THE FOUNDATION FOR SCIENCE AND TECHNOLOGY

DINNER/DISCUSSION SUMMARY

Merging diverse datasets can produce new insights.
What new applications are possible and are there new privacy and regulatory issues?

Held at The Royal Society on 9th May, 2006

We are grateful to the following for support for this meeting:
Department for Transport (Research and Technology Division)
City and Guilds of London Institute

Chair: **The Earl of Selborne KBE FRS**
Chairman, The Foundation for Science and Technology

Speakers: **Professor Frank Kelly FRS**
Chief Scientific Adviser, Department for Transport
Dr Mike Lynch OBE
CEO, Autonomy
Professor Mark Walport FMedSci
Director, The Wellcome Trust

PROFESSOR KELLY outlined DTP databases – MIDAS (Motorway Incident and Accident Survey), Transport Direct, Accession and Speed limit Database. He explained how the archive of data could be exploited and fused to predict journey times and other information to improve services – e.g. to predict bus arrival times at a bus stop, or vary bus routes to improve access to hospitals for frequent users. Further sophistication led to the development of Demand Response Transport, such as the Wiltshire "Wiggly Bus" and Lincoln's "Call Connect". New sources of data were continuing to appear – mobile telephones, smart cards etc. But there were IPR problems in using certain data – e.g. using maps derived from the Ordnance Survey base. The potential uses of merged data were great, but a balance had to be struck between convenience, privacy and personalization of data. The barriers to striking a successful balance were fears about confidentiality, security and the accuracy of the data. The challenges were not primarily scientific or technical, (although there were some) but social and regulatory. The public had to be convinced of the benefits – both direct and indirect – of data uses, even although it was not possible to forecast how data would be used in future developments; it was also important to evaluate carefully the economic costs and benefits of data use. Government had a difficult role in reconciling the different priorities and interests of Departments, developing public/private partnerships, reducing unnecessary fears about confidentiality, continuing technical work and, most important, setting a stable regulatory and legal framework.

DR LYNCH explained the importance of mining the vast amount of data which was intelligible to human beings, but, without modulation, to computers - "unstructured" data, of many kinds, oral or written, in different languages and formats. He instanced the search for the "Yorkshire Ripper" as an example of how, if "memory information" had been utilised, success would have been much quicker. He illustrated how an incident could be described by a variety of people in different terms; the key to using the information was to retrieve it in natural language, conceptualise it, sort it into hierarchies, and search out related groups (clustering), so that changes in circumstances can be identified, without having to be previously defined. Usage then depended on personalising the information, alerting users to features to be examined, and linking different bases and systems. Practical difficulties lay in persuading owners of different sets of data to work together, which often depended on political will (e.g getting the 21 different U.S. Security agencies to collaborate) and ensuring that lessons were learnt from failures as well as successes. The benefits, in health services and security, lay, not only in forensic inquiry, but also in foreseeing trends and opportunities.

PROFESSOR WALPORT said that health and medicine was an area where data collection and use aroused particular fears and controversy. But it was essential that health policy was underpinned by data which exposed real problems and trends. It was also vital if service delivery to individuals was to be improved. At present, because of the many points of contact between government and the citizen, and the fragmentation of management it was difficult to bring together sets of personal data – demographic, geopolitical, environmental, housing, and health history which would inform successful patient care. He instanced the work of the Small Area Statistics unit in analysing the relationship between the rates of still

births of those living close to landfill sites and others (a small excess but further study needed). But the risks in using information from dispersed databases must not be ignored if public opinion were to be won over. The risks were confidentiality and loss of privacy; unauthorised use of the material; exploitation of data for commercial gain; statistical discrimination; poor quality or inaccurate data; and cyber terrorism. The Council for Science and Technology had recommended that certain principles should underlie data access – the data should be anonymized, it should be facilitated if needed for research or genuine statistical analysis; and there should be safeguards and transparency in any usage. Much more R&D was needed in partnership with the private sector; and a clear regulatory framework be established which defined the limited areas – law and order, research and service delivery - where personalised data was acceptable, and other areas, such as traffic management, where it was not and must be aggregated. Gaining the public's trust in the use of data was crucial; this meant promoting understanding and erecting a framework which clearly identified where accountability and responsibility lay.

A leading theme in the following discussion was the confidentiality of data. Many speakers underlined the concern that data use would be tightly constrained unless people felt that their privacy was not compromised and any personal data was kept secure and used only for purpose of which they were aware and accepted. This raised the fundamental problem – how to reassure people about the future use of data when the benefits of merging databases opened up the possibility of new, unexpected and (in the eyes of those in charge) beneficial uses. A number of suggestions were made which could ameliorate the problem – the government should accept an obligation to make any data held on an individual open to him (as with credit agencies); it should be made a crime to release information (and retribution should not be limited to any loss suffered by the individual); the concept of reciprocity should be developed so that people can understand what they would actually gain from the use of data. Nevertheless, there was a deep-rooted suspicion that government would misuse – or at least use for unexpected purposes - any data that it held. It might well be more acceptable if data was held in a private company, which could be held more accountable for its misuse, and be subject to audit or withdrawal of contract if it did not meet accepted standards. There was a strong case for an independent body, which would regulate and oversee data collection and use. But even so, the danger remained that a hacker could extract information from a database; how could this be prevented? Possibly this concern was overblown, because it was assumed that paper records were secure – when they evidently were not – but it had to be acknowledged that personal information on a central database could make information more widely available, if compromised.

More needed to be done to identify which aspects of data collection were unacceptable to the public, and which could, possibly over time, could become acceptable. For example CCTV cameras might at one time have been thought to be an invasion of privacy, but, now, because of their evident success in improving safety, were accepted. There was evidence that public rhetoric about confidentiality was more words than reality and could be overcome by some modest financial inducement. Fingerprinting might be unacceptable in a society where large numbers had been fingerprinted because of suspicion of criminal offences, but not in others. But it was important to accept that there was already so much data already accessible that it was unrealistic to focus on controlling it. The aim must be to look at the end use of the data, which could be either beneficial or not, and establish the regulatory framework that promoted the first and punished the second. Only on such a clear basis could difficult questions such as the ability of individuals to withhold data (thus possibly biasing the data collection) be settled. Many legal questions, involving, for example, human rights, would undoubtedly arise but, it was, in the first instance, a political decision when public benefit should override personal preference just as it was a political duty for Ministers to campaign for the enormous benefits that data use could make to individual lives.

This emphasis on political will and decision raised the question that, in government, was, or should be, the driver for implementation of the use of IT and data. All Departments were involved, and, indeed, so should be the devolved administrations who could provide useful areas for experiments – both successful and unsuccessful. Pressure was building up for such a central point, and there was, indeed, a Cabinet Committee which co-ordinated IT, but there was some road to go. In an ideal world, there should be some common EU input on standards and safeguards, although, given the very different views about autonomy and confidentiality in different countries, this would be difficult.

Sir Geoffrey Chipperfield KCB

Useful Web Links:

Autonomy:
www.autonomy.com

City & Guilds of London Institute:
www.city-and-guilds.co.uk

Council for Science and Technology:
Council for Science & Technology Report: Better use of personal information: Opportunities and Risks:
www.cst.gov.uk/cst/reports/#10

Department for Transport:
www.dft.gov.uk

Google Maps, Google Earth and Google Scholar:
maps.google.com, earth.google.com and scholar.google.com

The Wellcome Trust:
www.wellcometrust.ac.uk