

Data Stewardship Program

DS026 - Automated Collection of Data from the Internet

Policy and Governance

PURPOSE

This document establishes a uniform vocabulary, basic policy, and an Automated Internet Data Collection Review Board (AIDCRB) to provide guidance and review and approve automated Internet-based collection activities that do not involve self-response instruments. These methods include *automated web scraping*, *web crawling*, and other emerging internet-based collection technologies. These methods may also involve the use of Application Programming Interfaces (APIs) provided by external websites.

BACKGROUND

The advent of new data collection methods such as automated web scraping and web crawling may provide opportunities for the Census Bureau to improve the statistical products we produce for the American public, reduce the burden we place on our business, government, and individual respondents, and lower the cost of data collection.

The Census Bureau should therefore explore ways to leverage these technologies in a manner that is consistent with our commitment to scientific integrity, ethical and legal responsibilities, and our Privacy Principles.

We understand that this is an emerging field and that the big data landscape, the technologies to navigate that landscape, and the public's attitudes toward privacy are rapidly evolving. Promoting the development of these methodologies while respecting the rights and preferences of our respondents will be key to keeping the public's trust while maximizing our value to the American taxpayer.

DEFINITIONS

Automated Web Scraping – The use of software or code to collect or “mine” specified information from one or more websites. Web scraping is also called *Web data extraction*, *screen scraping*, or *Web harvesting*.

Web Crawling – Crawling uses a program to automate Web indexing, a process of mapping the available resources on a website. Web crawlers access one page at a time through a website, identifying any available links on that page and following them until all pages have been indexed. Web crawlers also help in validating HTML code and hyperlinks. Web crawlers are also known as a *Web spiders*, *automatic indexers*, *bots*, or simply *crawlers*.

Application Programming Interface (API) – APIs are a tool offered by some data providers that produce structured data outputs in response to queries inputted by the end user. APIs may be open to the general public or may require the use of an authenticator, such as a username/password combination or an *API Key*, to access the data.

Public Information – Data that the Census Bureau can collect and use to support its mission without purchase or entering into of an agreement, and without violation of any applicable access rules, terms of use, or intellectual property rights.

MISSION and POLICY

Mission

The AIDCRB's mission is threefold. First, it will monitor ways that government, private sector, and academia are using these technologies and any applicable policies and regulations that govern them. The Board may use this information to develop Census Bureau-specific guidance and best practices. The Board may also review and recommend tools that program areas can use to conduct these activities in a way that is effective and consistent with statutes, regulations, and Census Bureau policies. Based on this work, the Board may raise policy issues for consideration by the Data Stewardship Executive Policy committee.

Second, the Board will take stock of current and proposed uses of these methods. The Board will review current and future projects to ensure the methodology adheres to the standards laid out below, and limits exposure to any potential risk from using these methods. This aspect of the Board's work will be coordinated with the Disclosure Review Board (DRB), the Office of General Counsel (OGC), and components of the Office of the Chief Information Officer (OCIO) to assure that automated data collection activities do not inadvertently compromise confidential information.

Third, the Board will work on developing guidance for storing, retaining, and making these data available to other programs for reuse. This function is vital to ensuring we maximize the value of these data, and limit the burden we place on providers and respondents.

Policy

Automated collection of data from the Internet is generally permissible if all of the following are true:

1. The data being collected are public information or the Census Bureau has received explicit informed consent from the respondent and/or data provider to collect them;
2. The data collection is consistent with the Census Bureau's mission and done in a way that is legal, ethical, transparent, and does not present a risk to the reputation of the Census Bureau;
3. Collecting the data does not constitute a disclosure risk for Title 5 data, Title 13 data, or Title 26-protected Federal Tax Information (FTI).

4. The systems and applications used for collection and analysis have a Census Bureau Authority to Operate (ATO) that supports storing Title 13 data.

The Board will review current and future collection methods to ensure proper consideration has been given to the potential implications of these activities. Use of these methods without approval from the Board is prohibited.

Once the Board determines that proposed collection methodology satisfies the review criteria, it will produce a certification. The project using that particular methodology should retain a copy of this certification with its files. The certification will be valid for three years' time or until there is a substantial modification to the methodology, at which point the activity must be re-certified. If a proposed methodology is not consistent with these standards, and in the view of the Board presents a risk to the reputation of the Agency, it may refer the issue to DSEP for adjudication.

SCOPE

The AIDCRB will only review and provide guidance and oversight on the legal, ethical, and policy considerations associated with collecting data from the Internet via automated methods.

The review Board will not comment on the "suitability for use" of data collected via these methods. The determination on whether or not these data are suitable for use is a programmatic decision that should be made consistent with the Census Bureau's Statistical Quality Standards and guidance from the Office of Management and Budget.

Online survey instruments that take direct input from respondents are not in this Board's purview. Nor will the Board review situations where staff are manually searching for data to support the Census Bureau's collections.

This policy also does not apply where APIs are used as the mechanism by which the Census Bureau will receive data from a data provider or agency through a data acquisition or other formal agreement.

MEMBERSHIP

The AIDCRB shall consist of the following members:

- A Chair appointed by the Data Stewardship Executive Policy Committee (DSEP)
- A Vice-Chair, also appointed by DSEP
- A representative appointed by the Chair of the DRB
- A representative appointed by the Chief Privacy Officer
- A representative appointed by the Chief Information Officer
- A representative appointed by the Associate Director for Demographic Programs
- A representative appointed by the Associate Director for Economic Programs
- A representative appointed by the Associate Director for Research and Methodology

- A representative appointed by the Associate Director for Decennial Census Programs
- A representative appointed by the Associate Director for Communications

Board members shall be selected based on their ability to represent their program areas interests and their aptitude for analyzing policy issues related to automated data collection, and should be able to make management decisions for their directorate. Members shall serve until they resign, are reassigned, or their replacements are nominated and approved. If a member is unable to attend a meeting, he or she shall designate an alternate to represent his or her position. The Office of General Counsel (OGC) may attend meetings and advise the Board as appropriate. The Chair may also invite additional individuals, such as Program Area Subject Matter Experts (SME) and technical experts, to attend meetings, as needed, to assist the Board with its work.

MEETINGS

The Chair will schedule a standing monthly meeting or special meetings as needed.

The Chair will circulate an agenda before each meeting. Any member may schedule topics for discussion, in addition to the collection methodologies that are scheduled for review.

Decisions are made by consensus. If consensus cannot be reached, the Board may escalate the project to DSEP for adjudication.

STANDARDS AND REVIEW

The Census Bureau will only use automated data collection methods that are legal, ethical, respect the rights of our respondents, are consistent with our Privacy Principles, and do not present unacceptable risk to the reputation of the agency. Furthermore, the Census Bureau will only use these methods to collect data in support of its mission to be the nation's leading provider of quality data about its people and economy.

To uphold these standards, the Board will review proposed automated data collection methods to ensure the following considerations have been addressed:

- Privacy of the Respondent
- Rights of the Provider
- Protection of Confidential Information
- Policy and Sensitivity Considerations

Project leads are responsible for presenting artifacts and evidence to facilitate the Board's review, and working with the Board to address concerns or modify their work accordingly.

Rights of the Respondent

Generally, the Census Bureau will only use these automated data collection methods to capture publicly available information. The Board will assess whether the researchers employing a particular methodology have exercised due diligence on behalf of the Agency to ensure that the data they intend to collect were not made public in a way that is unintentional, illegal,

unethical, or contrary to the wishes of the respondent. Publicly available data should not be collected or used in a manner that is contrary to the posted terms-of-use for the given website.

The Board will also review whether the proposed methods sufficiently limit or ideally eliminate the collection of data that is not essential to the research or stated objectives of the project. Furthermore, the Board will review whether there is a plan in place to account for and dispose of non-relevant data that is consistent with the sensitivity of the data collected (e.g. Personally Identifiable Information or Business Identifiable Information) and applicable records schedules.

Where these data collection methods do not involve direct interaction with the respondent, obtaining explicit or implicit informed consent is generally not possible. In these cases, the Board should verify that the use of these methods to collect data is sufficiently transparent to the public. This may include reviewing whether the automated collection methods are properly accounted for in all of the documentation associated with the data collection including the:

- Paperwork Reduction Act (PRA),
- Federal Register Notice (FRN),
- Privacy Threshold Analysis (PTA),
- Privacy Impact Assessment (PIA),
- National Archives and Records Administration (NARA) record schedules, and
- System of Records Notices (SORNs) that apply to the data being collected.

It is the responsibility of the project lead to provide this information to the Board when requesting a review.

In cases where a methodology involves the collection of non-public information, the Board will review whether the informed consent obtained by the data provider is appropriate, obtained from the correct respondent or representative of the respondent, and sufficiently protects the agency from liability.

The AIDCRB will also explore the feasibility and appropriateness of instituting broad enterprise-level informed consent procedures that support these methodologies. This could include broad transparency measures such as FRNs, or “opt-out” procedures. However, until these solutions are in place the onus will remain on the program area using these methods to achieve sufficient informed consent and transparency.

Rights of the Provider

The Census Bureau will use automated data collection methods that are respectful and transparent to the data's host, limits their burden, and honors their intellectual property rights.

The Board will ensure that the use of automated methods for a data collection is transparent and respects providers' terms of use. It is the responsibility of the project lead to provide this information to the Board when requesting a review.

The Census Bureau will respect any restrictions imposed by the robot.txt file of a targeted website provided it is current or has been updated within 5 years from the date that the automated collection will take place.

The Census Bureau will take reasonable steps to locate and respect Terms of Use or Conditions of target websites. The AIDCRB will review the proposed project to ensure they are employing sufficient measures to do so based on the scope and nature of the collection. For example, a project that proposes to scrape a large volume of PII from one or two target websites might merit a manual review of those websites' terms of use. Conversely, if the project will scrape a small volume of comparatively non-sensitive information from a large number of websites, the project might employ machine learning to look for key phrases in the terms of use that might prohibit the collection and respond accordingly. The steps taken may also depend on whether the project will involve a one-time collection, or multiple collections over an extended period.

To ensure transparency, the AIDCRB will ensure the proposed method leaves AIDCRB approved text in a target website's log files or user agent string to clearly identify that the data is being collected by the Census Bureau, and to direct them to more information on our website.

The AIDCRB will be responsible for maintaining the webpage referenced above to provide information to targeted websites as well as the general public and stakeholders. This webpage will include a link to this policy, as well as an explanation about how the use of these methods reduces respondent burden, improves the quality of our statistical products, and how we protect the data we collect from unauthorized disclosure.

As part of its review, the Board will verify that the collection method is structured in such a way that it does not overburden the provider's resources. Project leads should be prepared to speak to their plans to mitigate this burden including building delays in queries, conducting activities at off-peak hours, etc.

The Board will ensure that the collection method does not attempt to subvert any limiters on scraping/crawling activities such as attempting to solve a CAPTCHA, clicking a box that states "I'm not a Robot," or evading IP address blocking through the use of proxy servers or other technical means.

The Board will work with the project lead to engage the Office of General Counsel where necessary to review the collection method to ensure that it respects any intellectual property claims made by the owner of the information being collected. The Board may also review any terms of use that the researcher will need to accept to acquire the data, and may refer the

issue to the Office of General Counsel if there is a question on whether it is appropriate for the researcher to accept those terms on behalf of the government.

Some third party data aggregators offer a limited number of free searches to demonstrate the value of the data products they sell. The proposed method must not subvert any limits they place on these free searches. If the data offered by these data providers is found to be useful, it may be appropriate to purchase them through the Census Bureau's data acquisition procedures.

If the proposed method is not permissible under a website's terms of use the Census Bureau may need to reach out to the data provider directly and ask them explicitly for an exception. The Board, in consultation with the Policy Coordination Office, will advise on any informed consent or authorization that may need to be obtained from the data provider, and will work with the project lead to engage the Office of General Counsel to ensure the documentation is sufficient and limits any liability to the agency.

Disclosure Considerations

The Census Bureau will use automated data collection methods that do not compromise the confidentiality of any Title 5, Title 13, or Title 26-protected data.

The disclosure considerations associated with using these methods are twofold. First, the Census Bureau must collect new data in a way that does not compromise the confidentiality of data already in the agency's possession. Second, the Census Bureau must ensure the confidentiality of any new data we collect using these methods if necessary.

Protecting existing data - The sophistication of the threats to the confidentiality, privacy, and integrity of our data have advanced significantly in the digital age. It is therefore important to take a very broad perspective on what types of activities could potentially threaten the confidentiality and privacy of data already held by the Census Bureau.

Merely collecting Personally Identifiable Information (PII) or Business Identifiable Information (BII) does not generate a disclosure risk. Instead, the Board will ensure that the project lead has approval of the Disclosure Review Board (DRB) to perform automated data collection where the methods involve targeting based on any of the following:

- Sampling Frames
- Response Data
- Administrative Data

The DRB will be able to provide guidance on how to conduct these activities without compromising confidentiality, and may recommend ways to mitigate this risk such as through seeding or salting.

Protecting new data - The AIDCRB will work with the project lead and, if needed, the Office of General Counsel and Privacy Compliance Branch to determine which, if any, legal protections apply to the data once it comes into the possession of the Census Bureau.

Generally, any data collected by the Census Bureau for use in its statistical programs that corresponds to individual persons, establishments, or addresses are collected under the authority of Title 13 U.S.C. and will be protected in accordance with the confidentiality provisions of 13 U.S.C. section 9. These confidentiality protections are in effect once the Census Bureau is in possession of the data or they are stored on our network. There is no distinction made whether the data are collected for research, testing, experimental, or production purposes, nor does it only apply if the data are comingled with confidential data already in the Census Bureau's possession. Provisions and protections of the Privacy Act may be likewise applicable if the data collected correspond to individual persons.

Crawling methods require the use of parsed information from acquired web content. Parsed information, including PII and BII, may be used for web crawling if projects use strategies approved by the DRB to minimize the risk of disclosure of Title 13 information.

There are, however, certain types of information the Census Bureau may collect using these methods such as information about governments or aggregate population or economic data published by other federal or state agencies, private entities, academic institutions, etc. that have been made available to the public and do not carry confidentiality protections. The AIDCRB will help make a determination on whether the data to be collected require confidentiality protections during its review.

The data collected through these activities must be stored on systems that are authorized to store and process confidential information (except information about governments and other types of non-confidential data as defined above). Generally, this work will be done on production systems; however, other environments may be authorized for prototyping new tools and methods.

Policy and Sensitivity Considerations

The Census Bureau will use automated data collection methods that do not present undue risk to the reputation or mission of the agency.

The Board will assess whether the proposed collection method raises any additional policy, privacy, or sensitivity concerns. This may include for example, scraping activities that target sensitive topics or populations. It may also include activities that pose a reputation or perception risk for the Census Bureau.

Though the project approval process already mandates policy, privacy, and sensitivity reviews, since the use of these automated methods is new and evolving this additional layer is intended to capture any new or novel concerns with their use.

LOOKING FORWARD

The Census Bureau recognizes that many websites' terms of use do not allow for web scraping/crawling and many APIs may not permit the types of activities that would enable data collection. The websites' terms of use may prohibit these activities because they are generally perceived to be against the host's commercial interests.

Web scraping can also be limited by robots.txt exclusions. In certain cases, there can be conflicts between terms-of-use and robots.txt, and there may be no clear path forward. The Census Bureau desires to be part of the W3C (World Wide Web Consortium) to participate in standards making process, which would harmonize and streamline future web scraping activity.

The Board will explore ways to begin outreach to promote the legitimate activities conducted by the Census Bureau and other Federal Statistical Agencies, so that data providers can see how these passive collections can benefit them through reduced burden, and the American public through better, more cost effective statistics.

EFFECTIVE DATE and IMPLEMENTATION

The policy outlined in this document is effective upon signature. The creation of the AIDCRB and the steps by which program areas will come into compliance are as follows:

- After approving this policy, DSEP will nominate the chair and co-chair of the review Board.
- Along with this policy, DSEP will issue a data call to divisions to capture the current automated data collection activities that are in scope for this policy, and any proposed activities planned for the near future. Divisions will be expected to respond within 2 weeks.
- Within 2 weeks of signature of this policy, the relevant Associate Directors will nominate representatives from their directorates to serve on the review board.
- Once the committee is fully staffed, they will have 4-6 months to develop any additional guidance required, board procedures, preliminary outreach materials, a catalogue of available tools, and a submission process for projects to receive board approval for the automated methods they may be using.
- Once the Board has established procedures, projects using these automated methods will have 6 months to come into compliance with the terms of this policy. In addition to responding to the data call, researchers heading these projects should immediately begin to consider whether their methodology adheres to this policy, begin pulling together artifacts to demonstrate compliance, and approach the Board after procedures have been put in place.

LEGAL AUTHORITIES

Title 13 U.S.C.

Title 26 U.S.C.

Title 5 U.S.C.

SIGNATURE

By Direction: _____ Date: _____

Ron Jarmin
Chair, Data Stewardship Executive Policy Committee

Summary Information	
Charter Title:	DS026 - Automated Collection of Data from the Internet
Date Signed:	
Last Reviewed:	
Intended Audience:	All Staff
Charter Owner:	Policy Coordination Office
Office Responsible for Implementation:	Policy Coordination Office
Office Responsible for Dissemination:	Policy Coordination Office
Stakeholder Vetting:	ADDC, ADRM, CAT, CED, CODS, CSvD, CTO, EID, EMD, ERD, ESMD, OGC, PCO, SEHSD