An Approach to Tiered Access in the Department of Veterans Affairs

Michael Schwaber

September 21, 2020

Note: The information in this presentation pertains to the Office of Data Governance and Analytics within the Department of Veterans Affairs Office of Enterprise Integration. It does not to apply to the Department of Veterans Affairs as a whole.

Overview

- Office of Data Governance and Analytics (DGA)
- USVETS Database
- Tiered Access
- The Five Safes
- DGA’s Vision of Tiered Access for USVETS
- DGA’s Future Plans Regarding Data Access

Office of Data Goverance and Analytics (DGA)

- DGA is an office within the VA Office of Enterprise Integration

- Is VA's authoritative clearinghouse for collection, analysis, and dissemination of information about Veterans and VA programs

DGA Products

- Veteran Population Projection Model (VetPop)
- United States Veterans Eligibility Trends and Statistics (USVETS) database
- Reports about veterans
  - Income, poverty, health education, employment, utilization of VA benefits and services
- Maps
  - Veteran population distribution by geography
- VA geographic distribution of expenditure tables

USVETS Database

United States Veterans Eligibility Trends and Statistics (USVETS) database

- Is one of the motivating factors for DGA’s interest in tiered access
- Contains record level data for approximately 39.3 million living and deceased veterans
- Integrates data from 35 VA, non-VA federal, and commercial data sources

USVETS Database

United States Veterans Eligibility Trends and Statistics (USVETS) database

- Contains record level data for approximately 39.3 million living and deceased veterans
- Integrates data from 35 VA, non-VA federal, and commercial data sources
- Stored in SAS datasets

USVETS Database

United States Veterans Eligibility Trends and Statistics (USVETS) database

- Contains data regarding veteran identity and veteran demographics

- Demographics data includes information on veteran age, race, gender, socioeconomics, geography, military history, and utilization of VA benefits and services

- Can be used to perform statistical analyses on the veteran population, and the results of these analyses which can lead to more informed decision-making regarding policies that affect veterans

USVETS Database History

- Created by DGA in the year 2009
- Initially had few data fields, and now has 58 data fields available to organizations external to VA
- Since 2009, the number of data sources has increased from 6 to 35

Current Usage of USVETS Database

- USVETS database is one of several data sources that are used as inputs to the Veteran Population Projection Model
- Used a data source for several reports on veterans including the Women's Veterans Report and the Minority Veterans Report
- Is shared with the Veterans Health Administration (VHA) and the Veterans Benefits Administration
- Used as data source for some data requests from offices within VA and organizations external to VA

USVETS Database

Desire to Expand Usage of USVETS Database

- Initially in the year 2009, there were data quality issues with the USVETS database
- Over time, the quality of the data improved
- At this time, DGA has a high level of confidence in the data
- DGA now wants to increase the availability of the data to more offices within VA and organizations external to VA, while at the same time protecting the data

Desire to Expand Usage of USVETS Database

- DGA thinks the data might be useful in some areas of medical research, operational analyses, and programmatic analyses, including homelessness and suicide prevention
- Desire to expand usage of USVETS data is reinforced by the recent Foundations for Evidence-Based Policymaking Act and the recent Geospatial Data Act
- VA does not have a principal statistical agency or unit, and USVETS database is not covered by the Confidential Information Protection and Statistical Efficiency Act (CIPSEA). As a result, the USVETS database does not have the sharing restrictions that fall under CIPSEA. However, in DGA's USVETS data sharing agreement, it is specified that USVETS data is to be used for statistical purposes only.

What Is Tiered Access?

- According to the Commission on Evidence-Based Policymaking’s report of 2017 entitled “The Promise of Evidence-Based Policymaking,”
  - “tiered access is an application of data minimization”
    - “data minimization means giving access to the least amount of data needed to complete an approved project”
  - “tiered access structures set data access and security requirements based on an assessment of dataset sensitivity”
  - “A well-designed and properly implemented data minimization strategy like tiered access can reduce the risk of unauthorized use and unintended harm to individuals.”

Tiered Access

Model of Sensitivity Levels for Federal data from CEP report

Level	Sensitivity	Description
5	Crimson	**Maximally restricted.** Highly sensitive. Identifiable records from data collected with a promise of confidentiality.
4	Red	**Restricted.** Sensitive. Identifiable records from data collected with a promise of confidentiality.
3	Yellow	**Restricted.** Crimson or Red datasets modified by technologies that mask individual records (e.g. data query tools, differential privacy).
2	Green	**Minimally restricted.** Not sensitive. Data files made available to the public but subject to procedures designed to raise accountability by users, such as registration before accessing.
1	Blue	**Public data.** Most safe. Open data.

Based on the Harvard Model.

CEP final report “The Promise of Evidence Based Policy Making” www.cep.gov/report/cep-final-report.pdf, p. 40.

The Five Safes

What is the "Five Safes?"

- The Five Safes is a framework that can be used to assess and manage disclosure risk
- Safe People, Safe Projects, Safe Settings, Safe Data, and Safe Outputs
- According to the Australian Bureau of Statistics, the Five Safes framework is used by
  - Australian Bureau of Statistics
  - Office of National Statistics (UK)
  - Statistics New Zealand

Source: Australian Bureau of Statistics https://www.abs.gov.au/

The Five Safes

Elements of the Five Safes Framework

- **Safe people:** Is the researcher appropriately authorized to access and use the data?
- **Safe projects:** Is the data to be used for an appropriate purpose?
- **Safe settings:** Does the access environment prevent unauthorized use?
- **Safe data:** Has appropriate and sufficient protection been applied to the data?
- **Safe outputs:** Are the statistical results non-disclosive?

Source: Australian Bureau of Statistics https://www.abs.gov.au/

DGA’s Vision of Tiered Access for USVETS

How to approach the dimensions for tiered access?

- **Safe projects**
  - Data use considerations: laws, ethics, consent, for profit motive
- **Safe people**
  - Organizational Affiliation: VA employees, employees of other federal agencies, academics and non-profit, private industry with a for-profit motive, the public
  - Expertise: statistics, analysis, or medical, social or behavioral science
  - Training on data stewardship
- **Safe settings**
  - IT systems and physical environment: VA servers, secure enclaves, remote access with credentials; access and audit controls; secure file transfer and communication protocols
- **Safe data**
  - Spectrum of sensitivity: data elements such as PII, income
  - “Need to know”: which data elements, how much data, how many datasets linked/integrated
- **Safe outputs**
  - Statistical disclosure control: on data or output, cell suppression, data swapping, differential privacy

Which combinations of criteria have acceptable risk?

Model of Sensitivity Levels (preliminary)

Red : Any file or database that includes direct Personally Identifiable Information (PII) or Protected Health Information (PHI), especially

- SSN (or part), Name (any part)
- Date of Birth, Date of Death, Address – in conjunction with SSN or Name

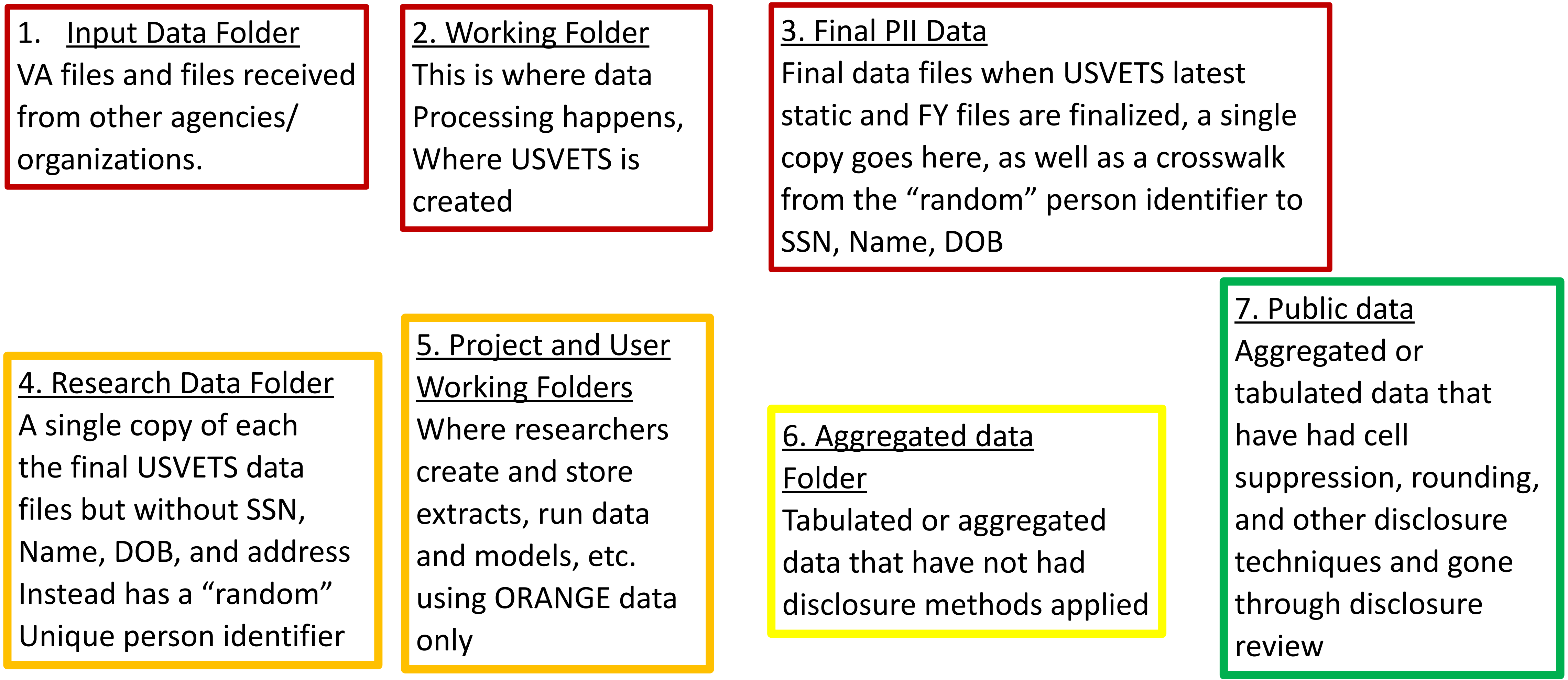
Orange : Any microdata file or database that does not include SSN or Name, and instead has a unique person identifier that does not carry information in the value, i.e. the values are random, NOT created from a sort or index using personal or household information or from a canned hash

Yellow : Tabulated or aggregated data that have not had disclosure techniques applied, e.g. cell suppression, rounding

Green : Public data that have been disclosure-reviewed and approved for dissemination outside of the DGA or VA firewall, maybe also sensitivity reviewed

DGA's Vision of Tiered Access for USVETS

Concept of Server/Folder/File Separation (each box represents a separate server or folder)



1. Input Data Folder
VA files and files received from other agencies/ organizations.

2. Working Folder
This is where data Processing happens, Where USVETS is created

3. Final PII Data
Final data files when USVETS latest static and FY files are finalized, a single copy goes here, as well as a crosswalk from the "random" person identifier to SSN, Name, DOB

4. Research Data Folder
A single copy of each the final USVETS data files but without SSN, Name, DOB, and address Instead has a "random" Unique person identifier

5. Project and User Working Folders
Where researchers create and store extracts, run data and models, etc. using ORANGE data only

6. Aggregated data Folder
Tabulated or aggregated data that have not had disclosure methods applied

7. Public data
Aggregated or tabulated data that have had cell suppression, rounding, and other disclosure techniques and gone through disclosure review

DGA’s Vision of Tiered Access for USVETS

Access Control: Red, Orange

Red

- Only a limited number of people have access
  - People who work on the data build for the USVETS (employees, contractors)
  - Researchers who need to access a copy of the finalized data and crosswalks in order to match with their own data
    - The researchers should have time-limited access to Red folders
    - DGA should have an MOU in place with the researchers’ agency or unit

Orange

- Researchers who only need access to USVETS, do not need to link with other data, AND the cubes don’t have the information they need for analysis

DGA’s Vision of Tiered Access for USVETS

Access Control: Yellow, Green

Yellow

- Researchers who can use the content in the Cubes for their analysis

Green

- Anyone with a BISL account
- Could make copies of this available on an FTP site

DGA’s Vision of Tiered Access for USVETS

Additional Access Criteria

EACH user of Red, Orange, or Yellow data must **sign**

- A **document** such as Rules of Behavior or Acceptable Use Policy
- Attest to applying DGA’s **Disclosure Rules**/Cell Suppression/Rounding standard operating procedure
- Certify **destruction** of data extracts at end of projects

DGA maintains **Access Control Lists** (ACLs)

- Conduct regular **audits of ACLs** – do users still need access, activities
- When a user is added to an ACL, **automate reminders** of when to remove access
- Conduct regular reminders to users to **review and delete files** that are no longer needed – this is a privacy issue as much as a storage issue

DGA’s Future Plans Regarding Data Access

In an evolving data access environment in which updates to data policy and new laws encourage more data access across agencies and to the public:

- DGA intends to expand access to USVETS data within VA and to organizations external to VA

- DGA will seek to refine and enhance its use of tiered access and other data security measures and will continually seek out new ways to protect its data

Contact Information

Michael Schwaber
mike.schwaber@va.gov

Office of Data Governance and Analytics webpage for data products and reports:
www.va.gov/vetdata
(National Center for Veterans Analysis and Statistics webpage)