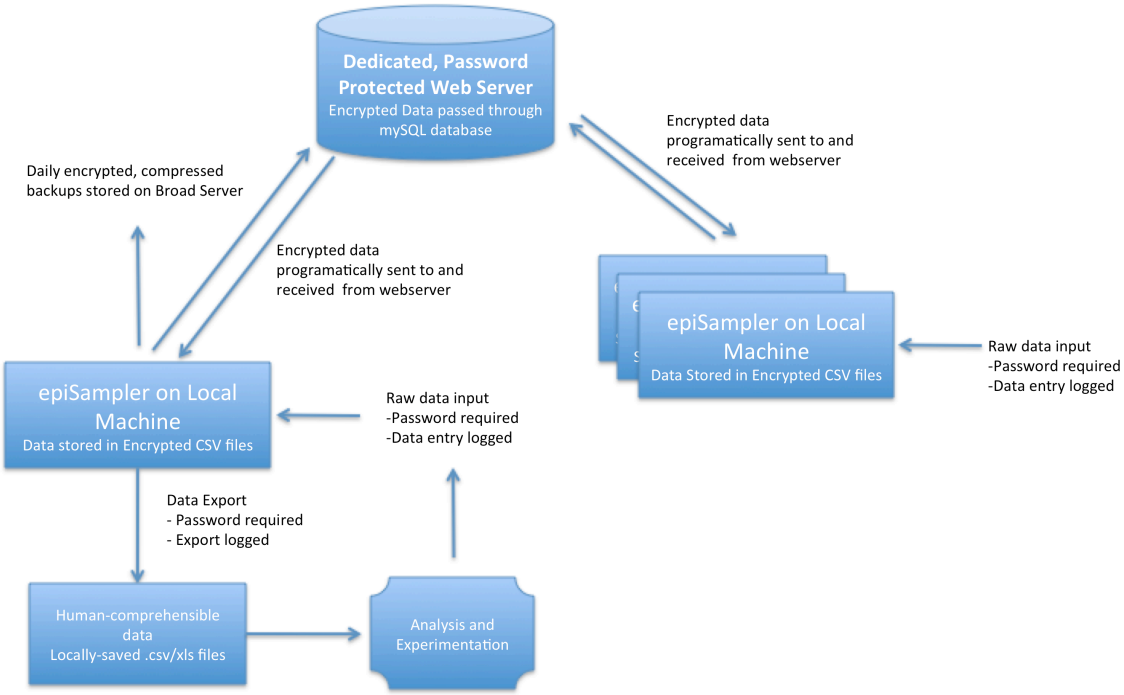
Harvard University
October 2010

Note: This document represents only a snapshot in the evolution of epiSampler with a focus on data privacy. The data storage and transmission techniques evolved beyond the Peer-to-Peer flat file model described herein to take advantage of the reliability and efficiency of JDBC ('Java Database Connectivity').

Data Storage Overview

Introduction

epiSampler was designed for use in resource-poor settings. In particular, epiSampler must function when there is no Internet access, but must also be linked to other users for data sharing whenever a network connection is available. To solve this problem, epiSampler stores copies of the entire dataset on each computer, synching over a networked mySQL database whenever possible.



Local Data Storage

Data is stored locally in text files using comma-separated values ('csv') format. Because these files could be opened and read by any text reader, like notepad, or a spreadsheet program such as excel, the data is stored encrypted. A custom algorithm that takes advantage of the structure of csv data is used for encryption. The data remains encrypted while epiSampler is running and is only rendered sensible upon display to users in epiSampler. It is frequently necessary for users to export data for use in analysis; only registered users may export data, and a log is kept of every export, recording the date/time, user, and variables exported. Similarly, if any data values are changed, a record of the change, and previous values is made.

Network
When epiSampler detects an Internet connection, it requests a list of new data to be downloaded and creates a list of data to upload. It then proceeds to transfer only new data between the server and local installation – if there is an interruption in the data transfer, the incomplete transfer is stored, and restarted when a new connection is established. The data remains encrypted during transfer and in the mySQL database on the server. After all known installations of epiSampler have interacted with a data point, that data is removed from the server. Thus, the server acts only as a pass-through between the local installations. When a user is working in epiSampler, he or she is interacting only with the locally stored data.

Backups
Daily backups are made of the mySQL database, and will be pruned such that only one week of backups exist. As with all other data, the backed up data is also encrypted. Users are able to create backup files from their local file systems. The entire local database, which consists of over 50 files, is reduced to a single file, which maintains encryption, and is compressed using the java deflate algorithm (a wrapper around the commonly-used zlib library). The single file can then be emailed between epiSampler users and the developers for diagnosis and data integrity checks. The .esf ('epiSampler file') must be unpacked by epiSampler, only after which the files can be rendered readable by the epiSampler 'dataTable viewer'.

Data Samples
Normal Data:

ID	Value 1	Variable 2	Datapoint 3	Var 4
1	This	is data	encrypted	by epiSense
2	This is	data	encrypted by	epiSense
12	As you	can see,	it is	hard to read
212	As	you can	see, it	is hard to read
2121	Fish	Fishery	First	Fi

Encrypted Data:

VK	tlruPuy	6aeo3Nre8z	slba6XDkb 4	6aeuT
y	0Vim	D4 da23	eZE6BYtPr	RB e6o7:HsP
Z	9uvs8oz	oQta	eZE6BYtPr1Nm	:YiFso4h
5A	AmuXl5	OQn8gh:,	vt8oz	haT8Dbo8Thlc
22z	Amu	Xl5 nvo	z:h,8o0	imuKlkd82pDkear
6M22	pvsU	lvLhPTX	lieg0	pv

Backed-up Data* :
xœì½{sªJ³?þVÌ:~k?çWO\Ü/Öª•B#4#/^×,"QŠŠB#ÿ#âj#;û9Uç½WÒa>ÓÝÓÓÓ3Ó3ókºÐlð#EÑÄq\}]˜¬
½_?ß¨?Ö†Õÿñï#^øßîÇ¯Ÿÿ##ëÿûýÃ#%goÿ;#ÉRHÀt;EFEîDþþ^>¢#¥#•{#¤¨Ñg€HSÑ0EÓT,$ 8&ã###Ó
ä##U„4#‰Ø#ƒ4YÈ¸œ"#<'#§ÎúH~KÄ#å9SObCÔ"Ø#õ$6D=‹LH"c]#vš#«BŸ¦ÉH¬!1MŽ#NÓLsØ›-
¤Ùæ°¸I#ú#=V1#¤èó"Jè„åÌO*¨'•@Ô"J êY%#ù¬#^|VD>«#"ŸU'$—#*ég•Àô³J'ôÅI%¡#ÎÔ"J
êI%#õ¤#^zVD>«#"ŸU#'Ï*Èg•$Éå„J`úY%0ý¬'$ýé¤#!áwžN*¨'•@Ô"J êY%#ù¬#^|

* Illustrative only – represents portion of .esf file, which includes data and file structure information. If the file were intercepted, this is how it would appear in a text reader.