

DATAVEILLANCE BY GOVERNMENTS: THE TECHNIQUE OF COMPUTER MATCHING

[Roger Clarke](#)

Principal, [Xamax Consultancy Pty Ltd](#), Canberra

Visiting Fellow, [Department of Computer Science](#), [Australian National University](#)

Version of 8 July 1993

© [Xamax Consultancy Pty Ltd](#), 1993

This paper was published in Information Technology & People 7,2 (December 1994) 46-85

This document is at <http://www.anu.edu.au/people/Roger.Clarke/DV/MatchIntro.html>

Abstract

Computer Matching is a mass surveillance technique involving the comparison of data about many people, which has been acquired from multiple sources. Its use offers potential benefits, particularly financial savings. It is also error-prone, and its power results in threats to established patterns and values. The imperatives of efficiency and equity demand that computer matching be used, and the information privacy interest demands that it be used only where justified, and be subjected to effective controls

This paper provides background to this important technique, including its development and application in the United States and in Australia, and a detailed technical description. It is contended that the technique, its use, and controls over its use are very important issues which demand research. Computing, telecommunications and robotics artefacts which have the capacity to radically change society need to be subjected to early and careful analysis, not only by sociologists, lawyers and philosophers, but also by information technologists themselves.

Contents

- [Introduction](#)
- [Historical Description](#)
 - [Background](#)
 - [Origins](#)
 - [Contemporary Use](#)
 - [Control Measures](#)
 - [The United States](#)
 - [Australia](#)
 - [Interpretation](#)
- [Technical Description](#)
 - [Definition](#)
 - [Identification-Based Computer Matching](#)
 - [Content-Based Computer Matching](#)
- [Conclusions and Prospects](#)
- [References](#)

1. 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 Australia and Canada. Although the latter term is perhaps more usefully descriptive, the United States term is more common in the literature, and is used in this paper.

Computer matching became economically feasible in the early 1970s, as a result of developments in information technology (IT). The technique has been progressively developed since then, and is now widely used, particularly in government administration and particularly in the three 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 jeopardise the information privacy of everyone whose data is involved, and even significantly alter the balance of power between consumers and corporations, and citizens and the State.

Computer matching has been the subject of many government publications, and its impacts have been discussed by sociologists and to a lesser extent lawyers. It has, however, attracted remarkably little attention in the computer science and

sociologists and to a lesser extent lawyers. It has, however, attracted remarkably little attention in the computer science and management information systems communities (see, however, Kusserow [1984a], Shattuck [1984] and Clarke [1988a]). This paper is intended to make good the lack of computing literature on the topic. Its purposes are to provide a history of the use of computer matching, and a clear description of the technique, and thereby provide a springboard for future, more critical works.

The paper's scope is restricted to computer matching in the public sector. This is because firstly the technique's primary applications are in these areas; secondly access to both primary and secondary sources is even more difficult in the private sector than it is in relation to government; and thirdly the issues which arise in the private sector are somewhat different, and justify separate treatment elsewhere. Throughout the paper, reference is made to circumstances in the United States and Australia. There is good reason to believe that 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 to date less actively applied there. Because of differences in the legal context including civil rights, and the patterns of use of IT in personal data systems, the application of this paper outside the four countries mentioned requires care.

The research underlying the paper has been undertaken during a protracted period. Literature search and analysis have been supplemented by considerable field work, undertaken primarily in the United States and Australia, but also in Europe. A full report on this research is provided in Clarke [1992e]. Background is provided concerning the context in which computer matching arises. Its origins and contemporary use are discussed, and the history of controls presented. The next sections discuss definitional matters and provide technical descriptions both of conventional computer matching and of the more sophisticated form which is emerging.

2. HISTORICAL DESCRIPTION

This first section provides descriptive material relating to the rationale underlying personal data systems, the origins and contemporary use of computer matching, and the history of regulation of the technique in the United States and Australia. This establishes the context within which the subsequent sections can analyse computer matching's technical aspects, and future papers can assess its impacts.

2.1 BACKGROUND

There are many economic relationships in which the parties do not necessarily identify themselves. These include barter, and cash transactions ranging from fishing licences, through gambling (in both illegal and legal establishments), to quite large consumer items, including expensive cars and boats. Electronic analogues to cash, which share with it the characteristic of anonymity, are increasing in importance (e.g. Chaum 1985). In recent years, concern about the ease with which organised crime 'washes' its illegally gained cash has resulted in some governments setting maximum limits on cash transactions, and/or requiring reporting by financial institutions of large cash deposits.

In respect of many kinds of transactions, however, organisations of all kinds keep records, both for their own needs, and in order to satisfy various forms of regulation, particularly audit and taxation. In addition, where ongoing relationships exist between an organisation and other entities, such as those between supplier and customers and between employer and employees, it is normal for organisations to establish and maintain standing records about them. Where the other party to a transaction is a private individual, records containing personal data result.

There has been a tremendous increase in the 'information-intensity' of administration during the twentieth century [Rule 1974, Rule et al 1980]. One cause has been the increasing scale of human organisations, making them more remote from their clients, and more dependent on abstract, stored data rather than personal knowledge. Other factors have been an increasing level of education among organisations' employees, and more recently the brisk development in information technology. The tendency appears to be still growing for organisations to record more data about more transactions, and about more entities with whom they have dealings.

In both the United States and Australia, there are large numbers of organisations in federal, state and local government, and in the private sector, which hold large amounts of data about individuals. These organisations generally have an interest in ensuring reasonable quality of the data they hold. In many records systems, the data subject shares that interest. For example, employees and government benefits recipients have an interest in ensuring that the data about the account into which payments are made is accurate, and updated in a timely manner.

There are many circumstances, however, in which the interests of the data-subject may not coincide with those of the record-keeper. Many people expect to be able to exercise control over data concerning themselves; they do not wish to have what they perceive to be sensitive information available to other people ostensibly in their roles as employees but in practice in their many other roles as well. In addition, there are circumstances in which a person can gain from cheating, e.g. by receiving a higher pension if their *de facto* marital relationship is not known by the relevant government agency; paying a lower rate of

marginal tax if the Internal Revenue Service or Taxation Office is not aware that the person has two jobs; and receiving a loan if the lender is unaware of the person's existing commitments.

In such circumstances, organisations are likely to seek confirmation of the accuracy and completeness of data they collect. To protect their interests, they seek confirmatory data from sources other than the person concerned. The term 'verification' is in common usage for this purpose, but since that word implies a higher standard of truth or correctness than is generally possible, the word 'cross-checking' is more descriptive.

A great deal of data-handling is internal to a single organisation. Cross-checking, however, generally involves disclosure, since data is passed to and from staff and/or computer applications which are concerned with another function and/or within another organisation. Cross-checking may be done in an *ad hoc* manner, or under standing arrangements between organisations. The checking may be done with or without the knowledge or consent of the individual, and with or without legal authority.

Many cross-checking activities are triggered by an application by a person, for example for a job, a pension or a loan, in which case they are commonly referred to as 'front-end verification'. The inverse arrangement involves an organisation which has a transaction with a nominated individual automatically notifying another organisation of particular data, e.g. the person's new address. There appears to be no term in general use for this, but it might be described as 'front-end notification'.

Front-end verification and notification are forms of 'data surveillance'. This is a set of techniques whereby one or more individuals are monitored not by direct physical means, but rather through their data. The instances discussed so far, where the monitoring is of a specific identified person and arises as a result of some transaction involving that person, are forms of 'personal dataveillance' [Clarke 1988a], and the entity which is subjected to dataveillance may be referred to as 'the digital persona' [Clarke 1993].

Cross-checking may also be undertaken in the absence of a trigger provided by a specific transaction with a specific person. The motivation may be instead a generalised belief or suspicion that some people with whom the organisation deals may be transgressing standards, or that data concerning some people with whom the organisation deals may be incorrect and that the organisation's interests may be thereby harmed. In addition to assisting in personal dataveillance, this approach is a facilitative mechanism for 'mass dataveillance' of a whole population, for a reason related to some suspected but as yet unidentified portion of that population [Clarke 1988a].

Computer matching is one of the most important mass dataveillance techniques. It is the comparison of personal data which relates to a large number of people, and which has been sourced from two or more data systems. The primary differences from front-end techniques are that it is undertaken some time after the event, and applied to large numbers of people and records. The technique is used to detect people who may be of interest to the organisation conducting the match, or to its clients. For the most part, matching has been undertaken by government agencies, for the purpose of identifying people who may have (intentionally or unintentionally) received excessive benefits, or failed to pay appropriate taxes.

There are other, closely related facilitative mechanisms for mass dataveillance. One technique is 'data-linkage', by which is meant the storage in an individual's record on one file of that person's identifier in one or more other files, to enable prompt and reliable inter-relationship of data in the future. A second, 'data concentration', involves the merger of databases, or creation of new databases, to support a number of functions. This has been referred to as the 'national databank' issue, particularly in the United States. A third technique is the use of a 'common, multi-purpose identifier'. This has given rise to debates over national identification schemes such as the Social Security Number (SSN) in the United States, the Social Insurance Number (SIN) in Canada, and in Australia the withdrawn Australia Card proposal and the Tax File Number (TFN). These other facilitative mechanisms are discussed in Clarke [1988a], and are not further dealt with in this paper.

The significance of the growth of data surveillance is immense: "[Historically], America has struck the balance more in favor of individual freedom and diversity than organizational demands for control and efficiency ... I call this new world a 'dossier society' - the other side of the information economy ... From the individual's point of view, the most significant characteristic of the dossier society is that decisions made about us as citizens, employees, consumers, debtors, and supplicants rely less and less on personal face-to-face contact, on what we say, or even on what we do. Instead, decisions are based on information that is held in [remote, large-scale] systems, and interpreted by bureaucrats and clerical workers in distant locations. The decisions made about us are based on a comprehensive 'data image' drawn from diverse files. From a technical and structural view, the central characteristic of the dossier society is the integration of distinct files, serving unique programs and policies, into more or less permanent national data bases ... From a political and sociological view, the key feature of the dossier society is an aggregation of power in the federal government without precedent in peacetime America" [Laudon 1986b, pp.3-4].

For general reviews of data surveillance, see Rule [1974], Laudon [1974], Smith [1974 et seq], Westin & Baker [1974], Kling [1978], Rule et al [1980], Burnham [1983], Rule [1983a, 1983b], Marx & Reichman [1984], Marx [1985], Roszak [1986], OTA [1986b], Laudon [1986b], Clarke [1988a], Flaherty [1989] and Bennett [1992]. The theory of dataveillance provides important aspects of the context within which this paper examines computer matching.

2.2 ORIGINS

Some precursors to computer matching as it is currently conceived can be detected in the 'income matching programs' which have long been used in taxation administration in many countries. Another ancestor is the Parent Locator System, authorised by the United States Congress by amendment to the Social Security Act in 1974, and originally intended to enable parents who had violated agreements relating to the maintenance of their children to be located and collection to be enforced [Laudon 1986b, p.329].

The first computer matching program is generally identified in the literature as 'Project Match', conducted in 1977 by the then Department of Health, Education & Welfare (HEW), now the Department of Health & Human Services (DHSS). In fact there were a small number of earlier projects: "before 1976 only two benefit program-related projects were conducted" [GAO 1985]; and "The first significant match of benefit programs was conducted in 1976. The Federal Bureau of Investigations matched the employee records of various governmental employers in the Chicago area with Welfare files" [Kusserow 1983, p.1]. No evidence has been found of matching having been undertaken earlier in any field other than benefits control.

Project Match compared "the records of roughly 78% of all recipients of Aid to Families with Dependent Children (AFDC) with the payroll records of about 3 million federal employees". It identified 33000 raw hits, later reduced to 7100, resulting in 638 internally investigated cases, of which 55 resulted in prosecutions [OMB 1986b p.18; see also Fischel & Siegel 1980, Weiss 1983, OTA 1986b p.42, Early 1986]. It was claimed to be a great success. It appears, however, that these prosecutions resulted in only about 35 convictions, all for minor offences, with no custodial sentences and less than \$10,000 in fines. This paradox of a project being hailed as a great success when the measurable financial costs are high and the measurable financial benefits very low, has been a feature of matching programs from the very beginning. With a few significant exceptions, it continues to be so.

In Australia, the history of use is nearly as long. A Report by the Law Reform Commission referred to a complaint dealt with by the Commonwealth Ombudsman in 1979, which related to matching by the Department of Social Security (DSS) of data received from a State superannuation authority [ALRC 1983, p.175], and of payroll data received from State government bodies as early as 1978 [p.176]. Matching between DSS and Veterans' Affairs files has subsequently been stated to have "commenced in the early 1970s" [PCA 1990, p.22] and "in the late 1970s" [DSS 1992, p.30].

By 1982, it was estimated that U.S. state and federal agencies routinely carried out about 200 programs [Cohen 1982]. The Reagan Administration instituted a drive for efficiency in government, and the President's Council on Integrity and Efficiency in Government (PCIE) "has been the strongest proponent of computer matching as a management tool" [Flaherty 1986, p.344]. The Congress' Office of Technology Assessment estimated a tripling in use between 1980 and 1984 [OTA, 1986b, p.37]. Laudon estimated 500 programs in 1986 [1986b, p.383].

In 1984-86, a GAO study found that the "current climate or environment surrounding computer matching" was a primary determinant of growth in its use. In particular, the report noted "(a) a rising concern about erroneous payments, (b) technological developments that make computer matching easier or more feasible ..., (c) reports of successful matches with large cost savings or cost avoidances, and (d) endorsement and recommendations by key oversight organisations" [GAO 1986c, p.2]. These are more consistent with a 'fashion' model of decision-making than a 'rational management' model. Unsurprisingly, growth continued. This was particularly so in the States, as a result of federally imposed requirements [GAO 1990, p.25].

For more detailed historical accounts of computer matching in the United States, see Smith [1974 et seq], Kirchner [1981], Cohen [1982], Azrael [1984], Laudon [1986b, pp.328-335], Greenberg & Wolf [1985, 1986], OTA [1986b] and Flaherty [1989, pp.344-358]. For a review of the history in Australia, see Clarke [1992d].

The next sections describe the purposes to which the technique is being applied, and the history of controls which have been imposed on its use.

2.3 CONTEMPORARY USE

This section analyses the manner in which computer matching is being used in government agencies, firstly on the basis of the various purposes to which it is being put, and secondly in relation to its application in particular areas of government administration.

Objectives

Computer matching is being used, or is capable of being used, for many different purposes. A list is provided in Exhibit 1 (after IPCO [1991], but significantly extended).

The majority of purposes have to do with social control and efficient government administration. There are of course indirect benefits to the public. in the form of lower taxes or. more realistically. the potential for more services in return for the same

benefits to the public, in the form of lower taxes or, more realistically, the potential for more services in return for the same tax. The sole direct advantage for data subjects (the identification of individuals who are eligible for a benefit or a higher benefit) is something of a cinderella. In one instance, it was discovered that the enabling legislation failed to include authority to use the results of matching programs for such a purpose [DSS 1991, p.42].

The following sub-sections document specific applications, each of which addresses one or (more commonly) several of the above objectives.

Taxation and Financial Institution Data

Income Matching Programs have been operated by the taxation authorities of both the U.S.A. and Australia since before computers began to be applied to administration during the 1950s. In evidence presented by the Australian Taxation Office (ATO) to the Senate Estimates Committee C [Hansard, 18 Oct 1988, pp.C154-170], the ATO stated that it received from financial institutions about 15 million reports annually relating to interest paid on investment accounts. About 65% of the financial data was successfully matched to taxpayers (later revised upwards to 70% [PCA 1990, p.17]). Since financial institutions supplied very little identification data, matching errors may have been a major problem, but no information is available on the incidence of error. The 65-70% match-rate was contrasted to the 95% match-rate claimed in respect of the Prescribed Payments Scheme, which is based on the Tax File Number rather than names. The ATO stated that between 8 and 12 percent of the interest income that should have been reported by individuals was not being reported, resulting in a loss of \$300-500 million in tax revenue. In 1990, the ATO stated that, in 1988, the program "raised \$140m in revenue, with an average staff level of 500" [PCA 1990, p.18].

In 1988, legislative approval was given for a major enhancement to the Tax File Number (TFN) scheme, including a requirement on all employers and financial institutions to collect and store their employees' and investors' TFNs, and include them on reports to ATO of salaries and interest paid and tax deducted. This, it was argued, would significantly lower error rates, and thereby increase collection levels [Clarke 1992c].

Exhibit 1: Objectives of Computer Matching Programs

- primary purposes:
 - **detection of errors** in program administration (e.g. erroneous assessment of benefit amounts, multiple invoicing);
 - **confirmation of continuing eligibility** for a benefit program, or compliance with the requirements of a program;
 - **detection of illegal behaviour** by taxpayers, benefits recipients, government employees, etc. (e.g. fraudulent or multiple claims, unreported income or assets, impersonation, omissions, unauthorized use, improper conduct, conflict of interest);
 - monitoring of grant and contract award processes;
 - **location of persons** with a debt to a government agency;
 - identification of those eligible for a benefit but not currently claiming;
 - **data quality audit**;
 - **updating of data** in one set of records based on data in another set;
- contributions to additional purposes:
 - support for actions with beneficial **financial** effects:
 - cancelling of incorrect payments;
 - reduction of excessive payments;
 - recoupment of incorrect agency payments, underpayment of tax or delinquent debts, including the withholding of benefit payments or taxation refunds in order to repay debts;
 - recoupment on behalf of other government agencies;
 - avoidance of future erroneous or excessive payments;
 - deterrence of future fraudulent behaviour;
 - construction and maintenance of **databases** for:
 - social control purposes;
 - research and statistical purposes;
 - improvement of program policy, procedures and controls.

The ATO conducts other matching programs, including one between data from the the Department of Social Security's Unemployment and Sickness Benefits register and its own Prescribed Payments System [PCA 1990, p.20], and another using data from the Australian Customs Service regarding sales tax returns [p.34].

Social Welfare

Social Welfare

In 1990, the United States Social Security Agency (SSA) conducted 42 matching programs, 10 of which used only internal data-sources (SSA 1991b, Clarke 1992a). Exhibit 2 provides a list classified according to type. SSA provided data for a further 13 programs conducted by other agencies, including the Veterans' Administration, the Internal Revenue Service, the Selective Service System agency, the Department of Labor, and State government agencies (SSA, 1991b).

The primary purposes of SSA matching programs are:

- comparison of income, as recorded on other files, against self-projected or self-declared income, in order to detect under-reported income and resultant over-payment of benefits;
- comparison of resources or assets, as recorded on or implied by the contents of other files, against self-declared resources, in order to detect under-reported resources and resultant over-payment of benefits. The primary example of this is interest-income data from the Internal Revenue Service (IRS), from which cash-investments can be inferred; and
- location of individuals, particularly those who have defaulted, e.g. on family support payments or recovery of overpayments.

Exhibit 2: SSA Matching Programs, 1991

- **with other benefit-paying organisations** (5 agencies, 16 programs)
 - Department of Defense (1.5 million retirees, annually);
 - Office of Personnel Management (2.5 million retirees, continually, bi-monthly, monthly and quarterly);
 - Railroad Retirement Board (1.1 million people, frequently);
 - Veterans' Affairs Administration (4.8 million recipients, bi-monthly and annually; 2 million income dependents annually);
 - Private Pension Employee Master (7.1 million private sector employees, monthly, based on data collected periodically from employers);
- **with the Internal Revenue Service** (5 programs):
 - 8.3 million people re non-wage income, three times p.a.;
 - 145 million people, annually;
 - address changes, monthly;
 - new SSNs, weekly;
- **external, with other federal agencies** (8 agencies, 12 programs, 2 common with State programs):
 - Immigration and Naturalisation Service (123,000 aliens absent from the country, frequently; 90,000 deported aliens, daily);
 - Department of Labor (130,000 Black Lung patients, monthly; 48,000 Employees Compensation recipients, pilot in 1989);
 - Selective Service System Registry (16 million people, quarterly);
 - Parent Locator Service (thousands of people, weekly);
 - Bureau of Public Debt (68 million Savings Bond Register entries, annually);
 - HCFA (37 million Medicare enrollees, annually; millions of Medicaid Statistics records, pilot);
 - Prisons (continually);
 - an unidentified set of agencies (deaths, daily);
- **external, with State Agencies** (scores of agencies, 7 programs):
 - prisons (continually);
 - Deaths Registers (daily);
 - Workers' Compensation (intended pilot with Alabama);
 - Wages - (3 separate programs, each with 3 different States, in pilot);
 - Unemployment Compensation (pilot with 9 States);
 - Title XIX (Medicaid Facilities) (six States, continually);
- **internal** (10 programs):
 - Master Beneficiary against Supplemental Security (daily);
 - Master Beneficiary against Master Earnings (3 times yearly);
 - Master Beneficiary against Black Lung (on request);
 - Recovery of Overpayment Accounting and Recording (ROAR) file against Supplemental Security (daily);
 - Master Beneficiary against itself (to detect duplications, semiannually);
 - Supplemental Security against Master Earnings (semiannually);
 - Numident file against Supplemental Security (daily);
 - Numident file against Master Earnings (annually);
 - ROAR file against Master Beneficiary (annually);
 - Master Earnings against own Payroll (variable).

At least some matching programs are used to update SSA records, and some are also used to update the source agency's records. This presumably relates particularly to address data.

In 1990, SSA provided data for matching programs undertaken by other agencies in the following cases:

- State/Local Agency External Matches (to many agencies, 6 programs);
- Pennsylvania Lottery File (monthly);
- Medicaid Eligibility (against NY City Deaths, monthly);
- SSA Benefits data to States (fortnightly);
- SSA Supplemental Benefits data to States (monthly);
- SSA client data to States (13 million, frequently);
- SSA client data on-line ('Wire to Wire', to 10 States).

In Australia, very little about the practices of the Department of Social Security (DSS) has been publicly knowable until quite recently. A 1987 request by the author under the Freedom of Information Act was rejected on the basis of various exemption clauses. The Department provided a general answer, which stated that computer matching programs were undertaken, using data from Departments administering laws concerning student assistance, taxation, veterans and illegal immigrants. Similarly vague information was provided by the Department in evidence to the Senate Standing Committee on Legal and Constitutional Affairs in its hearings relating to the Tax File Number in October 1988.

No mention was made on either occasion of the exchange of data with a wide range of organisations which are listed in one earlier and four subsequent documents [ALRC 1983, O'Connor 1990, DSS 1990a, PCA 1990, DSS 1992]:

- the Health Insurance Commission (HIC - in relation to Medicare clients);
- the Department of Finance (in relation to Commonwealth employees);
- the Child Support Agency within the ATO (in relation to maintenance default);
- the Australian Electoral Office;
- the Department of Immigration (regarding students leaving the country permanently);
- the Department of Industrial Relations (regarding pre-apprenticeship allowances);
- the electronic version of the telephone directories Australia-wide;
- State Registrars of Births, Deaths and Marriages;
- State Land Titles offices;
- State Departments of Corrective Services (regarding prisoners);
- employers in the State government and private sectors;
- banks;
- insurers;
- superannuation funds; and
- educational institutions.

DSS has the power under its legislation to demand data from any organisation, provided that the data is relevant to the question of benefit entitlement (Social Security Act, s.164). It seems, however, that this is only a fallback: "These powers are used infrequently for data-matching purposes because it has been the Department's experience that most non-government organisations approached for identity/payment/other circumstances information supply it without legal enforcement" [DSS 1990a, p.6]. It would have been expected that this would have changed since commencement of the Privacy Act 1988, almost 2 years prior to the date of the quotation

During 1989 & 1990, the Commonwealth Government stated a policy commitment to increased use of computer matching in the control of social welfare abuse. In particular, in its Economic Statement of February 1990, it proposed further sources of data for computer matching, including:

- State Government Land Titles Offices, to detect undisclosed real estate assets;
- organisations which make compensation payments, including the Workers' Compensation Authority;
- local government;
- overseas governments, commencing with New Zealand, to detect incorrect payments to 'non-citizens' participating in each country's social security programs;
- international arrivals, departures and visa records;
- Child Support Agency records, to detect pensioners who are not fulfilling their statutory obligation to take court action to collect unpaid maintenance;
- Australian missions overseas, regarding student benefits;
- Telecom; and
- Australia Post.

A report on the front page of the Sydney Morning Herald of 22 March 1990 also stated that DSS was to undertake a match with insurance company records, expected to result in 2400 cases saving \$98m over 2 years. See also PCA 1990, pp. 25-26.

with insurance company records, expected to result in 2400 cases saving \$98m over 3 years. See also PCA [1990, pp.35-36].

In December 1990, the Data Matching Program (Assistance and Tax) Act 1990 authorised the hitherto largest-scale Australian data surveillance program and its constituent database maintenance and computer matching programs [Greenleaf 1991]. All four major client-oriented agencies are involved as both source and recipient agencies, and the taxation, electoral office and health insurance agencies as source agencies. The program is subject to a set of controls [DSS 1991, PCA 1991, Kelly 1991]. "The objectives of the data-matching program are:

- to detect instances where persons are possibly receiving incorrect payments from an income support agency ...;
- to verify, independently and automatically with the Australian Taxation Office, the accuracy of figures of ... income disclosed to agencies which make income support payments ...; and
- to detect instances of tax evasion" [DSS 1991, p.12].

There are many other likely participants in such schemes. Additional Commonwealth agencies include:

- the Health Insurance Commission (in respect of the Pharmaceutical Benefits Scheme, which it recently acquired from the Department of Health);
- the Commonwealth Employment Service;
- the Department of Defence;
- the Department of Foreign Affairs & Trade (in respect of the Passports Office); and
- the Department of Primary Industries & Energy (in respect of its many registers of participants in the agricultural, pastoral and mining industries).

At State Government level, agencies which would seem likely to become involved include the regional public water, electricity and gas utilities.

Housing

The U.S. Department of Housing and Urban Development (HUD) provides rental assistance to low-income families (Clarke 1991). It does not provide the funds to the families, but instead either directly to landlords or via public housing agencies (PHAs) which own and maintain housing facilities. PHAs collect a contribution from the family, commonly about 30% of the family income, excluding the incomes of children under 18 and full-time students, and use HUD funding (and perhaps some other sources) for the remainder.

There are some 3000 PHAs, organised on a regional basis throughout the country. Some of these are agencies of State, County or City governments, while others are independent not-for-profit organisations. They vary greatly in size. HUD rental assistance amounts to about \$9 billion annually.

Applicants for rental assistance apply to the appropriate PHA, and provide data concerning their household, including the Social Security Number (SSN) and income of each member. Once approved for assistance, they are required to provide income data annually. The extent to which the rent is subsidised depends on family income, and beneficiaries under these schemes stand to gain by under-stating their income. Front-end verification can be undertaken with some sources, particularly state wage information collection agencies, and directly with organisations nominated by the applicant as a source of income. There is no certainty that all sources will be detected, nor that the level of income will remain the same. HUD does not require PHAs to actively search for unreported income. Moreover, many PHAs lack the legal authority to gather data from all appropriate sources, especially Federal government agencies.

Aggregated income data is subsequently collected by a number of sources:

1. most States gather data, quarterly in arrears, from all private sector employers concerning income paid to their employees;
2. the Federal Social Security Agency (SSA) gathers similar data from all employers, annually in arrears; and
3. the Federal Office of Personnel Management collects personnel data, including individuals' pay grades, from Federal agencies.

HUD estimates that "about 12% of HUD-assisted households are either ineligible or are receiving more assistance than is allowable" [HUD 1988, p.1]. It therefore considers it important that PHA data be matched against at least some of these income data sources, in order to detect cases of apparent under-statement of family income and hence over-subsidy. HUD undertakes this matching, because it has the necessary legal authority to gather the necessary data and the technical capability to plan, implement and manage matching programs, whereas PHAs generally do not.

Because there is at present no standardisation of formats among the various PHAs, even those within the same State, it has been hitherto necessary for each match to be planned and implemented as a separate program. This is gradually changing, as HUD's standard reporting format becomes increasingly widely used, its central database of PHA tenant data is developed, and software suppliers sell standard packages to PHAs.

software suppliers sell standard packages to PHAs.

Health Insurance

Australia has had a national health insurance scheme since 1983. It was reported in the press that, when it established the register, the Health Insurance Commission (HIC) acquired and matched personal data from a large number of government agencies. Because the scheme is universal in its coverage, and centrally administered, the database contains data on virtually every person in Australia.

HIC proposed to repeat the computer matching exercise in more sophisticated fashion in 1985-87, in order to create a national register to support the then-proposed Australia Card scheme [Greenleaf & Nolan 1986, Clarke 1987a, Graham 1990]. In 1986, it conducted a pilot match with the intention of demonstrating the feasibility of producing a highly accurate and complete register, based on a number of largely inaccurate, inconsistent and incomplete files from different agencies. It expropriated and matched data from several government agencies, relating to all inhabitants of the small island-state of Tasmania [HIC 1986]. The agency reported the 70% hit-rate across the databases as a good result, confirming its belief that a National Identification Scheme could be based on such a procedure. The report ignored the implication that, across the national population, the records of nearly 5 million people would remain unmatched, and they failed to apply any tests to establish what proportion of the 70% were spurious matches and what proportion of the 30% were failures of the algorithm used. There is a popular mythology that everyone in Tasmania is related to everyone else. For this reason alone, the agency might have been expected to recognise the need for such testing.

It is known that HIC provides data to at least the Department of Social Security for various purposes, including matching programs. It is not publicly known whether the Commission itself conducts other programs, but given that its functions include protection of its very large cash flows against error and 'medi-fraud', there are clearly a number of potential applications for the technique. Recently the Commission gained responsibility for the Pharmaceutical Benefits Scheme and changed its identification basis from the family-level to the client-level. It then proposed to dramatically increase the amount of data about prescriptions which is collected and stored centrally. The proposal was rejected following a joint review by the Auditor-General and the Department of Finance, which concluded that the scanty financial justification over-stated gross savings by a factor of between 2.3 and 3.8 times [ANAO 1991]. As and when the proposal is resuscitated, it is reasonable to presume that it would provide a further source for computer matching.

Law Enforcement

It is customary in many countries to suppress information about the enforcement of the criminal law, and even to exempt such applications from the controls applied to other agencies' use of privacy-invasive techniques such as computer matching. It is therefore very difficult to gather evidence concerning actual usage for such purposes, but it would be naive to assume other than that it is used, perhaps extensively, by various branches of the police, national security, customs and defence communities.

Other Federal Government Applications

In Australia, the Review of Systems for Dealing with Fraud on the Commonwealth [SMOS 1987, pp.92-115 and in particular 98-106] stated that "information ... is exchanged for a wide variety of purposes and by formal and informal means", and identified a number of explicit instances of computer matching which were already undertaken, implying that they were done on a regular basis. Apart from those already discussed above, these included matches between [p.103]:

- social security and veterans data (re dual benefits payment and inconsistency of declarations);
- health insurance and veterans data (re claims for medical expenses);
- immigration and social security data (re visa overstayers);
- education and tertiary institution data (re enrolment of recipients of student benefits), which began in the mid-1970s, and is conducted twice yearly [PCA 1990, p.28];
- education and taxation data (re income of recipients of student benefits); and
- education and immigration (re visa overstayers).

The Privacy Commissioner's partial review [PCA 1990] unearthed two further programs:

- veterans' affairs and superannuation fund data, claimed to have been commenced in the early 1970s [pp.23-24]; and
- student allowances and employer returns, proposed for implementation in 1990 [p.27].

It was implied by Government statements in February and August 1990 that DSS-style monitoring procedures were to be applied in other agencies also, in particular to student assistance programs.

An example of an application of computer matching to a national policy measure related to the drafting of young men into the

United States national service during the Vietnam War. Social Security records, which contain the dates of birth of the overwhelming majority of young people, were used to establish a file of eligible persons. This file was matched with files from the Department of Defense to select out those already in the military, and the remainder were matched against Internal Revenue Service files to obtain mailing addresses [Laudon 1986b, pp.331].

State and Local Government Level

The U.S. Federal Government has not only applied the technique within and among its own agencies, but it has also imposed requirements on State government agencies. A succession of statutes, culminating in the 1984 Budget Deficit Reduction Act, required state administrations to match data from a variety of their own data systems and those of other agencies of the same State, agencies of other States, Federal Government agencies, and the private sector. These requirements are supported by fiscal sanctions. As a result, some hundreds of interstate government matching programs are also in place, as are programs involving interchange between different, particularly adjacent, local government areas. See Kusserow [1983, 1984b].

Federal Matching Programs Drawing on Private Sector Data

It is clear that matching of private sector with public sector data is feasible, and may be beneficial. A recent survey of United States agencies identified six which currently source data from the private sector, but no information was provided as to what and from whom [GAO 1990c, p.30]. The Department of Social Security uses 'Australia on Disk', a machine-readable list of all telephone subscribers [DSS 1991, p.9]. Given the limited requirements for public notification and the preference of many organisations to undertake such programs surreptitiously, many other uses of private sector data may already occur.

The Private Sector

Computer matching is also used by the private sector for its own purposes. One example is in the construction of consumer profiles through the merger of mailing lists. Companies in the consumer credit and insurance industries exchange a considerable amount of data about their clients. In most cases this is triggered by transactions involving identified individuals (e.g. an application for a loan, or failure to pay instalments on an existing loan), but large-scale sifting operations across complete sets of files have been undertaken, particularly in the insurance industry. A recent survey of United States federal agencies identified 13 which provided data for computer matching to a private sector organisation, although what data and to whom was not shown [GAO 1990c, p.30]. This paper's focus is, however, restricted to the use of computer matching by government agencies.

This section has documented the very wide range of computer matching activities currently undertaken in the United States and Australia, by a wide range of agencies, using data from a wide range of sources, for a wide range of reasons. The next section traces the history of measures used to exercise control over computer matching.

2.4 CONTROL MEASURES

Agencies have seldom pretended other than that the use of computer matching involves serious invasions of privacy. The nature of these risks is examined in detail in section 4. This section reviews the history of the processes whereby the uses of computer matching is subjected to controls, to ensure that information privacy is appropriately balanced against other interests, and the invasiveness of applications is limited to that which is justifiable.

2.4.1 THE UNITED STATES

The U.S. Congress acted on a number of occasions between 1970 and 1983 to alter the balance of power between citizens and the State. Laudon [1986b, pp.372-4] lists fourteen major pieces of legislation in such areas as credit reporting, criminal justice information systems, education, taxation, banking, electronic funds transfer and debt collection. All were created within the framework set by Westin [Westin 1967, Westin 1971 and Westin & Baker 1974]. Westin had found no problems with extensive surveillance systems as such, only with the fairness of the procedures involved.

The centrepiece of this legislation, the Privacy Act of 1974, nominally precluded government agencies from transferring personal data among themselves without explicit legislative authority, or the consent of the data subject. In practice, the Act was quickly subverted. Publication of uses in the Federal Register proved to be an exercise in bureaucracy rather than control. The 'routine use' loophole in the Act was used to legitimise virtually any use within each agency (by declaring the efficient operation of the agency to be a routine use), and then virtually any dissemination to any other federal agency (by declaring as a routine use the efficient operation of the federal government as a whole). See PPSC [1977], Marx & Reichman [1984, p.449], OTA [1986 at pp.16-21] and Berman & Goldman [1989].

The original Bill had proposed to establish a permanent Privacy Commission, to ensure compliance by agencies with the law. Under threat of Presidential veto, this was replaced by a short term study commission, and oversight of the Act made the

responsibility of the Office of Management and Budget. Unsurprisingly, given its primary responsibilities, OMB has consistently interpreted the Act in a manner unsympathetic to information privacy concerns [see, for example, Kirchner 1981, Laudon 1986b, p.374-5 and Flaherty 1989, pp.346-9].

'Guidelines to [Federal] Agencies on Conducting Automated Matching Programs' were first created in 1979 [OMB, 1979a, b]. A much shorter, revised version was later issued [OMB, 1982a, b], which weakened some aspects of the guidelines, especially in relation to cost/benefit justification and demonstration that alternative means were insufficient. This appears to have resulted from submissions by agencies to the effect that such analyses were expensive and/or that programs were difficult to justify on the basis of quantifiable benefits, and pressure from the President's Commission for Integrity and Efficiency. Subsequent regulatory documents include OMB/PCIE [1983] and OMB [1985].

At no stage does the Privacy Act or the OMB Guidelines appear to have provided any significant form of restraint on computer matching [Flaherty 1989, pp.349-350], and there is no mechanism whereby the Guidelines are enforced [Flaherty 1989, p.357, quoting a report by a Committee of the House of Representatives]. Very few of the matching programs which have been, or are being conducted, appear to have ever been the subject of explicit congressional approval or review. Laudon concluded that "as it stands now, the reigning principle is that virtually any federal system may be developed in the absence of a clearcut injury to individuals (the 'bodies floating in the river' criterion)". This colourful expression refers to a reported quip from the Secretary of the agency responsible for the original Project Match in 1977 [Laudon 1986b, pp.382].

There are also significant doubts about whether organisations are liable in law in the event that they fail to assure data quality. For example, in his study of criminal justice systems, Laudon concluded that "under existing interpretations, police officers are generally not thought liable for false arrest in situations where the arrest was based on erroneous information ... The FBI, state and local systems do not appear liable for disseminating false information ... It is not clear if law enforcement agencies are liable for disseminating erroneous ... criminal records under the [Constitutional] guarantees of due process" [1986b, p.266].

Bills to regulate Computer Matching were introduced into the Senate in 1986 and 1987 by Senator Cohen (Republican, ME). After many amendments, the Computer Matching and Privacy Protection Act 1988 became Public Law 100-503 [S.496] of October 18, 1988 and amends Section 552a of title 5 of the United States Code. Progress with implementation of the Act was slow. It was originally scheduled to commence operation on 1 July 1989, and OMB issued guidance interpreting the law on 19 June 1989 [54 Fed Reg 25818]. In mid-July, the date of commencement was postponed to 1 January 1990 [HR 2848]. On 1 September 1989, OMB published a list of existing matching activities [54 Fed Reg 38364].

The Law contains manifold exceptions and exemptions. It has been subjected to restrictive interpretation by OMB, and that agency has blatantly disregarded its responsibility to provide Guidelines relating to Cost/Benefit Analyses. The Annual Reports under the Act have been very slow in appearing, and very limited in scope. The primary conclusion of the first report was that agencies had underestimated the impact of the verification and notice provisions. In due course a Congressman was found to sponsor an amendment Bill, and H.R. 5450 of 1990 significantly weakened the verification and notice provisions. The law is further discussed in section 5.3 below. Meanwhile, agencies of State governments have in many cases simply ignored the inconvenient provisions.

2.4.2 AUSTRALIA

Australia has debated the need for information privacy legislation since the late 1960s, but until very recently successive Federal Parliaments had done nothing, and State Parliaments had done almost nothing. Until the beginning of 1989, only ineffective credit reporting laws existed, plus incidental protections such as breach of confidence, telephonic interception, trespass and official secrecy laws.

It appears that no guidelines regarding the conduct of computer matching were ever established by any Australian Government agency with general responsibility. There was discussion of regulation of computer matching in the Law Reform Commission's Report on Privacy [ALRC, 1983], but no concrete conclusions were reached.

During the period 1985-87, the Government attempted to gain acceptance for a national identification scheme it called the Australia Card by offering an extremely weak Privacy Bill. The proposal was eventually withdrawn in the face of overwhelming public opposition [Clarke 1987a]. In its place, in 1988, the Government sought Parliamentary approval for less contentious enhancements to the Taxation Office's identification scheme. It offered a far less inadequate Privacy Bill as a means of placating public concerns, and the package was passed.

The Privacy Act 1988 created a permanent Privacy Commissioner, and required him to, among many other things, 'research into and monitor developments in ... data-matching' [s.27(1)(c)]. Related provisions provide him with the function of examining a proposal for data-matching on request by a Minister or agency (k), and the power to do all things necessary or convenient to be done for or in connection with the performance of these functions (s.27(2)). The Privacy Commissioner also has computer matching responsibilities in relation to the Tax File Number (s.28 and Schedule 2).

The Privacy Commissioner is explicitly required by the Privacy Act 1988 to have regard to 'the protection of important

The Privacy Commissioner is explicitly required by the Privacy Act 1988 to have regard to 'the protection of important human rights and social interests that compete with privacy' (s.29): privacy concerns about computer matching are not to be considered in isolation, and the benefits must be taken into account in assessing computer matching practices, and in preparing guidelines. The Privacy Act applies only to agencies of the Commonwealth Government, and those which are exempt from the Freedom of Information Act 1982 are also exempt from the Privacy Act [Clarke 1989, Greenleaf 1989a, 1989b].

In 1987-88, FOI requests to, and discussions with, a variety of Commonwealth Government agencies elicited the following information:

- despite the Freedom of Information Act's instructions to government agencies to be more open in their dealings with the public, many retained a strong preference for undertaking their work without publicity;
- all requests under the Freedom of Information Act for information concerning computer matching policies, programmes and guidelines were refused, variously because:
 - the relevant documents were exempt because they were 'documents affecting enforcement of law'; and/or
 - no document could be identified which contained the desired information.

Meanwhile a report by a committee representing government agencies, on a Review of Systems for Dealing with Fraud on the Commonwealth [SMOS 1987] firmly recommended "that the matching of data is an effective means of preventing and detecting fraud, and that its wider use should be considered and publicised" [p.5]. It was very supportive of the Australia Card proposal as a means of improving the reliability of computer matching [pp.100-102]. There have been other reports within the Commonwealth public sector which have supported or urged the increased use of matching, e.g. ANAO [1990], but these have seldom contained discussion of the protections necessary as a concomitant to the powers.

In late 1990, the Privacy Commissioner published an Exposure Draft of a set of Guidelines, with the intention of communicating the Privacy Commissioner's intended interpretation of his responsibilities and powers under the Act, and the standards which he is considering applying in their execution. In preparing the Guidelines, the Commissioner had regard to these explicit powers, but also had in mind the need for them to provide practical guidance on computer matching programs. Agencies sought and gained lengthy extensions to the submission date; their submissions are not publicly available.

During late 1990, a separate line of development saw the Government introduce a Bill to authorise the Department of Social Security to conduct a series of large-scale computer matching programs. After considerable negotiation and re-drafting, the Bill was passed by Parliament in December 1990, incorporating a set of controls related to, but different from, the Exposure Draft of the Commissioner's general Guidelines [DSS 1991, Kelly 1992].

The Privacy Commissioner issued revised draft guidelines in September 1991 [PCA 1991b]. These contained major weakenings in the positions adopted in the original draft. A set of Final Proposed Guidelines was completed in December 1991. These corrected some of the weaknesses which the previous draft had created. They were sent to agencies with a request that they voluntarily adopt the Guidelines for a period of 18 months [PCA 1992]. The document's existence and the procedure adopted were not publicly known until mid-1992, however, even by public interest groups and the computing professional body which were on the Commissioner's mailing lists and had submitted comments on the two drafts. Very few agencies accepted the invitation to adopt the Guidelines voluntarily, and the Commissioner's hope that the delay in implementation would result in valuable experience being gained was forlorn. It is easy to interpret the process as evidencing the capture of the 'watchdog' by the government agencies it is nominally required to control.

After a very late start, Australia has created a control mechanism in relation to information privacy generally which is subject to the critical weaknesses inherent in the Westin philosophy, and applicable only to some government agencies and in very restrictive ways to the private sector, but which within those limitations is of considerable consequence. The extent to which computer matching will be controlled will be limited, however, because of the considerable respect which the Privacy Commissioner feels constrained to accord to government agencies, the failure to consult with and to take account of the submissions of other interested parties, and the weakening in the regulatory proposals evident in successive drafts of the Guidelines.

2.4.3 INTERPRETATION

There has been a succession of failed attempts by U.S. Federal agencies to gain Congressional approval for a unified national information system. Laudon [1986b, pp.352-8] notes the following:

- the National Data Centre "designed in 1968 by Lyndon Johnson's great society planners to combine in a single data base all of the information on American citizens held by the federal government, to support obscure goals such as 'better social planning' and 'statistical analysis'";
- FEDNET, a 1972 proposal to create a network of federal data banks;
- the Internal Revenue Service's proposal to develop a Tax Administration System, brought forward in 1974, and finally laid to rest in 1978;

- the Social Security Administration's Future Process, proposed in 1978 and withdrawn in 1979; and
- the long-running saga of the national computerised criminal history system (CCH).

The explosion of computer matching in the United States can be regarded as a response to these failures.

In the eyes of federal public servants charged with the administration of large and complex systems, and the handling of a vast quantity of public money, computer matching programs act as a proxy for a comprehensive population data surveillance system. It is therefore unsurprising that agencies should have sought to neutralise attempts to regulate their use of the technique.

A wide range of methods are available whereby agencies can counter attempts at regulation. Class-based and agency-based exemptions can be invoked; high costs and gross inconvenience to government administration can be asserted; legislation can be phrased in such a way that a labyrinth of exceptions exists; amendment legislation can be created to add to the labyrinth; when the time is propitious, 'amendment' legislation can be created which effectively revokes large segments, or simply the heart, of the control regime; internal, self-regulatory mechanisms can be created which are subject to little or no external control; and an agency unsympathetic to privacy concerns can be proposed to administer the legislation. Many of these methods have been evident in both the United States' and Australian experiences.

Where a separate 'watchdog' agency is forced on them, agencies have additional measures at their disposal. Legal uncertainties about the scope of the watchdog's powers can be exploited; the watchdog's resources can be soaked up by onerous bureaucratic responsibilities such as registration and public notification; privacy-protective functions and the associated costs can be passed to the watchdog instead of being performed and expended by agencies; more of the watchdog's capacity can be absorbed by providing parallel communications from many agencies rather than coordinated submissions; external interests can be excluded from deliberations by marking all submissions as confidential and stressing the security aspects of the information provided; key appointments can be made from within government; and the 'watchdog' agency can be progressively 'captured' as a member of the fraternity of government agencies. In Australia, most of these measures have been apparent.

It is clear that research is necessary into the political economy of computer matching. The above history of the process of regulation provides a backdrop to a future evaluation of control regimes. The remaining sections of this paper analyse the technical features of computer matching.

3. TECHNICAL DESCRIPTION

This section commences with definitional matters, presents a model of the process, and identifies key aspects of the various steps involved. In addition to the conventional approach, referred to in this paper as 'identification-based' matching, an overview is provided of an emerging, more sophisticated approach based on content.

3.1 DEFINITION

No definition of the term 'computer matching' is to be found in computing dictionaries. Nor does the term appear in such landmark documents as Westin [1967, 1971], HEW [1973], Westin & Baker [1974], the U.S. Privacy Act 1974, FACFI [1976], PPSC [1977], NSWPC [1977], Lindop [1978] and OECD [1980]. Indeed in most of these references it is difficult to even trace the emergence of the concept.

The technique has only become practicable as a result of the developments in information technology since the mid-1970s, and the term appears to have come into currency following publication of descriptions of Project Match, conducted by the Department of Health, Education & Welfare (HEW, now DHHS) in 1977. The series of important documents through which the subsequent line of development can be traced includes OMB [1979a, b], OMB [1982a, b], OMB/PCIE [1983], HHS [1983a, b, c], HHS [1984], LRC [1983], Laudon [1986b], OTA [1986b, pp.37-63], SMOS [1987], the U.S. Computer Matching and Privacy Protection Act 1988, PCC [1989], O'Connor [1990] and PCA [1990]. See Exhibit 3 for definitions of computer matching used in authoritative documents.

Exhibit 3: Definitions of Computer Matching

A procedure in which a computer is used to compare two or more automated systems of records, or a system of records with a set of non-Federal records, to find individuals who are common to more than one system or set [OMB 1982]

The comparison of different lists or files to determine if identical, similar or dissimilar items appear in them [Kusserow 1983, p.1]

The comparison of the whole or part of one system of records with the whole or part of another ... They are generally computer-based operations [ALRC 1983, p.173]

The identification of similarities or dissimilarities in data found in two or more computer files ... containing information on names or qualifications of interest to the government [OTA 1986b, pp.10-16]

information on persons or organizations of interest to the government [GAO 1986b, pp.10,16]

The comparison of two or more sets or systems of computerized records to search for individuals who may be included in more than one file [OTA 1986, p.37]

A computerized comparison of two or more automated systems of records held by two or more Federal or other Government agencies [U.S. Computer Matching and Privacy Protection Act 1988]

An activity that involves comparing personal data obtained from a variety of sources, including personal information banks, for the purpose of making decisions about the individuals to whom the data pertains [Canadian Treasury Board/Privacy Commissioner - TBC 1989, p.14]

Any automated comparison of personal information obtained from two or more systems or sets of records for the purpose of taking administrative action or making decisions affecting the individuals concerned [Draft Australian Data Matching Guidelines - PCA 1990 at 18.1]

Activities that compare (electronically or manually) information/data contained in two or more data files to find common elements and: (1) generate a workload that requires additional processing, development, or followup; (2) address the quality/integrity of SSA administered programs (i.e., detecting ineligible, incorrect payments, fraud); and (3) have the potential capacity of reducing human, machine or dollar resources used to operate/administer these programs [SSA 1991b]

The large scale comparison of personal information on the basis of a generalised assumption or hypothesis with a view to identifying records of interest. [Revised Draft Australian Data-Matching Guidelines - PCA 1991b, p.4]

The large scale comparison of records or files of personal information, collected or held for different purposes, with a view to identifying matters of interest. [Final Proposed Australian Data-Matching Guidelines - PCA 1992, p.6]

There is considerable diversity among these sources, and many deficiencies and imprecisions in the definitions. In addition, computer matching needs to be distinguished from other forms of data-handling described in section 2.1 above, such as internal use and external disclosure, front-end verification and notification, and data concentration and data-linkage. The following paragraphs identify the key issues, and propose a set of definitions on which the remainder of the paper is based.

One concern is that the definition of a technique should be based on what is done and how it is achieved, and not on interpretations of the motives of the organisation which does it. The phrasing of any constraint such as **'for the purpose of ...'** must therefore be fairly general. Kusserow's definition is admirably broad, but it may be worthwhile to restrict the term to personal data rather than, say, information about inventory-items. The words "in order to identify cases of interest" have the advantage of breadth and clarity.

An important issue is what it is that is compared. The term 'system of records', used in both the 1974 and 1988 U.S. Acts, is inappropriate, because the term **'(information) system'** is conventionally used to refer to a collection of interacting elements, including not only records, but also computer-based processes, manual procedures and man-machine interfaces. This paper therefore avoids the term 'system of records' and suggests that legislative draftsmen do likewise.

The term 'set of records' is reasonably descriptive, provided that its use is kept vague. It could be construed to refer to all tuples in a single relation, all tuples in all relations in a single database, all tuples in all relations maintained within a single application, all tuples in all relations maintained within a single computer system or all tuples in all relations across a collection of networked computer systems. Over that range of possibilities must be laid the organisational question, i.e. a 'set' could be all tuples in some or all relations maintained by or on behalf of a single organisation, or could be unconstrained by organisational boundaries.

The standpoint adopted in this paper is that terms should be used which have adequately precise meanings, and which are in conventional usage by information technology professionals, but whose meanings are readily grasped and remembered by non-specialists. Moreover, the discussion must proceed at the logical rather than the physical level, i.e. without concern as to how the data is expressed on a storage medium. Accordingly, this paper uses the term **'record'** to refer to a collection of data items each of which represents an attribute of the same real-world entity, the terms **'file'** and **'data-set'** to refer to 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, and the term **'database'** to refer to a collection of files or data-sets. In general, formulations using the term 'record' are preferred.

The questions arise as to whether 'within-file', 'within-agency' and 'within-system' activities should be included in the term 'computer matching'. By **'within-file matching'** is meant comparison of records against all others in the same file or database, in order to locate, for example, instances of the same person claiming a benefit multiple times, a person splitting income with himself in order to take advantage of lower levels of taxation at lower levels of income, or pensioners of the opposite sex living

at the same address, who may be being paid pensions individually when they should be paid at the (lower) married rate. In the United Kingdom, the technique was applied to detect individuals seeking to buy twice as many shares in public enterprise privatisation sell-offs as they were entitled.

Multiple files may be maintained within a single system, to support the performance of a single function or to achieve the same set of purposes. Application of matching techniques to such files is referred to in this paper as '**within-system matching**'. It generally involves data which is fairly comparable because the purposes, definitions and procedures are generally the same or closely related. Many aspects of the information privacy concerns are therefore likely to be lower than would be the case in '**inter-system matching**'. Note that 'within-system matching' may cross organisational boundaries, where the system is inter- or multi-organisational in nature.

The concept of '**within-agency matching**' refers to matching of data sourced from within a single organisation. In many cases such activities will also be 'within-system', but many agencies perform multiple functions, and matches across, for example, different welfare applications, are 'within-agency' but not 'within-system'. '**Inter-agency matching**' is generally more likely to involve different data definitions, different software and different purposes. One leading author saw no need to apologise for the inclusion of within-agency matching within the definition: "Given the huge size of the seventeen [U.S.] federal agencies that are most involved with the Privacy Act, it is presumptuous to suggest that internal matching programs do not pose challenges to privacy" [Flaherty 1989, p.349].

It is important that all computer matching activities be recognised as such, and hence this paper adopts the approach that **computer matching includes any matches between files or within the same file, irrespective of whether data does or does not cross boundaries between organisations or systems, and whether the purposes of the data are or are not the same**. It is stressed that this does not imply that control measures for all classes of matching should necessarily be the same.

A confusion arises in some of the literature as to the difference between computer matching and front-end verification (FEV). The essential differences are that:

- FEV involves the comparison of data on an individual basis, rather than *on en masse* ;
- FEV is applicable prior to the administrative action, rather than afterwards; and
- FEV's purpose is to prevent errors and fraud, rather than to detect errors and fraud which have already occurred [GAO 1986b p.16, Clarke 1988b, IPCO 1991, Clarke 1992a].

For economic and practical reasons, FEV may be undertaken in batched runs rather than online, and may even be performed after a provisional decision has been made. In such cases, FEV has considerable similarity to computer matching.

Reflecting the above considerations, this paper adopts the following definition:

***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.*

It is further necessary to distinguish between:

- a '**computer matching run**', which is an event in which one file is compared against one or more files. In the 1990 Australian legislation and the proposed guidelines, the term 'cycle' is used in a similar manner; and
- a '**computer matching program**', which 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 organisation or set of organisations in addressing a single set of objectives.

The following section provides a model of conventional computer matching, and a discussion of the key aspects of each of the steps involved.

3.2 IDENTIFICATION-BASED COMPUTER MATCHING

This section provides an overview of computer matching based on an explicit identifier or set of identifiers for each data subject. It then discusses technical issues arising at various stages of the process.

The actual organisational structure and processes vary significantly between agencies and programs. This section presents a model designed to accommodate all of the important variations, and in particular GAO [1986a p.9], HUD [1988], GAO [1988 p. 57], PCA [1990 pp.4-5] and DSS [1991 p.15].

The **participant organisations** in computer matching can be classified into three groups:

- a **matching organisation** , which undertakes the matching step and perhaps also other steps;
- one or more **source organisations** , which provide input data and perhaps also perform other steps; and
- one or more **client organisations** , which receive data, in order to make decisions and take actions and/or to maintain databases.

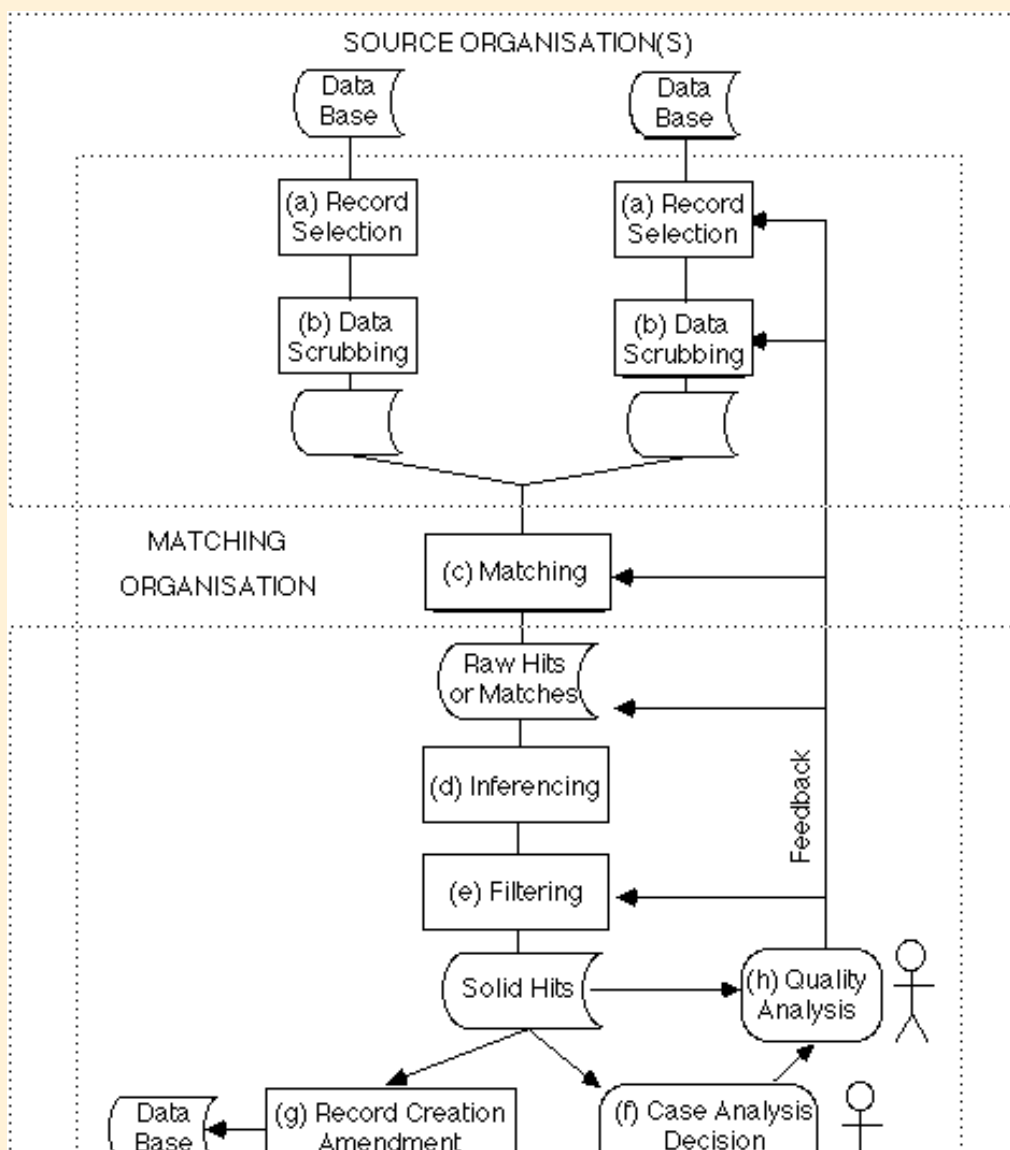
databases.

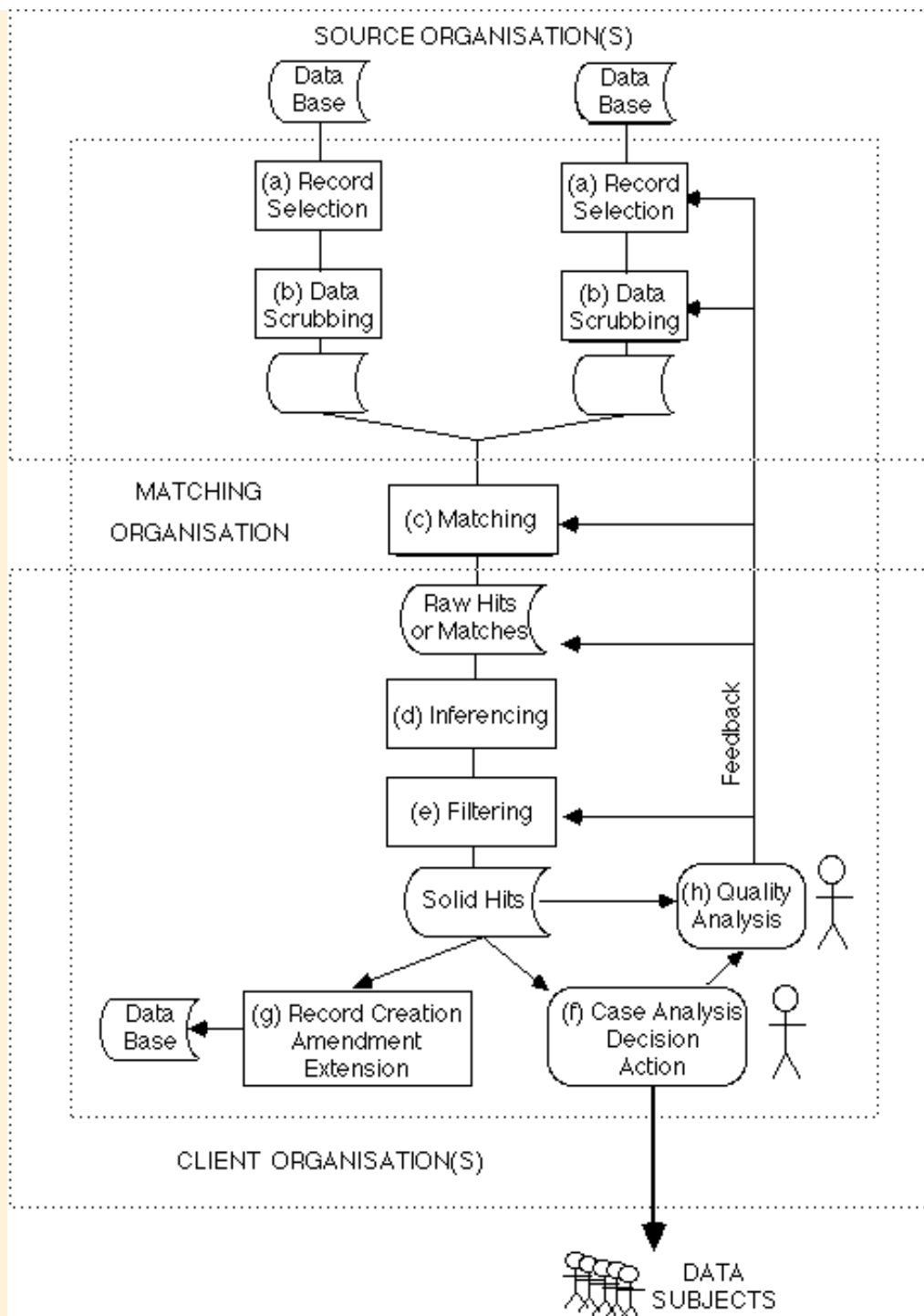
In any particular computer matching programme, participants may include government agencies, corporations, or (less usually) unincorporated associations or individuals; and a single government agency may be active in more than one of these roles. The scope of this paper extends to computer matching in which any of the participant organisations is a government agency.

Computer matching is predicated on the possession by source organisations of at least one, usually two, but perhaps more databases containing personal data records. The **steps in computer matching**, depicted in Exhibit 4, are:

1. **selection** from each database onto a file all or a sub-set of the records;
2. optionally, '**data scrubbing**' operations, to change the organisation, format and/or content of one of more of the files into a form suitable for the matching step;
3. '**matching**', whereby a 'matching algorithm' is applied to the file(s) 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, although the algorithm could involve a search for records on one file for which a match on the other **cannot** be found;
4. '**inferencing**', which involves an 'inferencing procedure' being applied to the outcome of the matching process (i.e. to the contents of the matched pairs of records or to the existence or non-existence of matches), to draw conclusions about the person to whom the data purports to relate, or to his or her behaviour, actions or proclivities;
5. '**filtering**' of the 'raw hits' to produce 'solid hits', and thus ensure efficient use of investigative resources and avoid insufficiently justified privacy-invasiveness;
6. **analysis** of the resulting information, and **decisions and action** arising;
7. optionally, **creation of new records and/or amendments or extensions to existing records** ; and
8. optionally, **quality analysis** activities may generate feedback from any and all stages back to earlier steps.

Exhibit 4: Conceptual Model of the Computer Matching Process





The steps preceding and following the matching step may be undertaken by a source or client organisation, or by the matching organisation. The following sub-sections discuss key aspects of this model.

(a) Record Selection

Depending on the source organisation's data schema, extraction of the records may be trivially simple, or require significant processing. It may be appropriate to select only a sub-set of the records available; for example, the Australian Department of Social Security supplies the primary identifier of its benefits recipients (the Tax File Number) to the taxation agency, and requests income details for that selection of people [DSS 1991, p.14]; and the U.S. Department of Housing and Urban Development provides SSNs of public housing tenants to several agencies and receives data in respect of only that sub-set of the source agency's files [Clarke 1991].

(b) Data Scrubbing

In many cases, the data may be in a form unsuitable for immediate processing. Pre-processing may therefore be advisable, in

In many cases, the data may be in a form unsuitable for immediate processing. Pre-processing may therefore be advisable, in order to avoid complexity in the matching, inferencing and filtering algorithms and perhaps to cope with weaknesses in the software development environment. This is commonly referred to as 'data scrubbing' [McDonald, 1990].

Data scrubbing may involve activities at several levels:

- at file- or dataset-level, headers and trailers may need to be removed, multi-record structures converted to flat files or mapped onto third normal form data structures, or sub-records for multiple members of a family exploded into a record for each member;
- at record-level, the sequence of data-items may need modification to bring intended key-items together; and
- at data-item level, alphanumeric fields may need to be internally reorganised (particularly name and address fields which contain multiple discrete components); the formats of identification codes may need to be edited; dates may need to be reformatted or imputed (a major area of difficulty in many programs); the content of null or empty fields may need to be inferred (e.g. gender or year of birth) from other data-items, or from file-header data or from the general context such as the source-organisation; and missing fields may need to be created and filled (e.g. the number of dependants based on a count of the number of sub-records).

Some of these massaging activities will result in data of demonstrably higher quality. In some cases, however, the nature of the processing may be such that some decrease in data quality may be necessary in order to achieve a desired form of increase in data quality; for example, a data-item containing full given and surnames and which has not been subjected to careful editing at source and which therefore in some cases contains the sequence givenname-surname and in others surname-givenname, will inevitably meet unresolvable problems with words which can be both given and surnames (e.g. Clark, Craig, James, Russell and Thomas).

(c) Matching

The matching step involves the application of a matching algorithm to the available data. 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**. It is important that the qualifications in that definition not be overlooked; for example, one commonly used definition of a 'hit' is "the identification through a matching program, of a specific individual" [OMB 1982a, my emphasis].

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.

The most straightforward matching algorithm, a simple test of equality between the contents of two data-items, is possible where the sets of data use a **common, unique person identification code**. Candidates for this role include in the United States the Social Security Number (SSN), in Canada the Social Insurance Number (SIN), in Australia the Tax File Number (TFN), and in many countries the passport number, person, driver or similar near-universal registration scheme. In some countries, there are legal provisions which prevent some organisations from using some of these codes, including Australia, Canada and (nominally) the United States.

The U.S. Social Security Agency issues the SSN, and all of its computer matching is based on that code. All of its programs commence by verifying the source data against the master file, using the SSN as the key, and comparing lastname, firstname, gender and date of birth as an integrity check. The Department of Housing and Urban Development also uses the SSN, and relies on investigators to check that the secondary identification details are identical or at least satisfactorily similar. The Australian Department of Social Security's programs have recently become heavily dependent on the (now misleadingly named) Tax File Number.

Even where a unique multi-purpose identifier is available, however, some complexities may arise. For example a system from which data is being extracted may define the data-item in a non-standard fashion, may not allow for very old or for very new codes, may drop the check-digit, or may extend the code, e.g. to indicate some characteristic of the record or the data subject (e.g. gender, citizenship or benefit entitlement).

Almost all identification codes actually in use are assigned more or less arbitrarily, and imposed on the person concerned. This creates significant difficulties for organisations in ensuring that a person, and transactions between the organisation and the person, are reliably associated with the correct record. For example, a token may be issued (such as a card with a magnetic stripe or embedded micro-chip containing the identification code, and perhaps a photograph), and services denied unless the token is produced. An alternative to such an imposed identifier is to encode into the record a representation of some physiological characteristic (such as fingerprint, voice-print, retinal print or DNA print), and use a sufficient match between a new measurement and the recorded id as the basis for recognition of the individual and association of new data with old [Clarke 1987b].

A less straightforward matching algorithm arises where the data subject's **name** is used as the primary identifier. Names are

A less straightforward matching algorithm arises where the data subject's **name** is used as the primary identifier. Names are much lengthier than artificially imposed codes, and are subject to spelling variants (e.g. 'John' and 'Jon'; 'Brown', 'Browne' and 'Braun'), presentation variants (e.g. 'John Pinkerton Brown', 'John P. Brown' and 'J. Pinkerton Brown'), nicknames ('Kate' and 'Kathy' for 'Katherine'), addendum variants (e.g. with and without 'III' or 'the 3rd', 'J.P.', 'MP' or 'PhD'), titles (e.g. 'Mr' or 'Rabbi'), multiple names in different orders and with multiple titles (e.g. 'Dr & Mrs Oliver and Margie Twist and Mr Martin Dobbs'), ambiguous structure (e.g. inclusion or exclusion of a regional or village name, or the name of a saint or prophet), localised names (e.g. 'Soh Chien Goh', or 'Lucy Soh'), change (e.g. anglicisation of the spelling of a name originating in another culture, alternative transliteration of a name originating in a language which uses a different alphabet or a Roman alphabet with diacritics such as 'ü', 'é' and 'ç', adoption by a woman of her husband's name on marriage, re-adoption by a woman of her maiden name on divorce, or change of name with a minimum of formality). To commit fraud through the use of aliases or aka's ('also-known-as' names) may be illegal, but in most countries their use is in itself not only entirely legal, but is also commonly used, e.g. by women in the professions, by people in occupations which involve security risk, and by artists of all kinds.

Beyond the complexities of names, mis-spellings are very common. Common errors include letter transcription (swapping of one or more adjacent letters, as in 'Smih' for 'Smith'), letter omission ('Barsley' for 'Bardsley', 'Schmid' for 'Schmidt'), letter substitution ('Cleverdale' for 'Cloverdale'), letter insertion ('Smith'), and phonetic substitution ('Filan' for 'Phelan').

A variety of sophisticated tools have been developed to cope with many of these variations in the presentation of names on files. For example, 'phonex' comparison techniques might enable the consideration of Clack, Clare, Clarke, Clark, Clague, Clay, Clear, Cleary, Clegg, Clerc, Clery, Cloake, Cloke, Klaic, Klug, Kluken and Klerk as being equivalent. In recent years, expert systems technology has been used to directly apply the heuristics which humans have developed over time [McDonald, 1990]. These techniques can achieve fewer 'false misses'. They do not, however, necessarily improve integrity, because decreasing the proportion of false misses may result in an increase in the number of false hits. Due to the complexity of the problem, there generally remains an appreciable risk that a pair of matched records do not relate to the same person.

An alternative to id-code or name-based matching is the use of similarities between the contents of **several fields** in each of the two files of personal data records. Data items which are commonly used in conjunction with **name** include **date of birth** and one or more elements of **address**, such as postal code (in the United States the Zipcode). With such data-items, a more refined matching technique than simple equality is generally needed; for example with postal codes some indication of proximity is desirable, because although many people change their address (e.g. in Australia in any given year, about 20% move at least once), the majority re-locate close to their old address.

A variant of multi-field matching has been used for many years in fingerprint processing. Master copies of fingerprints are encoded in a manner which reflects the presence, location and nature of their key features (such as 'whorls' and 'ridges'). Incoming prints (e.g. from the scene of a crime, or an unidentified body) are encoded in the same way, and the codes compared against the codes of the master copies. Because the act of encoding is imprecise, a set of putative hits results. This **pattern-based matching of encoded identifiers** has been directly automated in the computer-based fingerprint matching systems which have been installed since the mid-1980s. The approach is capable of application to computer matching more generally.

The proportion of records which result in raw hits varies considerably. In one HUD program in St Louis, 8308 records resulted in 778 hits, or 9.3% (although those were subsequently filtered down to a small number of cases for prosecution). On the other hand, even in the Australian DSS's initial runs, programs for other agencies generated only 3% raw hits, and those for the DSS itself only 1.1% [DSS 1991, p.49]. Later, following refinements in data scrubbing and matching algorithms, about 0.16% were surviving the automated filtering step, 0.04% were passing the manual filtering step, and less than 0.01% were resulting in action [DSS 1992, pp.89,107].

(d) Inferencing

The 'inferencing procedure' is the means whereby the contents of the matched records or the existence or non-existence of a match are analysed, and conclusions are drawn about the person to whom the data purports to relate .

This may be undertaken by a human, but because of the large volumes of data and hits involved, there is an economic incentive to apply information technology wherever practicable.

There are several broad ways in which an inference can be drawn. These are:

- the existence of a match where none would be expected to exist;
- inequality between apparently identical data-items in matched records;
- logical inconsistency among apparently related data items in matched records; and
- the failure to find a match where one would be expected.

Most commonly, the existence of a match may be regarded as establishing *aprima facie* case that, for example, an error has occurred or an offence has been committed. The complexity of life is such that the unrefined use of this form of inference is naive. Although it may seem surprising to find the same person on two seemingly mutually exclusive files (e.g. unemployment

benefit recipients and employees), there are several reasons why this may be legitimate; for example, the period in which the benefit was paid may not have overlapped with that in which income was earned; and some welfare schemes involve payments to beneficiaries even though the person is employed (e.g. some forms of disability pensions), and in some cases the amount of the benefit may even be independent of the earned income, and even of unearned income (e.g. some forms of veterans' pensions). In order to deal with these and other, more specific complexities which arise in respect of any given pair of files, additional tests generally need to be built into the inferencing mechanism.

The inferencing procedure can be considerably assisted by ensuring that there is consistency of data-item format and definition among the sets of records being matched.

(e) Filtering

Where it is impracticable to process all of the raw hits, there is a need to filter or screen them and thereby enable effective use of the available resources. Those hits can be removed from the list which are assessed to have the highest likelihood of being spurious, e.g. those which have fewest matched data elements or least convincingly matched data elements. Further, those hits may be selected out which offer low payback and/or low likelihood of payback and/or have been recently selected for review as a result of a previous matching run and/or for other reasons appropriate in the particular circumstances. This gives rise to what may be referred to as 'solid hits' [GAO 1986b, pp.21,101]. In addition to removing the least interesting cases, the filtering process can order the hits in descending sequence of assessed payback potential.

In U.S. Housing and Urban Development programs, cases are screened out on the basis of the difference between self-reported and apparent income. The threshold may be set at low as \$3,000 or as high as \$10,000, depending mainly on the resources available for case investigation. In the U.S. Social Security Agency, 'tolerances' are set, for each program, on the basis of a preliminary study. Factors considered in setting tolerance levels include whether other control mechanisms exist to detect erroneous self-reporting. An extreme instance was a match with Internal Revenue Service interest-income data, in which an initial 800,000 matched items was screened to 300,000 when the tolerances were applied (37.5%). In another instance, 9372 hits were reduced to 1265 (13.5%). Typically, it appears that 85-90% of 'raw hits' are filtered out [Clarke 1992a].

Considerable sophistication can be designed into the filtering process. For example, rather than a pre-determined threshold, the standard may be computed during the matching run, perhaps using artificial intelligence techniques; and the measure can be a code assigned corresponding to the highest-ranking matching criterion which was successful [McDonald 1990], or it can be the net of positive and negative points awarded for each inferencing criterion.

(f) Analysis, Decision and Action

It appears from the available literature that decision-making based on matched data, and action arising from those decisions, have generally not been automated. For example, "the computer cannot automatically decide who is receiving erroneous payments and who should be terminated from the program. It is a tool to identify conflicting information between what was reported by the individual and another source of information ... These cases must be further investigated ..." [HUD 1988, p.1]; and the data relating to hits "are sent to DSS staff in Area or Regional Offices either in hard copy form or on-line. The case is then reviewed in the normal way" [DSS 1990a, p.7].

The size of the effort involved in this phase is attested to by a study of HUD's procedures. The computation of excess assistance provided to the household is a non-trivial task, and HUD's regional staff use a pre-designed spreadsheet model to assist in this process. The cases are reviewed by HUD's specialist criminal investigators, who decide which, if any, are to be the subject of prosecutions. This depends on the amount of money involved, the nature of misrepresentations made, the number of occasions on which misrepresentations were made, etc. Cases proceed only after further investigation, including interviews with employers and consultation with officers of the U.S. Attorney.

The remaining cases are referred back to the appropriate PHA, with a requirement that the tenant be called for interview. Unless circumstances dictate otherwise, HUD requires that appropriate action be taken, such as increase in the rental payment made by the tenant, departure of the tenant from the premises and/or making of an agreement to repay the excess subsidy provided. Notice must be given to the tenant before adverse action is taken, and appeal rights exist [Clarke 1991].

In respect of the new programs approved by the Australian Parliament in December 1990, "action in each Agency ... will involve ... paper file examination, ..., letter to client ... indicating the information which the Agency has in its possession, that it came from the computer matching process, that it has been confirmed by the other Agency, that the person should contact the Agency ..., and the person's appeal rights ... Subject to the views of Agency Secretaries, action will not be taken to stop current payments without consideration of a client's response to the review letter mentioned above" [DSS, 1990b, p.11].

(g) Record Creation, Amendment and Extension

Data which results from computer matching programs may be used in several additional ways:

<http://www.anu.edu.au/people/Roger.Clarke/DV/MatchIntro.html>

Data which results from computer matching programs may be used in several additional ways:

- **new records may be created** from the results, particularly in the case of agencies which have no direct relationship with the data subject and therefore lack a flow of transaction data; and/or whose function is dataveillance;
- **existing records may be amended** on the basis of the matched data; and
- **existing records may be augmented**, by appending the data gleaned from other sources.

The resulting database can then be used as a reference point both for future computer matching runs and for key-based access.

Several agencies routinely conduct matches aimed at maintaining data quality, by which is meant using the contents of the database of another agency to make changes in their own records or initiate contact with the data subject. In some cases a scheme may involve multiple client organisations which were not themselves sources of data for the matching process; for example, the proposed Australia Card register included a feature whereby apparent changes of address would have been automatically notified to participating agencies.

(h) Quality Analysis

In addition to the mainstream activities described above, analysis of the quality of the results may be undertaken, as a control mechanism and to enable improvement in the various procedures and algorithms. This has been claimed to be common in large-scale and complex (and inevitably, therefore, error-prone) programs, and especially in the early stages of such programs, when an iterative learning or prototyping process is necessary in order to establish a moderately stable and routinised scheme [GAO 1986b, p.21, DSS 1991].

The following sub-section outlines a further level of sophistication which obviates the need for an explicit identifier.

3.3 CONTENT-BASED COMPUTER MATCHING

In this alternative approach, the match is based not on one or more explicit identifiers but on the content of any and all data-items deemed relevant by an investigator, such as address, physical characteristics such as height and weight, or association with a place and time, other individuals or groups, or corporations. It is possible in principle to support such an approach with an inverted-file system, provided that a secondary key-file is created for every data-item which may at some time be regarded as relevant. In practice, in the vast majority of systems, this is prohibitively expensive in terms of processing time, disk space and even access time. The only practicable environment in which the technique can be employed is with some form of associative memory.

A leading example of such a product is ICL's Content-Addressable File Storage (CAFS). This takes the form of a parallel searching mechanism incorporated into an otherwise conventional disk controller. It supports a wide variety of powerful logical operations, including precise comparisons (using '>', '<' and '=' operators) and 'fuzzy' comparisons (using variable-position masking, to enable a string to be found wherever in the data-item it may be stored; fixed-length and variable-length wild-cards; and bit-level masking for, for example, case-sensitive searching). Complex boolean conditions and quorum logic can be performed on one or several data-items, including exclusion conditions. Data format conversion (e.g. between textual and numeric representations of quantities, values and dates) is supported. Null values (i.e. empty data-items) are catered for. Comparison can be performed against a threshold value of the sum of weighted measures. Comparisons may be between a single pair of data-items or across two differently-structured groups of data-items, as is commonly the case with address data. An example of a relatively simple search enquiry is 'Which Post Office boxes receive more than three unemployment benefits, where the first initial of each name is the same, and the gender is the same, and the Dates of Birth are within 10 years of one another?'.

Because all of this is performed in silicon, the most complex possible evaluation takes less than 4 microseconds, and hence operates at about the same speed as disk rotation and channel transfer, and most such analysis can therefore be performed as fast as the machine can access the data. A further advance over conventional, key-based techniques is that CAFS-based content-addressable access applies as well to text as it does to structured data-items [Young 1988, Carmichael 1990].

CAFS, and its associated software development tool INDEPOL, were originally conceived to free key-based searches from the constraints imposed by files and databases in which each record had one primary key and several, but not too many, non-unique secondary keys. In addition to providing complete flexibility in key-based searching, however, it also offers a new dimension to computer matching.

A variety of approaches is feasible, but time will be required for applications to mature. One way in which it could be used would be to specify, for each data-item, the comparisons which should be performed between each pair of records, and the score associated with each successful match. The scores for each record would be summed. There would be no need to nominate an arbitrary score as the threshold for recognising a match. It would be sufficient to have the software present the most likely hits to the investigator in reverse order of scores. Frauds, disappearance techniques and multiple identities might remain effective in private life, but with imagination, skill and content-based computer matching, they would become highly

remain effective in private life, but with imagination, skill and content-based computer matching, they would become highly visible to investigative authorities with access to personal data from multiple sources.

The emergence of this technique is important. Until now, the preferred approach to dataveillance has been based on the assignment and use of a single identifier for each individual. The efficiency of dataveillance has been hampered by the difficulties involved with such identification schemes. Content-based matching obviates the need for an explicit human identifier. Provided that significant volumes of data are available, these new techniques makes it possible to associate data with a person with a considerable degree of confidence, based on a multi-dimensional correlation of whatever data is available. This offers the prospect of greater efficiency and effectiveness in surveillance, and brings with it new and enhanced threats to civil liberties which must be addressed.

4. CONCLUSIONS AND PROSPECTS

Computer matching is a powerful dataveillance technique, capable of offering in some circumstances benefits to the efficiency and effectiveness of government business greater than its financial costs. It is also a highly error-prone and privacy-invasive activity. Unless a suitable balance is found, and controls imposed which are perceived by the public to be appropriate and fair, its use is liable to 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 organisations which perform it, and perhaps against the technology which supports it.

The history of computer matching has been surveyed, and a technical description provided. 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.

Computer matching is a critical test of the resolve of information technologists to accept responsibility for the impact of their body of knowledge on people. It is not, in itself, an evil; but it is capable of being used evilly, or so insensitively that it will do significant harm to individuals, to groups, and to society as a whole. It shares that feature with many other techniques which are not yet empirically researchable, such as profiling, public networking, voice recognition, virtual reality in entertainment and education, the substitution of digital simulation for physical experimentation and intelligent robotics. It is vital that research be undertaken on such topics, and that that research be reported on in journals which reach the wide spread of academics and professionals, and not just discussed among a small clique of 'socially aware' fringe-dwellers (Clarke 1988). This paper has laid the foundation for serious discussion of computer matching.

5. References

ALRC 'Privacy' Aust. L. Reform Comm., Sydney, Report No. 22 (1983)

ANAO Report No. 24, Australian National Audit Office, 1990

_____'Pharmaceutical Benefits Scheme: Review of Estimated Savings from Proposed System for Eligibility Checking' Joint Review by The Auditor-General and the Department of Finance, Canberra, Dec 1991

Azrael M.L. 'Lost Privacy in the Computer Age' The Law Forum, University of Baltimore School of Law (Spring 1984) 18-26

Berman J. & Goldman J. 'A Federal Right of Information Privacy: The Need for Reform' Benton Foundation Project on Communications & Information Policy Options, 1776 K Street NW, Washington DC 20006, 1989

Bennett C. 'Regulating Privacy: Data Protection and Public Policy in Europe and the United States' Cornell University Press, New York, 1992

Bezkind M. 'Data Accuracy in Criminal Justice Information Systems: The Need for Legislation to Minimize Constitutional Harm' Comput. / Law J. 6,4 (Spring 1986) 677-732

Burnham D. 'The Rise of the Computer State' Random House, New York, 1983

Carmichael J.W.S. 'INDEPOL: A Software Package Exploiting CAFS' Paper available from the author, ICL Defence Technology Centre, Winnersh, U.K., 1990

Chaum D. 'Security Without Identification: Transaction Systems to Make Big Brother Obsolete' Commun. ACM 28,10 (October 1985) 1030-1044

Clarke R.A. 'Just Another Piece of Plastic for Your Wallet: The Australia Card' Prometheus 5,1 June 1987a. Republished in Computers & Society 18,1 (January 1988), with an Addendum in Computers & Society 18,3 (July 1988)

- ____ 'Human Identification in Records Systems' Working Paper available from the author (1987b)
- ____ 'Information Technology and Dataveillance' Comm. ACM 31,5 (May 1988a) Re-published in C. Dunlop and R. Kling (Eds.), 'Controversies in Computing', Academic Press, 1991
- ____ 'Economic, Legal and Social Implications of Information Technology' MIS Qtly 12,4 (December 1988b) 517-9
- ____ 'The Privacy Act 1988: Interpretations and Annotations from the Information Systems Perspective' Paper for the Australian Privacy Commissioner (March 1989)
- ____ 'The Expansionary History of the Enhanced Tax File Number' Aust. Privacy Found., Sydney (1990)
- ____ 'Computer Matching Case Report: Rental Assistance for Low-Income Families' Working Paper available from the author (December 1991)
- ____ 'Computer Matching in the Social Security Administration' Working Paper available from the author (January 1992a)
- ____ 'Computer Matching and the Office of Management and Budget' Working Paper available from the author (January 1992b)
- ____ 'The Resistible Rise of the National Personal Data System' Software L.J. 5,1 (February 1992c)
- ____ 'Computer Matching in Australia' Working Paper available from the author (May 1992d)
- ____ 'Computer Matching in Government: A Normative Regulatory Framework' (110 pp.), Working Paper, Dept of Commerce, Australian National University, available from the author (August 1992)
- ____ 'Computer Matching and Digital Identity' Proc. Conf. Computers, Freedom & Privacy, Ass. Comp. Machinery, San Francisco, March 1993
- Cohen 'Oversight of Computer Matching to Detect Fraud and Mismanagement in Government Programs' Hearings Before the Sub-Committee on Oversight of Government Management, Senate Committee on Governmental Affairs U.S. Govt Printing Office, Washington DC (Dec 15-16, 1982)
- DSS 'Data-Matching Arrangements Within The Department of Social Security' Department of Social Security, Canberra, Australia Paper prepared for the Commonwealth Privacy Commissioner, October 1990
- ____ '1990 Budget Decision - Computer Matching' Department of Social Security, Canberra, Australia, October 1990
- ____ 'Data-Matching Program (Assistance and Tax): Report on Progress' Department of Social Security and the Data Matching Agency, Canberra, Australia, October 1991
- ____ 'Data-Matching Program (Assistance and Tax): Report on Progress' Department of Social Security and the Data Matching Agency, Canberra, Australia, October 1992
- Early P. 'Big Brother Makes a Date' San Francisco Examiner, 12 Oct 1986
- FACFI 'The Criminal Use of False Identification' U.S. Federal Advisory Committee on False Identification, Washington DC, 1976
- Fischel M. & Siegel L. 'Computer-Aided Techniques Against Public Assistance Fraud: A Case Study of the Aid to Families with Dependent Children (AFDC) Program' Prepared by the U.S. Law Enforcement Administration by MITRE Corp. 1980
- Flaherty D.H. 'Protecting Privacy in Surveillance Societies' Uni. of North Carolina Press, 1989
- GAO 'Eligibility Verification and Privacy in Federal Benefit Programs: A Delicate Balance' General Accounting Office, GAO/HRD-85-22, 1985
- ____ 'Computer Matching: Assessing Its Costs and Benefits' General Accounting Office, GAO/PEMD-87-2, Nov 1986, 102 pp.
- ____ 'Veterans' Pensions: Verifying Income with Tax Data Can Identify Significant Payment Problems' General Accounting Office, GAO/HRD-88-24, Mar 1988, 100 pp.
- ____ 'Computers and Privacy: How the Government Obtains, Verifies, Uses and Protects Personal Data' General Accounting Office, GAO/IMTEC-90-70BR, Aug 1990, 68 pp.
- Graham P. 'A Case Study of Computers in Public Administration: The Australia Card' Austral. Comp. J. 22,2 (May 1990)
- Greenberg D.H. & Wolf D.A. 'Is Wage Matching Worth All the Trouble?' Public Welfare (Winter 1985) 13-20
- Greenberg D.H. & Wolf D.A. (with Pfister J.) 'Using Computers to Combat Welfare Fraud: The Operation and Effectiveness

Greenberg D.H. & Wolf D.A. (with Pfister J.) 'Using Computers to Combat Welfare Fraud: The Operation and Effectiveness of Wage Matching' Greenwood Press Inc., Oct 1986

Greenleaf G.W. 'The Privacy Act: Half a Loaf' 63 Aust.L.J. 116 (1989)

_____ 'The Privacy Act: Enforcement and Exemptions' 63 Aust.L.J. 285 (1989)

_____ 'Can The Data Matching Epidemic Be Controlled?' Austral. L. J. (March, 1991)

Greenleaf G.W. & Nolan J. 'The Deceptive History of the Australia Card' Aust. Qtly 58,4 407-25 (1986)

HIC 'Planning Report of the Health Insurance Commission' (26 February 1986)

HEW 'Records, Computers and the Rights of Citizens' U.S. Dept of Health, Education & Welfare, Secretary's Advisory Committee on Automated Personal Data Systems, MIT Press, Cambridge Mass., 1973

HHS 'Computer Matching in State Administered Benefit Programs: A Manager's Guide to Decision-Making' Dept. of Health and Human Services, 1983a

_____ 'Inventory of State Computer Matching Technology' Dept. of Health and Human Services, 1983b

_____ 'Nationwide Impact of Verifying Resources by Matching Recipients of Public Assistance with Bank Records' Dept. of Health and Human Services, 1983c

_____ 'Computer Matching in State Administered Benefit Programs' U.S. Dept. of Health and Human Services, June 1984

HUD 'OIG Communiqué', U.S. Department of Housing and Urban Development', Washington DC (September 1988)

IPCO 'Privacy and Computer Matching' Information and Privacy Commissioner / Ontario, 80 Bloor St West, Suite 1700, Toronto, Jan 1991

Kelly P. 'Australian Federal Privacy Laws and the Role of the Privacy Commissioner in Monitoring Data Matching' in Clarke R. & Cameron J. (Eds.) 'Managing Information Technology's Organisational Impact II' Elsevier / North Holland, 1992

Kirchner J. 'Privacy: A history of computer matching in federal government programs' Computerworld (December 14, 1981)

Kling R. 'Automated Welfare Client Tracking and Welfare Service Integration: The Political Economy of Computing' Comm ACM 21,6 (June 1978) 484-93

Kusserow R.P. 'Inventory of State Computer Matching Technology' Dept. of Health & Human Services, Office of Inspector-General (March 1983)

_____ 'The Government Needs Computer Matching to Root out Waste and Fraud' Comm ACM 27,6 (June 1984) 542-545

_____ 'Computer Matching in State-Administered Benefit Programs' Dept. of Health & Human Services, Office of Inspector-General (June 1984)

Laudon K.C. 'Computers and Bureaucratic Reform' Wiley, New York, 1974

_____ 'Dossier Society: Value Choices in the Design of National Information Systems' Columbia U.P., 1986b

Lindop 'Report of the Committee on Data Protection' Cmnd 7341, HMSO, London (December 1978)

McDonald K. 'The Matchmaker Expert Matching System' Paper available from MasterSoft International, 12-14 Malvern Ave., Chatswood NSW 2067, 1990

Marx G.T. 'The New Surveillance' Technology Review (May-Jun 1985)

Marx G.T. & Reichman N. 'Routinising the Discovery of Secrets' Am. Behav. Scientist 27,4 (Mar/Apr 1984) 423-452

NSWPC 'Guidelines for the Operation of Personal Data Systems' New South Wales Privacy Committee, Sydney, 1977

O'Connor K. Paper for the Communications & Media Law Association, available from the Privacy Commissioner, Human Rights & Equal Opportunities Commission, G.P.O. Box 5218, Sydney, 26 April 1990

OECD 'Guidelines for the Protection of Privacy and Transborder Flows of Personal Data' Organisation for Economic Cooperation and Development, Paris, 1980

OMB 'Guidelines to Agencies on Conducting Automated Matching Programs' Office of Management and Budget March 1979a

_____ 'Privacy Act of 1974: Supplemental Guidance for Matching Programs' 44 Fed. Reg. 23, 138 (1979b)

_____ 'Computer Matching Guidelines' Office of Management and Budget, May 1982a

_____'Privacy Act of 1974: Revised Supplemental Guidance for Conducting Matching Programs' 47 Fed. Reg. 21, 656 (1982b)

_____'Management of Federal Information Resources' Office of Management and Budget Circular A-130 (December 1985)

OMB/PCIE 'Model Control System for Conducting Computer Matching Projects Involving Individual Privacy Data' Office of Management and Budget & President's Commission for Integrity & Efficiency 1983

OTA 'Federal Government Information Technology: Electronic Record Systems and Individual Privacy' OTA-CIT-296, U.S. Govt Printing Office, Washington DC, Jun 1986

PCA 'Data Matching in Commonwealth Administration: Discussion Paper and Draft Guidelines' Privacy Commissioner, Human Rights & Equal Opportunities Commission, G.P.O. Box 5218, Sydney, Australia (October 1990) (56 pp.)

_____'Interim Report on Operation of Data-Matching Program under the Data-Matching Program (Assistance and Tax) Act 1990' Privacy Commissioner, Human Rights & Equal Opportunities Commission, G.P.O. Box 5218, Sydney, Australia (Sep 1991a)

_____'Data Matching in Commonwealth Administration: Revised Draft Guidelines' Privacy Commissioner, Human Rights & Equal Opportunities Commission, September 1991b

_____'Data-Matching in Commonwealth Administration: Final Proposed Guidelines' Privacy Commissioner, Human Rights & Equal Opportunities Commission, December 1991, published June 1992

PCC 'Data Matching Review: A Resource Document for Notification of the Privacy Commissioner of Proposed Data Matches' Privacy Commissioner, 112 Kent, Ottawa, Canada (July 1989)

PPSC 'Personal Privacy in an Information Society' Privacy Protection Study Commission, U.S. Govt. Printing Office, July 1977

Roszak T. 'The Cult of Information' Pantheon 1986

Rule J.B. 'Private Lives and Public Surveillance: Social Control in the Computer Age' Schocken Books, 1974

_____'1984 - The Ingredients of Totalitarianism' in '1984 Revisited - Totalitarianism in Our Century' Harper & Row, 1983 pp.166-179

_____'Documentary Identification and Mass Surveillance in the United States' 31 Social Problems 222 (1983)

Rule J.B., McAdam D., Stearns L. & Uglow D. 'The Politics of Privacy' New American Library, 1980

Smith R.E.(Ed.) 'Privacy Journal' monthly since November 1974

_____'Privacy Journal Compilation of State and Federal Privacy Laws' updated annually

Shattuck J. 'Computer Matching is a Serious Threat to Individual Rights' Comm ACM 27,6 (June 1984) 538-541

SMOS 'Review of Systems for Dealing with Fraud on the Commonwealth' Aust. Govt. Publ. Service (March 1987)

SSA 'SSA Matching Operations Inventory' Office of Program Integrity and Reviews, Social Security Administration, May 1991b

TBC 'Data Matching and Control of the Social Insurance Number' Canadian Treasury Board Policy Circular 1989-12, Ottawa, Canada (July 1989)

Weiss L.B. 'Government Steps Up Use of Computer Matching To Find Fraud in Programs' Congressional Qtly Wkly Report February 26, 1983

Westin A.F. 'Privacy and Freedom' Atheneum, 1967

_____(ed.) 'Information Technology in a Democracy' Harvard U.P., 1971

Westin A.F. & Baker M. 'Databanks in a Free Society' Quadrangle, 1974

Young M. 'ICL Investigative Systems: An Overview' Paper available from the author, ICL Australia Ltd, Canberra, 1988

ACKNOWLEDGEMENTS

My thanks to Bill Orme, then Executive Member of the N.S.W. Privacy Committee, who asked me to write a paper on the topic in 1976, when the term was only just coming into use and basic privacy protections seemed a higher priority. Sorry it's late, Bill. The assistance is gratefully acknowledged of fellow dataveillance researchers in Australia, the United States and

late, Bill. The assistance is gratefully acknowledged of fellow dataveillance researchers in Australia, the United States and Canada. So also is that of the Australian Privacy Commissioner, Kevin O'Connor, and his deputy Nigel Waters.

Staff of several U.S. government agencies, and, since 1989, of one Australian government agency, were very forthcoming. Their openness has enabled, for the first time, some amount of light to be thrown upon the practice and economics of computer matching by government agencies. Finally, several reviewers and the Area Editor contributed significantly to the tightening of the paper's focus. Responsibility for the interpretative comment in the paper is mine alone.

Navigation

Go to [Roger's Home Page](#)

Go to [the contents-page for this segment](#).

[Send an email to Roger](#)

Created: 18 October 1998

Last Amended: 18 October 1998



These community service pages are a joint offering of the Australian National University (which provides the infrastructure), and Roger Clarke (who provides the content).



[The Australian National University](#)

Visiting Fellow, Faculty of
Engineering and Information Technology,
Information Sciences Building Room 211

[Xamax Consultancy Pty Ltd](#) ACN: 002 360 456

78 Sidaway St
Chapman ACT 2611 AUSTRALIA
Tel: +61 2 6288 1472, 6288 6916