IESO Stakeholder Engagement
Foundation Working Group (FWG)

Meeting #3 Summary

Date held: June 17, 2015	**Time held: 8:30 AM – 3:15 PM**	**Location held:** Crowne Plaza Hotel, 33 Carlson Court, Toronto, ON
Working Group Members, Observers and Invited Guests	**Company Name**	**Attendance Status** (A)ttended; (R)egrets; (S)ubstitute; (P) Phone Participant
Adam White	Aitia Analytics	R
Jeff Evenson	Canadian Urban Institute	A
Rob Kerr	City of Guelph	R
Sarah Griffiths	EnerNOC, Inc.	A
Jennifer Gordon	Halton Hills Hydro	A
Brian Lennie	Horizon Utilities	A
Sally Barakat	Hydro Ottawa	A
Dean Dohring	IESO	A
Stuart Smith	London Hydro	A
Karen Carter	Ministry of Education	A
Guy Newsham	National Research Council	R
Leslie Goldsmith	Affinity Systems Limited	A
Jessica Webster	National Resources Canada	A
Marisa Uchin	Opower	P
Christine Dade	Horizon Utilities	A
David Craig	PricewaterhouseCoopers LLP	A
Gord Ellis	Soft Grid Analytics Corporation	P
Kevin Myers	Veridian	A
Brian Byrnes (Observer)	Ministry of Energy	R
Fei Chiang	McMaster University	A
Ann Cavoukian	Privacy and Big Data Institute	A
Khaled El Emam	CHEO Research Institute and the University of Ottawa	A
Janet Gore	Information & Privacy Commissioner	A
Renee Barrette	Information & Privacy Commissioner	P
Debra Grant	Information & Privacy Commissioner	A

Foundation Project Team	Company Name	Attendance Status (A)ttended; (R)egrets; (S)ubstitute; (P) Phone Participant
Lisa Barnet	IESO	A
David Barrett	IESO	A
Simon Geraghty	IESO	A
Bob Guberman	IESO	A
Ryan King	IESO	A
Julia McNally	IESO	A
Przemek Tomczak	IESO	A
Chris Tuff	IESO	A

Please note that the views represented in the summary below reflect the diverse views of members of the FWG and not necessarily those of the IESO. Links to the presentation materials are provided with each item.

Item 1 Introduction/Opening Remarks

Item 2 Discussion – Structures and Standards for Data Enhancement: MDM/R Requirements and Options

Address Format and Standards

Members of the working group representing LDCs shared their standards for address information and described the accuracy of premise data sent to the MDMR.

- LDCs' premise information is stored differently by different LDCs.
- Most LDCs use Canada Post information to correct/verify their postal code information; if a postal code is missing a dummy postal code is assigned and then corrected later by Canada Post information.
- One LDC confirmed all premise address information with the municipality's address data.

Options for specific structure of address data were discussed:

- Send the address as a single field using delimiters to separate the distinct components.
- Send the address as a single field that is "free form" i.e. does not necessarily use a delimiter to separate address components.
- Send the address using a separate field for each element of the address (i.e. street name, street number, unit number).

- All the LDC members of the FWG provide the address information to the MDM/R in a single field, although one has the information in separate fields in its own systems and converts it into a single field to send it to the MDM/R. The LDCs using a single field indicated it would be additional work were they to have to break the contents into separate fields. Also, the MDM/R is currently configured to receive the address information in a single field so additional work would also be required on its end. Therefore, the group converged on retaining the single field construct for the address information.

- Both single field options are possible for the MDMR to accommodate if it simplifies the work of the LDCs. If there is a desired construct, it could be converted outside the MDMR. Standards available are MPAC, MLS and Canada Post, with Canada Post being the most advanced.

- The group reached general agreement on Canada Post as the standard format for address information, regardless of the data structure.

- Most LDC members indicated they have some or all of the x-y coordinates or GIS information for service addresses. It was determined that if an LDC has this information the MDMR could be configured to accept it, but this would not be required information.

<u>Indication of Change in Occupancy</u>

- Members of the working group representing LDCs shared whether the data sent to the MDMR included a SDP ID to Account ID relationship that could be used to indicate a change in occupancy.

- A variety of answers were given including:

    - currently using the SDP ID to Account ID relationship in this way;
    - not using this relationship but indicating a change in some other way; and,
    - not providing information that could be used to determine this change.

- Some LDCs currently send this information to the MDM/R, whereas other LDCs previously sent this information but found that is was problematic and discontinued the process.

The FWG agreed on two options for LDCs to send occupant change information to the MDM/R:

- Use the existing MDM/R functionality of Account ID to SDP ID relationship to determine occupant change.

- Create a new date effective occupant change field for LDCs to send to the MDM/R that includes the SDP ID where the occupant change occurred.

It was also noted that for some LDCs, other information, already in the MDM/R, might be usable to determine an occupant change at a premise. The FWG recommended that the SME and LDCs explore this possibility further to potentially reduce the cost of implementation of providing occupant change information to the MDM/R.

Item 3 **De-identification: Overview of Techniques, Presentation and Analysis of Examples**

Sally Barakat – Hydro Ottawa

- Sally described her experience using a de-identification technique during a Time of Use (TOU) study conducted in 2013. The study was undertaken to quantify the change in energy usage by pricing period, estimate the peak period impacts using the OPA's summer peak demand definition and estimate the elasticity of substitution between the pricing periods and the overall price elasticity of demand. To conduct the analysis, Hydro Ottawa had to provide a third party (the OPA and its consultant) with electricity data associated with a single customer pre and post the installation of smart meters. To protect the privacy of the customer, Hydro Ottawa created a Study ID and provided this along with the electricity consumption data to the third party. Hydro Ottawa was the only party that could link the Study ID to a customer account and this link was not disclosed outside of the organization.

Jennifer Gordon – Halton Hills Hydro

- Jennifer described a technique of data aggregation that was used for a Conservation Achievable Potential Study. Halton Hills Hydro aggregated their consumption data by North American Industry Classification System (NAICS) codes and provided the aggregated data to a third party consultant to conduct the study. Lessons learned was that the manual process that was used was time intensive and the lack of a standardized form left the aggregated data open to interpretation.

Jessica Webster – National Resources Canada (NRCan)

- Jessica described de-identification techniques that were used during the TANDM project –an initiative sponsored by NRCan, BC Hydro, Fortis, and BC Assessment that looked at identifying a more granular, building-focused analysis for energy and Green House Gas

(GHG) inventories. Firstly, a de-personalization technique was used: the assessment data was screened and sensitive data such as the number of bedrooms in each household was removed. Secondly, an aggregation technique was used: energy and building data was matched by the utility behind a firewall at the parcel scale, and was then aggregated by a building type at the census tract level of geography. Lessons learned included:

- the importance of establishing a privacy threshold of aggregated data using no less than three accounts in one grouping;
- using a standard privacy impact assessment to vet the data;

Supplying the assessment data was at no cost with BC Hydro because it is a crown corporation.

Item 4 Panel Discussion –Experience with and Considerations for De-identification

Dr. Fei Chiang Presentation: Privacy Preserving Data Publishing

- Fei Chiang is an Assistant Professor in the Department of Computing and Software, Faculty of Engineering at McMaster University. Her research interests are broadly in the area of data management, with a focus on data quality, data analytics, data privacy, text mining, and information extraction.
- Fei gave a presentation on techniques that are used to de-identify and publish data while still protecting privacy. Three types of personal information identifiers were described: explicit identifier, which is used to uniquely identify a person or record; quasi identifier, where generally three quasi identifiers can be used together to identify a unique record (rule of thumb); and sensitive attributes, an attribute that contains sensitive/private information. "Attackers" use identifiers to infer sensitive attributes about an individual or record. For example, a name from a voter's list combined with aggregated health data could potentially be used to infer a disease the voter has. Health care patient data, which is highly sensitive, was used to illustrate various techniques. Anonymization, suppression, aggregation/generalization, bucketization, perturbation techniques were described, including K-anonymity and L-Diversity.

Discussion During Fei's Presentation

- No one technique was the most popular, rather techniques are applied based on the trade-off between protecting privacy yet maximizing the usefulness of the data; aggregation is a good compromise.

- Tools can be used to evaluate the usefulness of de-identified data, but evaluating the risk of re-identifying data is subjective.

- Data sensitivity can also be subjective; e.g. wage and salary info might be sensitive to some, but not to others. Electricity usage is sensitive if it can be used to derive activities within a particular home. However, monthly aggregated electricity usage that is disclosed at a point of sale is not considered sensitive. Electricity usage in a home is considered private information; electricity usage in a business is not private but could be considered confidential by the business and commercially sensitive by the public.

- To infer something from de-identified data, one would need tremendous resources that most people, governments, and companies do not have. A more realistic threat to privacy might be the number of different data sets that exist broadly and what can be inferred from them. There are a lot of open government data sets available and they have become easier to download and integrate or match. Companies are now doing the integration and selling the analysis. Re-identification might be difficult now, but in a few years the ability to access info could be very different. The techniques that are being applied to open datasets are difficult to discuss broadly. However, with open data sets, if one uses best practices, standards, recommendations, and guidelines, one is in a good position to protect oneself; similar to other professions, e.g. engineering.

- There was a court case involving sharing of electricity data that indicated criminal activity with the police. The court ruled that this was not a violation of privacy law. Criminal activity gives latitude to privacy, e.g. surveillance; there are times when there is a legitimate need for surveillance, but it must be warranted.

<u>Dr. Khaled El Emam's Presentation</u>

- Dr. Khaled has been working with de-identified health data in the USA and Canada for 12 years and also runs a company that provides solutions to de-identify data. Dr. Khaled's presentation gave a broad spectrum of techniques and focused on what techniques work and that are defensible.

- Statistics Canada has good methods for de-identifying commercially sensitive data.

- All health data is considered sensitive and each US state has unique laws that protect data. Some health data is more sensitive than others – for example mental health, drug use, STI, abortions, genetic, and youth data – and laws can be defined precisely for those data. Based on statistical surveys, all financial and health data is considered sensitive.

- Re-identification attacks are not as easy as some people say there are; it is very hard, data is unreliable and not easy to get a hold of. However, it is not impossible. With any dataset re-identification can be done for a small percentage; typically 1-2% of the dataset can be re-identified.

- Data usefulness can be determined by conducting the same analysis with the identified and de-identified data sets to see if the business or policy decisions that are produced as a result of the analysis are the same.

- An example of an inference from TOU data could include the type of appliances using electricity and also the age, gender, and number of inhabitants. However, if one starts to protect datasets from inferences the analytics capability drops dramatically. Inferences that can be made from data should be subject to review by an ethics committee and contracts can be used to put constraints of types of inferences and models that are allowed.

- There are good de-identification standards that exist today from reputable organizations. Standards exist for health care data of which some are more generic than others. No technique is perfect, but it important to follow best practices and existing standards.

- Location and date information are the biggest contributors to re-identification risks. For example, in Canada the combination of a postal code and date of birth can be used to identify an individual with 99% certainty. Almost all re-identification is done using these quasi-identifiers. Therefore, if you have these in your dataset it is possible to re-identify an individual.

- Measuring the risk of re-identifying data depends on the release context, contractual terms of use, existing controls, the chance of a deliberate attack or data breach. There are well developed and established metrics for determining this.

- There is a lot of precedence on what are acceptable thresholds for aggregating data and different organizations use different group sizes. The threshold of aggregation should be determined by the use of the data.

Dr. Khaled referenced some helpful links which are provided below

https://hitrustalliance.net/de-identification-license-agreement/

http://www.scienceadvice.ca/en/assessments/completed/health-data.aspx

https://www.privacybydesign.ca/content/uploads/2013/05/de-identifcation-developments.pdf

https://www.iom.edu/~/media/Files/Report%20Files/2015/SharingData/ElEmamandMalin%20Paper.pdf

Dr. Ann Cavoukian's Presentation

- Dr. Ann Cavoukian is the former Information and Privacy Commissioner of Ontario and currently serves as Executive Director of the Privacy and Big Data Institute at Ryerson University.
- Ann made the point that one must always de-identify data and take it as given that one can never use identified data. Ideally, a risk-based de-identification framework should be used.
- There is a growing movement amoung certain academics and in the press that any de-identifying technique is impractical because the risk of re-identification is too great. Access to data has become more restricted as a reaction to this. Ann disagrees with this movement because it is based on the premise that poor de-identification techniques were used in the first place. The risks that de-identified data can be re-identified are far lower if proper techniques are used.
- Ann referenced several articles – the article below in particular was noted as warranting being circulated to meeting participants.

  "Fool's Gold: an Illustrated Critique of Differential Privacy" by Jane Bambauer of the University of Arizona. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2326746

Item 5 Framework for Third Party Access

- The discussion considered the nature of data requests with respect to complexity, delivery timing, delivery channel, costs, and service constraints. The group was asked to take a broad view of all options, from which a proposed approach would be determined later. A list of all the main points raised by the group for each subtopic during these discussions is included in the Appendix.

- The discussion of third party access lends itself to the potential MDM/R Data Access Platform (MDAP) and/or other possible platforms beyond the Foundation project. The Foundation project scope does not include the implementation of third party access, but the FWG discussions will inform future work.

- The working assumption put forward for this discussion was that all third parties will be entitled to access de-identified data; however, some members of the FWG challenged whether this should be the working assumption. The definition of a third party is anyone other than the IESO, the LDCs and their respective authorized agents. The discussion was limited to de-identified data only.

Requests / Query Options

- Data requests of MDM/R can be classified along a spectrum: from standard requests to custom queries.

- Pre-determined reports could be used to fulfill standard requests, for example requests from municipalities for community energy planning or to fulfill the IESO's reporting obligations. Pre-determined reports would be limited in size and scope and could possibly be downloadable.

- A line by line full copy of the database would be considered a custom request; and implications for the delivery channel, cost, delivery timing and service constraints would have to be considered.

- Existing international standards, for example the green-button standard, could be used in fulfilling data requests.

Delivery Channels

- The discussion took a broad view and looked at all possible delivery channels; which channels to prioritize will be determined later.

- A public website could facilitate IESO data reports for download and a secure web-portal could enable authorized access and queries. It was noted that if downloadable data is open to the public, data requesters might try to download everything all at once; the channel bandwidth of both the IESO and the data requester would have to be considered. Also, public websites are more prone to random downloads, probes, and attacks. If data requesters must log into a website interface to download data, the IESO could track and control what they are allowed to do; however, any new delivery channels would require up front investment in new infrastructure.

- If query programming or a website GUI is needed, the technical implementation will need to be reviewed by the SME Steering Committee and the MDM/R Technical Panel.

- For large data requests, data could be supplied by using physical media (such as a hard-disk) and that the IESO has used this medium for some data requests. Since there is no additional infrastructure for delivery at the moment, initially physical media will be the only way to facilitate requests.

Constraints

- In an ideal world everyone would receive every request in real time. However, in practice there are constraints on the resources available to fulfill requests. The resources required and/or made available to fulfill requests can vary depending on the complexity and size of request, purpose or intended use of the data, how quickly the request must be fulfilled (eg. directives, regulation, laws, and deadlines), and the nature of the requestor.

- Data requests might come to the LDCs, as opposed to the IESO. Some LDCs might be willing to service these request, but others may not be able to because of the lack of resources or other constraints. It was clarified that the SME has no control over what data is requested of LDCs or what data LDCs can request.

- Centralizing requests would have several benefits including having one entity de-identify data as opposed to many entities using different techniques, dealing with one set of access rules rather than many, less re-identification risk, and having a one stop shop for the data requester.

- There exists a standard SME/LDC Agreement between the SME and each LDC that would have to be re-examined for any necessary modifications should third party access to data become a service. One such area might be the liability of each party associated with affording third parties access to de-identified data.

- A fee of service for a user could be a constraint and that cost for the IESO, which represents service hours, hours of operation, maintenance window, etc., should also be considered a constraint. Every type of request will require IT resources but for this brainstorming session costs should not be considered a limiting factor. The IESO operates as not for profit corporation and would need OEB approval to charge for data requests.

Prioritization

- Data requests might have to be prioritized if there are multiple requests, or if data requests are large and costly to execute. Priority could be determined based on use e.g. a higher

priority might be given to a regulatory requirement versus a general request. If an LDC requests data, they would be high priority as was the case with the TOU study conducted with the former OPA.

- An automated program that handles the data request could remove many issues, but this would require an upfront IT investment.

- The IESO does not want to build infrastructure to implement requests for which there is no demand. In the past the IESO has received custom data requests but has had to turn them down. Hopefully a business case from MDAP or another phase of this project would handle up-front investment, but for now, requests could be serviced manually using existing resources. Over time, after some experience with data requests has been gained, the IESO could determine if a specific request can be turned into a pre-determined report. Options that were discussed would allow the IESO to balance existing resources with the need to access data and will form the framework for what can be delivered now versus what can be implemented later as part of a greater vision.

Item 6 Wrap-up and Next Steps

Next Meeting Scheduled for July 22nd.

Appendix Chart Paper Notes: Framework for Third Party Access Group Discussion

The following chart paper notes were taken during the discussion about Third Party Access to capture the main points that were raised:

<u>Framework</u>

- Format options for disclosure
- Possible delivery channels
- Constraints and prioritization considerations for servicing requests
- Possible approach for fulfilling requests
- Others

<u>Request / Query Options</u>

- "Canned" reports
- Limited customizable requests
- Custom requests
    - Data dump

- Consider Green Button Initiative adoptions and standards

Delivery Channels

- Public website
- Service web portal GUI
    - Download ad-hoc queries
- Web services query
- File transfers (FTP, SETP, etc.)
- Physical media
- Geo-spatial web services
- RSS feeds

Constraints and Prioritization

- Complexity of requests
- Volume of requests
- Availability of resources
    - Personnel, systems, costs
- Size of data requested
- Use of data
- Deadlines, directives, regulations
- Requesting organization or individual
- Service hours
- Terms of agreement for services
- Ability to charge for access
- Inability of LDCs to fulfill requests