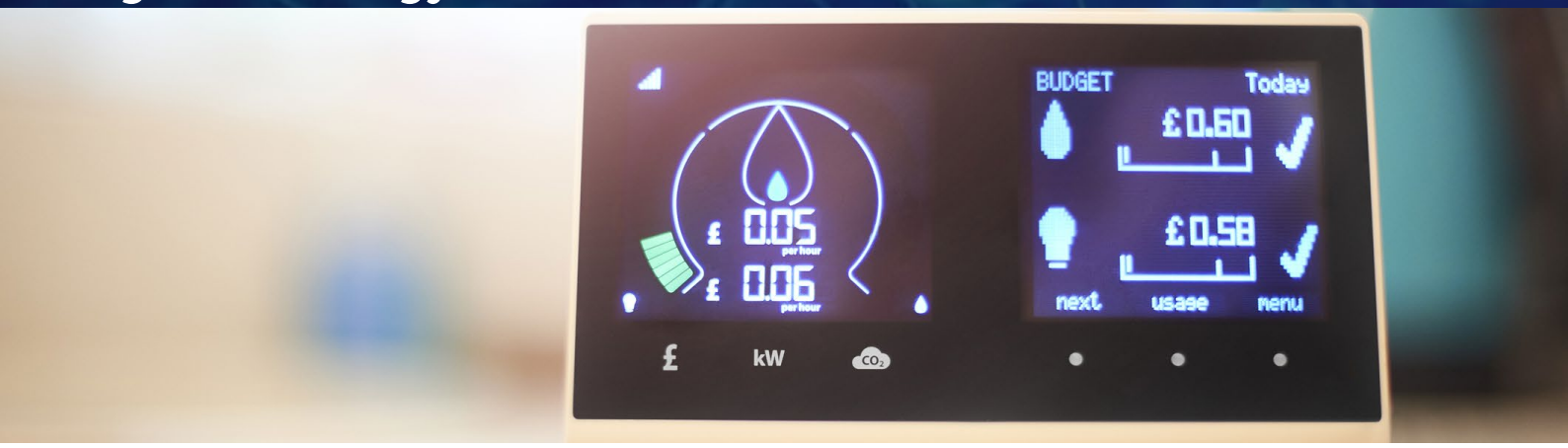
Legal and ethical challenges surrounding big data: energy data



The challenge

There are protocols for gathering informed consent for collecting data via social surveys, that address the increased disclosure risk that comes with linking these data to other information sources. However, non-research generated data from smart appliances or connected devices have distinct and different ethical and legal challenges. In some cases, various protections that are typically applied at several points in the research data life cycle may not have taken place. However, access to contexualised data from individuals or households, can be safely and successfully managed through a robust and trusted governance framework. Here we set out some of ethical challenges and consider the use of informed consent and the 'Five Safes' protocol to ensure safe access to data.

> We are faced with a tension between the privacy of peoples' or household data and the benefits of research using the data. And, as the risk of identification of a household increases so the need to enable safe and trusted access to linked data becomes essential.

Research example

The potential for significant public benefit derived from the use of smart meter data has been established by recent research (Elam, 2016). Two examples are:

- enabling time-of-use tariffs to help customers reduce bills and industry avoid unnecessary investment in 'peak load' power plants, which cater for demand peaks by shifting consumption to off-peak times
- undertaking detailed analysis of household energy consumption profiles that can lead to better identification and mitigation of fuel poverty

The ability to conduct innovative research is greatly enhanced if energy consumption data from meters can be linked to contextual data from a variety of sources. Smart meter data benefits from linking to:

- Energy Performance Certificate (EPC) data, which provide key data about the building such as the property type, size and insulation measures
- the English Housing Survey, which provides information on socio-demographics and household energy behaviours
- smart or connected devices that enable remote control and monitoring

Data and data issues

Existing research around privacy concerns point to (often largely unfounded) public concern about smart meters being used to:

- identify households' employment status or religion
- enable burglars to identify vacant properties
- glean information on life style and habits and subpoena this for divorce proceedings
- enable commercial organizations to use data for intrusive direct sales and marketing activities

UK Data Service

Issues for onward use of research are that:

- the nature of the consent process and consent form for energy consumers to allow access to their data from devices installed within their homes may be unclear and thus problematic
- data privacy notifications may be overly complex and thwart understanding
- protections applied when data are collected and processed (e.g., de-identification), may not be robustly implemented
- using the data for research may substantially differ from the original purpose for which it was collected (e.g., data to improve energy saving used later for research)
- linkage of smart meter data to additional data relating to the property or household greatly enhances the usefulness of smart meter data yet increases the risk of disclosure and hence privacy concerns.

We are faced with a tension between the privacy of peoples' (or household) data and the benefits of research using the data. And, as the risk of identification of a household increases so the need to enable safe and trusted access to linked data becomes essential.

Current legislation such as data protection requires a distinction to be made between identifiable and non-identifiable data and for it to be treated appropriately. At present, data custodians in both the public and private sectors rely on a number of practices in order to protect personal data including: anonymisation, or de-identification; obtaining informed consent; and regulation of access to data.

Privacy risk vs. benefits to society

But, despite the legal situation, recognition is growing that, in the era of big data, distinctions between what is identifiable and non-identifiable are actually becoming less tenable:

- Identifiability is increasingly being seen as a continuum, not binary
- Disclosure risks increase with dimensionality (i.e., number of variables), linkage of multiple data sources, and the power of data analytics
- Disclosure risks can be mitigated, but not completely eliminated
- De-identification remains a vital tool to lower disclosure risk, as part of a broader approach to ensuring safe use of data

And, it may not just be individual's identity in questions of privacy. If we consider analytics that discriminate on a group, i.e. a group gets an energy payment rebate, the current de-identification approach becomes inadequate. Despite any distinction between public and private, permitted uses with 'public benefit' outcomes or in the 'public interest' become useful. Alternatives to individual informed consent might be though a process of 'social consent', whereby sufficient protections are in place to ethically permit use of data.

In summary, it is helpful to move away from thinking about risk in black-and-white terms and focus instead on minimising risk and balancing risk with the benefits to society from the use of these data.

Data and data issues

Informed consent

A first solution is to reduce the data privacy concerns of participants who are allowing their household's smart meter data to be used for research. Data is collected on a strictly voluntary basis based with explicit consent of households who have agreed to provide their smart meter data for research. In the UK, smart meter data is protected by a robust Data Access and Privacy Framework incorporated into legislation via the Smart Energy Code. This mandates that, beyond monthly data for billing purposes, smart meter data can only be used with the informed consent of the energy consumer.

Safe access to data: robust governance

At the UK Data Service, our guiding principle is that we make data open where possible and closed when necessary. To do this, we provide access via a licensing framework that meets the needs of data owners and matches the risk level of the data collection to be licensed. Incoming data are classified according to their level of detail, sensitivity and confidentiality and appropriate data handling and access safeguards are in place.

De-identified datasets should be made publicly available where possible. Access to any identifiable data needs to be restricted to authorized persons who plan to undertake research which will lead to government, industry and organisations acting for public benefit delivering innovative products and services.

Safe access to data: robust governance (cont.)

A solid governance framework should ensure optimal ethical operation of data access. There are well- established protocols that govern ethical use of data, such as those developed by UK Data Archive, where only Approved Researchers who have been trained in safe data handling techniques obtain access to identifiable data, and whose projects are approved by a Data Access Governance Board. Access to data takes place via an approved Secure Lab environment. Trusted Third Parties are used for linking and de-identification where personal administrative data is used. This protocol is known as the 'Five Safes'.

The Five Safes

- Safe People – approved / accredited researchers
- Safe Projects – all projects must be approved by a Data Access Governance Board
- Safe Settings – identifiable or sensitive data will reside in approved Secure Lab environments where analysis will be conducted
- Safe Data – where possible data will de-identified before release to researchers. Identifiable data can only be analysed in secure environments.
- Safe Outputs – Results must comply with approved Statistical Disclosure Control protocols and results will be checked before they can be published

See the UK Data Service's Five Safes video

Our technical solution

As we start to roll out big data services, we can utilise our Five Safes model to roll out 'Five Safes at scale'. The UK Data Services Data Service as a Platform (DsaaP) is placing the Hadoop data ecosystem at its core with authentication and authorisation as key components. Security considerations must meet our spectrum of data access from Open, through Safeguarded to Controlled data. DsaaP's hybrid solution utilises Amazon Web Services cloud provider allowing us to perform on-demand scaling for our search infrastructure and web traffic, and an on-premises infrastructure to maintain tight governance controls on our local data stores.

The fully integrated nature of the combined platforms of HDP (Hortonworks Data Platform) and HDF (Hortonworks Data Flow) offer robust security features. Kerberos is used to authenticate users, processes running within and between clusters acting on behalf of the user and internal services running within the cluster itself, giving us confidence in the security of the data we hold. MS Active Directory is used as the domain controller for our Kerberized environment and auditing and governance components allow us to maintain comprehensive data lineage. All relevant actions and version information from all of the workflows are recorded as the data travels through the repository pipeline.

Further, privacy tools can help us consider when to release a non-disclosive information product, like an aggregated regional map of energy use, or when to lock down other more disclosive data products to authorised users.

See our case studies:

- Utilising smart meter data to enable Energy Demand research
- Scaling up digital data services for the social sciences

Authors:

Louise Corti and Libby Bishop, UK Data Service, and Simon Elam, CEE, UCL

DataFirst

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

NRF National Research Foundation

UK Data Service