MEETING SUMMARY

Issues Resolution Workshop

PARTICIPANTS (IN-PERSON): Steve Andrle, Aladdin Barkawi, Tim Barnett, Alan Blatt, Steven Buckley, Connie Citro, Charles Fay, King Gee, Alyssa Hernandez, Pam Hutton, Paul Jovanis, Skylar Knickerbocker, Jan Laaser-Webb, Yingeng 'Eric' Li, Kathleen Linehan, Timothy McDowell, Charles Meyer, Yusuf Mohamedsha, Jeffery Muttart, Nichole Oneyear, Miguel Perez, David Plazak, Chad Polk, R.J. Porter, Raghavan Srinivasan, Christian Richard, Peter Savolainen, Jen Smoker, Carol Tan, Zongwei Tao, Derek Troyer, Zhenyu Wang, Hao Xu

PARTICIPANTS (PHONE): Sandra Larson, Suzie Lee, Darren McDaniel, Omar Smadi

COPY TO: Aladdin Barkarwi, Pam Hutton, Kelly Hardy, David Plazak, Steve Andrle, Kathleen Linehan, Paul Jovanis

PREPARED BY: Chad Polk

MEETING DATE: Wednesday, April 27, 2016

MEETING TIME: 8:00 am – 4:40 pm EST

VENUE: Keck Center, Washington DC with Call-In Capabilities

Executive Summary

During the Issues Resolution Workshop, several important issues evolved from the active participant discussion:

- Concerning progress on Phase 2 of the Implementation Assistance Program (IAP), time is short and it is urgent for teams to move projects going so as not to let the funds lapse.
- Personally Identifiable Information (PII) continues to be a significant challenge, but there are mitigation measures that were identified in the IRW.
- It was broadly recognized that there is continued demand for additional access to PII data. Remote or project-driven enclaves, such as in use with census data, offer opportunities to meet this demand, while continuing to protect study participant identity.
- Data users benefit from knowing the process undertaken for data acquisition and the schedule risks
- Increasing communication between VTTI and data users will enable logjams to be broken more efficiently, and research schedules to be adjusted accordingly.

Background

The idea for a SHRP2 Issues Resolution Workshop (IRW) developed during a Transportation Research Board (TRB) Safety Data Oversight Committee meeting held in late October of 2015. The presentation summarized comments and questions collected from Implementation Assistance Program (IAP) researchers who had used the data during SHRP2 Phase 1 efforts. The SHRP2 Safety Task Force subsequently supported the idea as well.

The mission for the workshop was to discuss challenges encountered during previous efforts that utilized NDS and RID data, in order to identify enhancements or improvements to the process of data access and analysis. Specific goals included:

- Receive input from **users** of NDS and RID databases
- Receive input from **providers** about processes necessary to complete data collection requests
- Discuss ways to streamline requests and/or improve customer service after requests are initiated
- Arrive at "actionable resolutions" to improve the process for everyone moving forward
- Build stronger communication links between users and providers

The workshop agenda is included in this meeting summary as an attachment. This meeting summary is structured to provide a more cogent description of the issues raised and the discussions that ensued. As such it deviates slightly from the agenda.

Efforts to Date Addressing Known Concerns

TRB, Virginia Tech Transportation Institute (VTTI), and the Center for Transportation Research and Education at Iowa State (CTRE) opened the workshop with a presentation on issue resolution efforts already underway. Numerous charts and reports were referenced during the discussion (and may be included as an attachment at a later date, if permission is granted) to supplement the discussion.

Process of Data Acquisition

Considerable discussion focused on actions taken to provide accurate real time estimates for the time required to complete data acquisition. Notable discussion points included:

- Research teams seek greater transparency related to estimated turnaround times for data requests and are interested in the best ways to track requests. VTTI staff, particularly

Miguel Perez, was identified as the best source for this information and teams can feel free to call him at any point.

- Data requests are not processed as "first in, first out" due to the varied nature of each request. IRW attendees asked that data acquisition requests from IAP teams should be prioritized at or near the top of the list.
- After submitting data requests, the requestor should be contacted by VTTI within 48 hours. Participants agreed that auto-reply emails should be sent when data requests are submitted. These emails should include notification to contact within 48 hours along with contact information in case they are not contacted within that timeframe.
- The InSight webpage will be updated to include a clear set of Frequently Asked Questions (FAQ) that include tips for data requestors and for contracting. Additional documentation includes cost estimate guidance from the exemplar document (Shelton, et al., 2015),
- There are 4 separate Data Use Licenses (DUL. Each form has progressively more information required for the request. VTTI uses the highest level one that incorporates the nature of the request. VTTI will provide the correct form to the requestor based on specific user needs and will coach users through this requirement, if necessary. The four basic forms are:
  - Data available on InSight
  - Data in depth – beyond InSight
  - Using a Secure Data Enclave
  - Executing an Algorithm within the Enclave
- Users stressed the need to mitigate future delays by informing requesters of what to expect and what is required to make the process most effective. The list below was discussed and should be expanded upon as part of the InSight FAQ's for future data requests:
  - Expect to interact 1-20 times with VTTI/CTRE within the process, depending upon the thoroughness of the initial application and the complexity of the request.
  - Once the paperwork is finished, extraction of data can take as little as 7-10 days.
  - There are 4 steps in the data acquisition process. There is overlap in the 4 processes but they are also somewhat distinct. The steps should be initiated in parallel as much as possible. Researchers should work on data use licenses and contracts simultaneously.
  - It is imperative for the researcher to understand their institutional review board's operational needs and protocols.
  - Requests are assigned 1 or 2 analysts but the VTTI leader on data acquisition (currently Miguel Perez) can follow up for any concerns or lack of feedback. Copying the data acquisition lead on emails with the analysis helps track progress in cases involvement is needed at a later point.

  - Currently the human subjects lead at VTTI (Suzie Lee) has 2 staff working on data licensing; the data acquisition leader has 5 data analysts (1 administration staff, and 2 statisticians).
  - Future reliability will depend on balancing level of business with staffing constraints and ability to train. It is expected to be consistent but hard to forecast.

Data Processing and Analysis Enhancements for RID

- CTRE addressed what has already been initiated to address the quality of data available, mobile data collection, coordination with other state data sources and privacy limitations.
- Battelle has reviewed RID speed limit data, explaining that speed limits are based on where signs were located at the time of mobile van data collection. Approximately 70% of the links traveled by NDS drivers were captured by mobile data collection.

Discussion of Topics Previously Identified by Users

During the meeting, the facilitator led a discussion of issues raised during Phase 1 of the AASHTO Implementation Assistance Program (IAP). The following is a summary of those discussions. The structure of this section of the report is slightly different from the meeting agenda (see Appendix A) to facilitate comprehension of the discussions.

Structure of the Database and its Implications for Users

Factors that drive the cost of data acquisition

- Costs are derived from the time needed to address requests; they depend on what data are being requested. *InSight FAQ webpage will be updated to clearly explain this information to research teams.*
- Costs are also related to the impact on analysts' time. Once queries are set up, subsequent requests for similar data (i.e. larger request, same parameters) will cost less than the original request.

Structure of the database

- The best example of what is available in the NDS database is the open-source licensed training data set available for download on InSight. This information is a smaller data set without PII limitations that can be used to inform requestors prior to submitting their requests.
- Size limitations – to date, the largest data sets requested have included a maximum of 10,000 trips. Requestors should consider schedule when determining size of requests.

Larger requests require more processing time – such requests may require waiting for months and/or incurring premium charges to acquire these larger data sets.

- Changes to database are being documented to accurately track the source for any data set used in practice.
- VTTI clarified that using InSight allows one to see data but not download it. The original data is available with a DUL. InSight doesn't require a license based on its functions but a DUL is required to receive the raw data (this data has to be destroyed in 30-40 years according to the IRB-approved research protocol so it must be tracked.)
- Aggregated data is available to copy – but raw data requires a data use license.

Controlling User Costs

- Costs typically driven by the amount of time needed by VTTI staff.
- When using algorithms to process data, one alternative is to bring the algorithm to the enclave with the possibility of reducing costs. Testing algorithms in the data enclave was proposed as an alternative to sending larger data sets to a user and then have the user apply the algorithm at their site.

Understanding PII and Its Implications for Data Analysis

Personally identifiable information (PII) was discussed at length as well as its implications for data access and analysis. A pending Battelle report to TRB is anticipated to help clarify the risks of participant re-identification from a range of data sets.

Generally, re-identification was recognized as more difficult under several sets of circumstances including:

- Changing continuous variables to categorical, obscuring the ability to precisely identify an individual's characteristics (often termed "binning" data)
- Building more aggregate categorical variables; increasing the size of the pool of possible participants within a category
- Combining crash levels (near misses with crashes) to again increase the sample size of potential participants
- Providing attributes of crashes without identifying the precise GPS location. This strategy would be most effective when the combination of attributes led to a sample size sufficient to reduce the risk of re-identification by use of the attributes. It is important to recognize that such a sample size requirement must make sense from a statistical or analysis perspective.

Alternative Enclave Structures

The development of alternative data enclave structures evolved during discussions with the NAS IRB Chair. Precedents in other fields were identified in which data sets including PII would be resident at remote sites under specific limited conditions. These project-driven enclaves would contain PII, but for a specific problem or analysis goal, and strict conditions assuring protections on site. Enclaves are most relevant to NDS data concerns, but some of the re-identification threats are tied to the linking of NDS and RID information. Among the topics discussed during this broad-ranging exchange were:

- FHWA's STAC will open at Turner Fairbank in the summer of 2016. This enclave will have access to a very broad range of data through a secure link to VTTI. *Post IRW Note from TRB: It is likely that any other enclaves considered during Phase 1 will employ the secure data link approach as opposed to physically copying data.*
- Additional similar enclaves in the West or Midwest (with secure links to VTTI) would be very helpful in improving PII access while protecting subject identity.
- More information is needed regarding requirements and necessary costs for both project-driven enclaves and additional STAC-like enclaves in the West or Midwest.
- A committee (*Post IRW Note: Federal Advisory Committee Act*) stated during Phase 1 that cost and operational impact will limit the ability to make copies of portions of the raw data. This should be further articulated, particularly as it applies to these enclave discussions (can be communicated to the Federal government due to the nature of the committee).

Accessing crash location information

In general, this discussion provided researchers a better understanding of the IRB concerns associated with providing crash location data. These concerns focused on participant re-identification by linking specific SHRP 2 NDS crash location information (e.g. using GPS coordinates) with other data sources containing location, date and time of crashes.

- Re-identification risks were discussed at some length. Preliminary study results from TRB-sponsored study (SD03) indicate that removing variables doesn't always decrease the risk of re-identification.
- Creating data categories with a larger sample size of participants in each analysis unit generally reduces the risk of re-identifying a particular individual, but at the cost of specificity in the analysis,
- Adding near misses to similar crashes would also increase the sample size of participants, generally reducing re-identification risk. There is an emerging literature on the topic of identifying similar crashes and near crashes

- Attendees were informed of the prevalence of individuals skilled in using multiple disparate databases to re-identify individuals. These individuals pose a threat of re-identification to NDS and many other databases.
- Licensing remote enclaves could allow full use of data while maintaining privacy commitments. This includes project-driven enclaves with specific PII released to persons or institutions for a specific duration of time (ultimately to be returned). This offers an opportunity to access PII when other analysis options are not available. *Post IRW Note from TRB: The potential trade-offs to the convenience of this data are risk and cost.*
- Discussion of remote, project-driven enclaves vs. geographic enclaves (such as the STAC) included:
  - STAC is a secure enclave and is anticipated to have access to PII. *Post IRW Note from TRB: In the beginning, the access for the STAC will be electronic (API) as opposed to physically copying PII data.*
  - The issue comes back to risk. There is a need to balance the potential safety benefits of a particular analysis with the risk of re-identification. The discussion was generally favorable in support of the creation of these types of enclaves for NDS. Challenges remain in establishing the protocols of such data use and detailed requirements for such facilities.

Precision of Data within the NDS and RID

- Coders of original database were not highway design engineers, which led to terminology differences between database and common terminology used by highway practitioners. If costs permit, the database terminology (field names) may be able to be updated. Another suggestion was to create a legend that links the terminology used in the database to common practitioner terminology.
- It is not possible to reliably obtain lane usage from GPS but there may be systems available that can provide such analysis.
- CTRE provides whatever data dictionary they received from the specific state; they have not attempted to enhance those documents.
- Speed limits change periodically and are difficult to track accurately. The speed limit data in the RID needs to be static and represent the speed or speeds that were present at the time the NDS data were collected
- Concerning the RID, documentation now describes what is available – but users expressed a need for more information on sources of data and what elements are in each layer.

Data Processing and Analysis Enhancements

Responses to data errors

- The NDS and RID are large and complex databases; errors are to be expected. Periodic updates of database have been occurring and will continue based on concerns raised by users. June 30th, 2016 is the next scheduled update.
- A release document will be issued documenting database changes and what was affected. VTTI and CTRE want to know about the errors in the NDS and RID, respectively.

IAP Research Teams - Status Updates for Phase 2

- All IAP teams under contract with the FHWA
- Most teams are not fully contracted with their subcontractors yet
- Several teams are initiating NDS data acquisition and will be in contact with VTTI shortly.
- Aladdin Barkawi stressed the importance of the teams getting under contracts soon as possible:
  - Most teams' schedules for Phase 2 are 18-24 months (starting in January 2016)
  - **September 30, 2017 -** deadline to obligate funding for Phase 3.
  - **May 2017** – reports from teams on early findings. These reports will drive Phase 3 funding decisions in order to obligate funding prior to September 30.
  - Prior to **September 30, 2017** – FHWA obligates funds for Phase 3

Marketing Discussion

TRB led a discussion on how to most effectively market NDS data in the future. They were interested in the user perspective. Notable discussion points included:

- Use of NDS data shows how drivers interact with roadway infrastructure and the sequence leading to crash events
- NDS data provides insight into what is happening in addition to crash records – interactions of drivers at all times, including close calls and at crashes we wouldn't have otherwise known about.
- NDS data offers a new way to look at driving behavior. Past driver behavior insight has typically been self-reporting or based on law enforcement reports. NDS data contain objective measures of behaviors. It's a game changer to understand what drivers are doing and how frequently.

- Research offers face validity – it's what real life drivers have done. Simulated and test track studies are different. Now we have exposure data – we can see what happens when there isn't a crash. Can see exposure to danger, conditional on situations.
- We should strive to make sure the decision makers are well aware of what the NDS is and then what can be gleaned from it.
- Push possibilities created by NDS data at two levels – general benefits at the top, more specific examples of use to researchers.
- Use examples of data differences – distracted driving recording on crash records vs. video data of NDS.
- DOT's are guiding safety decision through use of crash modification factors (i.e. a factor that estimates the expected change in crash frequency when implementing a specific countermeasure). NDS may be useful in validating the mechanisms by which a specific safety investment reduces crashes.

Workshop Recommendations

- Provide extensive FAQs with tips on how to effectively navigate through the data acquisition process:
  - Managing the request process
  - Potential hurdles and time delays (and how to reduce or avoid them)
  - Typical time to receive data and costs
- Use the training data set to explore the structure of the NDS database and influence the selection of variables and their subsequent costs.
- Clarify the cost and processing implications of acquiring large data sets (10K trips or more). Discussions at the workshop revealed a bottleneck in processing NDS requests for large amounts of data (e.g. greater than 10, 000 trips). Existing configuration of hardware and software would experience long delays to other waiting users to process requests for large datasets.
- Enhance access to previously developed datasets
  - Encourage users to agree to share on Data Use License (a check box on the license) when they have completed their work.
  - Make available a catalogue of data sets from researchers for others to reuse or build upon (such as work zone, safer data set)
  - Provide contact information for the datasets

- Study-specific remote enclaves and project-driven enclaves hold promise for enhanced access to PII by researchers while respecting commitments made to retain the confidentiality of study participants. More work is needed to develop specific protocols for NDS project-driven remote enclaves, but there are examples in the social science field of such systems for data access and analysis.
- Attendees favored locating remote enclaves in the Midwest and/or West Coast to ease access to PII.
- Improve the interface between states, contractors and IRB's – through FAQs and other communications
  - Tracking lessons learned - questions researchers should ask themselves as they develop their data use license applications
  - Providing information concerning schedules and time frames,
- Modify language to align it with current highway design terminology (Glossary or modification to legends).

References

Shelton K., M. Perez, J Hankey, PHASE 1 OF SHRP 2 SAFETY DATA, IMPLEMENTATION & OVERSIGHT, National Academy of Sciences, Second Strategic Highway Research Program SHRP 2 SD-01, Research Question Categories, Virginia Tech Transportation Institute, Virginia Polytechnic Institute and State University, 3500 Transportation Research Plaza, Blacksburg, VA 24061, DECEMBER 21, 2015

Attachment A

Workshop Agenda

Safety Data Issues Resolution Workshop
Preliminary Agenda
April 27, 2016

Keck Center
500 5th St NW
Washington, DC 20001

8:00 – 8:15 AM	**Welcome and Introductions**	FHWA/AASHTO/TRB
8:15 – 8:30 AM	**Workshop Overview** • Purpose of the Meeting • Review of Rules and Procedures	Kathleen Linehan
8:30 – 9:00 AM	**Presentation of Efforts to Date to Addressing Known Concerns Including:** • Process of data acquisition. o How can research teams obtain information about the status of their request for data? o What is the expected time to obtain a response concerning their most recent request? o How many requests are ahead of a particular research team in the queue? • Data processing and analysis enhancements for RID: o Can vertical alignment information (e.g. Point of Vertical Tangency, Point of Vertical Curvature) be included as descriptors of a vertical curve? o What is the status of speed limit enhancements?	TRB/VTTI/CTRE
9:00 – 10:15 AM	**Discussion of Topics Pending:** • Structure of the data base and its implications for users: o What are the factors that drive the cost of data acquisition? o What is the structure of the database, particularly as it relates to analysis of tradeoffs between variables used for data analysis? • PII and its implications for data analysis: o What progress has been made concerning circumstances under which the location of crashes may be usable by teams in their research but not released publically? o What criteria are used to exclude vehicle traces from analysis because of potential PII concerns?	Kathleen Linehan

10:15 – 10:30 AM	**Break**	

10:30 – 11:45 AM	**Discussion of Topics Pending (continued)**
	• Precision of data within the NDS:
	  o Is it possible to add road design attributes to the crash descriptors in the NDS? The design attributes should be in common road design terminology.
	  o Is it possible to reliably obtain lane usage (e.g. left, center or right lane) from GPS or other data?
	• Data processing and analysis enhancements:
	  o Are requests for data managed on a first come, first serve basis?
	  o Is it possible to let researchers know approximately when their data requests may be completed?
	  o What is the process for responding to possible data errors within a data set (e.g. blank video clips, illogical variable values)?
	• Other?
		Kathleen Linehan

11:45 – 1:00 PM	**Lunch** - Cafeteria	
1:00 – 2:30 PM **2:30 – 3:30 PM**	**Breakout sessions for issue resolution and follow-up: in depth discussions** **Conclusion** • Summary of topics and results of discussions. Consensus of where efforts are currently implemented and expectations going forward.	Kathleen Linehan
3:30 – 3:45 PM	**Break**	

3:45 – 4:30 PM	**Marketing of Data**
	• Based on your user experiences, TRB is interested in input regarding marketing approaches. Participants will be asked to consider the following questions:
	  o What sorts of things are you able to consider/research with the SHRP2 Safety Data that you were not able to research previously?
	  o How would you tell a peer (e.g. a state official or a university researcher) about the data? What would the "elevator speech" sound like?
	  o What are some of the key advantages and disadvantages of using SHRP2 data?
	  o Would you be willing to provide a brief testimonial regarding the SHRP2 safety data?
		TRB/VTTI/AASHTO

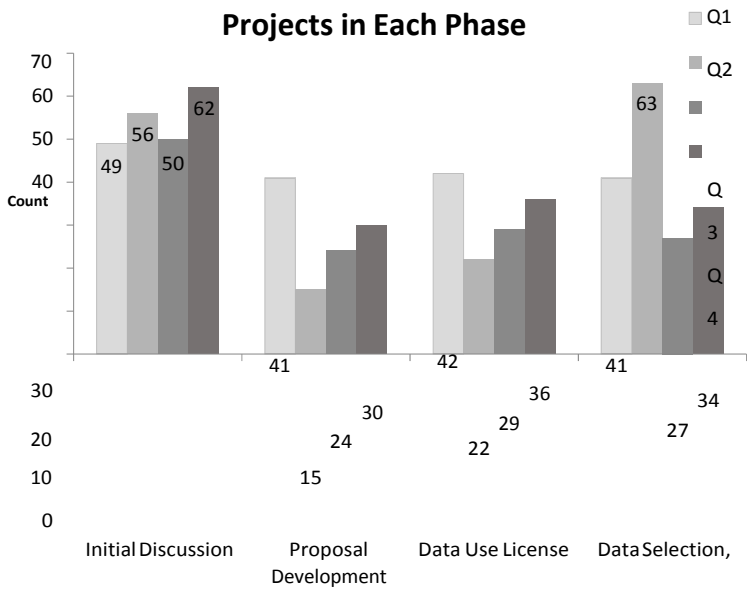
4:30 – 5:00 PM	**Wrap Up**	Kathleen Linehan

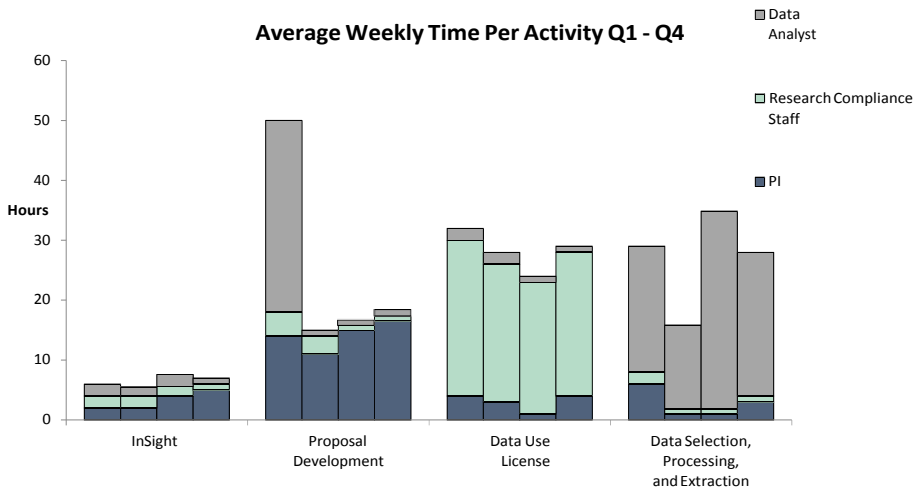
Attachment B

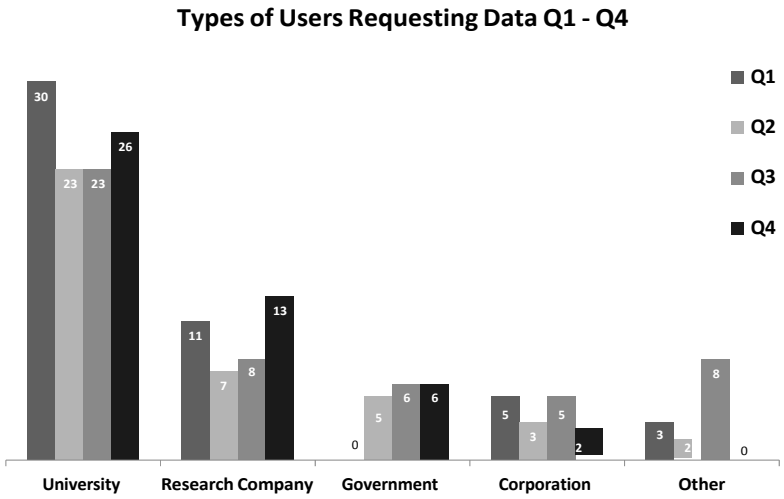
TRB Charts Shown During IRW

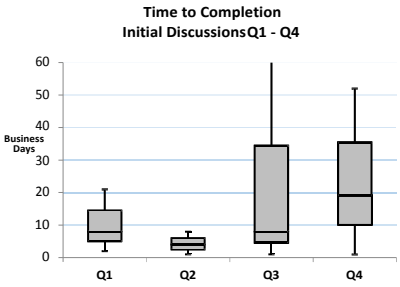
In Depth Update

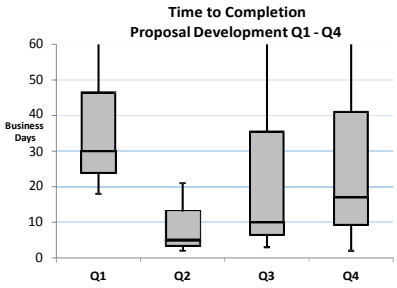
	Q1	Q2	Q3	Q4	Year 1 Total
Initial discussions with researchers	49	56	50	48	**203**
Quotes/proposals provided to researchers	41	15	24	29	**109**
Contracts through Phase 1 efforts	18	3	4	15	**40**
Data use licenses issued	33	14	14	10	**71**
Data use license addendums issued	26	23	5	20	**74**
Data sets provided to researchers	46	63	27	16	**152**

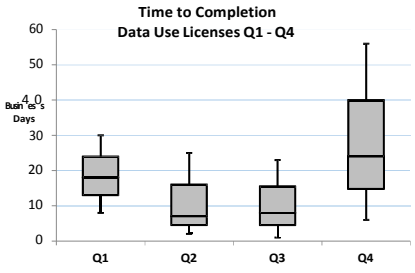


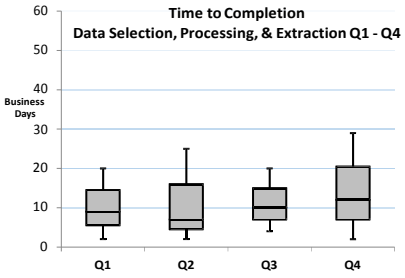

Average Weekly Time Per Activity Q1 - Q4
Data Analyst
Research Compliance Staff
PI
Hours
60
50
40
30
20
10
0
InSight
Proposal Development
Data Use License
Data Selection, Processing, and Extraction


Types of Users Requesting Data Q1 - Q4
Q1
Q2
Q3
Q4
30
23
23
26
11
7
8
13
0
5
6
6
5
3
5
2
3
2
8
0
University
Research Company
Government
Corporation
Other









InSight Update

- There are currently about 1700 registered users on InSight
- Nearly half the registered users (800) have taken the human research subjects training required to become Qualified Researchers and obtain access to the full functionality of InSight, including viewing forward video
- The first annual InSight users survey will be conducted soon to obtain user feedback and improved information about the market for InSight
- The most recent major addition in content to InSight is the downloadable Training Dataset which provides the same content (data and video) as the actual NDS data but without any PII issues
  - This dataset is treated as “open source” and uses a Creative Commons 4 license

Time Period	Q1	Q2	Q3	Q4
Last Week	68	77	25	85
Last Month	166	195	169	286
Last 3 months	327	454	350	518
Last 12 months*	819	939	983	1049

Individual Users on InSight and Page Hit Maximums Throughout 1st Year of Phase 1

	Q1	Q2	Q3	Q4
Individual Users	36	52	38	67
Individual Page Hits	1,517	3,800	5,754	5,478

Most Frequently Accessed Page Hits by Functional Area				
	Q1	Q2	Q3	Q4
Background	3237	1940	1997	9840
Forum	2314	1452	1613	1744
Home	2249	2001	3666	5934
Info	5109	4290	6695	6116
Main Page	2899	2562	3687	3741
Query	19734	15333	17908	20271
Video	-	25706	20801	15078

Most Frequently Accessed Page Hits by Graph or Dataset Name				
	Q1	Q2	Q3	Q4
Driver Demographic Questionnaire	39	201	220	201
Event Detail Table	3367	3411	2692	2657
Medical Conditions & Medications	624	18	55	102
Post-Crash Interview	52	156	201	187
Risk Perception Questionnaire	0	174	28	41
Trip Summary Table	390	410	633	730
Vehicle Detail Table	169	69	40	67
Vehicles by Model Year	143	87	162	119