CLEARINGHOUSE FOR INTEGRATIVE HUMAN ANALYTICS AND DATA SYNTHESIS (iHADS)

Comments to Commission on Evidence-Based Policymaking

Contributors

Marina Alberti: Department of Urban Design and Planning; Director, Urban Ecology Research Laboratory; University of Washington
Lilian Na'ia Alessa: Director, The Center for Resilient Communities, University of Alaska and President's Professor, College of Art and Architecture, University of Idaho; Program Lead, Department of Homeland Security Center of Excellence, the Arctic Domain Awareness Center (ADAC), University of Alaska
Rob Axtell: Department of Computational Social Science, George Mason University
Steven Bankes: Independent consultant
Michael Barton: School of Human Evolution and Social Change, and Director, Center for Social Dynamics and Complexity, Arizona State University. Director Network for Computational Modeling in Social and Ecological Sciences
Dave Bennett: Geographical and Sustainability Sciences, University of Iowa
Luís Bettencourt: Santa Fe Institute
Eduardo Brondizio: Department of Anthropology; Director, Center for the Analysis of Social-Ecological Landscapes, Indiana University
Daniel Brown: Interim Dean, School of Natural Resources and Environment; Director, Environmental Spatial Analysis Laboratory; University of Michigan
Lawrence Buja: Director, Climate Science & Applications Program, Research Applications Laboratory, National Center for Atmospheric Research
Giovanni Luca Ciampaglia: Research Scientist, Network Science Institute, Indiana University Bloomington
Emily CoBabe-Amman: Climate Strategies Group
Jessica Faul: Survey Research Center, Institute for Social Research, University of Michigan
Johannes Feddema: Department of Geography, University of Victoria, Canada
Peter Fox: World Constellation Chair, Earth and Environmental Science, Computer Science, Rensselaer Polytechnic Institute
Kathleen Galvin: Department of Anthropology and Natural Resource Ecology Laboratory, Colorado State University
Daniel Gaylin: President, NORC, University of Chicago
Robert Groves: Department of Mathematics and Statistics, Department of Sociology, Georgetown University
Ed Hackett: Vice Provost for Research Professor, Heller School of Social Policy and Management, Brandeis University
Sandra Hofferth: Maryland Population Research Center,University of Maryland; Co-Director, Social Observatory Coordinating Network
James S. Jackson: Department of Psychology; School of Public Health; Institute for Social Research,
Yushim Kim: School of Public Affairs, Arizona State University
David Lam: Department of Economics, and Director, Institute for Social Research, University of Michigan
Felicia LeClere: NORC Health Care Department, University of Chicago

Sander van der Leeuw: School of Human Evolution and Social Change and School of Sustainability; Director, ASU/SFI Center for Biosocial Complexity; Arizona State University; External Faculty, Santa Fe Institute
Marc Levy: Deputy Director, Center for International Earth Science Information Network, Columbia University
Jianguo Liu: Department of Fisheries and Wildlife; Director, Center for Systems Integration and Sustainability; Michigan State University; Director, Coupled Human and Natural Systems Network
Patricia Mabry: Executive Director, Network Science Institute, Indiana University Bloomington
Emilio Moran: Center for Global Change and Earth Observation, Department of Geography, Michigan State University; Co-Director, Social Observatory Coordinating Network
Gerald C. Nelson: Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign
Kenneth Prewitt: Carnegie Professor and Director, Scholarly Knowledge Project, Columbia University
William Rand: Poole College of Management and Director Emeritus, Center for Complexity in Business, North Carolina State University
Dan Rogers: National Museum of Natural History, Smithsonian Institution
Patricia Romero-Lankao: Climate Science & Applications Program, Research Applications Laboratory, National Center for Atmospheric Research
Katherine Smith Evans: Executive Director, Council of Professional Associations on Federal Statistics
James Syvitski: Institute of Arctic and Alpine Research, University of Colorado, Boulder; Director, Community Surface Dynamics Modeling System; Chair, International Geosphere-Biosphere Programme
Billy Turner II: School of Geographical Sciences and Urban Planning, Arizona State University
Shaowen Wang: Department of Geography and Geographic Information Science; Director, CyberGIS Project; University of Illinois, Urbana-Champaign

Overview

Comprehensive, reliable information has always been a key factor for generating effective public policy, and governments long have endeavored to collect data to inform information generation to help them carry out this fundamental responsibility. In the 21st century, the size and diversity of the United States population, the magnitude and changing structure of its economy, the range of interactions with the rest of the globe and the complexity of its economic and social institutions make it all the more imperative that information from accurate, intelligible, representative data be readily accessible to policy makers to ensure that enacted policies will have desired results.

We want to distinguish between two categories of federal data: data gathered as part of administrative actions and surveys whose release would impinge on privacy of individuals and national security, and all other data. We strongly endorse a principle of open, unfettered access to all federal data, including those that are not sensitive and those aggregated from sensitive data. Our answers to committee questions below focus on how to derive useful insights from sensitive and non-sensitive federal data while addressing privacy and/or national security concerns. Even

sensitive data can be powerful engines of innovation, insights for public policy making and economic value when appropriately curated. The American public has already paid with their tax dollars for the collection of the information in federal databases, and they should be able to derive multiple benefits from these data.

The sheer quantity of data compiled by the federal government alone is growing rapidly, the private sector is compiling much larger databases from sources like consumer purchases and healthcare decisions, and millions of American citizens are recording details of their personal lives and thoughts through social media. Data formats continue to proliferate as legacy data are retained and new information protocols are developed; local and regional sensor networks are expanding, autonomously collecting information daily on environmental conditions and human conditions and actions.

As numerous recent reports and publications have noted that we urgently need to restructure how we manage and use this cascade of data--to craft better policies, stimulate innovation and economic growth, improve security, and mitigate the impacts of natural disasters, while at the same time ensuring individual privacy and freedoms. Because so much of these data are now in digital form and resides on a global network of servers, it will require new information processing and data curation tools to provide access to and syntheses of these data, new organizations to build and manage this cyberinfrastructure for data, new social institutions to develop and promote common standards and best practices, and new approaches to converting these data into policy- and program-relevant insights.

This commission cannot hope to solve all of the many issues surrounding the range of data from all sources, and address the potential benefits and concerns for its use. However, data gathered and compiled by federal, state, and local agencies are central to policy making, and could be managed in more effective and secure ways. Unsurprisingly, much of the administrative and survey data collected by these agencies is of the actions of people in public and private life and their social, economic, and health conditions. We represent a growing community of scientists who study human social and behavioral systems. We strongly support the goals of this commission to make more effective use of administrative and survey data to "strengthen evidence-building to inform program and policy design and implementation". Integrative access to diverse government data combined with next generation analytics can lead to a better understanding the complex dynamics of our social world. To enable these goals in an efficient manner, we envision a national clearinghouse for *integrative Human Analytics and Data Synthesis* (**iHADS**) to:

- Increase the value of government data by facilitating access to and use of administrative and survey data from multiple sources;
- Promote community-wide standards and best practices for data management and access to reduce administrative costs and create incentive for data sharing across agencies;
- Harmonize diverse data definitions and formats to allow analytical syntheses across information from multiple sources;

- Develop cyberinfrastructure and security to enable next generation analytics, synthesis, and modeling that enables evidence-based policy making;
- Provide an environment that enables exemplary research that uses government data to support evidence-based policymaking to benefit the American public;
- Negotiate legal agreements between data managers and credentialed users to lower regulatory barriers to proper data use while ensuring their safety;
- Protect private, confidential, and identifiable information, when needed, through standards-based credentialing of qualified users and differential privacy safeguards matched with user credentials and analyze the risks associated with the release of that information to the credentialed user.

A national **iHADS** clearinghouse will make the current dispersed enterprise more effective and generate significant cost savings for agencies that must manage data (including sensitive data), identify qualified users of data, and provide data to those users while protecting sensitive information. It could also offer significant cost savings to potential users by reducing administrative and regulatory barriers, improving data discoverability, and facilitating linkage across multiple datasets. Data are the most powerful engines of innovation and economic value when they are transparently available and openly accessible. Transparency promotes innovative applications of data for diverse purposes across the private and public sectors and for scientific research that can contribute to policy-making. Facilitating access to and use of such data also increases the return on investment to the American public that has already paid with their tax dollars for the collection and management of the information in federal databases, improving their opportunity to derive multiple benefits from these data.

The creation of new cyberinfrastructure to facilitate the use and synthesis of diverse data sources about human systems--including federal administrative and survey data--has the potential for important benefits beyond better policy-making, including personalized health care, targeted economic development, and accelerating innovation. Because human decisions and actions are major drivers in many of earth's biophysical systems today, next generation science of human systems, supported by advanced data synthesis and analytics can offer new insights into 'natural' systems that can impact human life and well-being. As specifically noted in our response to Question #6, below, we recommend that the Commission consider a study by an independent scientific organization as a guide for possible organization and implementation of such a national clearinghouse.

Below, we respond to the Questions for Response #1-17, posed by the Commission.

Overarching Questions

1. Are there successful frameworks, policies, practices, and methods to overcome challenges related to evidence-building from state, local, and/or international governments the Commission should consider when developing findings and recommendations regarding Federal evidence-based policymaking? If so, please describe.

The US is a world leader in making key government-collected data freely available, with significant benefits to business, science, and the general public; many of which are available through the central *Data.gov* website. A few examples of individual federal data sources that provide a broadly useful public good, and facilitate science and evidence-based policy making include: national income accounts data (http://www.bea.gov/national/Index.htm), the Center for Disease Control and Prevention *WONDER Database* (https://wonder.cdc.gov) and *National Center for Health Statistics* (https://www.cdc.gov/nchs/), *EarthExplorer* of the US Geological Survey (http://earthexplorer.usgs.gov), national agricultural statistics (https://www.nass.usda.gov), *Data Tools* of the US Census (http://www.census.gov/data.html), multiple databases of the Bureau of Labor Statistics (http://www.bls.gov/data/), and *Envirofacts* (https://www3.epa.gov/enviro/) and *Developer Centra*l (https://developer.epa.gov/category/data/) databases of the Environmental Protection Agency. State and local level examples include the *EnviroStor* database of California (https://www.envirostor.dtsc.ca.gov/public/), and the New York City *Open Data Dashboard* (https://nycopendata.socrata.com/stories/s/es39-eesr).

However, to fully benefit from the many data sources we have, it is necessary to be able to integrate data and make links across different datasets. For example, two types of linkages are particularly important for connecting across administrative and survey databases: 1) linking individual demographic data drawn from a survey with administrative data, and 2) linking individual and administrative information to geographic locations. An example of a successful linkage across government agencies is the Health and Retirement Study (http://hrsonline.isr.umich.edu). With respondent consent, individual data are linked to CMS (Centers for Medicare and Medicaid Services), NDI (National Death Index), and SSA (Social Security Administration) data. Although an invaluable resource for successful policymaking for the rapidly growing population of seniors in the US, the kinds of data linkages that make the Health and Retirement Study possible are very difficult to overcome by individual researchers due to the need to negotiate data access separately with each administrative entity, the need to identify and reformat potential linking fields so that the databases can 'talk with each other', and the need to remove IDs from any linkages to protect individual privacy.

A successful example of the second type of linkage (i.e., to geographic locations) is seen in research by the University of Chicago's Urban Center for Computation and Data and Chapin Hall, a research and policy center focused on a mission of improving the well-being of children, families, and their communities. These groups have negotiated agreements across different governmental agencies within the city of Chicago to link administrative and survey data with

information on places people live. One use of these data is to identify geographic areas where resources can be targeted to improve family well-being and reduce dependence on public services. This type of use does not require linkage to individual information, thus having much lower risk of disclosure. Many data sets lack links to individual demographic data on individuals but contain geographic identifiers on the location of the individuals--if they can be linked with other datasets, regulatory barriers to use can be overcome, and different ways of representing geographic location can be translated (e.g., latitude/longitude vs. street addresses).

Recent advances in cyberinfrastructure offer the potential to link federal, state, local, and private sector data in new ways for information syntheses that can significantly multiply the current broad benefits of public access to publicly funded information sources. However, there are significant administrative, regulatory, and technical barriers to doing so that greatly limit the ability of most researchers to accomplish this. And these barriers are multiplied by the need to anonymize potentially identifiable information and simple discoverability of relevant datasets among the multitude of sources. For these reasons, we suggest a national clearinghouse for *integrative Human Analytics and Data Synthesis* **(iHADS)** to facilitate the discovery of relevant data, negotiate access to multiple databases, credential potential users, manage security and privacy concerns in a unified way, and create cybertools to facilitate translation and linkages across different ways of representing information (e.g., location). In addition, it could advise on improvements in data collection techniques and statistical practices, review and suggest best practices for transforming data into information, undertake cost-benefit assessment of the public goods created, and point out inappropriate uses of data because of sampling methodologies used.

2. Based on identified best practices and existing examples, what factors should be considered in reasonably ensuring the security and privacy of administrative and survey data?

The great diversity in management practices, protocols for access, and implementation of security protocols across the many federal agencies that hold data raises concerns for the security and privacy of administrative and survey data. A set of commonly and equitably applied *standards to vet and credential qualified institutions and users* would be an important step in improving security and reducing administrative load on individual data managers across the federal system. Such common credentialing would also help identify which institutions and users are qualified to have what level of access to what kind of data.

Likewise, a *common set of data access portals* to federal administrative and survey datasets, physically housed in multiple agencies, would facilitate implementing state of the art authentication with cost savings to individual agencies while improving access by qualified researchers.

In some cases, administrative and survey data do not contain any information that could be linked back to an individual or specific business. And for databases that do contain identifiable information, only a limited portion of the data in a database is potentially linked to individuals or

businesses in a way that could compromise privacy or security. It is important for metadata about datasets to clearly identify which (if any) fields are of particular security and privacy concern because they provide identifiable information (e.g, Social Security Numbers, personal names, house addresses). *Common and widely applied standards for metadata* would facilitate sharing of non-sensitive data and help ensure that potentially identifiable information is carefully protected and not released to unqualified individuals or institutions. Likewise, standardized ways of representing information about which data have received informed consent, and the conditions under which such consent is applicable can guide data access that maintains security and privacy.

With standards for credentialing, data access portals with standardized authentication, and sensitive information clearly identified according to metadata standards, it should be possible to apply *differential privacy* controls on data to ensure that only individuals or institutions with the proper level of credentialing have access to sensitive information. More importantly, differential privacy protocols potentially allow for linking information across multiple databases using sensitive identifiable information as linking (i.e., "key") fields--but then not including those sensitive fields in the joined information tables returned to users requesting data.

Implementing such a suite of standards for accessing and integrating data across the many datasets held diverse federal agencies would be desirable but also a monumental task from a practical standpoint. Hence, we recommend focusing initially on a national facility, an *integrative Human Analytics and Data Synthesis* (**iHADS**) clearinghouse, to provide a visible and accessible entry point to federal administrative and survey data. Such a national facility could develop and apply a common set of state-of-the-art standards for vetting, credentialing, authentication, and differential privacy that represent community-wide best practices. Providing a national suite of data portals to the many administrative and survey databases across the federal system, all accessible with a single credential and authentication, with the possibility of querying information from multiple, linked databases would encourage widespread use of such a system. Federal data managers could continue to maintain data as appropriate for their agencies and direct potential users to the national **iHADS** portals, with savings in administrative costs while improving security and privacy. Similarly, it would be more efficient and less costly to maintain and update these standards in a national **iHADS** clearinghouse than to do so across the many agency data management offices.

Data Infrastructure and Access

3. Based on identified best practices and existing examples, how should existing government data infrastructure be modified to best facilitate use of and access to administrative and survey data?

The full value of government data as a public good can only be realized if those data are indeed used to provide benefits to American citizens. In order to facilitate use of and access to

administrative and survey data, existing government data infrastructure should be modified so as to both reduce barriers and create incentives to widespread and diverse use.

Barriers to use and access include those that inhibit the sharing of data by data providers and those that make access and use more difficult for data users. From the perspective of data providers, it would be helpful to reduce legal, bureaucratic, political, and jurisdictional barriers to data sharing. For federal databases, a common set of standard protocols, in compliance with federal privacy protection statutes, for identifying which individuals and institutions are qualified to access data of different degrees of sensitivity with regard to personal and other identifiable information would make it more straightforward for individual data managers to decide what data should and should not be shared with which users (see also Question #10). This spans data that should be broadly accessible by all members of the public to highly sensitive data that should be restricted in raw form to a limited number of most qualified users, but could be used by a much broader audience if identifiable information can be removed or anonymized through the application of differential privacy protocols (see comments for Questions #5 and #11).

The use of cyberinfrastructure to provide single sign-on authentication across multiple federal databases, for users vetted and credentialed with a common standard would reduce financial costs and administrative burden on agencies when making data accessible while ensuring statutory protection from disclosure. Such consistent access protocols, especially if administered through a national data clearinghouse charged with this responsibility would also make it much easier to access and integrate different datasets across agency jurisdictional silos. It would also be of considerable benefit to promote a data management culture that views administrative and survey data as ultimately belonging to the US citizenry that pays for is collection and upkeep. Such a culture would reward agencies and managers who are able to ensure that the federal data for which they are responsible is available most broadly for public use. Reducing administrative overhead and needless duplication of effort of making such data accessible across multiple agencies can help to promote such a culture of data sharing, and treating such data as a public good that can lead to transparency, reproducibility, and accountability in government. To the extent that efforts to reduce these administrative barriers to sharing data can be extended to state and local governments, it would further facilitate use of and access to data--adding important value to information collected and stored at these more local levels (see also Question #5).

Reducing barriers to data sharing within agencies equally benefits potential users of that data. Additionally, investments in cyberinfrastructure increase accessibility and ease of use from the user perspective. A key way to facilitate use is to significantly improve discoverability of government datasets. While some databases (e.g., those mentioned in the response to Question #1) can be found easily, many others have very limited exposure or are not exposed at all to Internet searches. Though Data.gov does provide a centralized search engine for identifying many federal datasets, it is still the case that many federal databases can only be found if users already know of their existence and where to look for them. New technologies to make data, their structures, and ontologies easily discoverable need to be implemented across the federal

data ecosystem. To the extent possible, it would be beneficial for protocols to improve discoverability be done in a systematic, standardized fashion to improve user ability to identify relevant information from multiple data sources in an equally standardized way. For example, a search on Google Scholar provides a unified query tool to locate many different kinds of documents, all of which employ a common set of meta-tags for exposure on Google (https://scholar.google.com/intl/en/scholar/inclusion.html#indexing). Because the number of federal databases is much smaller than the number of documents indexed by Google Scholar, a central index or 'library' of data sources that could be browsed or searched by title, topic, agency, etc is also feasible and would further facilitate access and use of such data. Such a library of federal data sources could be maintained and exposed to search by a national data clearinghouse.

A further technology-enabled modification to facilitate use and access to administrative and survey data is automated or semi-automated harmonizing of diverse data ontologies to allow seamless queries and linkages across multiple databases. As mentioned in the response to Questions #1 and #5, different datasets may be potentially linkable by information (i.e., fields) they hold in common, but the way that information is represented often varies from database to database. This requires a translation of the way information is represented (the text or codes in potentially common fields, and what they mean) so that different databases can be joined. This can be accomplished through translation dictionaries (when options are relatively few), sophisticated use of metadata (requiring that standardized metadata be present, of course--see Questions #2 and #4), user-assisted matching, or other means. At an even more fundamental, data may be stored in different file formats (e.g., SQL, DBF, CSV, fixed format, OGR, XLS, or many others). Linking different databases and integrating their information will require the ability to seamlessly read/import/export among these many file formats. While ontology translation and read/import/export capabilities could be implemented for each federal database, it would be more efficient to do so in a central data clearinghouse, such as the *integrative Human Analytics and Data Synthesis* (**iHADS)** facility proposed here, with significant costs savings to data managing agencies and greater accessibility for users.

***4. What data-sharing infrastructure should be used to facilitate data merging, linking, and access for research, evaluation, and analysis purposes*?**

The increasing availability of real time high-resolution social, economic, and environmental data provide unique opportunities to advance public policies to address complex societal problems such as human health, social disparities and the threats of climate change. Forging new ground in many policy areas requires expertise in multiple disciplines across the human sciences and the data and computational science. Billions of records are generated everyday by many Federal and local agencies from multiple resources including daily economic transactions and yearly tax reports, daily health reports, environmental monitoring, social media, and other sources. Many of these data can be geolocated and linked to detailed geographic and environmental features.

However, datasets do not have value in and of themselves, but only through their use--e.g., in research, analysis, modeling, and evaluation. It is what is then done with the data that creates value. As we note in the overview and in responses to many of the questions here, linkages among disparate government datasets can significantly increase their potential to provide useful information. One way to use such government data is to download a dataset or subset of linked datasets to a user's local computer to be analyzed with their preferred software. However, some powerful methodologies being developed for advanced data science require high-performance computing (HPC) environments that may not be available to individual users, small businesses, or even relevant departments in academic institutions. The ability to build on the (properly credited) work of others--a hallmark of modern science--allows for rapid innovation and generation of new knowledge. A national facility, such as the *integrative Human Analytics and Data Synthesis* (**iHADS)** clearinghouse, could provide an environment for advanced analytics and data synthesis (e.g., large scale modeling, machine learning, data mining, visualization), that cannot realistically be accomplished on even modern desktop computers available to many users. Advances in cloud computing, including software as services, mean that this environment could be deployed and used by numerous individuals over the internet, without needing to be physically present at a national facility. This would offer a visible exemplar of the potential uses and benefits of government data. Additionally, it could also carry out or contract advanced research, including look-ahead modeling, to support specific policy related requests.

Such an environment also presents the opportunity for research computing workflows, using government datasets, to be saved and published so that uses could better work in teams and learn from each other. In this respect, we envision an integrated technological infrastructure whose main goal is two-fold. The first is to let users share computational/analytic tools and products; the second is to share and reproduce all data synthesis workflows, for example by means of open source interactive computing environments (e.g., IPython/Jupyter and R notebooks) and shared metadata standards. Dedicated secure data enclaves connected to a grid-based computing platform can provide the necessary computational capabilities for analysis and processing of the data. These enclaves can be accessed in a secure way by means of Virtual Desktop Infrastructure (VDI) technology, so that users can interact with the grid, its data, workflows, and metadata from individual workstations at their home institutions. Secure data enclaves also mean that the inadvertent sharing of sensitive data is reduced since the actual data will be maintained in one location and not replicated to the individual researchers' machines. Establishing a national **iHADS** clearinghouse and research environment will enable scientists to unlock complex patterns across space and time to better frame problems and present policymakers with opportunities for solving them.

5. What challenges currently exist in linking state and local data to federal data? Are there successful instances where these challenges have been addressed?

Challenges to linking state and local data to federal data mirror the challenges of linking different federal databases, but are multiplied by increased diversity over management goals, practices, statutes and regulations, and technical issues across state and local jurisdictions.

There is variability in state laws regarding data confidentiality that facilitate or prevent access to government-collected data (e.g., access to vaccination registries varies from state to state). Related to this, there is lack of common shared standards for data protection and sharing across federal, state, and local levels. If a state or local government has statutes or regulations that serve to prohibit data access, it may be very difficult to attempt to link such data with relevant federal data. It would be beneficial to have a set of national guidelines and recommendations about data accessibility, coupled with protection of privacy. This does not guarantee that state and local governments would necessarily implement or follow those guidelines. However, the lack of common guidelines helps to ensure needless regulatory barriers to more effective use of information whose collection is publically funded.

Many state and local governments are inadequately staffed to curate and share their data, even where it may be allowed, or even required, by statute. This is particularly a problem for identifying, credentialing, and authenticating qualified users--magnified by inadequate technology that may require this to be done manually. Moreover, data may not even be digitized in a way that facilitates use across agencies (e.g., HIV diagnosis surveillance data is kept on paper note cards for the state of Massachusetts). Another critical technical issue is the lack of identifiers held in common by local, state, and federal databases that could serve to link across different information sources.

These many challenges mean that there have been few successful instances for integrating data across different levels of government. One example is *Medicaid Analytic EXtract* (MAX) system (https://www.cms.gov/research-statistics-data-and-systems/computer-data-and-systems/medicaiddatasourcesgeninfo/maxgeneralinformation.html), where Medicaid encounter records across different states have been linked and compiled into a more research-friendly set of Medicaid administrative files. This data product benefits both the federal government as it manages the Medicaid program and monitor the progress of the health care delivery system nationally and researchers studying ways to improve health care.

Another example, mentioned in the response to Question 1, involves University of Chicago's Urban Center for Computation and Data and Chapin Hall. They have negotiated agreements across different governmental agencies responsible for different aspects of the city of Chicago to be able to link administrative and survey data on people with information on places where they live. As noted above, these data are being used to identify geographic areas where resources could be targeted to improve family well-being and reduce dependence on public services. The New York City Department of Health also links data from surveys, from traditional public health

surveillance data, and from agency and administrative data (education, planning, police), and observational data (social media, sensors), to geographic locations in the city to improve public health and mitigate disease (e.g., flu) outbreaks.

The proposed **iHADS** clearinghouse could help to address some of the issues facing linking data from federal sources with that from state and local governments. It could help develop national guidelines and best practices to facilitate such intergovernmental data linkages. It could also further develop and deploy cybertools to overcome some of the technical incompatibilities created by different ways of representing data. For example, methods of probabilistic matching exist and could be employed to link records where there are no common identifiers to serve as key fields. More than anything, a national facility like the **iHADS** clearinghouse could provide a powerful exemplar of the potential for evidence-based policy making to encourage the adoption of best practices by state and local agencies.

Although we propose that a national facility like the **iHADS** clearinghouse focus initially on federal databases, we also think that state and local governments could potentially also become data portals, in much the same way that Data.gov provides links to city, state, and federal datasets. This would give state and local governments the ability to adopt and benefit from common standards of credentialing, authenticating, differential privacy protocols, and querying across multiple databases. Because an **iHADS** clearinghouse is envisioned to serve more as a portal to data stored elsewhere than a data storage facility, it would benefit greatly by economies of scale in expanding its services to state and local agencies. As noted above, this could add significant value to state and local databases while lowering costs and improving security.

6. Should a single or multiple clearinghouse(s) for administrative and survey data be established to improve evidence-based policymaking? What benefits or limitations are likely to be encountered in either approach?

There would be a high payoff to having a national clearinghouse for *integrative Human Analytics and Data Synthesis* (**iHADS**). Such a clearinghouse could deal with the many of the challenges involved in increasing and coordinating access to administrative data from federal, state, and local governments. This clearinghouse could assemble expertise in the legal, computational, analytical, and human subjects issues required to successfully increase access to administrative data that have thus far been considered too sensitive for use by researchers. The clearinghouse could develop standard protocols and data use agreements, informed by research on disclosure risk, that could be useful in creating an environment in which federal, state, and local agencies become comfortable with increasing access to administrative data. The clearinghouse could also work with agencies in developing protocols for linking administrative records. A single clearinghouse with this mandate could be much more effective and efficient in making these arrangements than the current situation in which individual researchers and institutions approach agencies on an ad hoc basis. Finally, a national **iHADS** clearinghouse could

also enable exemplar advanced research to support evidenced-based policymaking and demonstrate the value of federal data.

A number of research data repositories also exist, generally associated with universities, for survey data, census data, genetic data, health data, education data, etc. Most of these focus on data collected in the course of scientific research (often federally funded). Like federal databases (as discussed in Question #7), it would be counter-productive and cost prohibitive to physically incorporate them into a single, national clearinghouse. However, a national **iHADS** clearinghouse could collaborate with scientific research networks to establish convenient portals and links these research data repositories that address different types and levels of data than government data. There would be significant returns to efforts to encourage and facilitate such coordination, avoiding duplication in data management, enabling and facilitating linkages between federal data and research databases, and sharing knowledge and cyberinfrastructure for data analytics. Because it is no longer necessary to have data physically located in a single locale or coterminous with resources for analysis, the most important goals should be facilitating access to and use of the data currently collected, curated and disseminated in multiple data repositories.

We recommend that the Commission consider a study by an independent scientific organization such as the National Academies of Sciences, Engineering, and Medicine as a valuable guide for the organization and implementation of such a national clearinghouse like the proposed **iHADS**.This could coordinate with and build on an ongoing study by the Council of Professional Associations on Federal Statistics (COPAFS), supported by the Sloan Foundation and American Economic Association on how to improve access to federal administrative data for evidence-based policymaking.

7. What data should be included in a potential U.S. government data clearinghouse(s)? What are the current legal or administrative barriers to including such data in a clearinghouse or linking the data?

Survey data is taken to include data typically collected through the Federal Statistical System that generates information about broad population characteristics as well as data typically collected and associated with a specific evaluation or other study that may be started and developed directly through agency evaluation functions or indirectly by a funding recipient. Examples include U.S. Census Bureau data on individuals and businesses, and data from national and local surveys. In a broader sense, survey data can might also encompass other forms of observational data like US Geological Survey land-cover data, National Oceanic and Atmospheric Administration night lights, and satellite remote sensing data from NASA.

Administrative data refers to administrative, regulatory, law enforcement, adjudicatory, financial, or other data held by agencies and offices of the government or their contractors or grantees (including states or other units of government) and collected for other than evidence-building purposes. Administrative data is typically collected to carry out the basic administration of a program, such as processing benefit applications or tracking services received. Any of these

data can (but do not necessarily) relate to individuals, businesses, and other institutions. Examples include national data collected by the Centers for Medicare and Medicaid Services, National Death Index, Social Security Administration, Federal Bureau of Investigation and Department of Justice (uniform crime reporting), Housing and Urban Development (housing), Health and Human Services (health and social services), and Immigration and Naturalization Service (immigration).

A national clearinghouse would offer the greatest benefit to users and to agency data managers if it facilitated access to the broadest possible spectrum of government data, but it should not be necessary to physically house most of those data in the facility. Rather, a national clearinghouse should employ advanced cyberinfrastructure (e.g., grid computing and software-as-a-service) to provide online *portals* to government databases physically housed within the agencies responsible for their collection and management. Only federal data for which no responsible managing agency currently exists (e.g., legacy data) might need to be physically stored at the national clearinghouse. Serving as a *meta-portal* to many diverse datasets, a national clearinghouse could focus on providing open access to non-sensitive data, facilitating data discovery, authentication of potential data users, application of differential privacy to protect sensitive information, and harmonizing data ontologies and formats to facilitate sophisticated data use and analytics (see Question #3). It also could coordinate the development and promulgation of common standards and best practices for federal data access and use. This avoids duplication of the effort and cost of large data storage and management, and the need to constantly synchronize duplicate copies of dynamic federal databases, while allowing other agencies to focus on their core missions. Rather than investing in large and quickly outdated banks of data servers, a national clearinghouse could focus its resources on the development and deployment of cyberinfrastructure to facilitate broad and secure access to federal data sources and a robust and comprehensive suite of next generation analytics for this data.

As discussed in more detail in Question #9, barriers to a national clearinghouse include diverse access protocols and interpretations of privacy protection statutes and regulations, jurisdictional silos, discoverability of federal data sources, diverse data ontologies and formats, and lack of pan-federal standards for metadata and other best practices for data management and accessibility. However, these are more effectively addressed at lower cost in the context of a national clearinghouse for *integrative Human Analytics and Data Synthesis* (**iHADS**) envisioned here than attempting to restructure the entire federal data ecosystem.

8. What factors or strategies should the Commission consider for how a clearinghouse(s) could be self-funded? What successful examples exist for self-financing related to similar purposes?

We strongly endorse a principle of open, unfettered access to all federal data, including non-sensitive data and those aggregated from sensitive data. Data are the most powerful as engines of innovation and economic value across both private and public sectors when transparently available

and openly accessible. Transparency and accessibility also facilitates research that can contribute to more effective policy-making. The American public has already paid with their tax dollars for the collection and management of the information in federal databases; thus, earning the opportunity to benefit from the shared national value of such data and arguing for broad public support.

A national **iHADS** clearinghouse for government data could generate significant cost savings while increasing return on investment for agencies that must manage sensitive data, identify qualified users of data, and provide data to those users while protecting sensitive information. It could also offer significant cost savings to potential users by reducing administrative and regulatory barriers, improving data discoverability, and facilitating linkage across multiple datasets. As discussed in responses to the other questions, it could significantly leverage the inherent value of federal data by creating technical capacity and best practice standards to link it with state, local, and research databases further increasing the returns to the American public.

The creation of new cyberinfrastructure to facilitate the use and synthesis of diverse data sources about human systems--including federal administrative and survey data--has the potential for important benefits beyond better policy-making, including personalized health care, targeted economic development, and accelerated innovation. Because human decisions and actions are major drivers in many of earth's biophysical systems today, next generation science of human systems, supported by advanced data synthesis and analytics can offer new insights into 'natural' systems that can impact human life and well-being.

As cost savings and benefits of coordinated and open access to government data grow, we expect that states and communities that can benefit by more accessible, integrative data will seek to participate in data integration initiatives. Many cities are already investing in data infrastructure of the type suggested here (see Question #1). Over the long term, a consortium model of producers and users, including data agencies, research institutions, the private sector, and non profit organizations could contribute to further development of a new national data infrastructure.

9. What specific administrative or legal barriers currently exist for accessing survey and administrative data?

Data privacy and confidentiality as well as data ownership are important barriers to all data exchange and access. All forms of data are governed by some type of access rule or restriction that are defined in the context of data collection. For instance, survey data are usually protected by informed consent and require that a respondent's confidentiality is ensured in data distribution. The methods for protection include limiting identifying information and protecting against deductive disclosure in public use data files and/or developing restricted use agreements and contracts to gain access that limit access through a legal agreement.

Federal agencies, in particular, have different processes to help data analysts gain access to the data, in part, because agency surveys are often covered by different privacy legislation and may require different legal agreements. Administrative data such as Medicare claims, social

security files, and other federal records are protected by different data use agreements to protect individuals' privacy and again vary substantially by agency and data source. The Center for Medicare and Medicaid Services, for instance, centralizes access to all of their data files through a secondary vendor at the University of Minnesota but the Census Bureau requires an application to use data in a decentralized virtual enclave in university settings. Similarly, privately held data such as social media or private organizations' records from vendors such as health care institutions and insurers require each analyst negotiate separate legal agreements and limit the amount of information that analysts can access.

The idiosyncratic and variable nature of the legal agreements necessary to access federal data alone often deters use. A centralized source of guidance and advocacy such as a clearinghouse for *integrative Human Analytics and Data Synthesis* (**iHADS**) would expedite use and linkage. Moreover, the ability to centrally credential individuals access to data would expedite the greater use of a wide variety of data sources, since currently individuals have to separately credential to each data source that they want to access. It would be especially valuable if the **iHADS** could act as an institutional agent for individual researchers because data use agreements often require an analyst to be in an institution that is willing to take on liability for violation of those agreements and many researchers do not or cannot make that commitment.

10. How should the Commission define "qualified researchers and institutions?" To what extent should administrative and survey data held by government agencies be made available to "qualified researchers and institutions?"

Researchers and research institutions are fundamental drivers of economic and technological innovation. Well-managed access data serves as a research and innovation accelerator. To accomplish equitable and informed vetting and credentialing of institutions and individual researchers, it is desirable to have a common set of standards and terms of use agreed on by the community of stakeholders, including data managing agencies and potential data users--including academic and research institutions that are already vetted and credentialed to carry out research by nationally recognized scientific and scholarly agencies like the National Science Foundation, National Institutes of Health, and National Endowment for the Humanities. Criteria for identifying qualified researchers and institutions should apply to quality of the potential user and not the ability to pay for or provide services in exchange for data access. These standards need to be applied fairly to all institutions and individuals who might seek to use administrative or survey data. We suggest that a central facility, like the proposed *integrative Human Analytics and Data Synthesis* (**iHADS**) clearinghouse could serve to develop and apply these standards to evaluate qualified researchers and institutions.

Peer review has been a powerful and widely successful model for evaluating research quality and for allocating research funding from science supporting agencies like NSF and NIH for institutions and individuals. A similar model would be an effective method to help define standards and procedures for applying them to identify qualified researchers and institutions.

This should involve assessment of the capacity of institutions, researchers within institutions, and independent researchers to follow best practices for secure data management, protection of privacy, ethical use, appropriate research protocols, and open dissemination of results so that they are available for public policymaking, as well as use by other researchers and the private sector. Different institutions and researchers may potentially be qualified for different levels of access to some data sources. Where sensitive information is involved, this can be managed through differential privacy protocols applied to vetted and credentialed users, which automatically give them appropriate level of access.

11. How might integration of administrative and survey data in a clearinghouse affect the risk of unintentional or unauthorized access or release of personally-identifiable information, confidential business information, or other identifiable records? How can identifiable information be best protected to ensure the privacy and confidentiality of individual or business data in a clearinghouse?

Currently, there are diverse, sometimes conflicting, standards--or even no standards at all--for management of federal administrative and survey data. This data ecosystem is even more chaotic when state and local governments are included. Some data are well-organized, carefully managed for appropriate access, and provided to vetted researchers. Other data sit on discoverable servers that make datasets available to anyone with the appropriate FTP address or URL. Yet other data are locked from any outside access, even though they do not contain sensitive information or are inaccessible to reputable researchers but stored on systems easily hacked by disreputable individuals.

A national facility, like the proposed *integrative Human Analytics and Data Synthesis* (**iHADS**) clearinghouse could develop and apply a consistent body of standards, access protocols, and credentialing tailored through differential privacy protocols to the privacy needs of each dataset and the credentials of each institution and/or researcher requesting access. By so doing, it could begin to establish and be an exemplar of best practices for governmental (and other) administrative and survey data. It would also be an active advocate for applying such best practices, and could help agencies identify places to improve data management protocols. Moreover, as discussed in the answer to question #4, by providing secure data enclaves for the analysis of the data, a clearinghouse, such as the one envisioned here and in other questions, would reduce the number of replications that exist of sensitive data and thus improve the overall security of that data since individual researchers would not be responsible for the security of the data, as they would be if they downloaded the data to their local machines.

That is, a central clearinghouse would actually provide more effective security for potentially identifiable information as an honest broker and responsible gatekeeper, than the current federal data ecosystem where each data manager for each data set within each agency at federal, state, and local levels independently determine how that data should be managed and protected. A

central clearinghouse like **IHADS** has the potential to help individual agencies better protect identifiable information than they do now, and with lower administrative overhead and cost.

12. If a clearinghouse were created, what types of restrictions should be placed on the uses of data in the clearinghouse by "qualified researchers and institutions?"

Data are a most powerful engine of innovation, economic growth, and effective policy when transparently available and openly accessible. The goal of a facility, like the proposed clearinghouse for *integrative Human Analytics and Data Synthesis* (**iHADS)** should be to make data collected and managed with public funds as accessible as possible in and enable the widest and most advanced possible use by qualified individuals and institutions institutions, while ensuring data integrity and preventing inappropriate or unauthorized access to identifiable information. That is, a clearinghouse should focus on reducing barriers to data access and use, rather than restricting use, so that government data can provide maximum value for US citizens, businesses, and research institutions—and can be most effectively used as a foundation for evidence-based policymaking.

As mentioned in the response to Question #10, clear standards and peer-review can help ensure that only qualified users access federal data with identifiable information through a national clearinghouse like iHADS. Moreover, differential privacy controls and behind-the-scenes linkages that subsequently remove identifiable information prior to returning it to data requestors can help make a much wider array of federal data openly accessible and useful than would otherwise be possible. As noted in the response to Question #9, centralizing legal responsibility to maintain data integrity and individual privacy in a national clearinghouse also makes publicly funded information more accessible and increases value to cost ratios. It would also be much easier for a centralized clearinghouse to update its policies with respect to how to best anonymize and privatize data to keep up with new scientific findings in this space; thus, it could ensure that the data was used in a responsible manner.

13. What technological solutions from government or the private sector are relevant for facilitating data sharing and management?

In addition to technologies discussed in response to other questions, there are a number of technological solutions are being applied in metadata standards, the private sharing of data, and even open access to private data that would be useful to build upon in facilitating data sharing and management. For instance, there has been considerable work done recently in metadata standards that attempt to provide standardized ways of describing data to make those datasets more useable and more searchable by scientists interested in working with those datasets. One recent group working in this space that seems relevant is the Open Collaboration Data Factories/Exchange (www.ocdx.io), an NSF-funded organization that is attempting to create

metadata standards, as well as query and analysis tools that work with those metadata standards to provide more transparent access to open online community data.

OCDX and other groups have also built upon the recent trend in cloud computing to create containerized space using tools like Docker and Kubernetes. These are self-contained computational entities that use operating-system-level virtualization to contain the work of a particular user. Utilizing these tools it is possible to create a separately contained virtual data enclave that secures data for one user, or group of users to use. This means that there is little to no risk of cross-contamination of data even when it is stored on the same physical machine.

There have also been advances in making private data available for social insight analysis. This has resulted in the creation of new standards and methods for anonymization. For instance, the JPMorgan Chase Institute, a new research think tank in DC, is planning to make detailed financial data available to researchers who are involved with the institute, and give them the tools to analyze the data for the advancement of science. These standards and solutions they develop could be most effectively developed and deployed across the federal data ecosystem in a national clearinghouse, like the *integrative Human Analytics and Data Synthesis* (**iHADS**) facility proposed here.

14. What incentives may best facilitate interagency sharing of information to improve programmatic effectiveness and enhance data accuracy and comprehensiveness?

Federal agencies ultimately are established to serve the American public. When the mission of an agency involves the collection of data, it can only serve the public if it also makes these data accessible to the extent possible, while still protecting sensitive components of those data. Hence, the obligation to share information must be built into the mission of any data-collecting agency—and especially so for improving public policy and program effectiveness. Federal agencies that support science require all recipients of federal funding to submit and then follow comprehensive data management plans. These plans must include details of how data collected with public funds will be sustainably curated and made accessible for broad public access, while protecting any sensitive information from disclosure. The same requirements should be extended to all federal agencies that collect and manage data.

In real-world settings, there are always limits and challenges to the accuracy and comprehensiveness of any dataset. Hence, agencies should be rewarded, rather than punished when efforts to make data accessible and transparent identify data gaps and inaccuracies so that they can be remedied. Conversely, agencies should be held accountable if they inhibit the opportunity to improve data quality by restricting access to those data.

A national clearinghouse, like the proposed *integrative Human Analytics and Data Synthesis* (**iHADS**) facility, could provide templates and guidance for agency data management plans. As a national portal to facilitate data discovery, linking across datasets, user credentialing, and next generation data synthesis it could provide significant and substantive help to agencies for

carrying out those data management plans and meeting mission obligations for data sharing (see overview and responses to Questions #2, 3, and 8). As discussed in these responses, this would both reduce costs and increase return on investment in data collection and management.

Data Use in Program Design, Management, Research, Evaluation, and Analysis

15. What barriers currently exist for using survey and administrative data to support program management and/or evaluation activities?

In many ways, the barriers to using survey and administrative data for program evaluation are the same as those to accessing this data, discussed in Questions #1, 3, 5, 7, and 9. If relevant data cannot be discovered, cannot be accessed by qualified researchers, cannot be made available due to insufficient agency staff, have incompatible ontologies or formats, or are restricted by bureaucratic or legal issues that data cannot be used to support programs or reliably evaluate their outcomes. Again, as discussed in responses to other questions, standards-based best practices should be used for making program activities and outcomes as accessible as possible. Program activities and outcomes are what generate new federal data. These data need to be transparently and openly available, while managing any relevant privacy and security issues as discussed in responses to other questions, in order for programs to be evaluated on the basis of factual evidence. Such transparency will also make it easier to identify any issues or gaps in data quality, completeness, or accuracy.

16. How can data, statistics, results of research, and findings from evaluation, be best used to improve policies and programs?

The development of policies and programs are ultimately the responsibility of policy and program makers, guided by elected officials who represent US citizens. Regardless of the policy or program enacted, it is vital to know whether the extent to which it is or is not meeting its intended objectives—which can only be accomplished by the collection and analysis of data that reports on the activities and outcomes of the policy/program. The evaluation and improvement of a policy or program should not be based on anecdotal accounts alone. Quantitative, statistical evaluation, carried out by qualified researchers, can be designed to minimize observer biases inherent in the observation and intuitive assessment of program activities and outcomes. Evaluating policies and programs in a way that can reliably identify both successes and failures requires information based on data that are collected using the best available analytical techniques, guided by the needs of policymakers and program designers. Answers to previous questions have suggested a range of activities that will improve program and policy evaluation.

To further ensure the reliability and confidence in program and policy evaluations, it is becoming possible to archive entire analysis workflows to provide a transparent to procedures and data used for this essential task. If questions arise with regard to such evaluations, the

relevant workflows can be independently reproduced and assessed. Such transparency also provides the basis for ongoing improvements in programs, policies, and the protocols used to evaluate their activities and outcomes. The potential for such open scrutiny can foster greater confidence in government. A clearinghouse like the proposed *integrative Human Analytics and Data Synthesis* (**iHADS**) facility could act as an independent mediator between information needs of policymakers and human systems analytics scientists using data.

17. To what extent can or should program and policy evaluation be addressed in program designs?

Evaluation cannot take place in an information vacuum. All new programs and policies should include sufficient resources to allow ongoing evaluation activities. As an honest broker and responsible gatekeeper of federal data, the proposed clearinghouse for *integrative Human Analytics and Data Synthesis (***iHADS)** could guide development of tools that would facilitate access to existing data, and develop and promote best practices for new data gathering and curation efforts to improve evaluation(see response to Question #16),

In addition to ongoing evaluation of existing activities, proposed new programs and policies should also be subject to evaluation. The improvements in data management proposed above would contribute to more robust program and policy design. An **iHADS** clearinghouse could also function as a clearinghouse for development and curation of evaluative tools that could be built into program designs (see response to Question #16).