DIME Analytics

Guide to sharing data securely

Overview

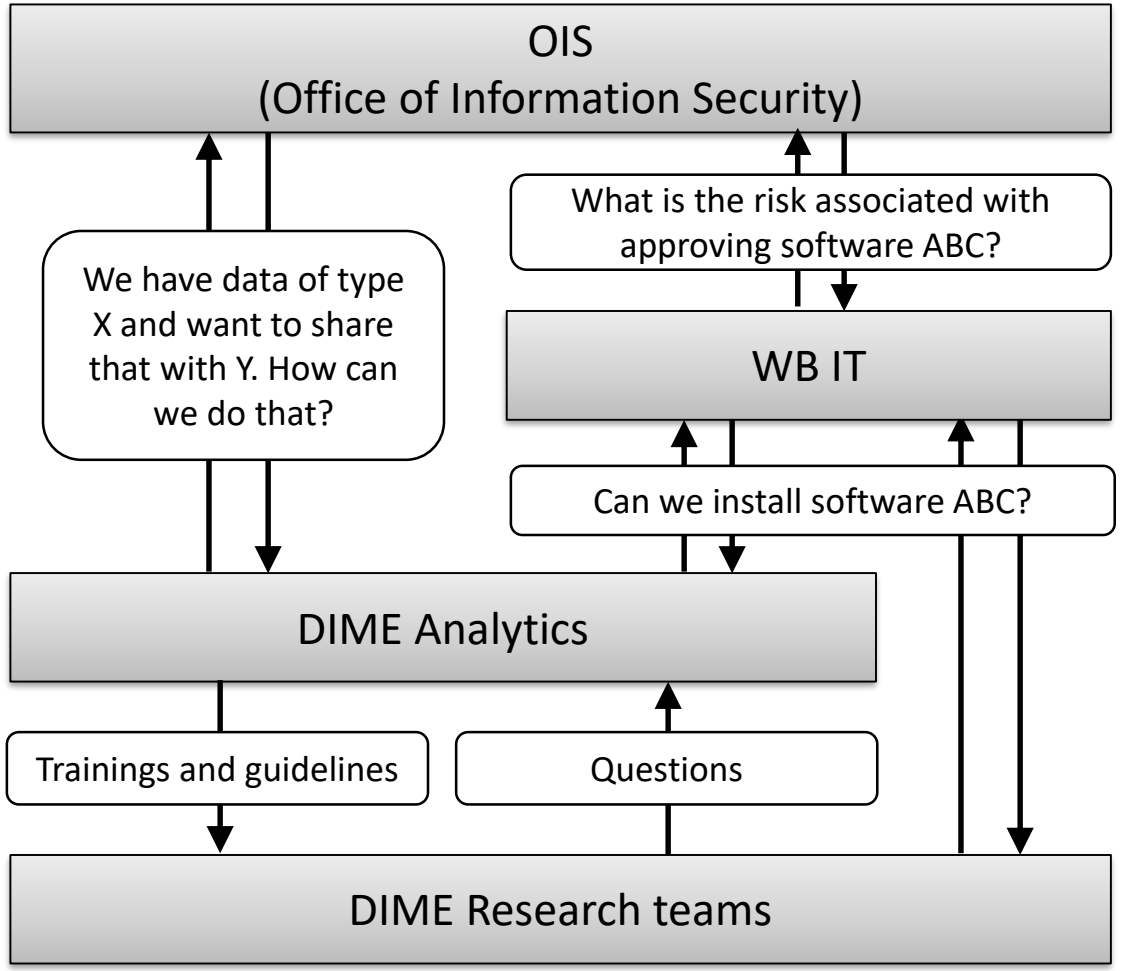
Prepared by **DIME Analytics**
dimeanalytics@worldbank.org

Presented by **Kristoffer Bjarkefur**
kbjarkefur@worldbank.org

3 main take-aways

- This guide is meant as a toolkit where you use different tools for different use cases
- De-identification is the best tool to reduce the extra work that proper data security unfortunately require
- An encrypted work flow is only as secure as its weakest link
- In the digital world, nothing is 100% secure, but we should mitigate the risk as much as possible

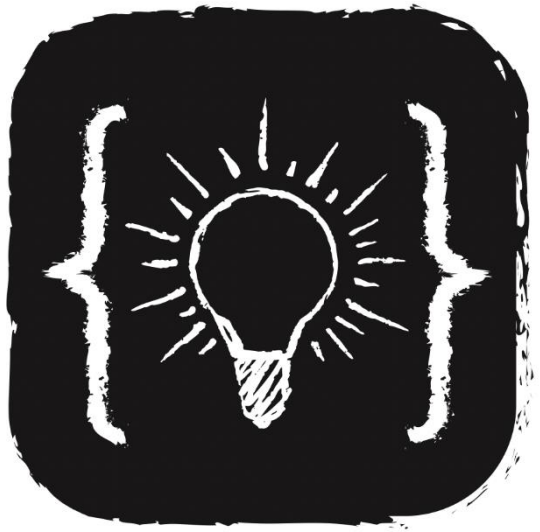
Who decides what we can do with data at the WB?



- OIS classifies personally identifying data as **Strictly confidential** which is the strictest classification
- Approvals by OIS related to *Strictly confidential* data **takes months or years**
- In the meantime, OIS's recommendation is to **simply not use identifying data and especially not share it with external collaborators**

DIME Analytics guide for *Strictly confidential* data

DIME Analytics Data Security Guide



- This is a non-public **pragmatic guide to navigating gray-zones** where OIS has yet to decide what solutions are approved for *strictly confidential* data
- All software recommended in this guide are approved for installation on WB devices
- To reduce the number of software tools teams need to learn, this guide uses the same software as much as possible across multiple solutions

Use cases with recommended solutions

Use case	Example	Sharing solution
DE-IDENTIFIED DATA		
1. Sync-sharing laptop sized (<10 GB) **de-identified** data outside WB	After data is de-identified it can be shared in any standard syncing service without encryption	Sync software (e.g. DropBox) without adding layers of encryption
IDENTIFIED DATA		
2. Sync-sharing laptop sized (<10 GB) **identified** data within WB	Sharing data and code for a research project with only WB collaborators, where data is automatically synced to the computer	WB OneDrive
3. Receiving smaller sized (<500 MB) **identified** data one-way from external users to WB	A government counterpart sends us a data set of identified admin data which we will use in analysis and then share results	WB OneDrive Browser
4. Sending single files with **identified** data (<10 MB) to external users	Sending a roster of respondents with names and ID to a survey firm	VeraCrypt + e-mail
5. Sync-sharing laptop sized (<10 GB) **identified** data outside WB	Sharing data and code for a research project with WB and external collaborators, where data is automatically synced to the computer	VeraCrypt + Sync software (e.g. DropBox)
6. Receive server sized (<1 PB) **identified** data from outside WB	An external organization or government counterpart shares very large data sets or data bases that are aggregated (or similar) on the server and then downloaded and shared using a different method	Secure file transfer (sftp) + cloud storage + cloud processing

Solutions for:
De-identified data

De-identification – the best way to simplify data security

- Data security involves encryption and encryption is inconvenient to work with
- Avoid encryption as much as possible by working with de-identified copies of the data on a day-to-day basis
- Less secure solutions are acceptable when sharing de-identified data
- Identified data sets should be shared with high level of security on a need-basis, while de-identified copies of the same data sets can be shared more conveniently with the full team
- De-identification is not anonymization. It is next to impossible to eliminate all risk that the data can be re-identified using cross-referencing and other tricks

De-identified data : Sync software (e.g. DropBox) without extra encryption

Use case 1: Sync-sharing laptop sized (<10 GB) de-identified data outside WB

- Data sharing **methods already used by most research teams**, like DropBox, can still be used to share de-identified data
- **De-identification is difficult** and data sets that are claimed to be *de-identified* are often still partially or indirectly identified
- De-identification is a scale and not binary, but the more you de-identify the more you reduce the risk
- While perfect de-identification is difficult, standard procedures like removing names, addresses, or reducing precision in age, geo-coordinates etc. mitigate a lot of risk

Pros:

- Based on methods most research teams already use

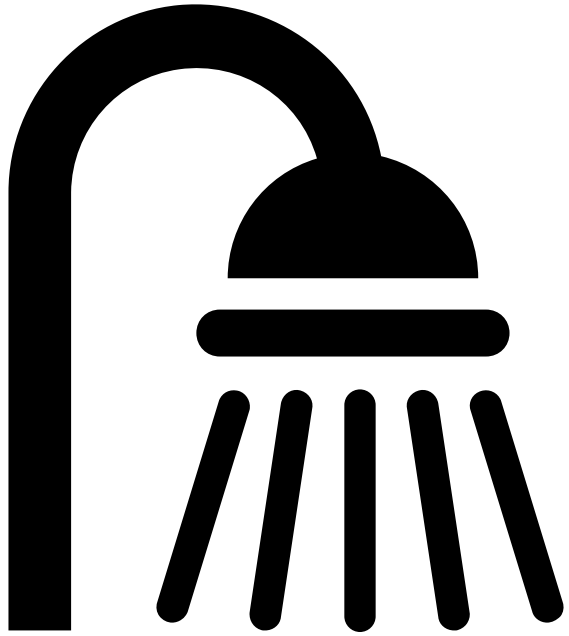
Cons:

- De-identification is difficult
- It is still the least secure of our recommended solutions, but it is acceptable

Solutions for:
Identified data

All de-identified data must be encrypted

- Proper encryption is a system not a single thing you switch on or off
- Even in in-secure connections, the data is likely to be encrypted at some point
- Think of it as plumbing. It does not matter if you have state-of-the-art pipes most of the way if one connection leaks
- An encrypted work flow is only as secure as its weakest link



Identified data : WB OneDrive

Use case 2: Sync-sharing laptop sized (<10 GB) identified data within WB

- All **syncing software are insecure**, including OneDrive, unless extra layers of encryption are added
- **OneDrive on WB devices has these extra layers of encryption** running seamlessly in the background. While it makes WB OneDrive very secure, it is also the reason why WB OneDrive cannot be synced to non-WB devices
- This is the best solution in terms of security and convenience when data is shared only among WB devices, however, this is often not the case
- WB OneDrive is installed on WB servers, so **your data is never stored on servers owned by private corporations**, which usually means that you give ownership of the files to those corporations

Pros:

- Secure and very convenient way to share files
- Syncs files automatically to computers

Cons:

- Syncing can only be done to World Bank devices

Identified data : WB OneDrive Browser

Use case 3: Receiving smaller sized (<500 MB) identified data one-way from external users to WB

- A secure and convenient way to **receive single files from any external user**
- Create a folder in WB OneDrive and share that folder with anyone who can then use a web browser to upload files to that folder
- You can **share a link to that folder with any email address**, and a code will be sent to that address each time the folder is accessed
- Best practice is to use this folder only as a way to receive files by moving the files to another storage location once they are received
- Technically you can also send data this way, but since it provides no safe storage place for the external user after downloading the file, **it is not a good solution for sending sensitive data**

Pros:

- Secure way to receive files
- Very simple setup for ad hoc data receiving
- Anyone with a browser can send data

Cons:

- It is only a good solution for receiving data
- Requires a manual step on both ends

Identified data : VeraCrypt + e-mail

Use case 4: Sending single files with identified data (<10 MB) to external users

- **E-mail is never a secure way to share un-encrypted files**, but e-mails is a convenient way to share information and single files
- We can **use VeraCrypt to encrypt the file with a password** and then securely send the encrypted file as an email attachment
- Even if the file is intercepted or sent to the wrong address, no one can read the file without the password
- **The password must be shared using a password manager** or another equally secure way. Sending the password in e-mail, WhatsApp etc. is like locking a safe and storing the key right next to it
- Works with single files or folders with sub-folders and files as long as it is small enough to attach to an email

Pros:

- Secure way to send files
- Files can be sent over e-mail

Cons:

- Requires that both the sender and the receiver use VeraCrypt
- Requires that both the sender and the receiver use password managers
- Requires a manual step on both ends

Identified data : VeraCrypt + Sync software (e.g. DropBox)

Use case 5: Sync-sharing laptop sized (<10 GB) identified data outside WB

- Syncing services, like DropBox, are not secure without extra layers of encryption
- We **use VeraCrypt to encrypt files before adding them to a sync-shared folder**
- After the file is decrypted **you can open it or access it from a do-file or an R-script**. If you make modification and save the file again, then an encrypted new version of the file will overwrite the old encrypted file
- Have **one encrypted folder for all identified data** in the synced folder that is shared with everyone but share only the decryption password to those who needs it. **Share de-identified copies of the data** in a non-encrypted location in the same synced folder
- **When the analysis requires identifying information**, then there might be no way to create a de-identified data set for the analysis and **then all data needs to be encrypted**, and encryption/decryption will have to be done on a daily basis
- You should only share your passwords using password managers. If the password is lost there is absolutely no way to decrypt your files

Pros:
- Secure way to sync-share files to non-WB computers
- Files are synced to computer

Cons:
- Requires that both the sender and the receiver use VeraCrypt
- Requires that both the sender and the receiver use password managers
- Requires a manual step on both ends

Identified data : Secure file transfer (sftp) + cloud storage and processing

Use case 6: Receive server sized (<1 PB) identified data from outside WB

- While easy to use, set up and maintenance of this solution **needs support from IT officers**
- Scripts that analyzes the data, or prepare the data for download (aggregation to make it smaller, de-identification etc.) can run in the cloud resource
- The **software used are open source or free, but cloud resources are not**. However, as long as the amount of data is not too large (<100 GB) or the processing of it is not too heavy it will not be a significant cost for a typical project
- Other than cost there is **no limitation** for **storage space** or for **processing power** needed
- Data is encrypted and sent from a *sftp-client* which looks like a file explorer window where you drag and drop files

Pros:
- Unlimited capacity

Cons:
- Advanced setup
- Cloud resources are expensive when data is extremely large
- Full access to cloud resources might require UPI authentication

3 main take-aways

- This guide is meant as a toolkit where you use different tools for different use cases
- De-identification is the best tool to reduce the extra work that proper data security unfortunately require
- An encrypted work flow is only as secure as its weakest link
- In the digital world, nothing is 100% secure, but we should mitigate the risk as much as possible

Thank you!