**ACCESS TIER: PROTECTED**

• Restricted-use data behind firewall with output controlled for disclosures;

• Automated output from SDL software with use restrictions (e.g. web-based query system)

• Licensing program (user controlled infrastructure)

Access to this data is not automatic but requires one or more additional steps such as, but not limited to, a data use agreement, license, system or website account, and automated SDL tools with built-in limitations on allowable output. The number of additional requirements for access is intended to be more than the public access, but less burdensome than the restricted tier. The lowered burden should improve access, timeliness, and other aspects of data quality in a meaningful way to the data customer.

**Web-Based Query Systems**

Query systems allow users to design queries to generate customized tabulations (WP 22). Also have predefined queries.

\*Data stored in query systems can be protected and restricted.

Pros

* Run queries on data more detailed than PUFs (varies by agency—could be restricted data or perturbed data from public use files)
  + Increases utility, improves data quality
* Permits a wider range of analyses than does releasing only data summaries and it provides results based on actual rather than simulated microdata (Gomatam, 2005, 164)
* Developer can build in SDL (see Gomatam, 2005, 167)

Cons

* No direct access to the microdata (may vary by agency)
  + Decrease utility
* May not prevent disclosures (table splicing; can do this in CDC Wonder)
* Built in disclosure technique can reduce utility and quality (Gomatam, 2005, 167-168)
  + Top coding, swapping, adding noise (WP 22)
  + Prohibit key identifiers (age, race, sex) as outcome variables but permit as predictors (Gomatam, 2005, 167)
  + Disallow any transformations (Gomatam 2005, 173)
* Functionality is limited to what the developer allows to be run in the query system
* Expensive to implement and maintain (Haggard, 2006, 189)

New queries vs. pre-computed queries

Examples of online query systems:

* NASS (Karr)
* Australian Bureau of Statistics, TableBuilder (Chipperfield, 2019)
* CDC’s Wonder
* BLS

**ACCESS TIER: RESTRICTED ACCESS**

• Licensing program (user controlled infrastructure)

• Virtual Data Enclave

• Physical Data Enclave

This is the default tier for containing data collected under a pledge of confidentiality. Critical to its modernization is to include more access options using newer technology such as commercial cloud computing and remote access/virtual enclave. This should also include viable options not constrained by the current system of Federal Statistical Research Data Centers (FSRDCs).

**Virtual and Physical Data Enclaves**

There are two types of data enclaves: 1) physical data enclaves; and 2) virtual data enclaves. Both types of enclaves allow researchers to access restricted use data under highly restricted conditions which reduces disclosure risks. Some data enclaves allow researcher to access the full data sets while others require researchers to prepare data dictionaries and limit access to only the variables the researchers need to complete their analysis.

The primary difference between the two types of data enclaves is the process by which the data are accessed. In physical data enclaves researchers must physically sit in a controlled environment at the data owner’s office or site where the data are stored. Virtual data enclaves allow researchers to access the restricted use data remotely over secure electronic lines via their personal computers while they sit in their own offices or homes. The output generated is returned to the researchers.

Both types of data enclaves increase the usability of the data. Researchers are permitted to access data not publicly available under controlled conditions. For example, [*list types of data that cannot be publicly accessed: genetics data, geocoded data, detailed geography, exact dates, detailed race, income*].

Conversely, both types of data enclaves can also decrease the utility of the data. The process for requesting access to restricted use data can be arduous and thus reduce the number of researchers who can access the data. Researchers must submit research proposals containing detailed information about the research project, the hypotheses to be tested, the data set and variables to be used in the analysis,

the empirical methods to be used, and the specific data outputs that will result from the project thus limiting exploratory analyses. Research proposals are reviewed and approved by a review committee which can take several weeks or months to complete. Additionally, users must agree to terms and conditions governing the access and use of the confidential data as well as sign nondisclosure affidavits. Some researchers are required to complete background investigations, Special Sworn Status, be citizens of the U.S., *…etc*. Furthermore, there are costs associated with accessing restricted use data via data enclaves [*include examples*]. Costs reduce the utility of the data because some researchers may not have funding to complete research. Most students completing graduate or doctoral level research, living on fixed incomes, may not be able to afford to access data in enclaves.

Both types of data enclaves allow researchers to improve the accuracy and precision of their estimates. Data available in enclaves are not subject to the statistical disclosure limitation methods that public use files are subject to prior to release. For example, detailed race/ethnicity and geography measures are typically not available on public use files due to disclosure concerns. These types of measures are available to researchers in data enclaves thus increasing the accuracy and prevision of estimates. [*Might include examples of research completed in RDCs that could not be completed using PUFs].*

Data quality can also be decreased when accessing restricted use data in data enclaves. Extreme values or values representing an individual are generally removed from analysis (e.g. minima, maxima, medians). These values might be useful to researchers doing sensitivity analysis. {*need to expand this section*}

**Licensing Program (user controlled infrastructure)**

Licensing agreements permit a researcher to use restricted data offsite, but under highly restricted

conditions, as spelled out in a legally binding agreement [text from Restricted Access Procedures]. Arrangements that place restrictions on who has access, at what locations, and for what purposes access is allowed normally require written agreements between agency and users. These agreements usually subject the user to fines, being denied access in the future and/or other penalties for improper disclosure of individual information and other violations of the agreed conditions of use. Users may be subject to external audits conducted by the agency to assure terms of the agreement are being followed. Users in violation may be required to pay fines or be subject to other legal penalties [text from WP 22].

Licensing agreements require:

- a demonstrated need for sensitive data;

- authorization for all users at the requesting institution;

- signature by a senior level official and key staff;

- a data security plan;

- agreement by researchers not to identify individual research subjects or to link data received with other microdata files; and

- review of all statistical output before publication.

[text from Restricted Access Procedures].

The license is for is a specified period of time and data files must be returned or destroyed. Some licensors require fees and/or approval by an institutional review board. Additional information on this method is provided here: <https://nces.ed.gov/FCSM/pdf/CDAC_RAP.pdf>

Pros for Licensing Program

* …
* …

Cons for Licensing Program

* …
* …

**Future Placeholder for Highly Restricted Access Tier**

(leave blank)

**References**

Chipperfield, James. John Newman, Gwenda Thompson, Yue Ma, and Yan-Xia Lin (2019) “Prospects for Protecting Business Microdata when Releasing Population Totals via a Remote Server,” Journal of Official Statistics, Vol. 35, No. 2, pp. 319–336, <http://dx.doi.org/10.2478/JOS-2019-0015>.

Gomatam, S. A. F. Karr, J. P. Reiter and A. P. Sanil (2005) “Data Dissemination and Disclosure Limitation in a World without Microdata: A Risk-Utility Framework for Remote Access Analysis Servers,” Statistical Science, Vol 20, No. 2, pp.163-177. DOI: [10.1214/088342305000000043](https://doi.org/10.1214/088342305000000043)

Haggard, Lois M. and Shandra J. Burnett “Measuring the Impact of aWeb-based Data Query System: The Logic Model as a Tool in the Evaluation Process,” (2006) Journal of Public Health Management and Practice, Vol 12, No. 2, pp 189-195. Retrieved from <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=ovfth&NEWS=N&AN=00124784-200603000-00009>.

Karr A.F., Lee J., Sanil A.P., Hernandez J., Karimi S., Litwin K. (2002) “Web-Based Systems that Disseminate Information from Databases but Protect Confidentiality.” In: McIver W.J., Elmagarmid A.K. (eds) Advances in Digital Government. Advances in Database Systems, vol 26. Springer, Boston, MA. DOI <https://doi.org/10.1007/0-306-47374-7_11>

WP 22: <https://nces.ed.gov/FCSM/pdf/spwp22.pdf>

Restricted Access Procedures: <https://nces.ed.gov/FCSM/pdf/CDAC_RAP.pdf>