

## Privacy in Cloud-Based Data Collection Practices for Commercial Dialogue Systems

Christine Doran<sup>1</sup> and Beth Ann Hockey<sup>1</sup> and Esther Horowitz<sup>1,2</sup>

Intel Corporation<sup>1</sup>  
2200 Mission College Blvd.  
Santa Clara, CA 95054  
christine.doran, beth.ann.hockey@intel.com

Speech Morphing, Inc.<sup>2</sup>  
1245 S. Winchester Blvd, Ste. 216  
San Jose, CA 95128  
esther@speechmorphing.com

### Abstract

There is a natural tension between obtaining the dream data for development purposes and addressing privacy concerns. Data is so crucial to the development of language technology that addressing this tension between data needs and privacy needs is unavoidable. We explore this issue from the perspective of dialogue system development. In this paper we discuss what is needed from data for dialogue system development, what is needed to protect privacy, where the tension arises, and put forth ideas about how to reach workable compromises.

The increased use of language technology, frequently implemented with data hungry algorithms, means that more and more of what users say and write is recorded, analyzed and retained by research, government and commercial organizations. Data is critical for development in language technology; we use it for training statistical algorithms, for guiding design in rule-based components, and for testing. Ideally, for language technology research and development we would:

- Keep all the demographic information tightly associated (Metadata attachment); and
- Have access to donors for additional demographic information, or more data (Donor access)
- Keep it in its original form (Authenticity)
- Keep the data forever (Retention)
- Have unrestricted use, including giving it to whoever we wish to share it with (Distribution)

These features that make data most useful directly challenge the conceptual core of privacy: keeping donors anonymous, informed consent, and restricting the use and distribution of their (potentially sensitive) material.

On a practical level, privacy is a legal requirement in many locations. Nations' individualized privacy controls and laws may affect what kinds of data can be collected from users. Countries' interests in their citizens' privacy are not limited to the import and export controls that govern international data transfer. When gathering data from users once a product has been released, said collection must adhere to the laws of the country in which the user is creating the data. There is a newly-created EU-U.S. agreement

that will require U.S. companies to adhere to stringent privacy protocols when dealing with the personal data of European citizens collected on European soil.<sup>1</sup> This agreement is based on a more cohesive understanding of what "personal information" means to the EU, an understanding which was compromised in October 2015 when the prior framework, known as Safe Harbor (2016.export.gov/safeharbor), was invalidated. Companies can begin the certification process for the Privacy Shield Framework in August 2016 (www.privacyshield.gov).

The interesting discussion is in the details of where compromise lies. How close can we get to the researcher and developer's wish list without significantly reducing privacy? We address this issue in the context of dialogue systems; an area of computational linguistics in which reaching data use nirvana is especially difficult.

### Dialogue Systems

Dialogue systems differ from other language tech systems. Dialogue systems are direct interfaces to the user rather than a behind the scenes process – interacting with users through language creates different expectations. Users are tempted to interact with a system as though it were another person. The ELIZA system (Weizenbaum 1966) was an early example of this phenomenon. Cliff Nass studied the effect and produced a large body of work e.g. (Nass and Brave 2005), (Nass and Reeves 1996) showing that people resort to their usual social and conversational behavior even with systems that are only mildly human-like. Dialogue systems have the potential to encourage more disclosure of sensitive information simply by virtue of talking. Some populations, such as children and lonely or cognitively impaired seniors could be especially vulnerable to this effect, and products such as talking toys and in-home companion systems are targeted directly at them. Although the current commercial offerings are simple and few in number, there have been many research projects in these areas. Increasing commercial interest in conversational toys and electronic home companions, has increased entries into these markets (Nichols 2016), and concern about their privacy (Fowler 2015).

Also, as dialogue systems have become more natural, the

<sup>1</sup>[ec.europa.eu/justice/data-protection/files/factsheets/factsheet\\_eu-us\\_privacy\\_shield\\_en.pdf](http://ec.europa.eu/justice/data-protection/files/factsheets/factsheet_eu-us_privacy_shield_en.pdf)

occurrence of sensitive data has become less predictable. System-initiative made it easy to predict when users were going to divulge sensitive information. When an IVR banking system asked a user for their bank account number, name, or social security number, the point of likely disclosure was clear, as was the nature of the information. Even commercial systems are now frequently user-initiative, e.g. assistant systems that handle search requests, or information retrieval. The user says what they want to say unprompted. In system initiative or mixed initiative systems, sensitive information could surface anywhere, and its type may be difficult to identify.

While data that is useful for building speech recognition systems, training parsers or POS taggers, or doing named entity extraction can be bought, dialogue data close enough to the target domain with necessary context is rarely available at the outset of a project, especially if the domain is novel. For dialogue systems to be able to have better conversations, researchers and developers will need data consisting of longer, more complex conversations with more context and more information about the user. It's often useful to know whether the user is male or female, where they live, where they grew up, educational background, age, areas of expertise. All these factors affect how people will interact with the system, and also add up to a lot of personal information.

Finally, dialogue systems share with other language tech systems the need for data for the development phase as well as data from actual users after deployment. Like many NLP systems, that data is uploaded to the cloud. This happens even if dialogue processing is local. Data is driven to the cloud, particularly on small platforms, by lack of storage and the desire of researchers and developers to improve the system for the particular user, for all users, and potentially to sell the data. Once in the cloud, distribution (or escape) of data is easy.

### Metadata Attachment

Data collection for dialogue system development necessarily involves collecting both demographic and administrative metadata about the donors. In the initial phases of development, some, if not all, data is likely to be collected from known, local users. Demographic information such as gender, age range, native-ness in the language **must** be linked with the data elements to make the data fully useful in development or research. These friendly early donors offer a opportunity that may not be available in later phases of development, when data may be collected from recruited/unknown subjects where it is not possible (logistically or legally) to collect the same amount of demographic information. Why not leave all the metadata tightly linked to the data elements? One of the basic privacy concepts is anonymity, and tight linking makes it too easy to identify people. In addition, it is a legal issue in most jurisdictions, even if the people, such as the friendly donors are willing. This is not restricted to commercial settings. Although the legal definition of "personal information" is country-specific, there is a shorthand formula for industry standards: if the use of a metadata element can, alone or in tandem with

any part of the data collected, expose the participant's identity or allow it to be exposed, it can be considered "personal information."

Once you are in the position of needing to make people non-identifiable, numerous procedures follow. For metadata the usual procedures are disassociation and anonymization which delink the sensitive PII metadata from language data collected. For speech data that is being sent out for transcription, transcribers typically get only recorded data and no metadata with PII. This is an opportunity for data to become more anonymous, because any remaining PII in the data will be less likely by itself to identify the speaker and third party transcribers are unlikely to know your initial local donors. The disassociation process may involve deleting unusable elements (e.g. address or phone number); obscuring them (e.g. replacing birth dates with age ranges); or anonymizing them (e.g. using a random alphanumeric code to identify participants in lieu of their name).

Later in a project when data is acquired "natively," through users under a Terms of Service agreement, the metadata collected and retained is easier to control; users' emails, phone IMEI numbers, or IP addresses can be "personal information" and are readily obtained via front-end user input or an API query without needing to interact with the user. Self-assigned user IDs can be completely anonymous ("User1234") or eponymous ("FirstNameLastName"), and as such they should not be used in place of disassociated alphanumeric identification.

From an ideal privacy perspective, no personal information would be retained alongside the data to minimize risk to the organization; but from a research and development perspective, data is useless without metadata required to troubleshoot, improve upon, or continue development of the device in question. Reducing the amount of data that requires protection reduces in turn amount of risk the commercial institution is undertaking (Steen 2015), (Columbia University). A compromise position, and the one we have seen in our commercial environment is to retain only those metadata elements that are required for analytic purposes, and only in their anonymized or disassociated form.

### Donor Access

Typically there is only one access to a donor. Once data is anonymized, the door to linking multiple collections from the same donor is closed. It would not have to work this way. As dialogue researchers and developers become more interested in applications such as electronic home companions or virtual personal assistants, system performance will be improved by the ability to capture interactions with a user over time, knowing that data is being collected from the same individual. There are models for handling this situation; longitudinal studies in epidemiology track individuals over many years. We could use a similar approach for language data.

### Authenticity

In addition to information in the metadata that could identify the donor, identifying features may reside in the data itself; some types will be more problematic than others. In spo-

ken data an example of this is people's unique voice quality. In our experience many people worry about this, including people in corporate privacy and people on IRBs. However, in practice, it is relatively rare to be able to identify someone in a data set by their voice unless they are very famous, or are part of a data collection from known locals. The known locals are sympathetic and only the original collection team would recognize them by voice. Other inadvertent disclosures are more of a concern.

Users are more likely to reveal PII in open-domain systems. It has been repeatedly shown that individuals can be identified not just from designated PII, but from combinations of **any** pieces of information that suffice to differentiate them. The logs released by AOL in 2006 made news across the popular press when individuals were identified from their search histories (cf. (Barbaro and Zeller 2006) and (National Public Radio February 24 2014)). Narayanan and Shmatikov (2008) list several other such examples and present their own work on the Netflix prize data set, from which they conclude that 99% of records can be uniquely identified based on eight movie ratings and a noisy value for the rating date. In closed domains, it is at least easier to predict when someone will present PII intentionally, and it may also be possible to flag elements that are out-of-domain as potential PII. In domains which necessitate disclosure of PII (healthcare, finance), protections/redaction can be built in.

One approach to the problem of PII embedded in the language data is to have the team that is transcribing the data also do redaction and anonymization. Post-processing and transcription guidelines should take into account the possibility that data may contain sensitive or personal information. The transcribers are already listening to everything, and, given clear guidance, are well-situated to redact either portions of audio or whole recordings. Additionally, they do not require access to the metadata. In a commercial setting, giving transcription teams and other third parties processing data NDAs provides some protection against disclosure of the data and any PII contained in it.

Another approach is to adapt techniques such as differential privacy and k-anonymity which are used to improve confidentiality over large structured data sets, such as medical or census databases, or even better, by linking multiple resources. K-anonymity (Sweeney 2002) relies on having enough records to noise up the data set with enough individuals sharing particular combinations of values for pre-identified 'quasi-identifier' attributes that it's harder to identify a single individual. This relies on what Narayanan and Shmatikov (2008) call "the fallacious distinction between 'identifying' and 'non-identifying' attributes...[which] is increasingly meaningless as the amount and variety of publicly available information about individuals grows exponentially." Differential privacy takes an algorithmic approach, also assuming larger structured data sets, where for each data set, an approximation function is determined, such that query results are still *useful* but are not precisely accurate (Dwork 2011).

An approach worth investigating is to automate the identification of PII and redacting. Techniques used to identify and filter profanity, or to de-identify PII in other domains

could potentially be adapted do a similar job for PII in open domains (cf. (Aberdeen et al. 2010) for one such system). More automation in this area would improve privacy and data processing efficiency.

## Retention

Dialogue data is costly and difficult to collect, so longer or indefinite retention is attractive because it allows for more potential use. The privacy conflict here is not evident until we consider that in most organizations, there is no realistic expectation that the data will have a long-term guardian. The authors have worked in academia, research labs, government and industry and observed this problem in all environments. When the last team member from a project turns out the lights, where does the data go? If there is no guardian, there is no control on whether the data continues to be used under the terms the donor agreed to. Although it feels wasteful, in these cases for privacy purposes short retention times could be a solution. There are exceptions to this pattern; organizations such as ELRA and the LDC have data handling as a core function, and continuity comes from the organization rather than depending on individuals. We would argue that a more data positive approach would be to establish an organizational level multi-option data disposal plan that considers privacy and reuse.

## Distribution

While academic institutions are typically able to release data (assuming appropriate consent, anonymization, etc.), and this is, in fact, a required feature of some government funded projects, releasing data for general distribution can be a problem for commercial organizations. The problems have more to do with competitive advantage than privacy although they can affect the consent procedures and data handling in commercial organizations. Typically a company has invested effort and money on dialogue data in order to have a competitive edge. Releasing data publicly can amount to sharing with competitors. As long as there is an advantage in having the data, public sharing is counterproductive. Data needs to be kept private at least from the start of development to product release, and potentially longer depending on the market space. On the flip side, giving away the data increases the number of keepers and therefore the lifespan of the data. Using consent forms wherever possible that would allow future release to LDC or ELRA would leave that open as an option once the business use has expired. Additionally, data can be shared in limited NDA arrangements with external partners

## Consent

Consent issues apply across the data features and warrant additional discussion. Various types of documents are used for consent including consent forms and Terms of Service.

### Consent Forms

In the development and pre-production stage, any person who contributes data must give their consent to be recorded

or for their data to be harvested. Contributors must be informed of all the ways in which their data may be captured or recorded. For systems with open microphones or with a Low Power Always Listening devices triggered by an activation phrase, consents can include warnings that these systems are either always recording or can be mistakenly triggered to be recording. From a research and development perspective, consent form should be as broad as possible — what development processes *could* the data be used for? — while remaining within the scope of the project and ensuring that the participant has enough knowledge to be giving *informed* consent. The consent form should, if possible, specify how the raw data will be retained, under what type and scope of access controls it will be kept, and how long it will be retained.

## Terms of Service

In the production stage of data collection, the Terms of Service serve as legal notice of the conditions of use — including, but not limited to, the collection of metadata and/or data resulting from the user's operation of the device — and consent is often a requirement of use. However, it is widely accepted that users frequently do not read terms or service or privacy policies. Gomez et al. (2009) present an excellent overview of consumer behavior and privacy policies on web sites, and conclude that people fail to read the terms and even having read them, fail to understand them. Gindin (2009) cites multiple passages from the FTC in which it is taken as given that users will not read privacy policies, even the 'better' kind (shorter, easier to read, with information most important to consumer up front). Whether a company is happy with meeting the minimum legal requirements or sets a higher standard is a matter of corporate philosophy.

## Conclusion

Dialogue data is crucial to dialogue development and research. and from that perspective, the best data arrangement would be to know everything about the speakers in our corpora. From a privacy perspective the best protection would be for the user to keep all their information to themselves. What is the solution?

- Separate PII and data, separate parts of the data during processing. Distribute across different storage and different people so no one place/person has all of the information. Automate the data separation.
- Think carefully about what is genuinely needed for a particular dataset. What demographic data is actually needed? Do multiple instances of a task by the same person have to be identified as the same person or could they be separated? Keep as little metadata as possible.
- Have good broad informed consent. Think about how the data could be used including eventual public release.
- Use epidemiological approaches for safely tracking users over time.
- Have as few people as possible handle the PII. Automate as much of the data processing as can be managed. If tran-

scribers are already listening to everything, have them do the anonymizing.

## References

- Aberdeen, J.; Bayer, S.; Yeniterzi, R.; Wellner, B.; Clark, C.; Hanauer, D.; Malin, B.; and Hirschman, L. 2010. The mitre identification scrubber toolkit: design, training, and assessment. *International journal of medical informatics* 79(12):849–859.
- Barbaro, M., and Zeller, T. 2006. A face is exposed for aol searcher no. 4417749. *The New York Times*. [www.nytimes.com/2006/08/09/technology/09aol.html](http://www.nytimes.com/2006/08/09/technology/09aol.html).
- Columbia University. Privacy and confidentiality. *Current Issues in Research Ethics*. [cnmtl.columbia.edu/projects/cire/pac/foundation/](http://cnmtl.columbia.edu/projects/cire/pac/foundation/).
- Dwork, C. 2011. A firm foundation for private data analysis. *Communications of the ACM* 54(1):86–95.
- Fowler, G. 2015. Talking toys are getting smarter: should we be worried? *Wall Street Journal online* Dec 17.
- Gindin, S. E. 2009. Nobody reads your privacy policy or online contract: Lessons learned and questions raised by the ftc's action against sears. *Nw. J. Tech. & Intell. Prop.* 8:1.
- Gomez, J.; Pinnick, T.; and Soltani, A. 2009. KnowPrivacy. *School of Information*.
- Narayanan, A., and Shmatikov, V. 2008. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, 111–125. IEEE.
- Nass, C., and Brave, S. 2005. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. Cambridge, Massachusetts: MIT Press.
- Nass, C., and Reeves, B. 1996. *The Media Equation: How People Treat Computer, Television and New Media Like Real People and Places*. Cambridge, Massachusetts: Cambridge University Press.
- National Public Radio. February 24, 2014. If you think you're anonymous online, think again. [www.npr.org/sections/alltechconsidered/2014/02/24/282061990/if-you-think-youre-anonymous-online-think-again](http://www.npr.org/sections/alltechconsidered/2014/02/24/282061990/if-you-think-youre-anonymous-online-think-again).
- Nichols, G. 2016. At your service eight personal assistant robots coming home soon. [www.zdnet.com/picture/at-your-service-8-personal-assistant-robots-coming-home-soon](http://www.zdnet.com/picture/at-your-service-8-personal-assistant-robots-coming-home-soon).
- Steen, M. 2015. Ethical uses of collected data. [www.scu.edu/ethics/focus-areas/business-ethics/resources/ethical-uses-of-collected-data/](http://www.scu.edu/ethics/focus-areas/business-ethics/resources/ethical-uses-of-collected-data/).
- Sweeney, L. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(05):557–570.
- Weizenbaum, J. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1):36–45.