed further with the matter, it should be deemed that a decision has been taken not to proceed further with the matter.

An organization must dispose of personal data used in and arising from a computer matching program in a manner which ensures the data's security.

5.2 *Quality Analysis*

Participating organizations must measure, analyze, document and feed back to the appropriate participating organizations information regarding the quality of the data, processing steps and outcomes of a computer matching program.

5.3 *Post-Program Evaluation*

A post-program evaluation must be undertaken shortly after the completion of the computer matching program, to evaluate it against the original (and, where appropriate, amended) Terms of Reference and Prior Cost/Benefit Analysis, and to make information available to assessments of subsequent proposals for computer matching programs. Such a post-program evaluation must:

_ contain a copy of the original and amended Terms of Reference and Prior Cost/Benefit Analysis;

_ be in a form which facilitates comparison with the Prior Cost/Benefit Analysis;

_ contain descriptive data about the program, including the participating organizations, and the number of files, records, hits and cases arising;

_ contain measures of financial benefits and costs;

_ contain measures of non-financial but quantifiable factors;

_ contain evidence concerning the extent to which non-quantifiable benefits and costs have been achieved; and

_ include a comparison with the Prior Cost/Benefit Analysis.

Where a computer matching program comprises multiple computer matching runs, post-evaluation must be performed at least annually.

6. General Requirements

6.1 *Overview by the Privacy Protection Agency*

All aspects of computer matching programs must be under the purview of the privacy protection agency. The agency must have the power to disapprove any proposed program or proposed variation to a program, and to impose conditions on a program's implementation.

6.2 *Public Visibility*

In general, all information about a computer matching program must be publicly available in advance of the first run, for example, through publications of the relevant organization and of the privacy protection agency. Where one or more participating organizations wishes to claim in respect of particular information (for example, the matching algorithm) that publication would be detrimental to the conduct of the program, or would be in breach of secrecy provisions, the argument and the information must be submitted in writing to the privacy protection agency. If it accepts the argument, the privacy protection agency must explicitly authorize the non-publication, partial publication or deferral of publication, as appropriate, making clear precisely what information is being granted confidentiality.

The privacy protection agency must report to the public periodically concerning dispensations granted, in such a manner that disclosure of information prejudicial to the conduct of approved programs is avoided.

6.3 *Consistency with the Terms of Reference*

The conduct of a computer matching program must be consistent with the Terms of Reference. Any material modifications arising during the conduct of the program must be reviewed by all participating organizations, agreed by all of them prior to the modifications being implemented, notified to the privacy protection agency, and not disapproved by that agency.

6.4 *Security and Confidentiality*

Participating organizations must take appropriate precautions at all stages to ensure that personal data available to, used in and arising from a computer matching program:

- are not subject to accidental or intentional modification;
- are not accessed by staff within the organization except where such access is necessary for the conduct of that computer matching program; and

_ are not disclosed other than in accordance with the Terms of Reference and relevant law.

6.5 *Staff Training*

Participating organizations must provide staff with training in all aspects of computer matching, and of these Guidelines, that are relevant to tasks they may perform in relation to a computer matching program.

6.6 *Documentation*

Participating organizations must ensure that all documentation relating to a computer matching program is maintained in a form which facilitates access and assessment by:

_ staff of the organization in which the documentation originated;
_ staff of other participating organizations;
_ members of the public and their agents; and
_ the privacy protection and other regulatory agencies, in relation to the exercise of their functions.

6.7 *Subject Access*

Participating organizations must arrange for convenient access by individuals to records concerning themselves, including records used in and generated by computer matching programs, in compliance with the provisions of all relevant laws, such as Freedom of Information legislation.

6.8 *Audit*

Participating organizations must maintain all documentation in a form which facilitates the audit of the organizations' compliance with these Guidelines by internal and external auditors. Participating organizations must give consideration to including audit of compliance with these Guidelines as part of the organization's internal audit programs.