OJPHI

ISDS 2013 Conference Abstracts

Data De-Identification Toolkit

Aaron Kite-Powell* and Kelly Moran

MIT Lincoln Laboratory, Lexington, MA, USA

Objective

To develop a robust, flexible, and easy-to-use data de-identification tool that makes it easier for data providers to create data sets that are sharable with external collaborators.

Introduction

Developing effective data-driven algorithms and visualizations for disease surveillance hinges on the ability to provide application developers with realistic data. However, the sensitivity of the data creates a barrier to its distribution. We have created a tool that assists data providers with de-identifying their data in preparation for sharing. The functions in the tool help data providers comply with the HIPAA "Safe Harbor" de-identification standard [1] by removing or obscuring information such as names, geographic locations, and identifying numbers.

Methods

This tool was developed in Java 7 using the Apache Commons, Java Crypto, and Java Security libraries. It allows a user to select a file in a number of formats and apply one or more de-identification filters to its fields. The filters provide a variety of options for de-identifying a given field, such as obscuring values, removing values, encrypting and decrypting, generalizing, offsetting, or filtering values. Specifically, the filters can (1) bin numbers such as ages into groups; (2) offset dates and times; (3) encrypt and decrypt values with a private key ; (4) filter out identifying numbers such as SSNs or phone numbers; (5) obscure values by systematically renaming distinct values; (6) remove values altogether; (7) filter out stopwords, both from off-the-shelf lists or user-provided lists; and (8) generalize zip codes or other numbers by removing the specified number of digits. This tool runs on any machine with the Java Runtime Environment.

Results

A software suite that includes the described algorithms and user interface was created, including a collection of tests that ensure the described behavior for each filter. Tests were conducted with epidemiologists from state and local health departments using data sets extracted from local disease surveillance systems. An initial version has been shared with data providers who have used it operationally including creating data sets shareable with external partners.

Conclusions

Creating tools that make it easier for data providers to share data with other public health jurisdictions or with researchers should facilitate various cross-jurisdictional and cross-disciplinary collaborations and increase innovation in the field. This tool provides a simple method and user interface for data providers to de-identify their data, allowing for easier and safer sharing with external collaborators. The tool provides a variety of options to users, allowing them to finely control their data de-identification process on a field-by-field basis, and even recover scrubbed values using private key encryption.

In the future, we plan on extending this tool to include statistical anonymization methods as specified in the HIPAA "Expert Determination" de-identification standard.

Keywords

Data sharing; de-identification; anonymization

Acknowledgments

This work is sponsored by the Defense Threat Reduction Agency under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, recommendations and conclusions are those of the authors and are not necessarily endorsed by the United States Government

References

U.S. Department of Health & Human Services, "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule," 2010. [Online]. Available: http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#standard. [Accessed 15 August 2013].

*Aaron Kite-Powell
E-mail: Aaron.Kite-Powell@ll.mit.edu