Sense-T Platform
Data Privacy Whitepaper

Contact:

Dr Paul Neumeyer
Program Manager – Technology and Infrastructure, Sense-T

Purpose of this document

The goal of Sense-T is to achieve social, economic and environmental benefits through harnessing knowledge derived from historical and real-time sensing data. Through the availability and analysis of data, Sense-T aims to help businesses, governments and communities to make better decisions.

The Sense-T Data Platform supports the objectives of Sense-T and will increase the availability of rich, diverse, high quality, measurement data that is vital to the public good, community functioning and well-being.

This paper is a document that describes a systematic approach to data privacy and confidentiality that will enable Sense-T to deliver as much value as possible to government, research, and business in a way that appropriately protects stakeholder rights to data privacy and confidentiality. This paper is informed by best practice taken from established data privacy practitioners and organisations such as the Australian Bureau of Statistics, but has been specifically tailored to meet the unique needs and objectives of Sense-T.

Scope and Challenges

Privacy has become an important debate worldwide as the collection, storage and analysis of large amounts of digital data increases. The key aspect to the global privacy debate has been regarding risks to an individual and the protection of private freedom that could be affected if a person can be identified and linked to activities and details in data.

The challenge for Sense-T is not only to manage the risks to an individual but also the broader issue of confidentiality of sensitive business information and the commercial secrecy of data that Sense-T has to consider in its data management framework.

It is essential for Sense-T to be aware of and put in place measures to correctly handle industrial collaboration including commercially sensitive data to fulfil its stated goals to achieve economic benefits. The Sense-T Data Platform could also be used in commercial operations that do not involve University of Tasmania research activities. The key question for commercial users of the Sense-T Platform is who can access their commercially sensitive data if it goes into the Sense-T platform and how access is controlled.

Sensitivity of data may also be non-commercial. For example, some data is hard to interpret or could potentially raise unnecessary concern or fears in the community. As an example of community concern, water quality measurements are available and if those showed levels of a measured value were outside the desired range for a period the raw data may on its own cause unwarranted fear regarding the use of that water resource. However, the community may be unaware of other aspects such as a management plan that could act to remove the risks to the community or factors that can affect the measurement such as a malfunctioning sensor.

Another aspect of data is that access to data that informs and may reduce concerns, such as identifying the reason behind an issue. As an example, in the case of water quality the considerable distress to users in the community who see discolouration of the water may be reduced if they have access to data that details the reason for that discolouration.

Sense-T must face these challenges with the knowledge that the protection of data must be balanced with mechanisms to encourage sharing of data. The approach presented allows the Sense-T Data Platform to support a wide spectrum of use from the purely commercial end, through the range of collaborative research and to the other end of the spectrum where data is open to public use.

To inform the discussion this privacy whitepaper outlines the Sense-T privacy framework, including privacy definitions developed by Sense-T that extended privacy and confidentiality beyond individual privacy and included commercially private data and the means to share sensitive data confidentially when there is agreement in place.

The approach detailed in this whitepaper supersedes any previous communications or proposals for the privacy and data management for Sense-T

Definitions

For this document we adopt the following definitions:

Data: means … any unfiltered facts, information and measurements collected from environmental and other sensor technologies and associated metadata, which

(a) is being processed by means of equipment operating automatically in response to instructions given for that purpose,

(b) is recorded with the intention that it should be processed by means of such equipment,

(c) is recorded as part of a relevant filing system or with the intention that it should form part of a relevant filing system,

Data Subject: means … any person, entity or observable property of the environment that is the subject of observation in the process of collecting data

Data Controller: means … any person or entity who (either alone or jointly or in common with other persons) determines the purposes for which and the manner in which any data are, or are to be, processed or disclosed to a data receiver, or has been authorized via a license or agreement by a data controller to have the right to determine those purposes on any data

Data Processor: means … any person or entity (other than an employee of the data controller) who processes the data on behalf of the data controller[1]

Data Receiver: means … any person or entity who has been authorized by any data controller to receive the data and to whom the data are disclosed, other than any data processor or other person authorised to process data for the data controller or processor

Metadata: means … any records, facts, information and measurements defining and describing data so that it can be interpreted correctly

Privacy: means … the right of individuals and entities to hold data in a private domain secret with the reasonable expectation of no intrusion to that domain.

Confidentiality: means … the assurance and obligation by a data processor or data receiver, that data intended to be confidential will only be disclosed if authorised and only at an authorised level of detail.

Confidentialise: means … to output a data with a change to the level of detail in one or more aspects so that confidential information cannot be easily or spontaneously identified in the output, either directly or indirectly.

Security: means … the mechanisms implemented to restrict and control the access to data as part of the means to maintain an acceptable level of Privacy and Confidentiality.

Private Domain: means … a domain around a person or entity which includes all the things that are a part of it and that includes seclusion, limited accessibility or the ability to control information flow[2]

[1] Influenced by terms used in the *Data Protection Act 1998*, UK Goverment

[2] Yael Onn, et al., *Privacy in the Digital Environment* , Haifa Center of Law & Technology, (2005)

Sense-T: is a collaboration between the University of Tasmania, CSIRO and the Tasmanian Government, based at the University of Tasmania. It is also funded by the Australian Government. Data agreements will be between the University of Tasmania and the data controller, unless otherwise indicated.

In most cases Sense-T acts as a data processor for data provided from other organisations, who are the data controllers. Sense-T, the University of Tasmania and CSIRO may collect their own data and be the data controller of that data on the Sense-T data platform or be a data controller because they have rights to sub license data from another data controller. Sense-T may also make data accessible (link) to other systems in which case it acts as a data proxy.

Core Principles

To facilitate our activities Sense-T will hold and measure our people, processes, policies and products to the following principles:

Data Management Principles

P1: Privacy by design
P2: Data is an asset to share
P3: Data usability through open standards
P4: Data integrity and attribution by design
P5: Data protection through security and monitoring
P6: Federated data preferred over duplication

Confidentiality Principles

C1: Disclose only if authorisation given
C2: Authorized level of detail
C3: In accordance with legal and regulatory frameworks
C4: Inform stakeholders on Confidentiality

Three levels of Confidentiality
L1: Open
L2: Controlled
L3: Research/Private

Principles in detail

P1: Privacy by design

Privacy by design is a term established by Ann Cavoukian in the 1990's. Privacy by design is characterised by **proactive** rather than reactive measures to build in privacy that must become the default mode of operation for an organisation, its people, procedures, policies and products.

One of the foundational principles is that **the initial status of data added to a data sharing platform is that access is only given to the data controller that added it**. The access can subsequently be opened up by a data controller in a controlled way including the creation of derived versions of the original data that may be Open data. It is our experience that when the initial default access to added data is closed both the trust of the platform and the volume of commercial data on a platform increase.

P2: Data is an asset to share

Data controllers need to be aware and encouraged to share their data. A systematic mechanism needs to be built to make discovering data and the types of available data easier for data receivers so that data that is useful can be accessed and analysed.

P3: Data usability through open standards

The use and promotion of open standards for the ingestion and consumption of data is essential for the technical openness of data. This relates to protocols used to access, query, and perform analysis of the data and format the results. Standards need to be well documented and where more than one standard is involved, because of the different aspects of the system, those chosen also be aligned so they complement each other.

P4: Data integrity and attribution by design

All data needs to have meta-data attached to ensure that it has attribution of the source and data controller, and if a modification occurs meta-data attached to indicate the form of change and reference to the type of modification. If data modification occurs in multiple steps a providence history should be recorded so that the modification steps and data dependencies can be understood.

P5: Data protection through security and monitoring

Security mechanisms need to be implemented for physical, transmission and storage security. Mechanisms to identify abnormal data access and potential security breaches is essential so that a response to a possible data security issue can be initiated as early as possible in order reduce the impact of a security threat.

P6: Federated data preferred over duplication

There is a wealth of public and private data being collected by organizations that if shared would benefit data receivers. Due to the volume of data stored in existing systems a mechanism to link systems is essential where those links can enable access to the data resources by proxy, and avoid making a copy of the data. The linking mechanism would need to automatically interpret if query for data will directly access data stored on a local system or if the request must be passed on to an external linked system where the data is stored. This dual model of access is a safeguard against the resources needed to duplicate the storage of data and also avoids out of date copies of data that are orphaned from the privacy or confidentiality mechanisms on the original system data is sourced from.

C1: Disclose only if authorization given

When the agreement between a data controller and a data processor requires protection of commercial sensitive then that data can be protected from disclosure using confidentiality mechanisms. This does not override or erode the legal imperative that there may be times where data is shared with others because of a legal requirement.

C2: Authorized level of detail

The data controller must be able to store their data with sensitive and confidential details included and have security mechanisms so that only they can access those sensitive and confidential details. Confidentiality mechanisms need to be implemented so that new versions of that data can created by a data controller and be shared. The confidentialisation mechanisms need to be automated transformations configured to meet the data controller's requirements to change the level of detail including exclusion, de-identification, de-identifiable statistical protection, and time embargo of their confidential data.

The original sensitive version of data will often be updated by a data controller as new data is added so that the other confidentialised versions of the original data must also be updated in sync with the changes to the underlying data.

This principal does not override or erode the legal imperative that there may be times where data is shared with others because of a legal requirement.

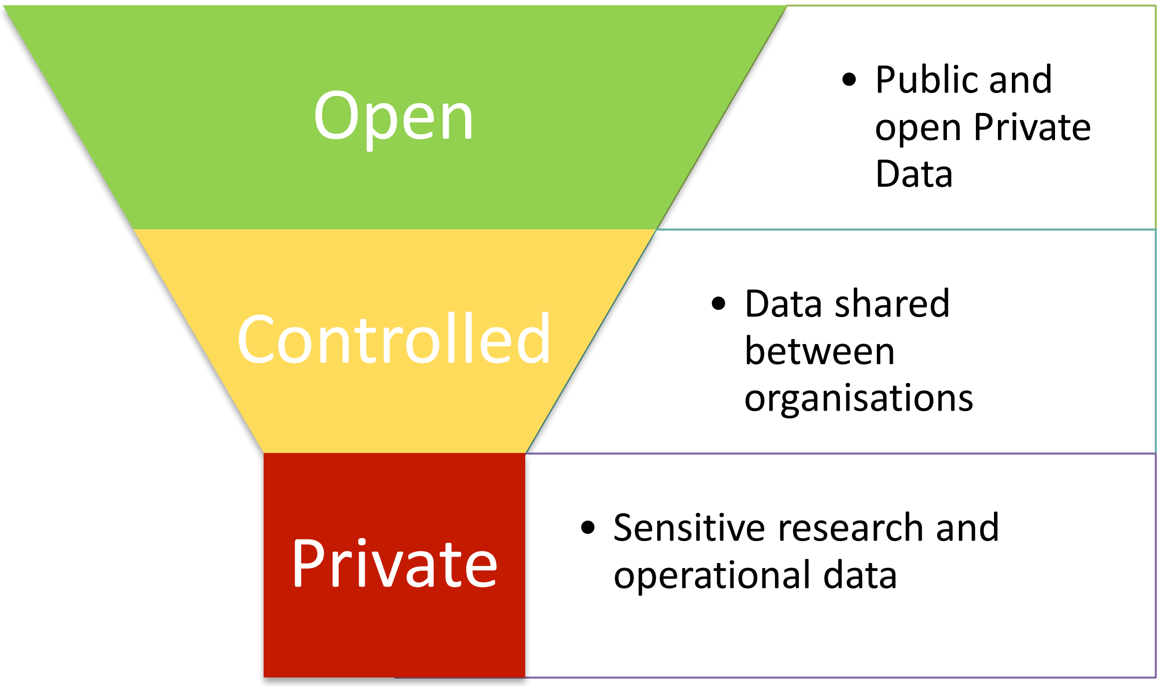
C3: In accordance with legal and regulatory frameworks

Data must be managed in accordance with the Australian and other external legal and regulatory requirements related to Privacy, Confidentiality and Security. The number of legislative and legal requirements and variation in terms between those makes it difficult to ensure all legal and regulatory requirements are met.

C4: Inform stakeholders on Confidentiality

Confidentiality and privacy requires the data controller, data processors and data receivers to understand, implement and manage confidentiality and privacy considerations correctly. It is a key principle to implementing education and communication initiatives.

Three Levels of Data Confidentiality


Open
Public and open Private Data
Controlled
Data shared between organisations
Private
Sensitive research and operational data

L1: Open

Data controllers will be encouraged to assign the open confidentiality level on as much of the datasets as possible. For these datasets, the value to the community for this data is very high, and the risk — commercial, regulatory or otherwise — is very low. In general, Open data will be de-identified and may have a variety of confidentiality processes, such as aggregation, applied in order to prevent obvious or spontaneous recognition.

The data controller must obtain all Open data provided to the public lawfully and the data controller of the Open data will be identified with the Open data. In addition, the data controller accepts responsibility for ensuring that they have the authority to make the data public so once approved to be Open by the data controller Sense-T will then follow a path of "public until requested to be private". That is, such Open data will be disclosed without Sense-T taking explicit responsibility to audit the privacy or confidentiality, but Sense-T reserves the authority to remove the data from Open access if an individual or organisation can demonstrate to a reasonable degree that such disclosure might be likely to enable the identification of that person or organisation. When such situations arise, Sense-T will reserve all rights to cease disseminating the data while Sense-T notifies the data controller and reviews the claim to determine whether to cease or continue to disseminate the data as Open.

The Sense-T Platform will support licensing terms for access to Open data sets by registered users of the Sense-T Platform.

.

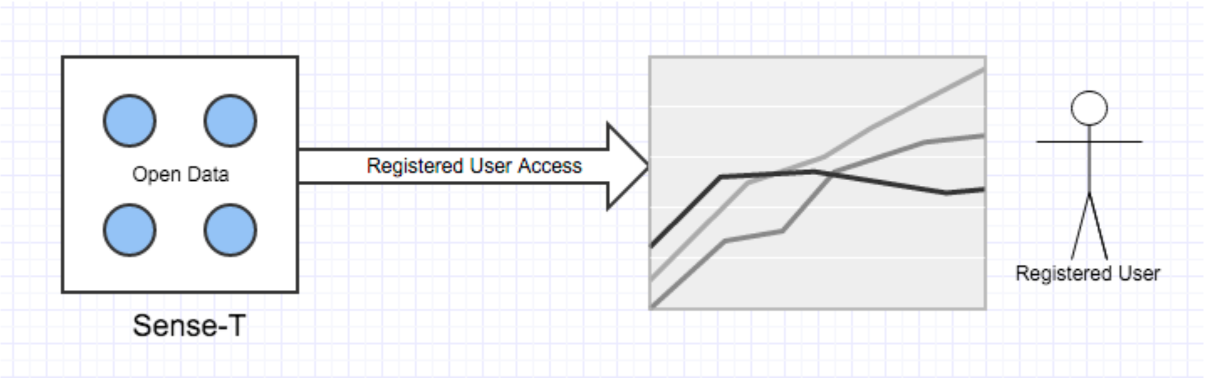


Figure 1 An Open data set will be accessible by registered users

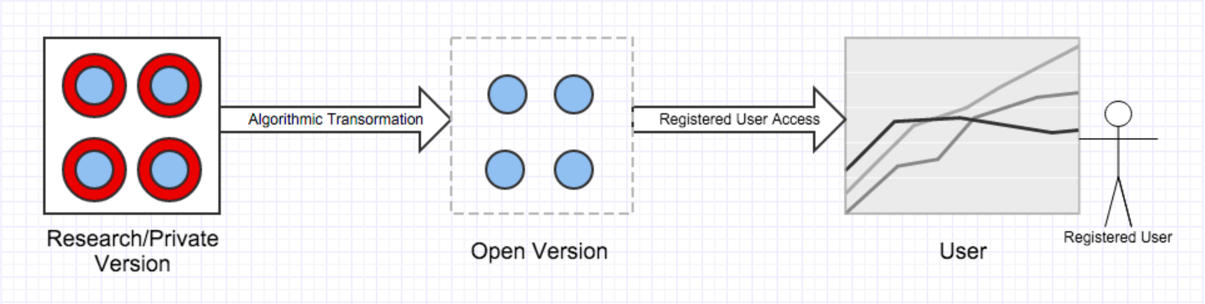


Figure 2 An Open data set can be algorithmically generated from private data in real-time as new data arrive leaving the Research/Private version protected

L2: Controlled

Data can be shared between organisations under controlled confidentiality[3]. In contrast to data that can be accessed in full detail by the data controller, the type of user and their level of authority dictate the detail that is appropriate in the controlled data they access. That is, it will be less protected than raw data, but more protected than Open data. A level of confidentialisation can be setup by the data controller.

Further protection will be enforced through a contractual agreement between the data controller and data receivers, however this does not override or erode the legal imperative that there may be times where data is shared with others because of a legal requirement.

Agreements should be made with authorised representatives of the organisations regarding what they can do, what they cannot do, as well as the potential consequences

[3] This environment is comparable to the ABS microdata platforms such as Remote Access Data Laboratories (RADL), and Confidentialised-Unit Record Files (CURFs).

if these conditions are breached. The existence and visibility of such arrangements are necessary to increase public and industry confidence that the data will be used appropriately when shared between organisations.[4]

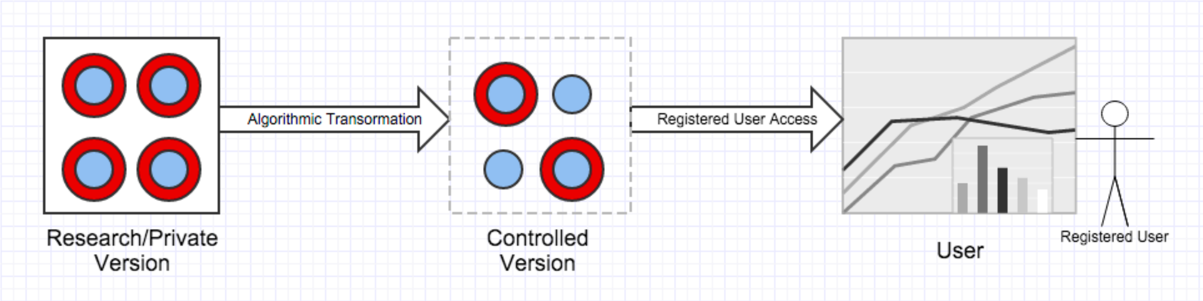


Figure 3 A controlled version of a data set can be generated from private data to share with other organisations

L3: Research/Private

The data controller will have access to the raw data they add to the system. Research questions often need the detail of the raw data to succeed. Only the data controller controls the permission for the group of stakeholders with access to the raw data because it may include personally identifiable data and sensitive metadata associated with the source of the data. The data controller will also restrict access to the raw data to only specific individuals from their organisation as required.

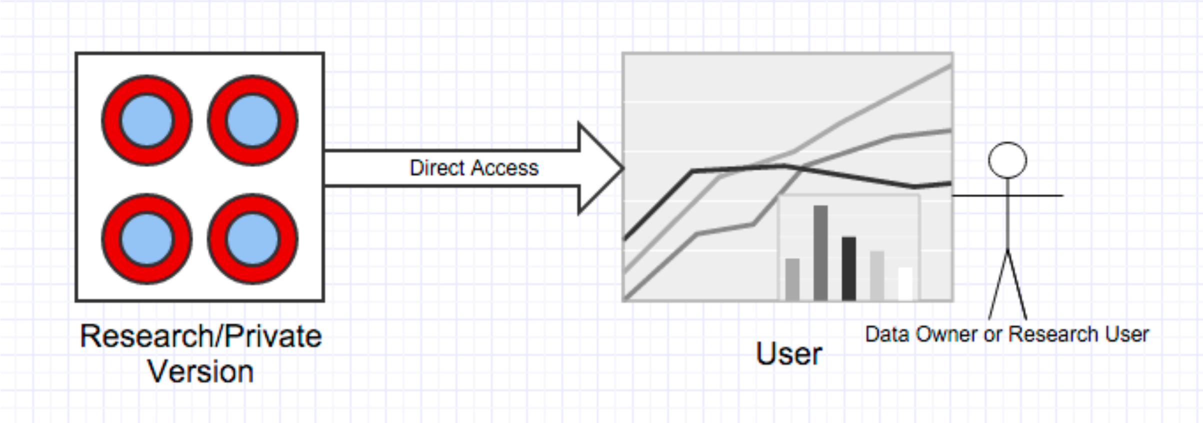


Figure 4 Research data can only be directly accessed by the research team

[4] For further information, see the definition and recommendations as outlined in "Principles and Guidelines for Managing Statistical Confidentiality and Microdata Access", Statistical Commission, UNECE March 2007

Confidentiality Framework

The Sense-T confidentiality framework consists of 8 parts.

Part	Description	Supported Principles
Restrict	Determine and enforce through access control, and by default limit, who can send, view, modify, copy, disseminate, annotate and agree to disclose particular data.	P1, P2, C1
Assess	Assess who would like to access information, what it will be used for and the level of detail that can be disclosed to them for that purpose. The objective is to minimise the risk that confidential information from an individual or organisation will be disclosed while maximising the usefulness of the data.	P2, P3, P6, C1, C2
Protect	Protect the integrity of incoming data, secure the stored data to the appropriate degrees using established contractual, physical, electronic and statistical processes; and disseminate using statistical confidentiality techniques. Furthermore, a Disaster Recovery Plan (DRP) will be in place so that all staff and stakeholders know precisely what actions to take in the event of a disaster in order to recover the system to a functioning status.	P4, P5, C2
Legislate	Understand the legal obligations to protect the confidentiality of and privacy of data. This provides the cornerstone of all confidentiality processes. Legislative and regulatory frameworks include and are not limited to: • Privacy Act 1988; • Personal Information Protection Act 2004 (Tas) • Right to Information Act 2009 (Tas)[5].	C3

[5] Reference to "Right to Information Act 2009 (Tas)" http://www.ombudsman.tas.gov.au/right_to_information/process

	- Australian Code for the Responsible Conduct of Research.	
Contract	Sense-T may put in place contracts as appropriate. In circumstances where legislation does not explicitly or comprehensively cover protection, contracts between relevant parties should be established. The principle of constant adaption and continual review is vitally important to adopt. It will also require all policies and processes relating to privacy, confidentiality and security to be periodically reviewed and assessed for its continued relevance, including this document.	C3
Inform	Establish a culture, both internal and with external stakeholders, that confidentiality is a central and shared value. Inform staff, stakeholders and users about the value and importance of privacy, security and confidentiality through both formal and informal sessions. Engage data stakeholders regularly about what is being collected, how it might be used, and how the privacy and confidentiality can be informed.	C4
Create	Create, implement and enforce policies within the organisation to ensure that confidentiality and security standards are maintained. Furthermore, a Data Breach Plan (DBP) will be in place so that all staff and stakeholders know precisely what actions to take in the event of a data breach.	P5, C1
Monitor	Log data usage, log operational and security metrics (as outlined in the UTAS Web Privacy statement), and periodically review policies and procedures. Monitoring will also include measures to identify security risks.	P5

Security Framework

The security policies and procedures are designed to protect the data from four categories of disclosure risk:

1. **Malicious attempts by external hackers to access and expose enterprise-wide confidential data.**

2. **Deliberate attempts by external parties to access or reverse-engineer confidential information about a particular individual or organisation.**
3. **Incidental, spontaneous or accidental occurrences of external users identifying some personally identifiable information.**
4. **Illegal or inappropriate attempts by internal staff or suppliers to use information**

Sense-T will use its best efforts to ensure security for data during the stages it has control of, including data ingestion, storage, analysis and dissemination. Sense-T will also encourage and inform internal and external stakeholders on how data can be protected through all phases of its lifecycle in the parts that Sense-T have no control over, including collection procedures, sensor security, telemetry equipment and remote data storage and during the process of disseminating data.

The fundamental principle of layered protection will also be applied and checked at multiple points within the Sense-T Platform in order to minimise the risk of a minor or systemic security breach. Such protective layers include legislative, contractual, physical, digital, and statistical processes. Generally speaking, digital protection secures data by tightly controlling access to the data, whereas statistical confidentiality processes confidentialise the data by removing detail.

“The degree of confidentiality should be proportional to both the assessed risk, and the likelihood that a person could obtain the same information through other means.”

Sense-T security includes physical security (including electronic surveillance, staff identification policies, security passes, electronic identification, etc.), as well as security of the computer systems themselves. Digital/electronic access will be on a need-to-know and need-to-use basis, with modern password procedures including password creation, verification and authentication, rotation and encryption. By combining encryption and digital certificates, Sense-T will implement authentication, integrity, encryption and token verification.

These protections are further supported by contractual agreements enforceable through legal proceedings.

Statistically Protecting Data

"98% of urban Canadians can be uniquely identified by the combined attributes of postal code, birth date, and gender."[6]

[6] El Emam, K, "The Re-identification Risk of Canadians from Longitudinal Demographics".

"In 2013, Montjoye et al. showed that 4 spatio-temporal data points of people's mobile phone usage are sufficient to uniquely identify 95% of the individuals."[7]

Sense-T proposes to make available a wide range of the techniques regarding the statistical protection of confidentiality through confidentialising algorithms. Applying a combination of confidentialising algorithms on the same data is also possible.

Aggregation: The value of a measurement will be based on a (weighted) average of observations within a defined dimension over a segment of that dimension.

Sampling: Sampling for confidentialising sensor data is that the statistic / data stream published corresponding to a dimension, will be a statistical representation of the sensor data over a segment that is centred at a point and large enough to include sensors from at least N distinct data streams. The statistic will be a random subsample of at least K observations in that dimension.[8]

Thresholding: This approach is a modification of the observed value so that values above an upper limit are set to an upper limit value, and any below the lower limit are set to a lower limit value. Values between those two limits are left unchanged.

Quantising: The observed values are modified in this approach so that it is represented by the nearest quantisation value.

Perturb: A random change is made on the observations based on a maximum change and distribution function of the changes.

Rearrangement: The observations are mixed randomly within a dimension but the distribution of values can be preserved.

Tokenisation: Identification information is replaced through the use of tokenisation and correlation to external systems can be made using those tokens.

Relative: Observed values are modified to be relative values which can be combined with an external reference point to get absolute values.

Embargo: Observations are made inaccessible if they are within the embargo windows. There can be an embargo on the most recent data or an embargo on trailing data.

Excluded: A fixed segment of data has a particular dimension excluded.

[7] http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html

[8] The architecture will also allow for an implementation of the dominance rule that can be described as "no individual data point should contribute more than x% of the total variation of the published statistic."

Unique considerations for Sense-T Privacy and Confidentiality

Real Time Data. Sense-T has a major focus on real time sensor data. This is in contrast to most statistics from government or scientific organisations where there is a delay between time of collection and time of publication due to a large number of intermediary manual processes. For Sense-T this means that human intervention must be eliminated from the critical path between the collection to storage and the dissemination of the data. Confidentialisation relies on having algorithmic processes applied automatically to the data as it is received and disseminated in real-time.

Spatiotemporal Series Data. Sense-T data pivots around data recorded at a position in space and time from sensors. This series of observations is often called a data stream. This implies that all data management practices, including confidentiality, need to treat time-series data and spatial data as a first-class statistical entity. This is in contrast to most existing statistical organisations that focus on the publication of data tables as the main mode of data dissemination often with months or years between new data being published.

A data eco-system. In contrast to systems centred on storing data from a single organisation, sharing of data provided into the Sense-T Data Platform will be encouraged and facilitated. Sense-T will put in place mechanisms to balance the privacy needs of the data controllers, against the details required to discover what data is available and how to get agreement to access it.

Data proxy or processor not data controller. Sense-T does not directly manage or own all the sensors providing data to the Sense-T Platform, but in some cases may have sensors and be both the data processor and the data controller. In most cases Sense-T acts only as a data processor and sensor data and data feeds from participating organisations may be stored and processed within Sense-T. Sense-T may also act as a data proxy and provide access to (link) to data in other systems and not store or process that data.

Highly localised. An important aspect of the value of shared data is to encourage the ingestion of sensor data from multiple sources to create a high spatial resolution of sensing data in critical regions so that analysis and decision support can source relevant and accessible data suitable for a location.

Real time analysis. Analysis of data ingested by the platform will be supported as the data arrives to avoid the need to download datasets to separate environments for analysis. Dissemination of the results of analysis, including the execution of associated confidentiality processes, must therefore be sufficiently automated.

Real-time Confidentiality Techniques: Sense-T will be using real-time data transformations to output and disseminate confidentialised versions of datasets as the central mechanism and those mechanisms implemented within the Sense-T platform will have satisfy the performance of a real-time system covering multi-dimensional data combining time, spatial and complex data formats that make this a challenging goal.