Key Issues in Confidentiality Research: Results of an NSF workshop

Julia Lane
The Urban Institute[1]

[1] This workshop was organized in conjunction with John Abowd, Cornell University and George Duncan, Carnegie Mellon University. Laura Zayatz, of the US Census Bureau, served as rapporteur for the technical session; George Duncan for the geospatial session

Overview

Data, and data access, lie at the heart of social science research. Billions of taxpayer dollars are spent in supporting the collection and dissemination of federal, state and local data, billions of dollars are spent in data analysis, and this, in turn, both informs scientific understanding of core social science issues and guides decision in how to allocate billions of dollars in social programs. Although an entire analytical infrastructure depends on the dissemination of high quality data, statistical agencies which have gone to great expense to collect such data, then deliberately destroy data quality –- often in ad hoc fashion -- in order to protect respondent confidentiality. Indeed, one statistical agency spends tens of millions of dollars, with concomitant respondent burden, to collect county level data on employment and earnings by industry, only to suppress over 60% of the resulting tabular output, and create tables with unknown statistical properties.

Although research in confidentiality protection has been supported by both statistical agencies and the National Science Foundation in the past, there are increasingly urgent reasons to further invest in such research. First, statistical agencies are not only increasing the degree to which they distort data, but also considering not releasing public use micro-data files in response to technological advances that increase the probability of respondent re-identification. Second, new ways of presenting data need new protection techniques. For example, the explosion of GIS software means that it is possible to present micro-data in new and useful ways, but the understanding of how to protect the underlying individual information is still in its infancy. Third, the legal and social context for data dissemination have changed – particularly after the events of Sept 11. Finally,

new institutions have developed – such as Institutional Review Boards – yet there is wide variability in their capacity to apply and disseminate confidentiality research.

In recognition of the importance of expanding research in this area, the National Science Foundation sponsored a workshop on data confidentiality May 12-13, 2003. Participants in the workshop, who came from a wide variety of disciplines, were charged with providing some broad insights into confidentiality issues across the social and behavioral sciences and identifying some important areas for future research. The workshop had four major themes -- rethinking the conceptual framework of confidentiality research, identifying promising new technical approaches for data release, developing and understanding the data dissemination context and identifying the key confidentiality issues associated with the release of geo-spatial data[2].

Rethinking the conceptual framework

"The National Zoo told a Washington Post writer that she cannot have some medical records on a dead giraffe because it would intrude upon that patient's privacy."
--- *News Media Update, May 6, 2002 Reporters Committee for Freedom of the Press*[2]

The fundamental tension faced by statistical institutes is disseminating data while at the same time protecting respondent confidentiality. While this precept sounds straightforward, it raises a number of complex philosophical, economic and legal issues. decision is the notion of protection respondent confidentiality. At the core of these issues is the notion of whether data are a private good, or whether they are societal. Some important areas for research in this area include:

[2] The agenda and presentations for the workshop are available at http://www.urban.org/nsfpresentations/index.html. The agenda is also available as an appendix to this document.

1. *Developing an analytical framework to guide data dissemination from the standpoint of applied ethics.*
2. *Estimating the social benefits and costs associated with data dissemination*
3. *Examining the current legal foundations that govern data dissemination and identifying potential conflicts.*

The philosophical issues are well summarized by Madsen[3]. He identifies a "privacy paradox" in confidentiality research – which occurs when data managers, in interpreting the right to privacy very narrowly, results in less social benefit, rather than in more. Two factors contribute to this paradox. One is the fear of a pan-opticon society, in which an all-seeing few monitor the behavior of many, which has been exacerabated since Sept 11, 2001. The second is a fundamental uncertainty about data ownership – whether data constitute private or public property. It is possible that the tension in the core paradox results from a framework which simply includes rights and responsibilities into the decision-making mix, rather than including social utility.

The second set of issues[4] is economic in nature. Given the clear public good aspects of data collection and dissemination, how can the costs and benefits of the social investment in data be tallied to identify the optimal level data confidentiality research? A partial list of the social benefits would include: improved decision making, avoidance of the moral hazard associated with monopoly government control of information, and improved data quality. A similar list of the social costs would include legal sanctions, the cost of breaches of confidentiality (which might substantially reduce data quality), and support costs.

---

[3] Peter Madsen "The Ethics of Confidentiality: The Tension Between Confidentiality and the Integrity of Data Analysis in Social Science Research", mimeo, Carnegie Mellon University, June 2003

[4] Julia Lane "Uses of Micro-data", miemo, The Urban Institute, June 2003.

The third set of issues is legal[5]. While a wide variety of laws govern data dissemination, it is unclear what the overarching legal doctrine should be. Are data one item, like land, to be governed by property law, multiple items, like wheat or sheep to be governed by commercial law, an idea or an invention, to be covered by patent law, or is copyright law more appropriate? Or should data be covered by license law, whereby the data owner permits others to use ideas? A large part of the lack of clarity in legal guidance stems from a lack of clarity in who owns the data – whether it is the person who is the subject of the information, the person/organization who collects data (the data custodian), the person who complies, analyzes or other wise adds value to information, the person who purchases interest in data, or society at large. Cecil and Eden are particularly concerned about the possibility of a collision between laws that treat data as property and those treating it as information; between individual respondents and those who collect the data; between individual rights and the rights of governments purporting to act for the greater good.

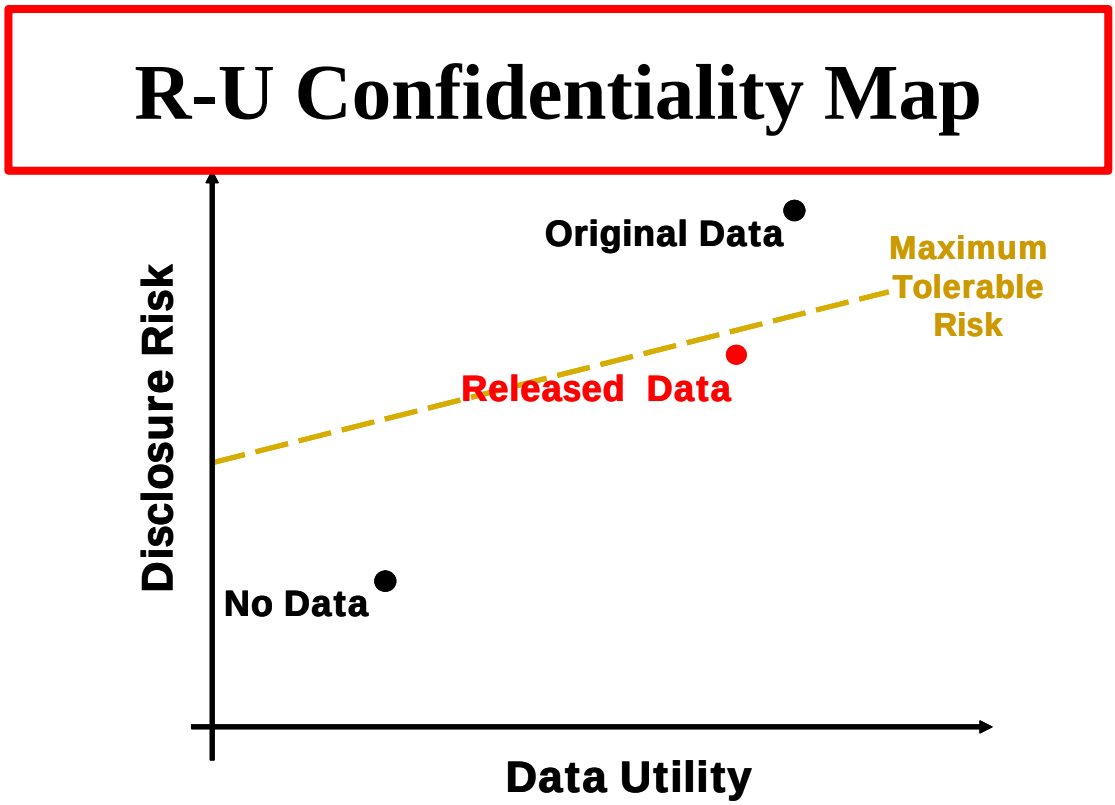
Identifying New Technical Approaches

"From a statistical user's perspective, a table modified for confidentiality reasons should yield statistical inferences that are the same as those that would come from the original, unmodified table" Steve Roehrig, May 12, 2003

Fienberg (2003) summarized the technical goals of disclosure limitation techniques as follows: (i) inferences should be the same as if we had original complete data; (ii) researchers should have the ability to reverse disclosure protection mechanism, not for individual identification, but for inferences about parameters in statistical models (e.g, likelihood function for disclosure procedure); (iii) there should be sufficient variables to

[5] Joe Cecil and Donna Eden "The Legal Foundations of Confidentiality", mimeo, May 12, 2003

allow for proper multivariate analyses and (iv) researchers should not only have the ability to assess goodness of fit of models but also be provided with most summary information, such as residuals (to identify outliers). The core guiding principle should be to generate released data that are as close to the frontier as possible (see Figure 1)



(Duncan, et al. 2001)

a) Tabular Data

The most common form of data release continues to be tabular data. A number of older approaches to protecting confidentiality in tables have serious statistical flaws. Expanding attribute classifications so that cell counts are "large enough" often results in the loss of key attribute information; cell suppression not only results in tables without clear statistical properties but "cheap heuristics" often result in an incorrect application

that actually does re-identify respondents; while data swapping has uncertain effects on goodness of fit measures. A very promising conventional approach is data rounding, which does allow users to know the range of possible values in each cell. The fundamental problem with all of these methods is that the mathematical properties of two-way tables do not hold for higher dimensional tables – because two-way tables have the structure of a mathematical network, while three-way and higher tables do not.

While newer techniques are being developed, substantial research issues remain. Synthetic tabular data (now known as controlled tabular adjustment or CTA, due to Dandekar and Cox), is a less damaging substitute for cell suppression, but the effects on statistical analysis are yet unknown, and the effect on the data quality of particularly sensitive cells is not clear. Cyclic perturbation techniques (Duncan and Roehrig) have the advantage of quantifying disclosure risk and the effect on data utility, but a number of technical issues, such as the definition of appropriate priors and the development of appropriate chi-square tests have yet to be resolved. Another approach is to publish only a restricted and carefully-defined set of lower-dimensional marginal tables derived from a high-dimensional base table (Dobra, Fienberg and Karr, cited by Roehrig, May 12, 2003). Finally, the perturbation of the underlying micro-data, rather than the actual cell values, which discussed in Giessing (2003), is an important alternative approach.

A major problem for users in deciding which approach to use to protect a particular set of tables is that it is often difficult to apply different techniques, as well as compare the impact on the number of cells suppressed. A package has been developed by European researchers, τ-ARGUS, which permits comparisons to be made directly (see Giessing, May 12 2003) – and allows researchers to determine exactly how much

resources need to be invested in improving the quality of output tables for a given level of information loss.

The discipline could most readily be advanced by focusing on the following

1. *Developing different ways of perturbing the underlying micro-data rather than perturbing the cells themselves.*
2. *Developing an understanding of the potential statistical use of the data BEFORE techniques are developed in order to maximize the data utility – leave the statistical properties intact*
3. *Furthering disclosure limitation approaches for high-dimensional tables*
4. *Developing effective, computationally tractable techniques and investing in the software to make effective comparisons*

b. Microdata

A great deal of attention has recently been paid to the potential of using synthetic data as an alternative approach to releasing public use data files. (see Muraldhiar and Sarathy (2003), and Abowd and Woodcock (2003) for reviews. One approach is to shuffle data; another is to develop samples composed of draws from the posterior predictive distribution of the confidential data, given some conventionally disclosure-controlled data The advantages of these approaches is that they are inference valid in that the synthetic data contain exactly the same statistical information as the micro data. In addition, the effect of disclosure protection on data quality can be measured. Finally, the multiple synthetic data implicates are not identical so the analyst can use the between implicate variation to measure the extent to which confidentiality protection made the inferences less precise.

In practical terms, an important additional value of such inference-valid synthetic data is that multiple public use files can be created from the same underlying data - targeted at different audiences. For example, some users of business data (such as transportation agencies) are particularly interested in geographic detail, while others are

interested in industry detail (such as industry analysts). Providing both levels of detail on the same data set immediately re-identifies important businesses. However, inference-valid synthetic data could be used to produce two separate data sets that can not be re-linked for such re-identification.

The discipline could most readily be advanced by focusing on the following

1. *Developing measures of data quality and risk for synthetic data*
2. *Developing more and better masking techniques for categorical data.*
3. *Developing masking techniques for both continuous variables and categorical variables that can be applied locally (to a subset of records with a high disclosure risk) rather than globally.*

c. Measuring Disclosure Risk

The quantification of disclosure risk goes to the heart of confidentiality research. Agencies need global measures of risk to be developed – to tell agencies when global masking techniques (whether in the form of reduction in detail or perturbation) need to be applied to a file and to which specific variables. In addition, local measures of risk should be developed - which would tell agencies which specific records have a high risk of disclosure and should be altered in some way.

The analytical calculation of disclosure risk for microdata is based on the uniqueness concept (see Greenberg 2003) and a number of different risk measures have been proposed (Skinner and Holmes, 1998; Singh, Yu and Dunteman, 2003). However, as the excellent review by Domingo-Ferrer and Torra (2003) point outs, when continuous variables are used, it is difficult to define uniqueness, and hence record linkage to identify reidentification risk – defined to be the proportion of records in the microdata file that can be re-identified with a sufficiently high probability against the population file using the best method or best combination of methods (Winkler, 2003, Domingo Ferrer and

Torra, 2003) – is an important area for future research. This approach has a number of advantages, both because it does not rely on restrictive matching assumptions and because multiple data sources can be considered in the matching exercise. However, much research needs to be done: to identify the datasets to which intruders are likely to have access (in consultation with the appropriate statistical agency); to identify the record linkage methods that could be used – e.g. distance-based, probabilistic, cluster-based, Bayesian network-based; and to agree on comparison criteria for the effectiveness of re-identification effectiveness, conditional on computational complexity. In addition, the continuing development of data-mining techniques means that an ongoing research investment is needed to systematically assess the risk of disclosure in micro-data. In particular, future research should focus on:

1. *Developing techniques that quantify the disclosure risk and data utility loss associated with the confidentiality protection – possibly by creating standard test cases*.
2. *Comparison of record linkage methods in assessing the value of microdata SDL methods.*
3. *Developing realistic intruder scenario with realistic, available micro data sources – which will typically require extensive computing capacity.*

Protecting New Methods of Disseminating Data: Confidentiality Issues with GeoSpatial Data

"Confidentiality is an excuse for providing bad data" Marc Armstrong, May 13, 2003

The expansion of research on the human dimensions of environmental change has increasingly meant that researchers want to include the contextual variables surrounding an individual-the schools they go to, the neighborhoods they live in, the firms they work for, etc, maybe also the people they interact with. As Rindfuss (2002) points out,

"Linking data on people and their environments is at the very core of IHDP" (p.1)

An entire literature has developed in spatially explicit analysis because location, pattern, and spatial structure all matter in understanding human behavior. As Entwistle points out, and as is evident from the figure below

"Land is stationary, continuous, and numerous aspects of land cover are measurable by remote sensing technologies. Linkage across different remotely sensed images is straightforward. Humans are mobile, discrete, and not easily captured by remote sensing technologies. The effect of a human on land use can range over considerable land territory. Linking humans to the land they influence is not straightforward" (Entwistle, May 13,2003)

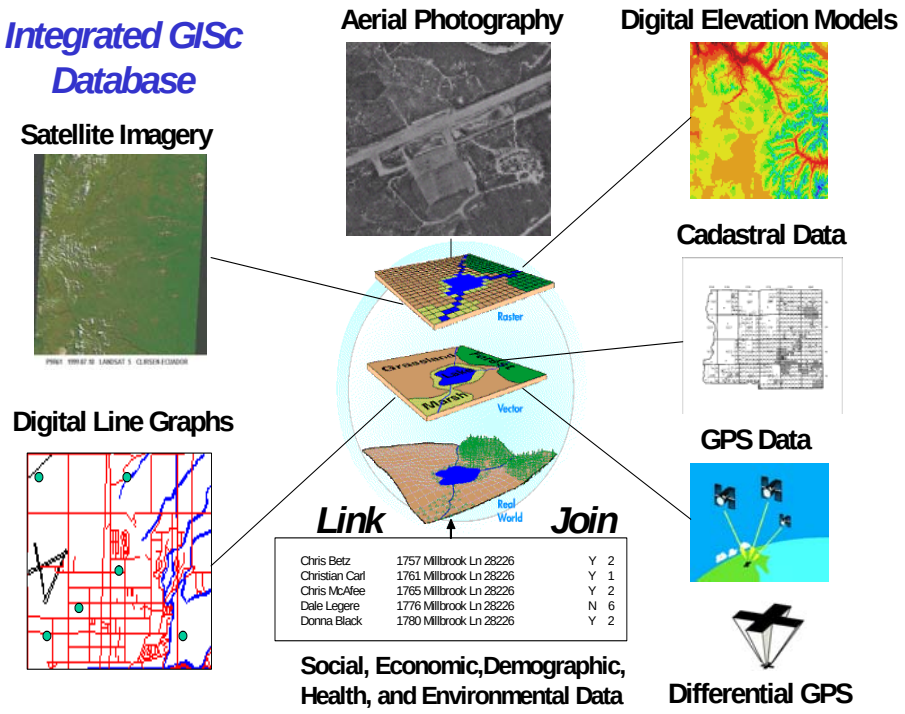


Figure (from Entwistle: May 13,2003)

The use of geo-codes (such as latitude-longitude coordinates) rather than addresses, political units can create risks to respondents because publicly available data

based on real property records—such as lot size, property tax maps – can lead to re-identification. As Balk (2003) points out, however, just as new administrative datasets have made it more feasible to link micro-data, so have technological advances such as global positioning system (GPS) instruments and satellite technology made it much easier to link location-specific data at the household or neighborhood level and re-identify individual respondents.

The most striking outcome of this section of the workshop was the degree to which approaches that have been used to protect confidentiality at the geo-spatial level mimic those that have been used to protect micro-data. In particular, researchers rely on geographical aggregation and removal of spatial context to protect confidentiality, but have similarly serious concerns about the impact of these measures on data quality (Balk, 2003). Other protection approaches, such as data masking are described in Armstrong, 2003 are very similar to the noise addition approaches in statistical disclosure limitation research – for example, locations are "offset" by a parameter that moves the geo-coded location off the centerline to a "plausible" (approx) location on the correct side of the street or "squeezed" by a compression factor that moves locations inward on block face to ensure they are on correct street. Similarly, inverse address matching approaches to measure the degree of re-identification risk are very similar to record linking approaches.

A core ethical question that was raised at the beginning of the workshop was echoed by Onsrud (May 13, 2003). He made it clear that the expanding use of spatial technologies in combination with communication technologies via location based services (LBS), poses a particular challenge to increase beneficial uses and grow the industry, while protecting users. The core assumption of the LBS industry

- that corporations and industry will own and control location and related information about individuals, individual choice limited to "opt-in" or "opt-out" of our services and boilerplate conditions (Onsrud, 2003) – leads to very different technical challenges and research questions than those that will be addressed by the market place. In particular, there is a very strong case to be made for research into the public goods aspect of protecting privacy - particularly development of a legal/ethical code of conduct.

The most obvious research foci should:

1. *Develop cross-disciplinary collaboration to address the intriguing new problems raised by geo-spatial research*
2. *Develop more expertise in geo-masking approaches and the address inversion technology – particularly to determine whether a HIPAA-compliant geo-mask can be developed to permit sub-state analyses*
3. *Examine the possibility of virtual enclaves to examine geo-spatial data*

Understanding the Data Dissemination Context

"Some people believe that the govt. has one big computer that houses all of their information" Eleanor Gerber, May 12, 2003

The risks associated with disseminating data give rise to a major concern on the part of data custodians that respondent perceptions might adversely affect their response rates. Research to understand this effect has primarily focused on demographic data, with a number of useful findings. Research by Eleanor Singer (2003) suggests that trust in survey organizations in general has decreased over time, and this has significantly adversely affected response rates (although size and explanatory power of the effect is quite small). Ethnographic research that the acceptability of different protection practices should not be decided in isolation from the respondents – since a major reason to protect confidentiality is to preserve response rates, finding out what resonates with individuals would seem sensible (Gerber, 2003). In both case, since both trust and laws and

regulations are likely to have undergone substantial changes post 9/11, the effect of these changes should be monitored. A second research issue is identifying just how much trust in different institutions exists, and how well confidentiality and nondisclosure practices are communicated to respondents. In particular, since "informed consent" is the mantra which often permits data release, what does "informed consent" mean?

A major problem for confidentiality research is that, despite its importance, it is relatively marginalized as a statistical discipline. Indeed, for many social science researchers, the first contact with confidentiality issues occurs during an institutional review board (IRB's) analysis of their project. The ubiquity of IRB's in social science research presents an potential opportunity to both disseminate confidentiality "best practices" as well as create new interest in the field in academic institutions. However, there is wide variation in how much IRB's know about basic confidentiality principles. Research in this field could provide a stocktaking of current approaches, and suggest ways in which IRB's could further researchers' understanding of confidentiality issues (Sieber, 2003)

Major areas for future research include

1. *A continuation of measurement of trends in public beliefs about privacy and confidentiality – both cross-sectional and panel in nature.*

2. *A continuation of measurement of public perceptions of disclosure risk and harm and the acceptability of different levels of risk and the communication of risk information to respondents.*

3. *Developing a better understanding of how trust in research and agencies that conduct can be increased.*

4. *Developing models at a few major institutions of a well-trained Disclosure Review Board to advise researchers and the IRB that involve the university's applied statisticians in interpreting, adapting and applying nondisclosure*

techniques. Develop a Data Research Center or enclave that can do in-house analyses when restricted access is required

References (All summaries from NSF workshop May 12-13 unless otherwise noted)

Abowd, J and J Lane "The Costs and Benefits of Microdata Access"

Abowd, J "Synthetic Data"

Armstrong, M "Geographic Information as a "Rosetta Stone" for Forging Individual-Level Data Linkages that Compromise Confidentiality"

Balk, D. "Confidentiality issues arising from integrating social and health behavioral data with geospatial data"

Cecil and Eden ": Legal Frameworks Governing Individually Identifiable Research Data"

Domingo-Ferrer, J and V Torra "Advanced Record Linkage for Disclosure Risk Assessment"

Duncan, G and S Roehrig "Protecting Confidentiality in Tabular Data"

Entwistle, B "Linking Data on People with Data on their Environments"

Fienberg, S. Allowing Access to Confidential Data: Some Recent Experiences and Statistical Approaches (presentation at Stockholm workshop on microdata access, August 21, 2003)

Giessing "Disclosure Control Methods for Tabular Data"

Gerber, E. "Privacy and Confidentiality: Ethnographic Approaches"

Greenberg, B "New approaches for assessing disclosure risk"

Madsen, P. "The Ethics of Confidentiality: The Tension Between Confidentiality and the Integrity of Data Analysis in Social Science Research"

Muralidhar, K and Sarathy, R. "Data Access, Data Utility, and Disclosure Risk are NOT Always Mutually Exclusive"

Onsrud, H. "Privacy in the Use of Spatial Technologies: Ethics as a Driver of Technological Research Priorities"

Sieber, J "Institutional Review Boards and Data Sharing"

Singer E "Public Perceptions of Confidentiality"

VanWey, L. “Challenges to Confidentiality in Spatially Explicit Research”

Winkler, ” Masking and Re-identification Methods for Public-Use Microdata” Seminar Italian National Statistical Institute, January, 2003

Confidentiality Workshop Agenda
NSF May 12-13

8:00 – 8:30	Breakfast and Coffee	
8:30 – 8:45	Opening Remarks	
Rethinking the conceptual framework		
8:45 – 9:05	The Ethics of Confidentiality	Peter Madsen, Carnegie Mellon
9:05 – 9:15	Floor Discussion	
9:15 – 9:35	The Economics of Confidentiality	John Abowd, Cornell University and Julia Lane, The Urban Institute
9:35 – 9:45	Floor Discussion	
9:45 – 10:00	The Legal Foundations of Confidentiality	Donna Eden and David Korn, HHS (TBC)
10:00 – 10:15		Joe Cecil, Federal Judicial Center (TBC)
10:15 – 10:25	Floor Discussion	
10:25 – 10:40	Coffee	
New Technical Approaches		
10:40 – 11:00	Data Access, Data Utility and Disclosure Risk	Krish Muraldhiar, University of Kentucky and Rathindra Sarathy, Oklahoma State University
11:00 – 11:10	Floor Discussion	
11:10 – 11:30	New approaches for assessing disclosure risk	Betsy Greenberg, University of Texas
11:30 – 11:40	Floor Discussion	
11:40 – 12:00	Advanced Record Linking Methods to assess risk	Josep Domingo-Ferrer, Universitat Rovira i Virgili and Vicenc Torra, Inst. Inv. Intel.ligencia Artificial
12:00 – 12:10	Floor Discussion	
12:10 – 1:10	Lunch	
1:10 – 1:30	Tabular Data Protection	George Duncan and Steve Roehrig Carnegie Mellon University
1:30 – 1:40	Floor Discussion	
1:40 – 2:00	Disclosure Control Methods for Tabular Data	Sarah Giessing Statistisches Bundesamt
2:0 0 – 2:10	Floor Discussion	
2:10 – 2:30	Synthetic Data	John Abowd
2:30 – 2:40	Floor Discussion	
2:40 – 3:00	Synthesis and Review	Rapporteur
3:00 – 3:15	**Coffee**	
Understanding the Data Dissemination Context		
3:10 – 3:30	Individual Perceptions	Eleanor Singer, University of Michigan
3:30 – 3:40	Floor Discussion	

3:40 – 4:00	Perceptions: Ethnographic Dimensions	Eleanor Gerber, Census Bureau
4:00 – 4:10	Floor Discussion	
4:20 – 4:30	Institutional Review Boards	Joan Sieber, UC Hayward
4:30 – 4:40	Floor Discussion	
4:40 – 5:00	Synthesis and Review	Rapporteur

8:00 – 8:30	Breakfast and Coffee	
Confidentiality Issues with GeoSpatial Data		
9:00 – 9:20	Challenges to Confidentiality in Spatially-Explicit Research	Leah Van Way, Indiana University
9:20 – 9:30	Floor Discussion	
9:30 – 9:50	Integrating social or health behavioral data with geospatial data	Deborah Balk, Columbia
9:50 – 10:00	Floor Discussion	
10:00 – 10:15	Coffee	
10:15 – 10:35	Geographic Information as a "Rosetta Stone" for Individual-Level Data Linkages that Compromise Confidentiality	Marc Armstrong, University of Iowa
10:35 – 10:45	Floor Discussion	
10:45 – 11:05	Linking social surveys with data sets containing environmental data	Barbara Entwisle, University of North Carolina
11:05 – 11:25	Floor Discussion	
11:25 – 11:45	Privacy in the Use of Spatial Technologies: Ethics as a Driver of Technological Research Priorities	Harlan Onsrud, University of Maine
11:45 – 12:15	Synthesis and Review	Rapporteur
12:15 – 1:15	Lunch	