Federal Register Notice 87 FR 35250, https://www.federalregister.gov/documents/2022/06/09/2022-12432/request-for-information-on-advancing-privacy-enhancing-technologies, June 9, 2022

# Request for Information (RFI) on Advancing Privacy Enhancing Technologies

## Richardson, Douglas

**DISCLAIMER:** Please note that the RFI public responses received and posted do not represent the views or opinions of the U.S. Government. We bear no responsibility for the accuracy, legality, or content of the responses and external links included in this document.

Response to the Office of Science and Technology Policy (OSTP) for
Information on Advancing Privacy-Enhancing Technologies

Submitted by
**Douglas Richardson, PhD**
**Center for Geographic Analysis**
**Harvard University**
July 8, 2022

# Innovative and Interactive Geospatial Virtual Data Enclave (GVDE) Technologies:

*Robust and reliable privacy-enhancing infrastructure for*
*accessing, sharing, and analyzing confidential geospatial data and locational privacy*

## INTRODUCTION

The generation, analysis, and protection of geospatial data is now at the frontier of many governmental, scientific, and private sector domains. Several contemporary trends are driving new privacy concerns: massive quantities of data streaming, data warehouses from global positioning system (GPS)-enabled devices and sensors, location-aware technologies, advances in web services and cyberinfrastructure, and new geoprocessing tools for analyzing, exploring, and visualizing large and multi-scale spatiotemporal data sets (Richardson, 2013).

These trends increase opportunities for exciting new use and integration of data sets to create multi-disciplinary and data-intensive collaborations. However, the unique confidentiality characteristics of geospatial and locational data present special challenges to such collaborations by governmental agencies and the public. Individuals are often identifiable when geospatial and locational data is presented in maps and other visualizations, or when combined with sensor data or other related geospatial data sets.

Our interactive GVDE technology identifies, integrates, and builds on **four interrelated components** required to create and implement a robust and reliable GVDE system for widespread use by governments, the public, and scientists conducting projects involving confidential geospatial data:

> **I. Develop the Interactive Geospatial Virtual Data Enclave and its Core Functions.** Our GVDE technologies, funded by the National Science Foundation (NSF), address challenges for working with geospatial data and to improve the user experience with regards to geospatial data. This component of the technologies evaluates and integrates software tools and procedures (e.g., data management, GIS, analytics, modeling, spatial statistics, etc.) to enable governmental agencies using confidential geospatial data to a) ***share,*** b) ***access,*** c) ***analyze,*** d) ***replicate***, and (e) ***build on*** projects **within** the GVDE.
>
> **II. Evaluate and Implement Masking and Encryption Capabilities for the GVDE.** We have evaluated and implemented multiple geomasking methods, encryption, and other processes to enable government agencies to anonymize and then ***export*** *maps, analyses, and visualizations* derived from their analyses of confidential geospatial data from the GVDE (after review) for use in public dissemination via public reports, presentations, or publications. This component of the technologies has researched and tested numerous anonymization methods and related disclosure risks for specific types of geospatial data (point, line, polygon, raster, vector, etc.) including rapidly growing new sources of confidential geospatial data such as GPS trajectories, crowdsourced, and social media data used by government agencies. We are also extensively engaged in ongoing research on **Differential Privacy** methods. Our research team is collaborating closely with the Harvard ***OpenDP Project***, which includes the Harvard Institute for Quantitative Social Science (IQSS), the Harvard Center for Geographic Analysis, and Microsoft.

**III. Develop the GVDE User Credentialing System.** Our research team has developed and implemented an innovative, robust, and reliable system to provide trained and with a durable digital identifier. This digital user passport creates an efficient mechanism for large numbers of credentialed government personnel to safely use the GVDE. This innovative system builds on research evaluating multiple access mechanisms at restricted data facilities around the world and improves on a decade-old on-line application systems for restricted data. This GVDE passport serves as a transferable and durable credential to allow governmental personnel to access restricted data at multiple agencies.

**IV. Sustainability.** To ensure the long-term sustainability of the GVDE technologies and its widespread usage by government agencies, researchers, and others, the GVDE is maintained as part of a portfolio of ongoing data management and stewardship services. To support the usage of the GVDE technologies for the broader governmental communities, we provide training, outreach and dissemination activities on the use of the GVDE technologies, data confidentiality ethics, credentialing requirements, and on policies and best practices.

Our GVDE team involves leading scientists with deep experience with privacy-enhancing technologies and confidential data protection, and is uniquely positioned to apply its expertise in data management and disclosure risk assessment to the development of the GVDE technologies for use by the broad range of users, including governmental agencies. The creation and use of geospatial data is becoming pervasive in many federal (and state) governmental agencies, but the unique privacy challenges of confidential geospatial data access, sharing, analysis, and safe dissemination are not yet fully understood by the public. To address these challenges, our scalable, robust, and reliable Geospatial Virtual Data Enclave (GVDE) privacy-enhancing technologies can easily be adapted for use by special needs of various governmental agencies, as needed.

Our GVDE technologies provide workable and sustainable solutions to key geospatial data confidentiality issues, in government agencies and the broader society. Projects and policies using geospatial data is transforming many governmental agencies. The ability to share and analyze these data securely advances public affairs broadly. The development of a new transferable user credentialing system also benefits governmental agencies, and significantly, at less well-resourced institutions, as their access to valuable data is no longer be limited by their agencies' ability to offer these services. Resolving practical issues related to access to confidential geospatial data benefits many other governmental agencies where progress is impeded due to limited data sharing. The creation of the GVDE technologies has created important new data infrastructure for governmental agencies to share data and to commit to data management plans that enable sharing confidential geospatial data.

## RATIONALE

In order to leverage investments in geospatial data creation and analysis and to share data, governmental agencies put data into a trusted analytic repository where data can be safely accessed and analyzed by other agencies and the broader society. Our Geospatial Virtual Data Enclave (GVDE) technologies enable governmental agencies to build on prior data collected, safely, securely, and at low cost. Below we provide the rationale and need for our technologies address key issues and challenges of (a) data sharing; (b) creation of a robust and reliable GVDE technologies for widespread use by the governmental geospatial community; (c) implementation of an innovative credentialing system for accessing, sharing, and analyzing confidential geospatial data; and (d) disclosure risks and methods for protecting confidential data and geoprivacy.

### *Challenges to confidential geospatial virtual data sharing and analysis*

In 2013, the White House issued an executive order establishing the Open Data Policy, which addressed data sharing practices. In response to the Open Data Policy and using NSF practices as a model, many federal agencies are now requiring data management plans (Adler, 2015). Confidential data was more frequent and was associated with a wide range of stakeholders, including governmental agencies, researchers, and private sectors (Bishoff and Johnston, 2015). Data sharing is especially challenging for governmental agencies using geospatial data because of the risk of revealing both subject identities and precise locations when data is visualized as a map or linked to other datasets.

### *The opportunity to build on successful GVDE technologies*

Our GVDE research teams developed an experimental, functional prototype GVDE that demonstrated proof of concept and was successfully tested for a small number of users. From 2014 to 2018 the experimental GVDE allowed authorized testers to access restricted and confidential geospatial data in a secure environment. In the prototype GVDE technologies, users in multiple locations connected to a virtual desktop to view and analyze data secured on a server. Within the virtual desktop users were able to access confidential data files as well as a wide range of statistical and other analysis software, including geospatial software such as ArcGIS and GeoDa. While they could view and analyze these data, they could not export them from the virtual desktop or download them to a local computer. Users were able to share a workspace, fostering collaboration; however individual projects were isolated from one another to maintain security. The prototype GVDE technologies design was able to replicate the performance and computing power of a standard desktop computer, so that users could perform the same geospatial analyses within the GVDE that they would have been able to do on their local personal computer, with minimal performance degradation.

### *A robust and reliable credentialing system for the GVDE*

An integral component to the GVDE technologies is a credentialing system that establishes transferable digital identities for trained and trusted users to expedite their access to restricted geospatial data. The creation of a trusted user passport builds on the GVDE research team's prior work examining the standards and processes for accessing restricted data at repositories around the world. In a recent white paper, the GVDE research team proposed the establishment of a user passport (Levenstein, Tyler and Bleckman, 2018). By reducing the time and paperwork necessary to access restricted data, the passport helps to overcome the justifiable concerns of governmental agencies and the broader community about working in a restricted environment. While there is a tradeoff between ease of access and confidentiality protection, this project moves the frontier of that tradeoff outward to provide *both* greater access and greater protection so that governmental agencies can achieve desired standards of analysis.

In the current environment, restricted data are often available to users only after a lengthy and complicated application process. This process usually requires the interested users to address the following:

- *Detailed data request*: The user must specify the requested datasets, and in some cases, particular variables, and may include specification of data requested from the provider (both restricted and public use) and other data to be used in the analysis.

- *Research topic and plan*: The user is usually required to provide an analysis plan explaining why the restricted data are necessary to complete the study and the project.

- *Computing environment and data security plan:* Restricted data requests often require that the user describe a particular computing environment that the user or the user's agency/institution provides. A required data security plan specifies the rules, process, and location for accessing and analyzing data. The security plan must be reviewed and approved by the data custodian; in some cases, this includes physical on-site inspections.

This process is burdensome both for those who try to make data available and for users trying to use data. It creates opportunities for people to hoard data and refuse to share, under the guise of protecting confidentiality, or to claim quite legitimately that it is simply too costly to share data safely. The GVDE credentialing system addresses these concerns. A system used to identify and credential users using the GVDE is essential and benefits governmental agencies in four primary ways.

> *First*, to implement the digital passport, the GVDE technologies standardize the vocabulary used to describe potentially disclosive data and its degree of sensitivity. The GVDE research team's analysis of repository practices around the world found that the language used to describe levels of data restriction, confidentiality, and access methods differs significantly both between and within restricted data repositories. This language inconsistency confounds the challenge of developing a transferable digital user identity, therefore the GVDE research team standardizes the terminology used to describe the elements of restricted data security and access. Establishing a common set of terms and definitions allows different repositories to understand and integrate shared standards and technologies into their own processes. The GVDE research team harmonizes language characterizing disclosure risk associated with the geospatial data (distinguishing, for example, between small area estimates and trajectories of individuals) that undergird the standards necessary for a user to access data of different levels of risk.
>
> *Second*, by creating a durable and transferable user ID that maintains a record of both responsible use and any prior breaches in handling confidential data, the user passport creates incentives that reduce risk and encourage data sharing. The passport reduces the risk of irresponsible user behavior, because there is a reputational consequence that affects future data access. The passport thus also increases the willingness of potential data providers to share because they can have more confidence that their data is be protected.
>
> *Third*, by establishing a common set of standards across restricted data custodians, the passport facilitates new, creative analyses of datasets of data held by multiple custodians; current inconsistencies in standards for access often make such analyses of multiple restricted datasets impossible.
>
> *Fourth,* because all analysis takes place in the secure GVDE, the local computing environment is less critical. This is particularly advantageous for users from less resourced governmental agencies and institutions who do not have the technical staff or facilities to establish a secure local environment (which, for example, often requires a dedicated computer and locked office space).

### *Disclosure risks of confidential geospatial data*

The GVDE research team develops and evaluates geomasking and encryption techniques, with particular emphasis on rapidly growing new geospatial big data sources including GPS trajectories, mobile GPS data with sensor inputs, and social media and crowdsourced data. We evaluate which types of geomasking and encryption tools are most appropriate for addressing multiple traditional types of geospatial data (e.g., point, line, polygon, vector, raster) as well as these new big data sources. The disclosure risks associated with these geospatial data are unique as these data can be highly identifiable when presented in maps, visualizations, or when combined with other related data (Richardson, 2015). To protect individual identities in confidential geospatial data sets, various methods and privacy protection metrics have been developed in the past two decades. These include several geomasking and geospatial encryption methods. They mainly seek to modify or hide the original location information in georeferenced data through adding statistical noise or including more data records when responding to spatial queries in order to render re-identification difficult. The GVDE research team investigated an expanded set of analytical tools and geomasking and encryption methods, and their applications to new confidential geospatial data sources and types. These techniques are useful in the analysis of confidential data and can also be used to anonymize results of analysis, maps, and other anonymized visualizations of confidential data within the GVDE so that they can be exported from the GVDE for use in presentations, publications, and other outlets for sharing results with scientists or public audiences.

## METHODOLOGY AND ACTIVITIES

The GVDE research team identifies four key areas of work required to create and implement a robust and reliable fully scaled-up GVDE resource for widespread use by governmental agencies involving confidential geospatial data. These key areas are:

- I. Develop the Geospatial Virtual Data Enclave and its Core Functions
- II. Evaluate and Implement Masking and Encryption Capabilities for the GVDE
- III. Develop the GVDE User Credentialing System
- IV. Ensure Sustainability of the GVDE

Our approach to these key interrelated GVDE technologies and implementation components is discussed below.

### I. Develop the Geospatial Virtual Data Enclave and its core functions

We evaluate and implement software tools and procedures (e.g., data management, GIS, analytics, modeling, spatial statistics, etc.) to enable users to a) *share*, b) *access*, c) *analyze*, (d) *replicate*, and e) *build on* confidential geospatial data within the GVDE. This part of the project tests the suitability, efficacy and efficiency of a set of analytical methods within the GVDE environment using different types of data and for different applications. This part ensures that the analytical tools made available to users in the GVDE enable them to perform analyses on a variety of data types and formats. Providing these analytical tools within the GVDE benefits the governmental agencies and the broader society, which may not have as much access or exposure to geospatial analytical tools—especially for less resourced governmental agencies and institutions. Specific examples of the datasets to be used for suitability testing are listed later in this section.

#### *Enhancing the core capabilities of the GVDE technologies*

To enhance the GVDE technologies prototype, we test and evaluate the system's efficiency, reliability, security, and the user experience. A diverse set of georeferenced datasets are used for testing in the GVDE.

- *User experience captures* the friendliness of the system to users. A system with excellent user experience is critical to successful adoption by users. The following four aspects of the user experience are evaluated: (a) GVDE set-up and login process; (b) system interface; (c) user control and input; and (d) display. Feedback is collected to ensure that the set-up and login process is as user-friendly as possible. The system interface is designed to replicate the standard Windows desktop environment that most users are familiar with. The system interface and user control and input can be affected by network lag if the user has a slow internet connection or is located at great distance from the host servers. The intention is to test the boundaries of these issues to minimize their effect on the user experience while maximizing system efficiency, reliability, and security.

- *System efficiency* is the computational and analytical performance of the GVDE. Comparison between the performance of a range of spatial analysis using the GVDE and local computers is conducted to compare analytical performance, operation time, and so on.

- *Software reliability* assesses whether the GVDE system is reliable. It addresses questions like: Are there any connection failures from remote desktops and how often do they happen?

- *System security* is the ability to secure geospatial data in the GVDE and minimize disclosure risk of confidential data. High system security is one of the most important features of the GVDE system, which must ensure the secure sharing of confidential data in a controlled and safe environment. System security must include data confidentiality, data accessibility, and data integrity and cover both access to the system (ensuring only authorized users can access given data, and that unauthorized removal of data is prevented) as well as proper vetting of data that is authorized to be removed from the GVDE system. The GVDE system is designed to meet Federal Information Security Management Act (FISMA) Moderate standards in regard to system security. No data or analytic output can be removed from the GVDE system without undergoing formal statistical disclosure control and approval of sponsor agency and confidentiality officers.

There are two authorization steps during the login process. After initially logging in to the GVDE Account, where digital credentials from the proposed user credentialing system resides, the two-factor authentication requires a second authorization step; users need to provide a passcode dynamically generated by a pre-assigned electronic device (smartphone app or hardware token), or respond to a push notification sent to their smartphone app. A new passcode is generated every 60 seconds and is specific to each individual user. This process can detect and stop unauthorized access in the event of a compromised password, as both the password and electronic device are needed for login. Only users who successfully pass these two independent authorization steps can access the GVDE system. The GVDE system prevents users from uploading or downloading data or files to their local computing environment (e.g., their PC). External files can be added to a user's secure computing space by staff. Output files are made available to the user outside the GVDE only after disclosure review by a GVDE staff with expertise in confidentiality protection. Finally, to ensure that users disconnect or lock the GVDE account when he/she intends to leave the connected computer, the remote server automatically disconnects or locks the users' account when no activity is detected for a defined period in order to prevent unauthorized access to the GVDE system. Results of our testing confirm that these security technologies and processes ensure high system security of the GVDE.

### *Evaluation and implementation of GIS and analytical software tools for the GVDE*

The GVDE research team evaluates and implements geographic information system (GIS), statistical and analytical tools for use within the GVDE. These tools are useful for analysis, sharing, and display of data within the GVDE, and also for visualizing data for geomasking and encryption methods so that output can be safely removed from the GVDE. Below is a brief description of some of the geospatial data management and analysis software tools which have been implemented and tested in the GVDE. We also monitor new and emerging techniques and methods for geospatial analysis and for confidential geospatial data protection and integrate them as they mature.

*Social statistics*: A suite of commonly used social statistical techniques is supported in the GVDE. These include multiple regression, principal component analysis, cluster analysis, factor analysis, discriminant analysis, contingency table analysis, general linear models, survival analysis, log-linear models, multi-level models, and structural equation models.

*Spatial statistical techniques for area-based data*: All major geospatial analytical techniques for area data are supported in the GVDE together with procedures to ensure that the aggregation level is adequate for geoprivacy protection. These techniques include various area-based measures of spatial association and spatial cluster analysis methods (e.g., Anselin's (1995) local indicator of spatial association (LISA), Moran's I, Geary's C, and Getis's Gi,), and a suite of spatial regression models.

*Geospatial methods for point-based and linear data*: Many geospatial methods for point-based and linear data can generate results, in most cases, that do not reveal the original locations of the records in the dataset. These methods include geographically weighted regression, kriging, spatial point pattern analysis, spatial cluster analysis, kernel density estimation, and the K function. However, in some data sharing situations, users may need to see the original point locations during the analytical process (e.g., to visually assess the spatial distribution of the points). We develop and test a set of procedures for visualizing point locations while preventing disclosure of the identity of the subjects, masking the point locations and evaluating cartographic output to examine the effect of different parameters (e.g., bandwidth and impedance functions that model the effect of distance decay) on disclosure risks, and testing the value of these procedures.

*Geographic Information Systems (GIS) tools for data management and analysis*: The suitability of GIS data management tools, including ArcGIS and standalone open-source GIS, is evaluated and implemented in the GVDE. In the ArcGIS environment, each geospatial data protection method is implemented through scripting tools. Each method is implemented in the form of a Python script that uses the ArcGIS Python package (ArcPy). Through ArcPy, a script can access the geoprocessing environment of ArcGIS and can be easily distributed and reused without programming knowledge on any computational platforms with ArcGIS installed. In addition, the geospatial data protection methods is developed using standalone Python code based on open-source mathematics and GIS libraries, such as NumPy (doing mathematical calculations), SciPy (including data processing, optimization and statistics), Shapely (manipulating and analyzing geometric objects), and GDAL (processing vector and raster data formats). These standalone open-source tools run in the Python environment with the open-source library installed in the GVDE (users working in the GVDE cannot directly call on web-based software). The geoprocessing environment of ArcGIS also offers many user-friendly functions such as selection set support, validation of inputs, error messaging, and recording of history, as do some open-source GIS environments.

## II. Evaluate and implement masking and encryption capabilities for the GVDE

Many users may wish to *export* the results of their analyses via maps or other visualizations from the GVDE, for use in publications or presentations. This component of the project examines appropriate anonymization methods and related disclosure risks for several types of geospatial data (point, line, polygon, raster, vector, etc.) as well as for rapidly growing new sources of confidential geospatial data such as GPS trajectories and geospatial data confidentiality issues pertaining to crowdsourced and social media data. The GVDE technologies have the ability to apply geomasking and encryption methods as well as other techniques to anonymize the data in order to protect the identities of human subjects when users need to extract maps or data analysis summaries and graphics from the GVDE.

### *Masking and Encryption Techniques*

The GVDE research team evaluates multiple geomasking methods, encryption, and other processes to enable users to anonymize and then export *visualizations, maps, or analyses* derived from confidential geospatial data from the GVDE (after review) for use in publications or presentations. Below are some of the geomasking and encryption techniques that we evaluate for inclusion in the GVDE.

*Geomasking* - Geomasking techniques modify and hide the original location in georeferenced data by adding statistical noise to the original data (Kwan et al., 2004; Armstrong and Ruggles, 2005; Leitner and Curtis, 2006). By masking the locations in a data set, users may still use illustrations that include the locations of subjects' homes or workplaces in their maps or geovisualizations when publishing their results, while protecting their geoprivacy. Various geomasking methods have been developed to date (Armstrong et al., 1999; Kwan et al., 2004; Chen et al., 2008; Zimmerman et al., 2008; Zandbergen, 2014; Zhang et al., 2017), and are described below.

In *aggregation*, data may be grouped by areal units (areal aggregation) or multiple individual records can be assigned to one point-location (point aggregation). Further, aggregate patterns can be used to make it impossible to identify individual subjects (pattern aggregation, e.g., hot spot maps). (b) An *affine transformation* translates, contracts, or expands a point pattern. For instance, the scale of the point pattern may be altered so that relative positions and orientations between locations are maintained while the location pattern's relation to the study area is modified (re-scaling). Alternatively, all locations may be shifted a determined distance and direction from their original locations (shifting). (c) A *random perturbation mask* allows both the amount and direction of spatial displacement to vary between points, thus altering the relative locations and orientation of the points in a particular point (or location) pattern. For instance, each point may be randomly placed along some line feature, such as a circle defined by a center at the original point and a chosen radius (circular masks). The size of the perturbation circle could be weighted by the population density at each point (weighted masks) in order to take into account its effect on the risk of disclosure (as lower population density in an area leads to higher disclosure risk). (d) A *donut mask* is similar to random perturbation within a circle, but in this method a smaller inner circle is also created inside a larger outer circle, creating the "donut," and the perturbed location is then placed outside of this smaller circle but inside the larger one (Zhang et al., 2017). This method thus sets the minimum and maximum distance of random perturbation. (e) In *Gaussian displacement*, the direction of spatial displacement is random while the distance follows a Gaussian distribution. The dispersion of the distribution may vary based on other parameters of interest, such as local population density. (f) *Bimodal Gaussian displacement* is a variation on Gaussian displacement, but it uses a bimodal Gaussian distribution for the random distance function. In effect, this is similar to donut masking, but with a less uniform probability of spatial placement (Zandbergen, 2014).

(g) The *location swapping* method "replaces an original location with a masked location selected from all possible locations with similar geographic characteristics within a specified neighborhood" (Zhang et al., 2017). (h) An extension of this method is *location swapping with donut*, which uses the same method as location swapping, but like donut masking, a smaller internal circle within which points cannot be displaced is employed. The radius for creating the inner and outer circles can vary based on local population density.

Geomasking techniques may be applied either to the data before analysis or to the products (e.g., maps) after analysis. Most users prefer post-analysis masking. Since all masking procedures change the data in some way, pre-analysis masking may affect the results of georeferenced individual-level data analysis, obscuring important geographic patterns. The analysis of confidential individual-level geospatial data is important for understanding critical social and policy issues. Our GVDE technologies provide users with access to detailed geographic data *inside the GVDE* so that they can conduct their analyses using this preferred approach, and then offers tools for making their output (e.g., maps and analytical tables) safe before removing it from the GVDE.

*Geospatial cryptography (Encryption)* - Using cryptographic techniques to protect confidential geospatial data can be achieved in many ways including "transforming all data to a different space using cryptographic techniques so that they can be mapped back to spatial information only by the user" (Andrés et al., 2013). For example, Clarke (2016) employs forward and inverse algorithms to mask point-based data by switching the digits of data coordinates. Jacquez et al. (2017) have proposed that, using cryptographic techniques, it is possible to design and implement geospatially encrypted geographic information systems (GEGIS) to promote the sharing and spatial analysis of confidential data. For instance, one may conduct geospatial analysis in an encrypted space using original geographic coordinates and report the results without revealing individual locations. These and other basic geospatial cryptography techniques are evaluated for feasibility and appropriateness for the GVDE.

### *Evaluation of confidential geospatial data protection methods*

The GVDE research team has evaluated the effectiveness of the confidential geospatial data protection methods by assessing how the disclosure risk and analytical utility of each analytical or visualization outcome has been altered. To do so, we conduct simulation experiments to generate estimates of disclosure risk and utility for various geomasking and encryption methods and analyze the trade-off between the quality of analytical results and preservation of confidentiality for sensitive geospatial data (e.g., comparing the results generated by masked and unmasked data). The GVDE team evaluates patterns of disclosure risk for maps after the application of a masking or encryption method to determine levels of confidentiality risks (Kwan et al., 2004; Zandbergen, 2014). In addition, raw and masked data undergo spatial statistical analysis (Haining, 1990; Bailey and Gatrell, 1995; Anselin, 2013; Chun and Griffith, 2013) to identify and compare geographic patterns that could be relevant to data confidentiality (e.g., spatial distribution and clustering patterns). Using a "baseline" model the GVDE research team compares statistical inferences generated from a subsequent "masked" model to see whether perturbed coordinates shift enough to result in different beta coefficients. Based on the evaluation results concerning the suitability, efficacy, and efficiency of each geomasking or encryption method (e.g., what types of geospatial data protection methods are suitable and effective for what types of data and analyses), guidelines in the form of a suitability matrix and guidebook are prepared and disseminated to the GVDE users, including governmental agencies and the broader community.

### III. Develop the GVDE user credentialing system

The development of an innovative credentialing system is an integral component of the GVDE technologies. The credentialing system provides a transferable digital identifier to trusted users of the GVDE, to ensure responsible data stewardship and the protection of confidential geospatial data. Our GVDE technologies build on a foundation of preliminary work by the research team to develop a standardized and broadly accepted system of user credentialing. That previous research identified significant discrepancies in rules and modes of access to confidential data, and even to the language and definitions to describe data confidentiality, modes of access, and user requirements. This GVDE project proposes standardized language and best practices regarding data confidentiality, user requirements, and modes of access for confidential geospatial data. It implements these criteria into a system of digital identities that control access to data in the GVDE, so that the system verifies, at login, that a particular user has the appropriate credentials for accessing the requested datasets. This system increases the willingness of potential geospatial data producers, including governmental agencies, to share data, because they can have confidence in its security, and the ability of users to more readily undertake creative analyses with the least possible risk to privacy and confidentiality.

The output from our GVDE technologies' component is threefold. The *first* is a recommended matrix system of user credentialing. These recommendations draw on interviews with organizations delivering geospatial data as well as survey, administrative, and other non-designed data. In order for a system of user credentialing to be successful, it must reflect the concerns of data providers and funders. In many cases, data providers distinguish citizens or residents of a particular country; "legitimate" users, journalists, and commercial entities; researchers at institutions with Institutional Review Boards; those subject to subpoenas or Freedom of Information Act requests; and those with the legal ability to submit to the requirements of the data custodian. The conditions under which users access data depend on the interaction between data and user characteristics. In addition to describing current practice and identifying areas of overlap and difference, the GVDE team identifies steps to move the field forward in terms of best practices and required elements for a system of standardized user credentials.

The *second* component is software and an interface that can be used to match user credentials (a passport) with dataset requirements (captured in a dataset-specific visa) in order to access the GVDE. When a user wishes to use a particular dataset, their information is compared with the requirements associated with those data. The GVDE team issues a "visa" when a user is approved to access a particular dataset. The GVDE technologies verify the passport and visa before providing the user access to the data in the secure computing space. In order for the system to work properly, the credentials necessary for the user, the user's institution, and the data themselves have to be clearly defined – which means that the interface has to be user-friendly and intuitive, and the software has to provide a bridge that performs "checks" for matches between the user's information and the data requirements. Our GVDE technologies may also provide a foundation for broader applications beyond the GVDE, which would allow user access to restricted datasets held at multiple repositories.

*Thirdly*, users need training in awareness regarding the availability of the user credential and in geospatial confidentiality protection itself. The GVDE research team offers training to users and other stakeholder groups, including governmental agencies, to build community understanding of the process and value of geospatial user credentialing. Users who successfully complete this training receive a digital badge identifying their knowledge in geospatial data stewardship. This badge is recognized by the user credentialing digital access system. This outreach and training ensure that the social and technical infrastructure changes described above make the process of accessing and using geospatial data with potential disclosure issues more conducive to data sharing.

## IV. Ensuring sustainability of the GVDE

To ensure sustainability and to build the usage of the GVDE technologies, we undertake extensive training, outreach, and dissemination activities. The GVDE technologies are maintained as part of the larger portfolio of data management and stewardship services that the GVDE research team provides to the user community, including governmental agencies. As the GVDE is scalable, the GVDE system has the capacity to meet expanded demand. The most efficient way to expand capacity is to move to the cloud. This also requires careful attention to security issues. of course. The GVDE research team is now largely operating in a secure Amazon Web Service environment which provides the GVDE with a FISMA compliant and ATO (authority to operate) FedRAMP compliant platform; that is, it is approved for uses by DoD, IRS, and Census. The GVDE research team has a similar contract with Azure, the Microsoft cloud environment, so it is not tied to a single provider.

### *GVDE training programs*

The GVDE research team is a leader in offering regular training to users regarding data management, handling, ethics, and analysis. The research team has conducted numerous training and workshops under grants from the NSF and the NIH and are highly experienced in developing training materials and conducting training directly related to these technologies.

The GVDE team also develops training and technology transfer modules for using the GVDE system, and on sharing and analyzing confidential geospatial data within the GVDE. These training modules support users in the use of the GVDE system (including user support and frequently asked questions), the GVDE credentialing process, articulating NSF data management plans, and ethics related to the handling and use of confidential geospatial data. All training activities are assessed based on the number and diversity of participants and trainee evaluations (e.g., anonymous feedback forms, group discussions, and post-event follow up). Evaluations cover both content understanding and retention, as well as the effectiveness of training materials and instruction. Trainees are asked to assess the quality of masking tools, usefulness of results, ease of use, and to comment on their general satisfaction with the GVDE system and information provided. Trainees are recruited from multiple sectors based on their needs for using geospatial data, and our selection process ensures diversity among our participants. The training materials and activities help to develop usage of the GVDE technologies.

Contact or Correspondence:

Douglas Richardson, PhD
Distinguished Researcher
Center for Geographic Analysis
Institute for Quantitative Social Science
Harvard University