The NCSES Data System

Kimberly Noonan

March 29, 2012

National Science Foundation
National Center for Science and Engineering Statistics
www.nsf.gov/statistics/

National Center for Science and Engineering Statistics (NCSES)

- **A federal statistical agency within NSF**

- **Charged with the mission to provide a central clearinghouse for the collection, interpretation, and analysis of data on scientific and engineering resources**

- **12 periodic data collections covering science and engineering**
  - **Research and Development**
  - **Education**
  - **Workforce**

- **Individual and establishment, surveys and censuses**

- **First data collection began in 1951**

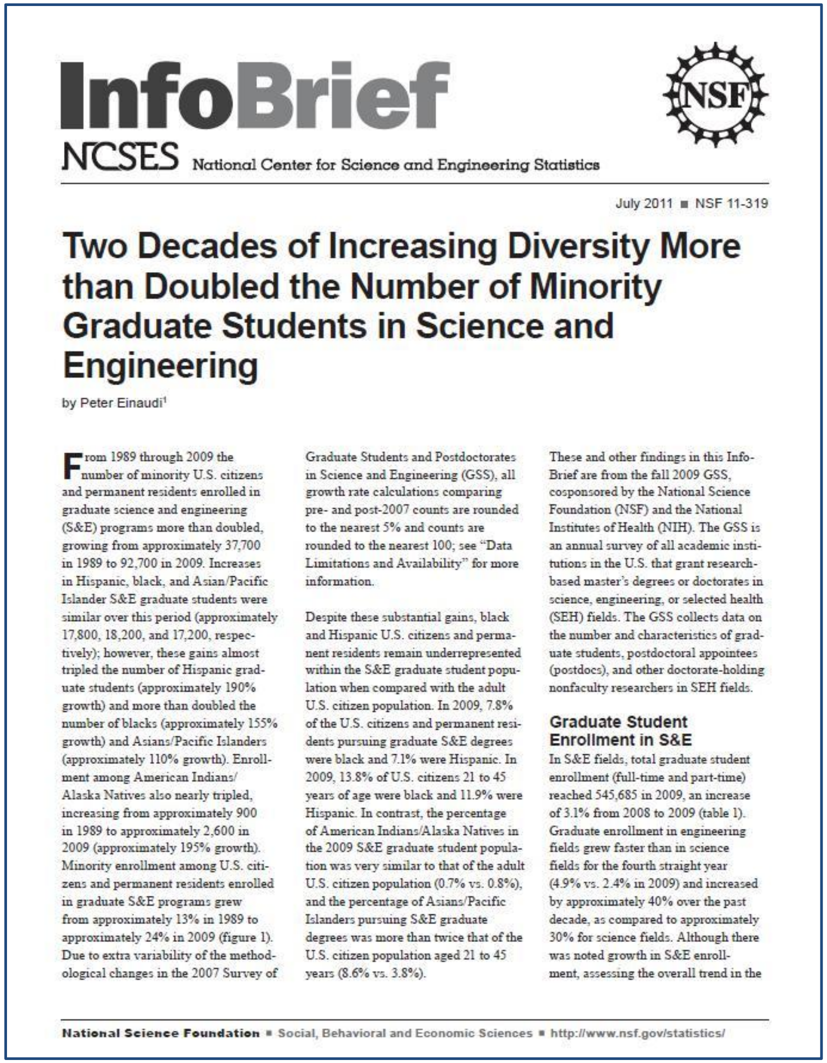
- **Over 7 contracts for external support**

NCSES Surveys

Select NCSES surveys:

- GSS = Survey of Graduate Students and Postdoctorates in Science and Engineering
- RDX = Survey of Research and Development Expenditures at Universities and Colleges
- FFS = Survey of Federal Funds for Research and Development
- FSS = Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions
- FAC = Survey of Science and Engineering Research Facilities
- SED = Survey of Earned Doctorates
- SDR = Survey of Doctorate Recipients
- NSRCG = National Survey of Recent College Graduates
- NSCG = National Survey of College Graduates

InfoBrief



InfoBrief
NCSES National Center for Science and Engineering Statistics

July 2011 ■ NSF 11-319

Two Decades of Increasing Diversity More than Doubled the Number of Minority Graduate Students in Science and Engineering

by Peter Einaudi[1]

From 1989 through 2009 the number of minority U.S. citizens and permanent residents enrolled in graduate science and engineering (S&E) programs more than doubled, growing from approximately 37,700 in 1989 to 92,700 in 2009. Increases in Hispanic, black, and Asian/Pacific Islander S&E graduate students were similar over this period (approximately 17,800, 18,200, and 17,200, respectively); however, these gains almost tripled the number of Hispanic graduate students (approximately 190% growth) and more than doubled the number of blacks (approximately 155% growth) and Asians/Pacific Islanders (approximately 110% growth). Enrollment among American Indians/Alaska Natives also nearly tripled, increasing from approximately 900 in 1989 to approximately 2,600 in 2009 (approximately 195% growth). Minority enrollment among U.S. citizens and permanent residents enrolled in graduate S&E programs grew from approximately 13% in 1989 to approximately 24% in 2009 (figure 1). Due to extra variability of the methodological changes in the 2007 Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS), all growth rate calculations comparing pre- and post-2007 counts are rounded to the nearest 5% and counts are rounded to the nearest 100; see "Data Limitations and Availability" for more information.

Despite these substantial gains, black and Hispanic U.S. citizens and permanent residents remain underrepresented within the S&E graduate student population when compared with the adult U.S. citizen population. In 2009, 7.8% of the U.S. citizens and permanent residents pursuing graduate S&E degrees were black and 7.1% were Hispanic. In 2009, 13.8% of U.S. citizens 21 to 45 years of age were black and 11.9% were Hispanic. In contrast, the percentage of American Indians/Alaska Natives in the 2009 S&E graduate student population was very similar to that of the adult U.S. citizen population (0.7% vs. 0.8%), and the percentage of Asians/Pacific Islanders pursuing S&E graduate degrees was more than twice that of the U.S. citizen population aged 21 to 45 years (8.6% vs. 3.8%).

These and other findings in this InfoBrief are from the fall 2009 GSS, cosponsored by the National Science Foundation (NSF) and the National Institutes of Health (NIH). The GSS is an annual survey of all academic institutions in the U.S. that grant research-based master's degrees or doctorates in science, engineering, or selected health (SEH) fields. The GSS collects data on the number and characteristics of graduate students, postdoctoral appointees (postdocs), and other doctorate-holding nonfaculty researchers in SEH fields.

Graduate Student Enrollment in S&E

In S&E fields, total graduate student enrollment (full-time and part-time) reached 545,685 in 2009, an increase of 3.1% from 2008 to 2009 (table 1). Graduate enrollment in engineering fields grew faster than in science fields for the fourth straight year (4.9% vs. 2.4% in 2009) and increased by approximately 40% over the past decade, as compared to approximately 30% for science fields. Although there was noted growth in S&E enrollment, assessing the overall trend in the

Published Data Tables

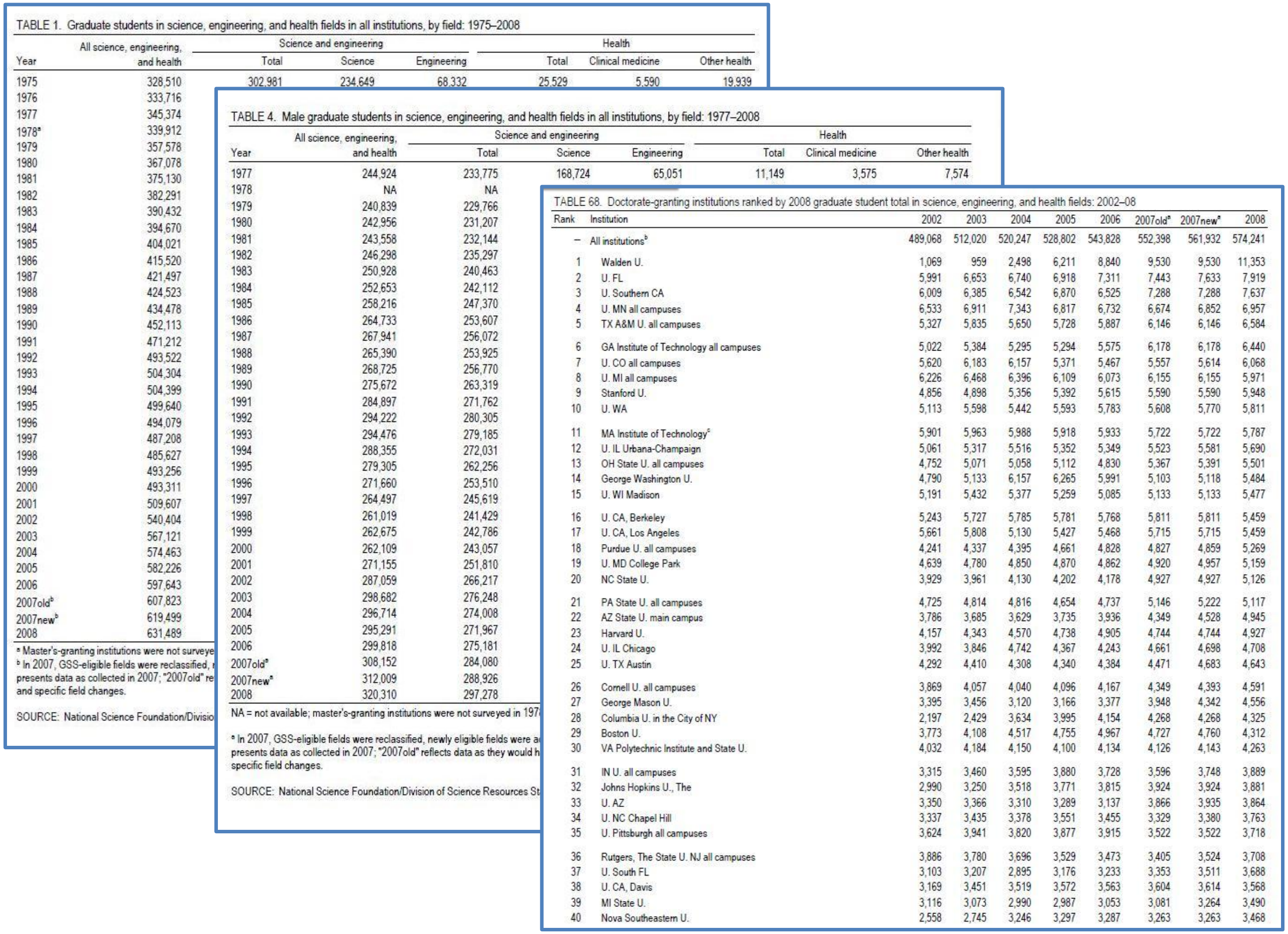


TABLE 1. Graduate students in science, engineering, and health fields in all institutions, by field: 1975–2008

Year	All science, engineering, and health	Science and engineering			Health		
		Total	Science	Engineering	Total	Clinical medicine	Other health
1975	328,510	302,981	234,649	68,332	25,529	5,590	19,939
1976	333,716						
1977	345,374						
1978[a]	339,912						
1979	357,578						
1980	367,078						
1981	375,130						
1982	382,291						
1983	390,432						
1984	394,670						
1985	404,021						
1986	415,520						
1987	421,497						
1988	424,523						
1989	434,478						
1990	452,113						
1991	471,212						
1992	493,522						
1993	504,304						
1994	504,399						
1995	499,640						
1996	494,079						
1997	487,208						
1998	485,627						
1999	493,256						
2000	493,311						
2001	509,607						
2002	540,404						
2003	567,121						
2004	574,463						
2005	582,226						
2006	597,643						
2007old[b]	607,823						
2007new[b]	619,499						
2008	631,489						

[a] Master's-granting institutions were not surveye

[b] In 2007, GSS-eligible fields were reclassified,
presents data as collected in 2007; "2007old" re
and specific field changes.

SOURCE: National Science Foundation/Divisio

TABLE 4. Male graduate students in science, engineering, and health fields in all institutions, by field: 1977–2008

Year	All science, engineering, and health	Science and engineering			Health		
		Total	Science	Engineering	Total	Clinical medicine	Other health
1977	244,924	233,775	168,724	65,051	11,149	3,575	7,574
1978	NA	NA					
1979	240,839	229,766					
1980	242,956	231,207					
1981	243,558	232,144					
1982	246,298	235,297					
1983	250,928	240,463					
1984	252,653	242,112					
1985	258,216	247,370					
1986	264,733	253,607					
1987	267,941	256,072					
1988	265,390	253,925					
1989	268,725	256,770					
1990	275,672	263,319					
1991	284,897	271,762					
1992	294,222	280,305					
1993	294,476	279,185					
1994	288,355	272,031					
1995	279,305	262,256					
1996	271,660	253,510					
1997	264,497	245,619					
1998	261,019	241,429					
1999	262,675	242,786					
2000	262,109	243,057					
2001	271,155	251,810					
2002	287,059	266,217					
2003	298,682	276,248					
2004	296,714	274,008					
2005	295,291	271,967					
2006	299,818	275,181					
2007old[a]	308,152	284,080					
2007new[a]	312,009	288,926					
2008	320,310	297,278					

NA = not available; master's-granting institutions were not surveyed in 197

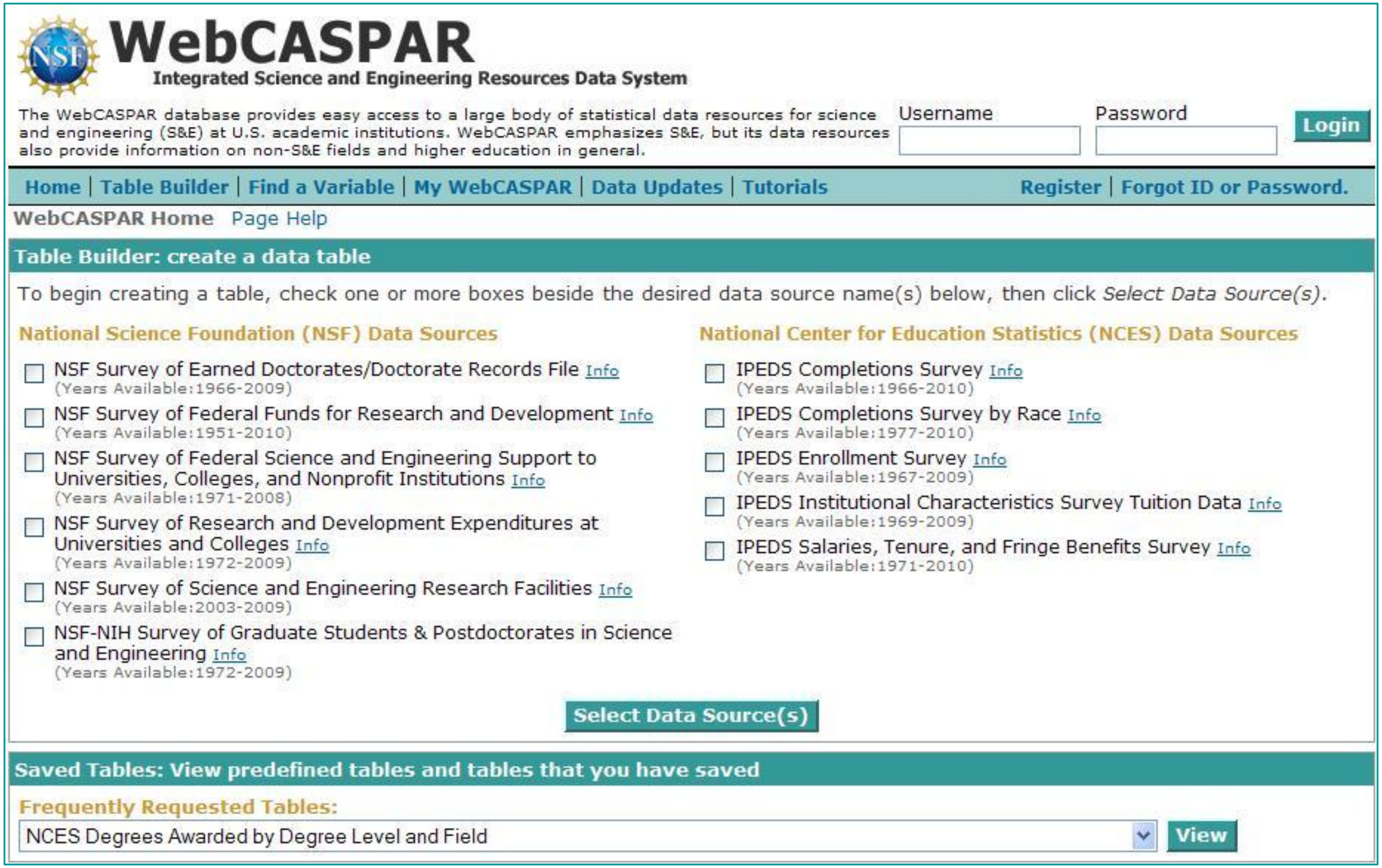
[a] In 2007, GSS-eligible fields were reclassified, newly eligible fields were a
presents data as collected in 2007; "2007old" reflects data as they would h
specific field changes.

SOURCE: National Science Foundation/Division of Science Resources St

TABLE 68. Doctorate-granting institutions ranked by 2008 graduate student total in science, engineering, and health fields: 2002–08

Rank	Institution	2002	2003	2004	2005	2006	2007old[a]	2007new[a]	2008
–	All institutions[b]	489,068	512,020	520,247	528,802	543,828	552,398	561,932	574,241
1	Walden U.	1,069	959	2,498	6,211	8,840	9,530	9,530	11,353
2	U. FL	5,991	6,653	6,740	6,918	7,311	7,443	7,633	7,919
3	U. Southern CA	6,009	6,385	6,542	6,870	6,525	7,288	7,288	7,637
4	U. MN all campuses	6,533	6,911	7,343	6,817	6,732	6,674	6,852	6,957
5	TX A&M U. all campuses	5,327	5,835	5,650	5,728	5,887	6,146	6,146	6,584
6	GA Institute of Technology all campuses	5,022	5,384	5,295	5,294	5,575	6,178	6,178	6,440
7	U. CO all campuses	5,620	6,183	6,157	5,371	5,467	5,557	5,614	6,068
8	U. MI all campuses	6,226	6,468	6,396	6,109	6,073	6,155	6,155	5,971
9	Stanford U.	4,856	4,898	5,356	5,392	5,615	5,590	5,590	5,948
10	U. WA	5,113	5,598	5,442	5,593	5,783	5,608	5,770	5,811
11	MA Institute of Technology[c]	5,901	5,963	5,988	5,918	5,933	5,722	5,722	5,787
12	U. IL Urbana-Champaign	5,061	5,317	5,516	5,352	5,349	5,523	5,581	5,690
13	OH State U. all campuses	4,752	5,071	5,058	5,112	4,830	5,367	5,391	5,501
14	George Washington U.	4,790	5,133	6,157	6,265	5,991	5,103	5,118	5,484
15	U. WI Madison	5,191	5,432	5,377	5,259	5,085	5,133	5,133	5,477
16	U. CA, Berkeley	5,243	5,727	5,785	5,781	5,768	5,811	5,811	5,459
17	U. CA, Los Angeles	5,661	5,808	5,130	5,427	5,468	5,715	5,715	5,459
18	Purdue U. all campuses	4,241	4,337	4,395	4,661	4,828	4,827	4,859	5,269
19	U. MD College Park	4,639	4,780	4,850	4,870	4,862	4,920	4,957	5,159
20	NC State U.	3,929	3,961	4,130	4,202	4,178	4,927	4,927	5,126
21	PA State U. all campuses	4,725	4,814	4,816	4,654	4,737	5,146	5,222	5,117
22	AZ State U. main campus	3,786	3,685	3,629	3,735	3,936	4,349	4,528	4,945
23	Harvard U.	4,157	4,343	4,570	4,738	4,905	4,744	4,744	4,927
24	U. IL Chicago	3,992	3,846	4,742	4,367	4,243	4,661	4,698	4,708
25	U. TX Austin	4,292	4,410	4,308	4,340	4,384	4,471	4,683	4,643
26	Cornell U. all campuses	3,869	4,057	4,040	4,096	4,167	4,349	4,393	4,591
27	George Mason U.	3,395	3,456	3,120	3,166	3,377	3,948	4,342	4,556
28	Columbia U. in the City of NY	2,197	2,429	3,634	3,995	4,154	4,268	4,268	4,325
29	Boston U.	3,773	4,108	4,517	4,755	4,967	4,727	4,760	4,312
30	VA Polytechnic Institute and State U.	4,032	4,184	4,150	4,100	4,134	4,126	4,143	4,263
31	IN U. all campuses	3,315	3,460	3,595	3,880	3,728	3,596	3,748	3,889
32	Johns Hopkins U., The	2,990	3,250	3,518	3,771	3,815	3,924	3,924	3,881
33	U. AZ	3,350	3,366	3,310	3,289	3,137	3,866	3,935	3,864
34	U. NC Chapel Hill	3,337	3,435	3,378	3,551	3,455	3,329	3,380	3,763
35	U. Pittsburgh all campuses	3,624	3,941	3,820	3,877	3,915	3,522	3,522	3,718
36	Rutgers, The State U. NJ all campuses	3,886	3,780	3,696	3,529	3,473	3,405	3,524	3,708
37	U. South FL	3,103	3,207	2,895	3,176	3,233	3,353	3,511	3,688
38	U. CA, Davis	3,169	3,451	3,519	3,572	3,563	3,604	3,614	3,568
39	MI State U.	3,116	3,073	2,990	2,987	3,053	3,081	3,264	3,490
40	Nova Southeastern U.	2,558	2,745	3,246	3,297	3,287	3,263	3,263	3,468

WebCASPAR


Home | Table Builder | Find a Variable | My WebCASPAR | Data Updates | Tutorials
Register | Forgot ID or Password.
Login
Select Data Source(s)
View

WebCASPAR

Integrated Science and Engineering Resources Data System

The WebCASPAR database provides easy access to a large body of statistical data resources for science and engineering (S&E) at U.S. academic institutions. WebCASPAR emphasizes S&E, but its data resources also provide information on non-S&E fields and higher education in general.

Username

Password

WebCASPAR Home Page Help

Table Builder: create a data table

To begin creating a table, check one or more boxes beside the desired data source name(s) below, then click *Select Data Source(s)*.

National Science Foundation (NSF) Data Sources

- ☐ NSF Survey of Earned Doctorates/Doctorate Records File Info (Years Available:1966-2009)
- ☐ NSF Survey of Federal Funds for Research and Development Info (Years Available:1951-2010)
- ☐ NSF Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions Info (Years Available:1971-2008)
- ☐ NSF Survey of Research and Development Expenditures at Universities and Colleges Info (Years Available:1972-2009)
- ☐ NSF Survey of Science and Engineering Research Facilities Info (Years Available:2003-2009)
- ☐ NSF-NIH Survey of Graduate Students & Postdoctorates in Science and Engineering Info (Years Available:1972-2009)

National Center for Education Statistics (NCES) Data Sources

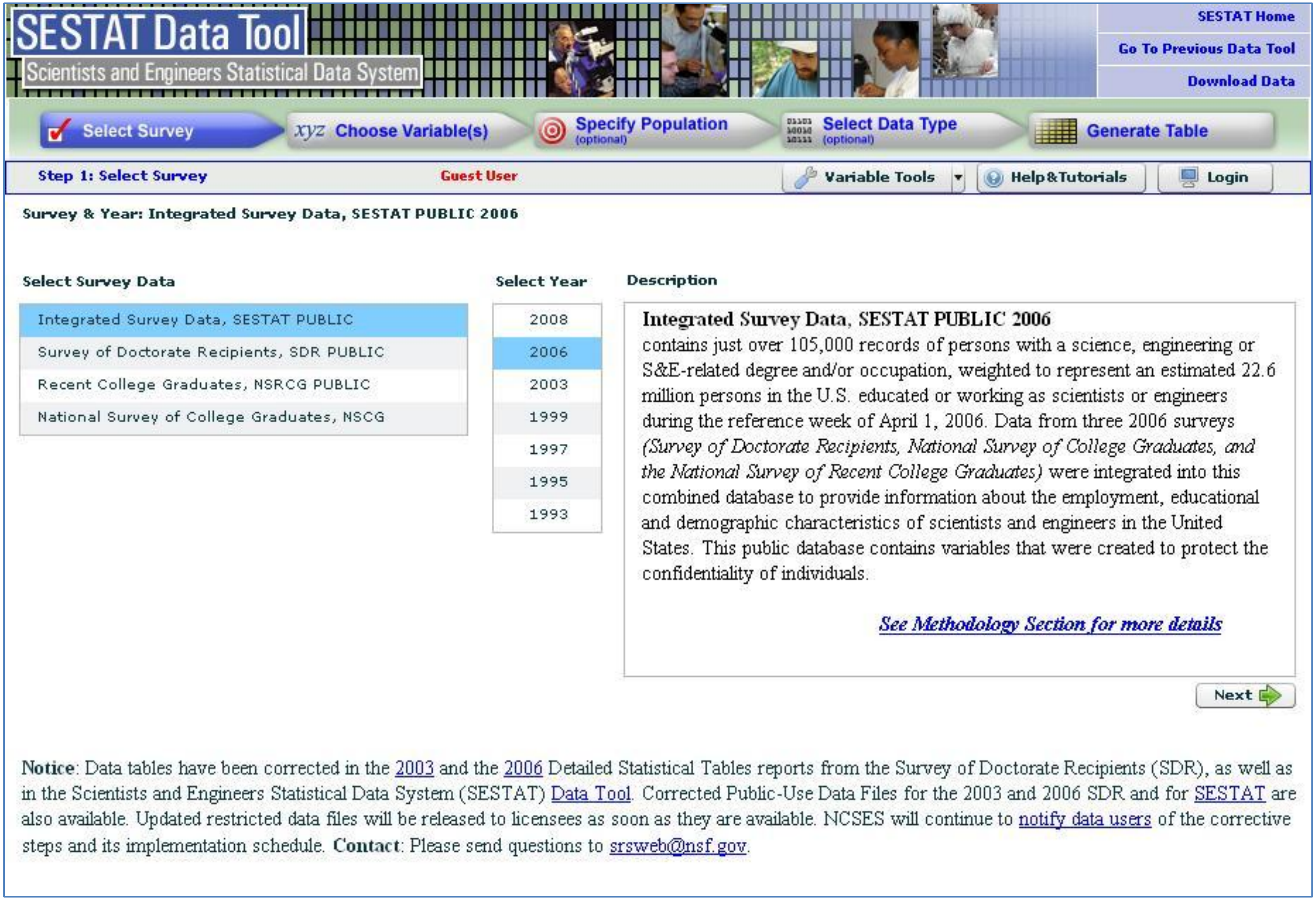
- ☐ IPEDS Completions Survey Info (Years Available:1966-2010)
- ☐ IPEDS Completions Survey by Race Info (Years Available:1977-2010)
- ☐ IPEDS Enrollment Survey Info (Years Available:1967-2009)
- ☐ IPEDS Institutional Characteristics Survey Tuition Data Info (Years Available:1969-2009)
- ☐ IPEDS Salaries, Tenure, and Fringe Benefits Survey Info (Years Available:1971-2010)

Saved Tables: View predefined tables and tables that you have saved

Frequently Requested Tables:

NCES Degrees Awarded by Degree Level and Field

SESTAT


SESTAT Data Tool
Scientists and Engineers Statistical Data System
SESTAT Home
Go To Previous Data Tool
Download Data
Select Survey
xyz Choose Variable(s)
Specify Population (optional)
Select Data Type (optional)
Generate Table
Variable Tools
Help&Tutorials
Login
Next

Step 1: Select Survey Guest User

Survey & Year: Integrated Survey Data, SESTAT PUBLIC 2006

Select Survey Data

- Integrated Survey Data, SESTAT PUBLIC
- Survey of Doctorate Recipients, SDR PUBLIC
- Recent College Graduates, NSRCG PUBLIC
- National Survey of College Graduates, NSCG

Select Year

- 2008
- 2006
- 2003
- 1999
- 1997
- 1995
- 1993

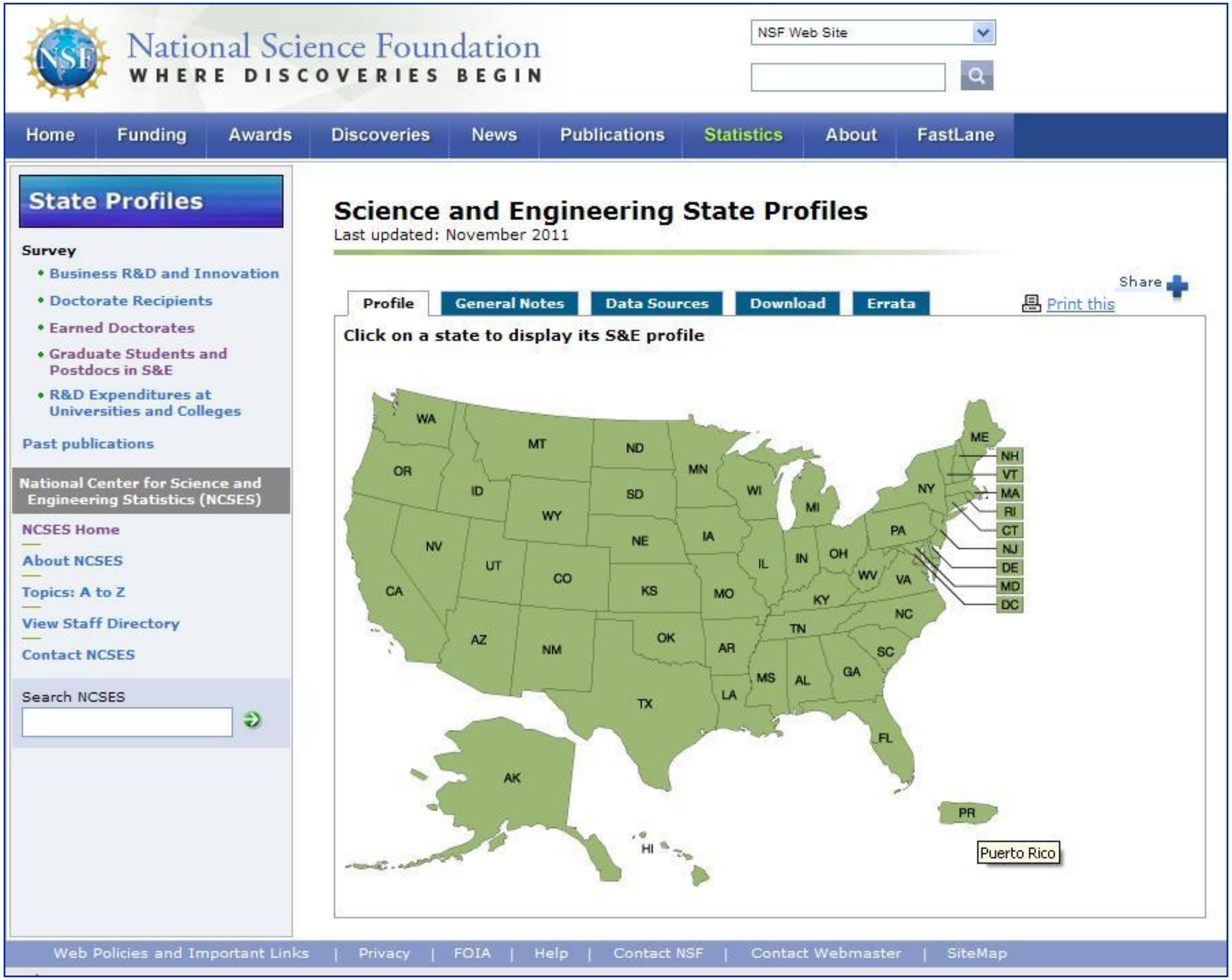
Description

Integrated Survey Data, SESTAT PUBLIC 2006 contains just over 105,000 records of persons with a science, engineering or S&E-related degree and/or occupation, weighted to represent an estimated 22.6 million persons in the U.S. educated or working as scientists or engineers during the reference week of April 1, 2006. Data from three 2006 surveys *(Survey of Doctorate Recipients, National Survey of College Graduates, and the National Survey of Recent College Graduates)* were integrated into this combined database to provide information about the employment, educational and demographic characteristics of scientists and engineers in the United States. This public database contains variables that were created to protect the confidentiality of individuals.

See Methodology Section for more details

Notice: Data tables have been corrected in the 2003 and the 2006 Detailed Statistical Tables reports from the Survey of Doctorate Recipients (SDR), as well as in the Scientists and Engineers Statistical Data System (SESTAT) Data Tool. Corrected Public-Use Data Files for the 2003 and 2006 SDR and for SESTAT are also available. Updated restricted data files will be released to licensees as soon as they are available. NCSES will continue to notify data users of the corrective steps and its implementation schedule. **Contact**: Please send questions to srsweb@nsf.gov.

State Data Tool


National Science Foundation
WHERE DISCOVERIES BEGIN
NSF Web Site
Home
Funding
Awards
Discoveries
News
Publications
Statistics
About
FastLane
State Profiles
Profile
General Notes
Data Sources
Download
Errata
Share
Print this
WA
OR
CA
NV
ID
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MI
IN
OH
KY
TN
MS
AL
GA
FL
SC
NC
VA
WV
PA
NY
ME
NH
VT
MA
RI
CT
NJ
DE
MD
DC
AK
HI
PR
Puerto Rico
Search NCSES
Web Policies and Important Links | Privacy | FOIA | Help | Contact NSF | Contact Webmaster | SiteMap

Science and Engineering State Profiles

Last updated: November 2011

Survey

- Business R&D and Innovation
- Doctorate Recipients
- Earned Doctorates
- Graduate Students and Postdocs in S&E
- R&D Expenditures at Universities and Colleges

Past publications

National Center for Science and Engineering Statistics (NCSES)

NCSES Home

About NCSES

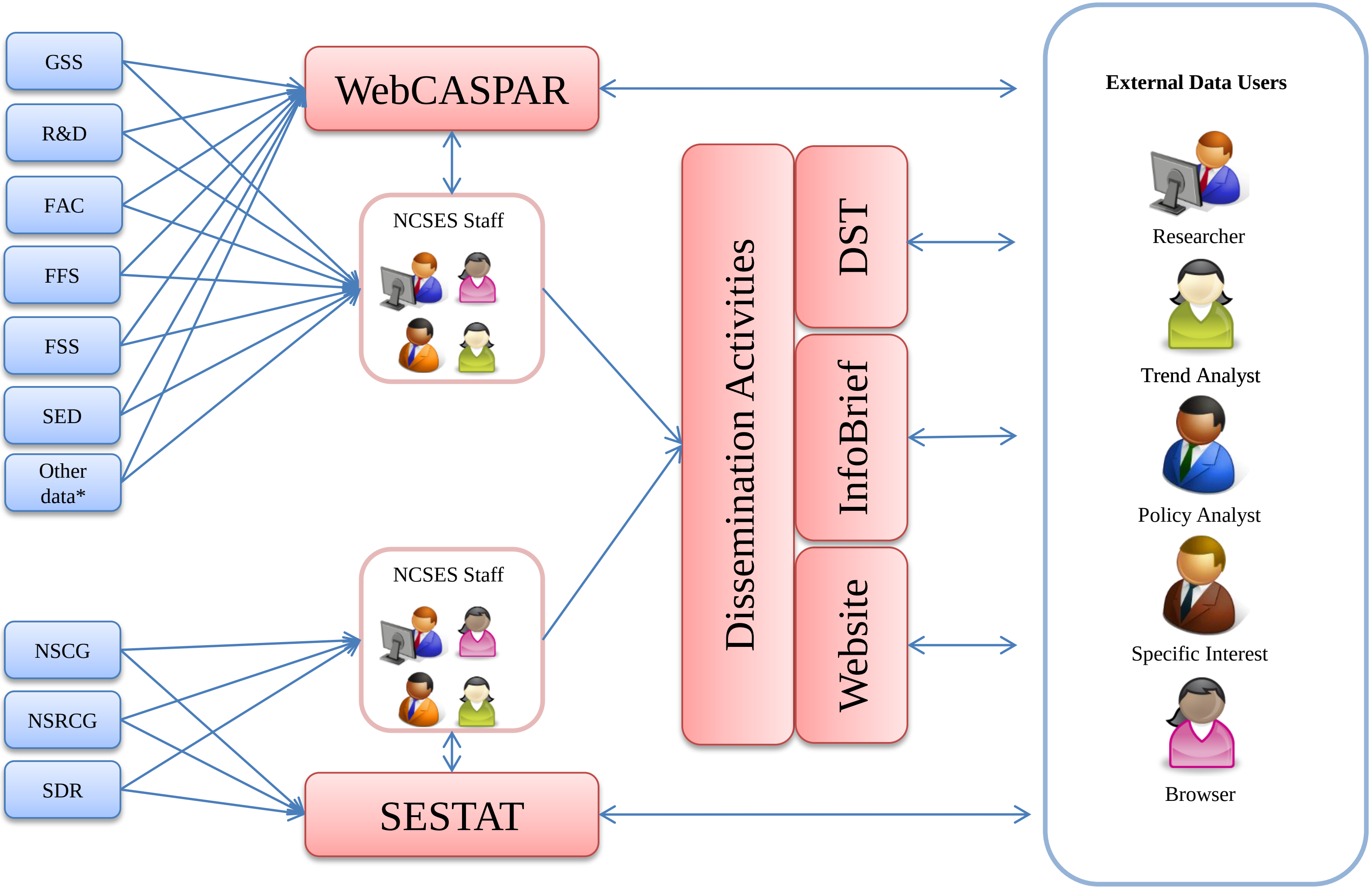
Topics: A to Z

View Staff Directory

Contact NCSES

Click on a state to display its S&E profile

Current Environment


GSS
R&D
FAC
FFS
FSS
SED
Other data*
WebCASPAR
NCSES Staff
NCSES Staff
NSCG
NSRCG
SDR
SESTAT
Dissemination Activities
DST
InfoBrief
Website
External Data Users
Researcher
Trend Analyst
Policy Analyst
Specific Interest
Browser

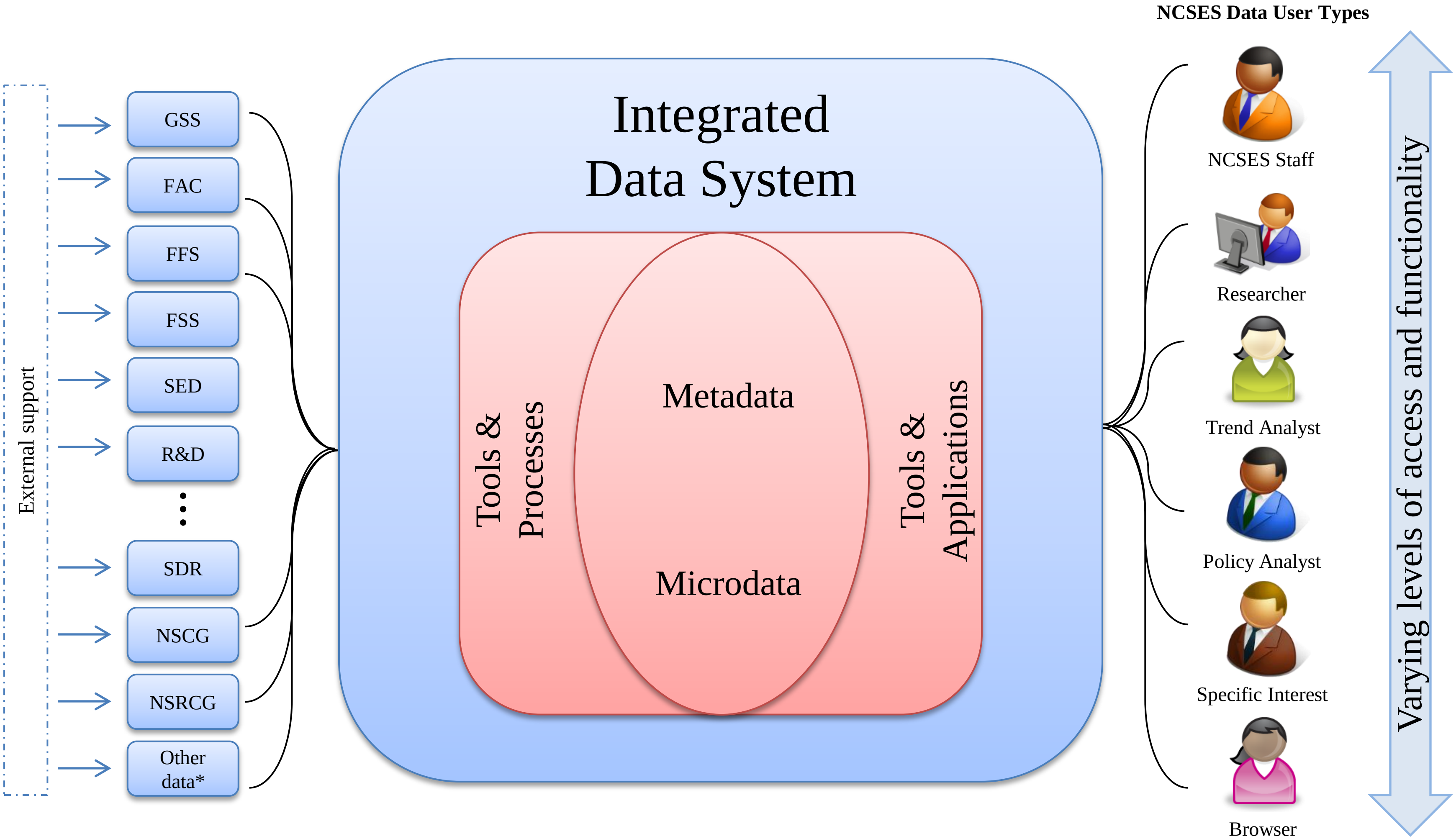
* Other data used regularly in NCSES publications

DST = Detailed Statistical Tables

Data System Vision

- **Objectives**
    - To meet transparency and reproducibility requirements of a federal statistical agency
    - To provide enhanced content and data management across multiple tiers of NCSES data
    - To enhance dissemination of these data to a broad audience
- **Main Goals**
    - To store and maintain NCSES survey data in one central warehouse
    - To integrate advanced data tools with the data management system
    - To deliver a broad range of content and analytical functionality via advanced data tools to a variety of users
- **Future process: integrated, dynamic, robust**

Data System Vision



External support

GSS

FAC

FFS

FSS

SED

R&D

⋮

SDR

NSCG

NSRCG

Other data*

Integrated Data System

Tools & Processes

Metadata

Microdata

Tools & Applications

NCSES Data User Types

NCSES Staff

Researcher

Trend Analyst

Policy Analyst

Specific Interest

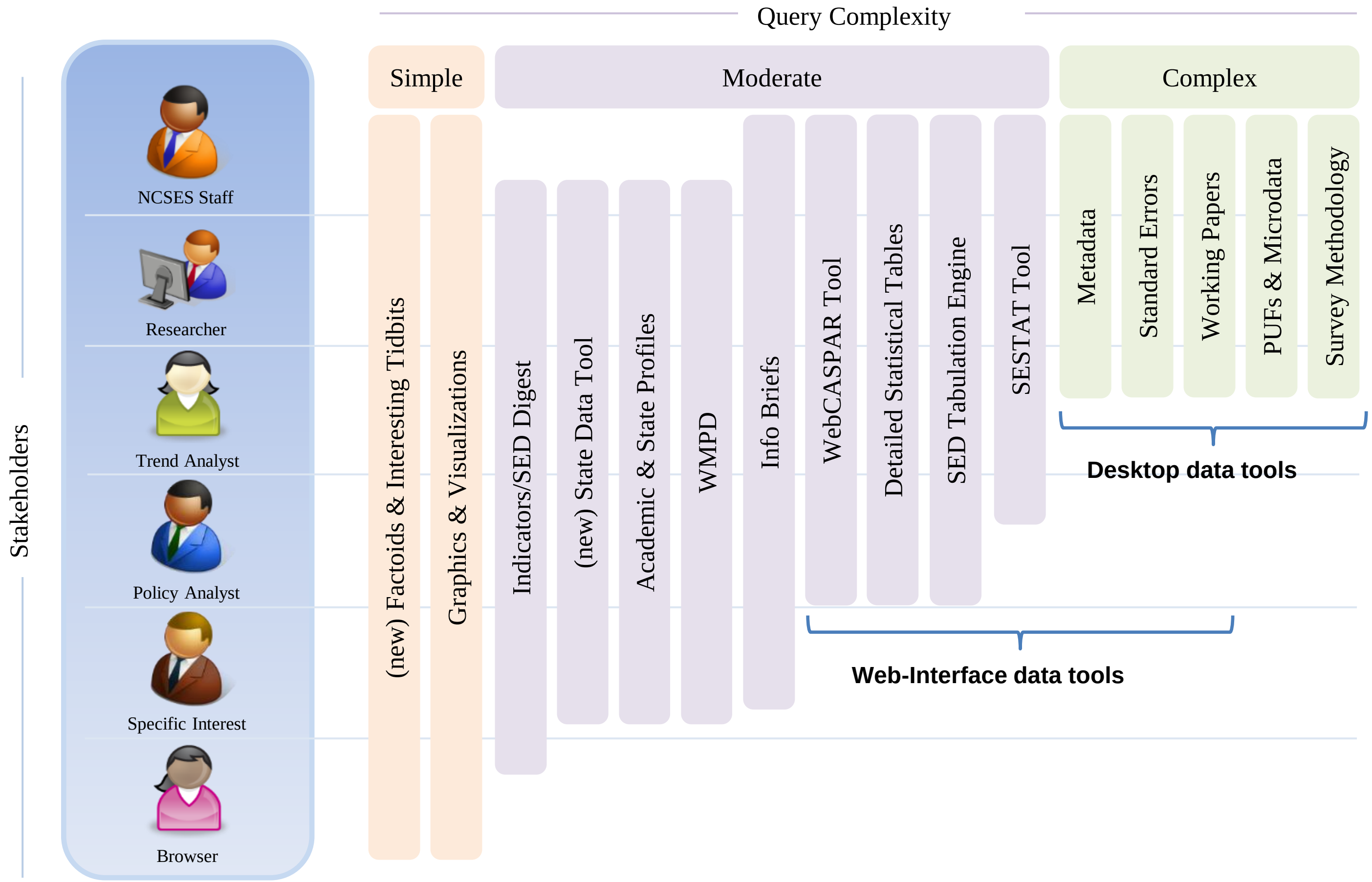
Browser

Varying levels of access and functionality

* Other data used regularly in NCSES publications

User Needs and Requirements

NCSES Data Tool Sectors



Query Complexity

Simple | Moderate | Complex

- Simple: (new) Factoids & Interesting Tidbits; Graphics & Visualization
- Moderate: Indicators/SED Digest; (new) State Data Tool; Academic & State Profiles; WMPD; Info Briefs; WebCASPAR Tool; Detailed Statistical Tables; SED Tabulation Engine; SESTAT Tool
- Complex: Metadata; Standard Errors; Working Papers; PUFs & Microdata; Survey Methodology

Desktop data tools

Web-Interface data tools

Stakeholders:
- NCSES Staff
- Researcher
- Trend Analyst
- Policy Analyst
- Specific Interest
- Browser

Examples:

Staff:
- Info Briefs
- White Papers

Research:
- PhD Dissertations
- MS Theses
- Journal Articles

Policy/Trend Analyst:
- Lobbying Interests
- Policy Associations

Specific Interest:
- Quick Facts
- Specific Questions

Browser
- No Specific Interest
- See what's available

Simple:
- Point estimates
- Ranges
- Graphics

Example queries:
- Tables
- Trends
- Simple Comparisons

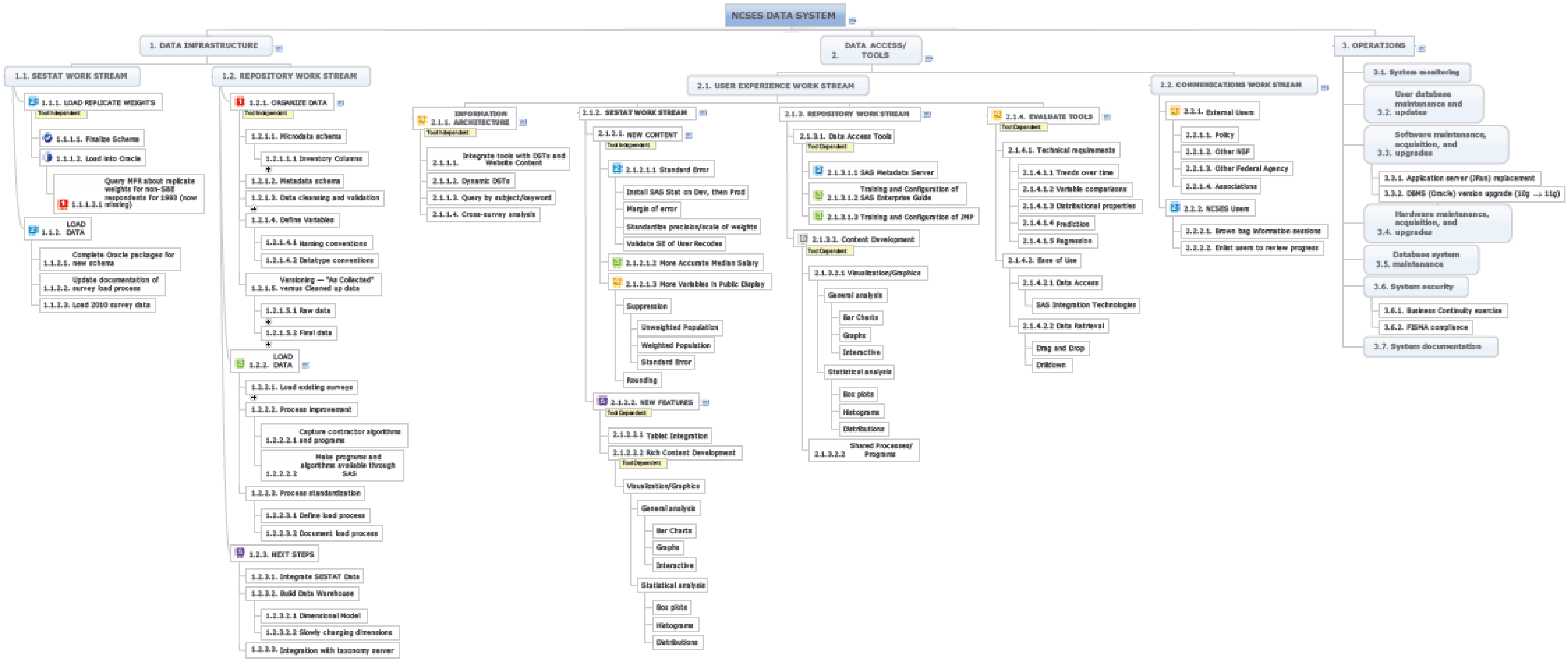
Complex:
- Regression
- ANOVA
- Data Mining
- Diagnostics

Data System Requirements

- Data
    - Provide consistent view of microdata across surveys
    - Provide consistent view of metadata across surveys
    - Real-time access to microdata and metadata

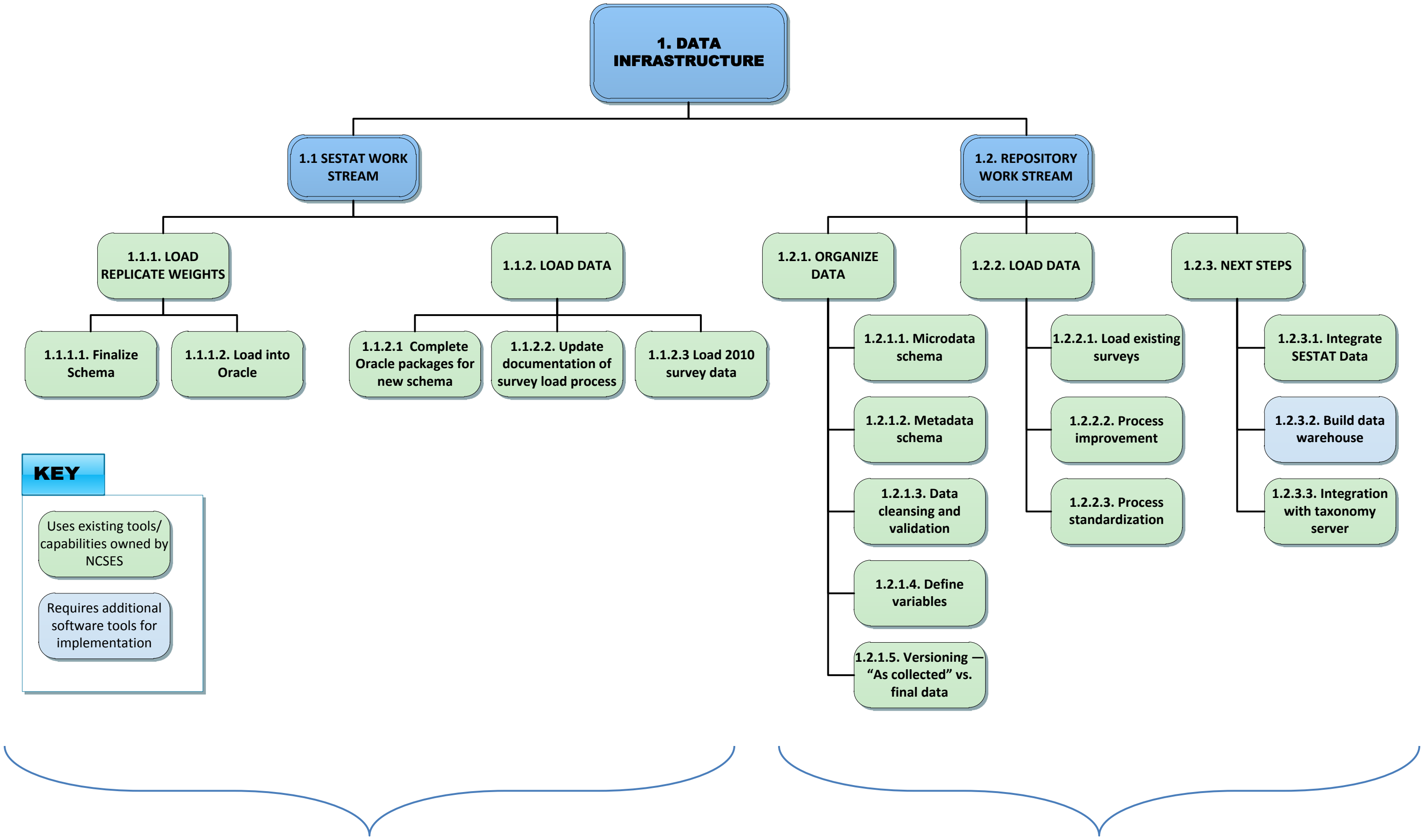
- Tools
    - Provide access to better tools
    - Enhanced capabilities include
        - Pre-generate tables (popular statistics)
        - Dynamic crosstabulations
        - Visualizations
        - Advanced statistical methods
        - Stored libraries and shared programs for internal staff

Overview of Work Streams



- NCSES DATA SYSTEM
  - 1. DATA INFRASTRUCTURE
    - 1.1. SESTAT WORK STREAM
      - 1.1.1. LOAD REPLICATE WEIGHTS
        - Tool Independent
        - 1.1.1.1. Finalize Schema
        - 1.1.1.2. Load into Oracle
          - 1.1.1.2.1 Query NPR about replicate weights for non-SAE respondents for 1993 (now missing)
      - 1.1.2. LOAD DATA
        - 1.1.2.1. Complete Oracle packages for new schema
        - 1.1.2.2. Update documentation of survey load process
        - 1.1.2.3. Load 2010 survey data
    - 1.2. REPOSITORY WORK STREAM
      - 1.2.1. ORGANIZE DATA
        - Tool Independent
        - 1.2.1.1. Microdata schema
          - 1.2.1.1.1 Inventory Columns
        - 1.2.1.2. Metadata schema
        - 1.2.1.3. Data cleansing and validation
        - 1.2.1.4. Define Variables
          - 1.2.1.4.1 Naming conventions
          - 1.2.1.4.2 Datatype conventions
        - 1.2.1.5. Versioning — "As Collected" versus Cleaned up data
          - 1.2.1.5.1 Raw data
          - 1.2.1.5.2 Final data
      - 1.2.2. LOAD DATA
        - 1.2.2.1. Load existing surveys
        - 1.2.2.2. Process improvement
          - 1.2.2.2.1 Capture contractor algorithms and programs
          - 1.2.2.2.2 Make programs and algorithms available through SAS
        - 1.2.2.3. Process standardization
          - 1.2.2.3.1 Define load process
          - 1.2.2.3.2 Document load process
      - 1.2.3. NEXT STEPS
        - 1.2.3.1. Integrate SESTAT Data
        - 1.2.3.2. Build Data Warehouse
          - 1.2.3.2.1 Dimensional Model
          - 1.2.3.2.2 Slowly changing dimensions
        - 1.2.3.3. Integration with taxonomy server
  - 2. DATA ACCESS/ TOOLS
    - 2.1. USER EXPERIENCE WORK STREAM
      - 2.1.1. INFORMATION ARCHITECTURE
        - Tool Independent
        - 2.1.1.1. Integrate tools with DSTs and Website Content
        - 2.1.1.2. Dynamic DSTs
        - 2.1.1.3. Query by subject/keyword
        - 2.1.1.4. Cross-survey analysis
      - 2.1.2. SESTAT WORK STREAM
        - 2.1.2.1. NEW CONTENT
          - Tool Independent
          - 2.1.2.1.1 Standard Error
            - Install SAS Stat on Dev, then Prod
            - Margin of error
            - Standardize precision/scale of weights
            - Validate SE of User Recodes
          - 2.1.2.1.2 More Accurate Median Salary
          - 2.1.2.1.3 More Variables in Public Display
            - Suppression
              - Unweighted Population
              - Weighted Population
              - Standard Error
            - Rounding
        - 2.1.2.2. NEW FEATURES
          - Tool Dependent
          - 2.1.2.2.1 Tablet Integration
          - 2.1.2.2.2 Rich Content Development
            - Tool Dependent
            - Visualization/Graphics
              - General analysis
                - Bar Charts
                - Graphs
                - Interactive
              - Statistical analysis
                - Box plots
                - Histograms
                - Distributions
      - 2.1.3. REPOSITORY WORK STREAM
        - 2.1.3.1. Data Access Tools
          - Tool Dependent
          - 2.1.3.1.1 SAS Metadata Server
          - 2.1.3.1.2 Training and Configuration of SAS Enterprise Guide
          - 2.1.3.1.3 Training and Configuration of JMP
        - 2.1.3.2. Content Development
          - Tool Dependent
          - 2.1.3.2.1 Visualization/Graphics
            - General analysis
              - Bar Charts
              - Graphs
              - Interactive
            - Statistical analysis
              - Box plots
              - Histograms
              - Distributions
          - 2.1.3.2.2 Shared Processes/ Programs
      - 2.1.4. EVALUATE TOOLS
        - Tool Dependent
        - 2.1.4.1. Technical requirements
          - 2.1.4.1.1 Trends over time
          - 2.1.4.1.2 Variable comparisons
          - 2.1.4.1.3 Distributional properties
          - 2.1.4.1.4 Prediction
          - 2.1.4.1.5 Regression
        - 2.1.4.2. Ease of Use
          - 2.1.4.2.1 Data Access
            - SAS Integration Technologies
          - 2.1.4.2.2 Data Retrieval
            - Drag and Drop
            - Drilldown
    - 2.2. COMMUNICATIONS WORK STREAM
      - 2.2.1. External Users
        - 2.2.1.1. Policy
        - 2.2.1.2. Other NSF
        - 2.2.1.3. Other Federal Agency
        - 2.2.1.4. Associations
      - 2.2.2. NCSES Users
        - 2.2.2.1. Brown bag information sessions
        - 2.2.2.2. Enlist users to review progress
  - 3. OPERATIONS
    - 3.1. System monitoring
    - 3.2. User database maintenance and updates
    - 3.3. Software maintenance, acquisition, and upgrades
      - 3.3.1. Application server (JRun) replacement
      - 3.3.2. DBMS (Oracle) version upgrade (10g → 11g)
    - 3.4. Hardware maintenance, acquisition, and upgrades
    - 3.5. Database system maintenance
    - 3.6. System security
      - 3.6.1. Business Continuity exercise
      - 3.6.2. FISMA compliance
    - 3.7. System documentation

1. Data Infrastructure Work Stream



- 1. DATA INFRASTRUCTURE
  - 1.1 SESTAT WORK STREAM
    - 1.1.1. LOAD REPLICATE WEIGHTS
      - 1.1.1.1. Finalize Schema
      - 1.1.1.2. Load into Oracle
    - 1.1.2. LOAD DATA
      - 1.1.2.1 Complete Oracle packages for new schema
      - 1.1.2.2. Update documentation of survey load process
      - 1.1.2.3 Load 2010 survey data
  - 1.2. REPOSITORY WORK STREAM
    - 1.2.1. ORGANIZE DATA
      - 1.2.1.1. Microdata schema
      - 1.2.1.2. Metadata schema
      - 1.2.1.3. Data cleansing and validation
      - 1.2.1.4. Define variables
      - 1.2.1.5. Versioning — "As collected" vs. final data
    - 1.2.2. LOAD DATA
      - 1.2.2.1. Load existing surveys
      - 1.2.2.2. Process improvement
      - 1.2.2.3. Process standardization
    - 1.2.3. NEXT STEPS
      - 1.2.3.1. Integrate SESTAT Data
      - 1.2.3.2. Build data warehouse
      - 1.2.3.3. Integration with taxonomy server

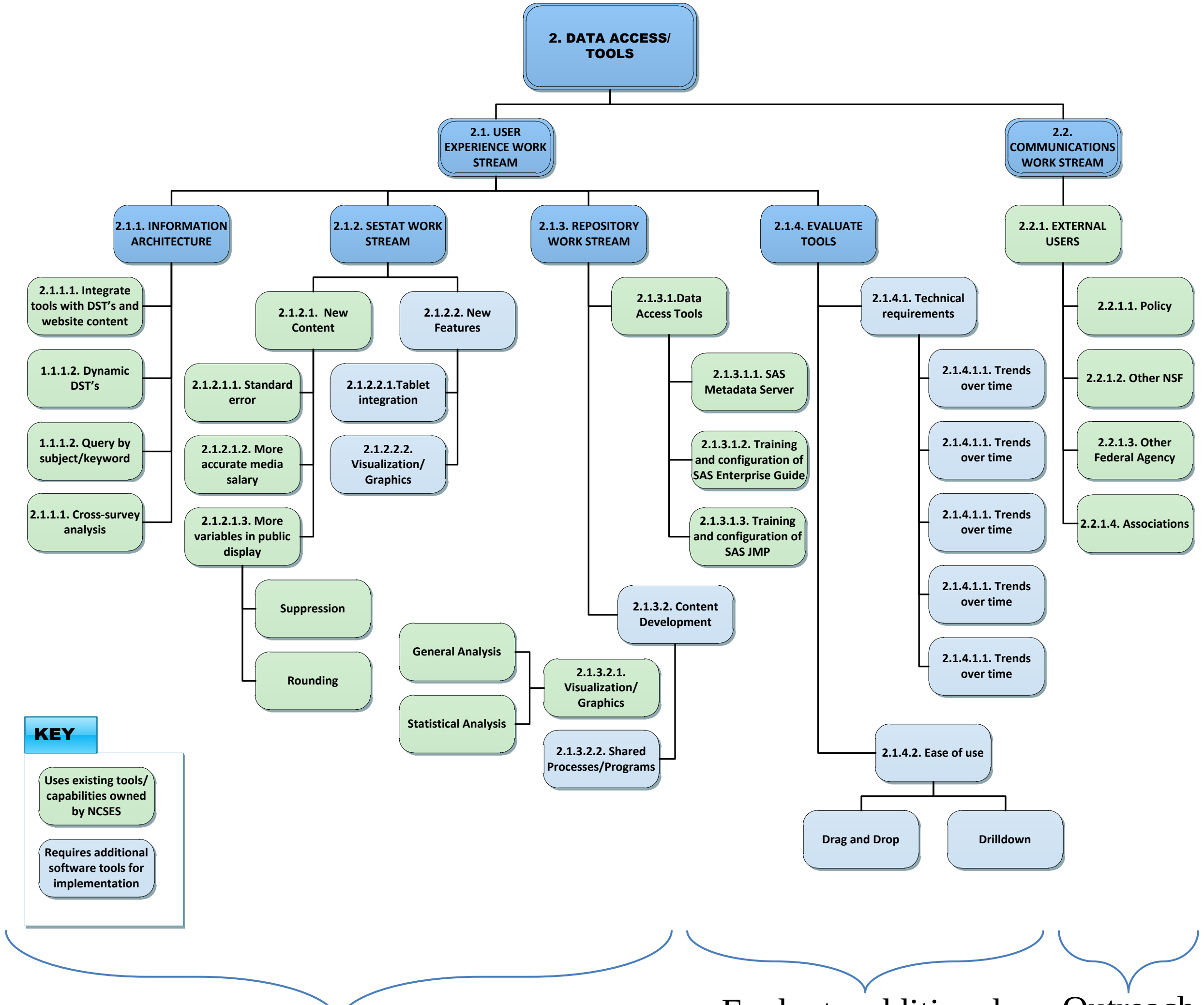
KEY

- 🟩 Uses existing tools/ capabilities owned by NCSES
- 🟦 Requires additional software tools for implementation

Maintain current functionality
Enhance transferrable functionality

Design and build new infrastructure

2. Data Access and Tools Work Stream



- 2. DATA ACCESS/ TOOLS
  - 2.1. USER EXPERIENCE WORK STREAM
    - 2.1.1. INFORMATION ARCHITECTURE
      - 2.1.1.1. Integrate tools with DST's and website content
      - 1.1.1.2. Dynamic DST's
      - 1.1.1.2. Query by subject/keyword
      - 2.1.1.1. Cross-survey analysis
    - 2.1.2. SESTAT WORK STREAM
      - 2.1.2.1. New Content
        - 2.1.2.1.1. Standard error
        - 2.1.2.1.2. More accurate media salary
        - 2.1.2.1.3. More variables in public display
          - Suppression
          - Rounding
      - 2.1.2.2. New Features
        - 2.1.2.2.1.Tablet integration
        - 2.1.2.2.2. Visualization/ Graphics
    - 2.1.3. REPOSITORY WORK STREAM
      - 2.1.3.1.Data Access Tools
        - 2.1.3.1.1. SAS Metadata Server
        - 2.1.3.1.2. Training and configuration of SAS Enterprise Guide
        - 2.1.3.1.3. Training and configuration of SAS JMP
      - 2.1.3.2. Content Development
        - 2.1.3.2.1. Visualization/ Graphics
          - General Analysis
          - Statistical Analysis
        - 2.1.3.2.2. Shared Processes/Programs
    - 2.1.4. EVALUATE TOOLS
      - 2.1.4.1. Technical requirements
        - 2.1.4.1.1. Trends over time
        - 2.1.4.1.1. Trends over time
        - 2.1.4.1.1. Trends over time
        - 2.1.4.1.1. Trends over time
        - 2.1.4.1.1. Trends over time
      - 2.1.4.2. Ease of use
        - Drag and Drop
        - Drilldown
  - 2.2. COMMUNICATIONS WORK STREAM
    - 2.2.1. EXTERNAL USERS
      - 2.2.1.1. Policy
      - 2.2.1.2. Other NSF
      - 2.2.1.3. Other Federal Agency
      - 2.2.1.4. Associations

KEY

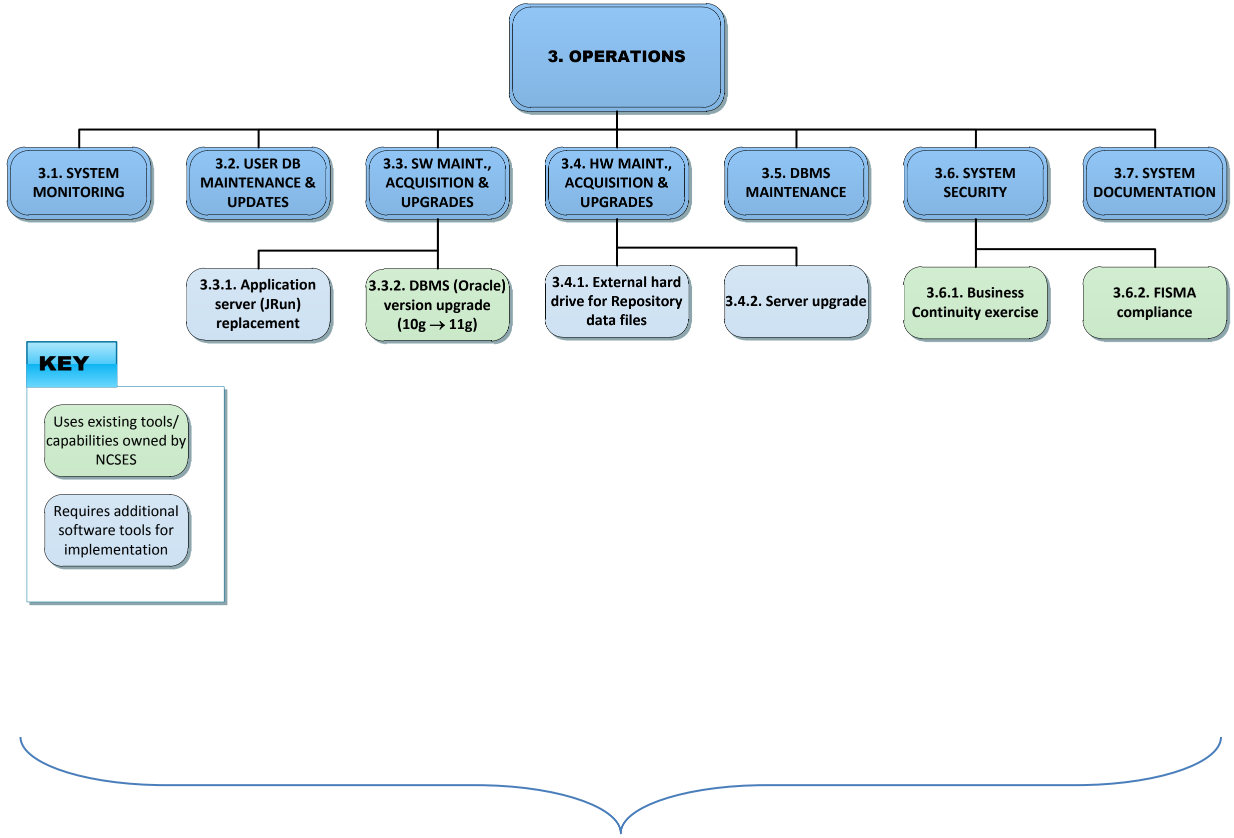
- 🟩 Uses existing tools/ capabilities owned by NCSES
- 🟦 Requires additional software tools for implementation

Expand use of current tools

Evaluate additional COTS tools

Outreach

3. Operations Work Stream



3. OPERATIONS

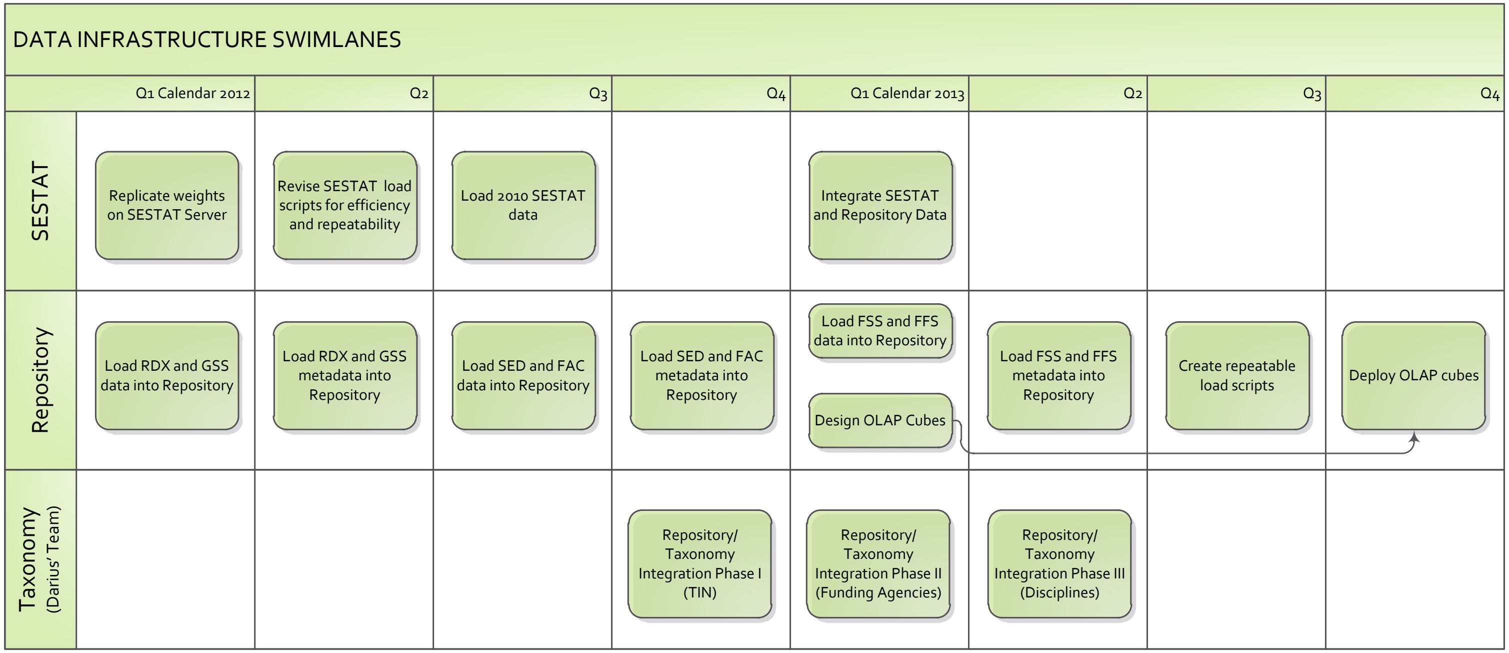
- 3.1. SYSTEM MONITORING
- 3.2. USER DB MAINTENANCE & UPDATES
- 3.3. SW MAINT., ACQUISITION & UPGRADES
  - 3.3.1. Application server (JRun) replacement
  - 3.3.2. DBMS (Oracle) version upgrade (10g → 11g)
- 3.4. HW MAINT., ACQUISITION & UPGRADES
  - 3.4.1. External hard drive for Repository data files
  - 3.4.2. Server upgrade
- 3.5. DBMS MAINTENANCE
- 3.6. SYSTEM SECURITY
  - 3.6.1. Business Continuity exercise
  - 3.6.2. FISMA compliance
- 3.7. SYSTEM DOCUMENTATION

KEY

- 🟩 Uses existing tools/ capabilities owned by NCSES
- 🟦 Requires additional software tools for implementation

Keep servers running and portals active

1. Data Infrastructure Timeline



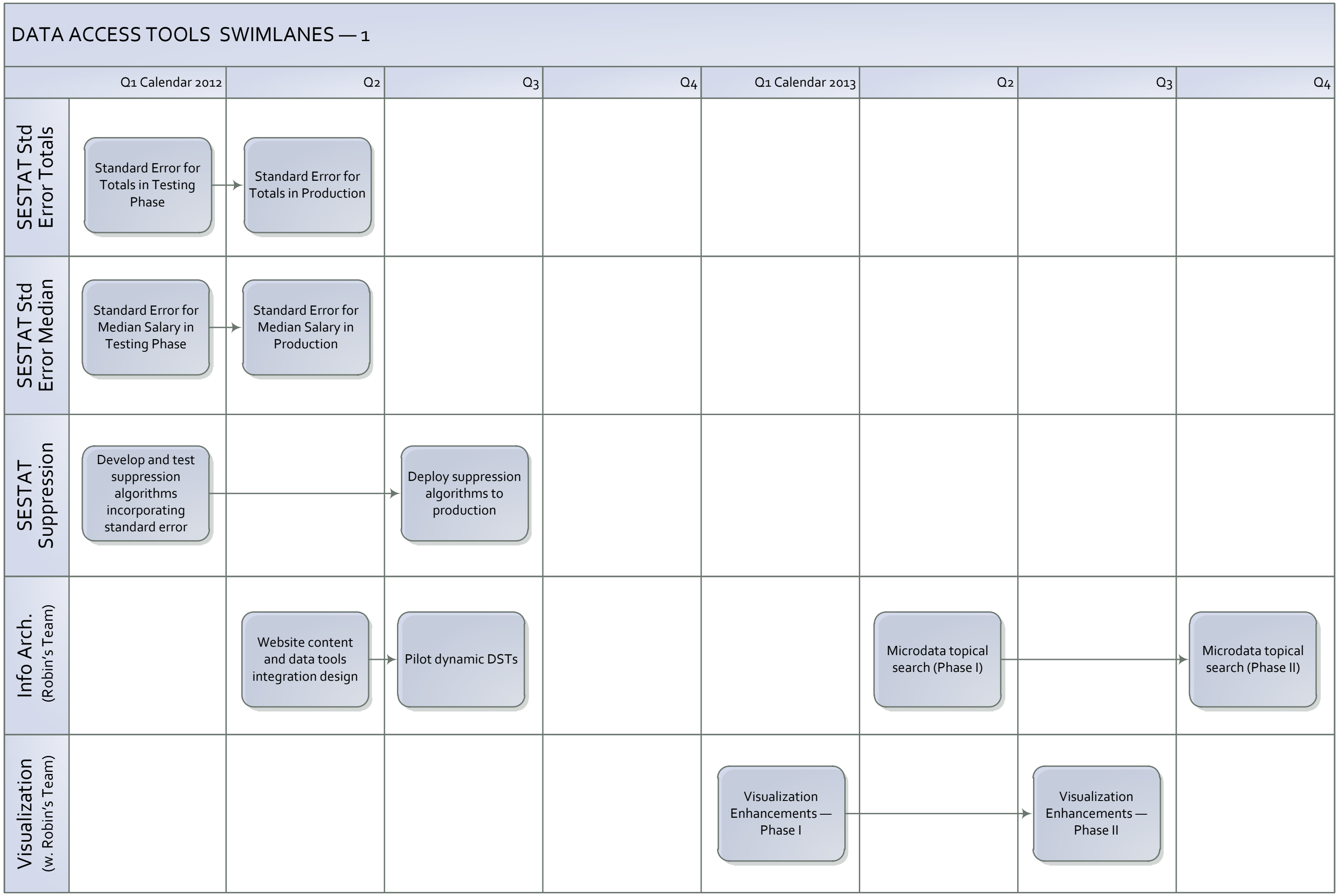
DATA INFRASTRUCTURE SWIMLANES

	Q1 Calendar 2012	Q2	Q3	Q4	Q1 Calendar 2013	Q2	Q3	Q4
SESTAT	Replicate weights on SESTAT Server	Revise SESTAT load scripts for efficiency and repeatability	Load 2010 SESTAT data		Integrate SESTAT and Repository Data			
Repository	Load RDX and GSS data into Repository	Load RDX and GSS metadata into Repository	Load SED and FAC data into Repository	Load SED and FAC metadata into Repository	Load FSS and FFS data into Repository Design OLAP Cubes	Load FSS and FFS metadata into Repository	Create repeatable load scripts	Deploy OLAP cubes
Taxonomy (Darius' Team)				Repository/ Taxonomy Integration Phase I (TIN)	Repository/ Taxonomy Integration Phase II (Funding Agencies)	Repository/ Taxonomy Integration Phase III (Disciplines)		

2012

2013

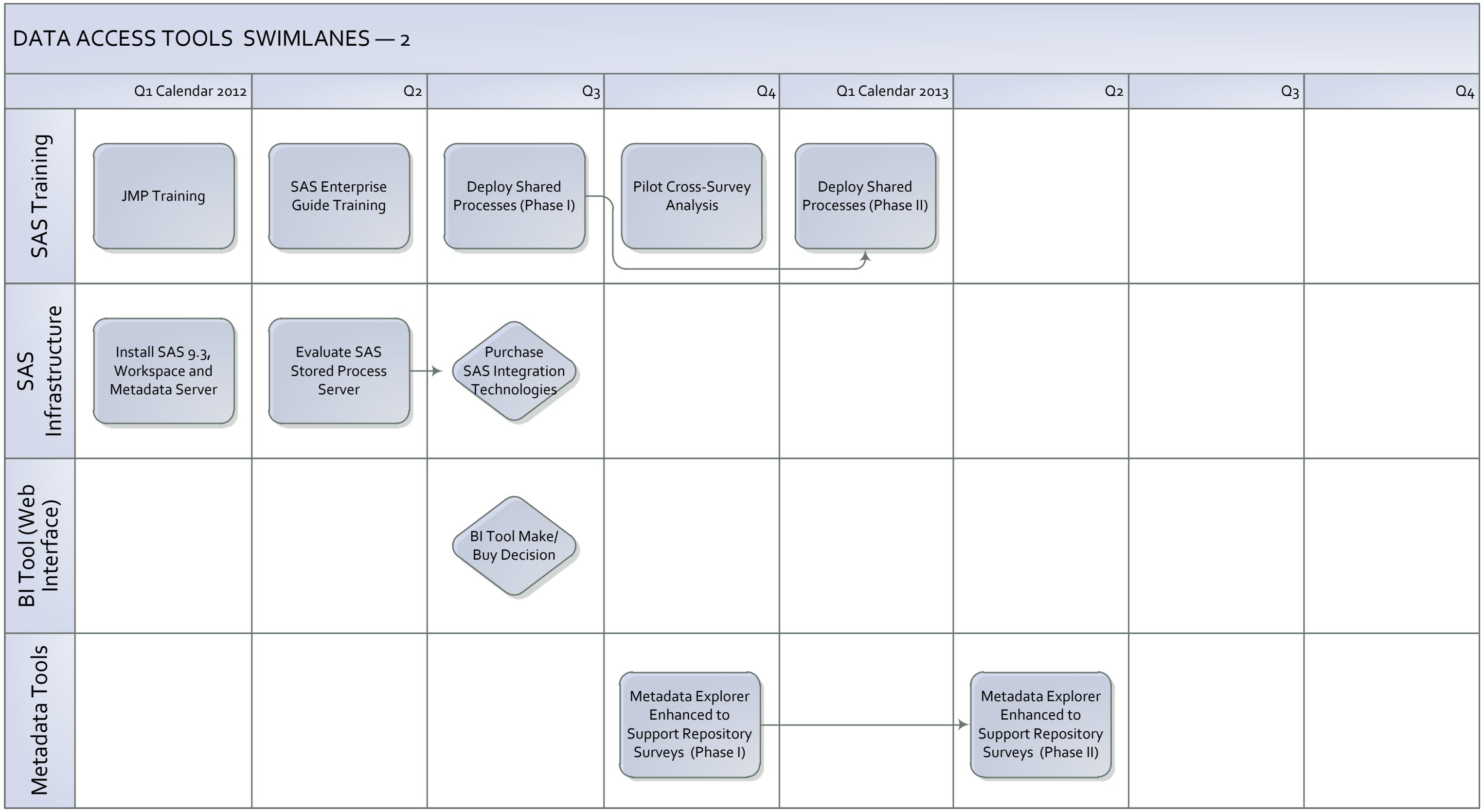
2. Data Access and Tools Timeline



DATA ACCESS TOOLS SWIMLANES — 1

	Q1 Calendar 2012	Q2	Q3	Q4	Q1 Calendar 2013	Q2	Q3	Q4
SESTAT Std Error Totals	Standard Error for Totals in Testing Phase	Standard Error for Totals in Production						
SESTAT Std Error Median	Standard Error for Median Salary in Testing Phase	Standard Error for Median Salary in Production						
SESTAT Suppression	Develop and test suppression algorithms incorporating standard error		Deploy suppression algorithms to production					
Info Arch. (Robin's Team)		Website content and data tools integration design	Pilot dynamic DSTs			Microdata topical search (Phase I)		Microdata topical search (Phase II)
Visualization (w. Robin's Team)					Visualization Enhancements — Phase I		Visualization Enhancements — Phase II	

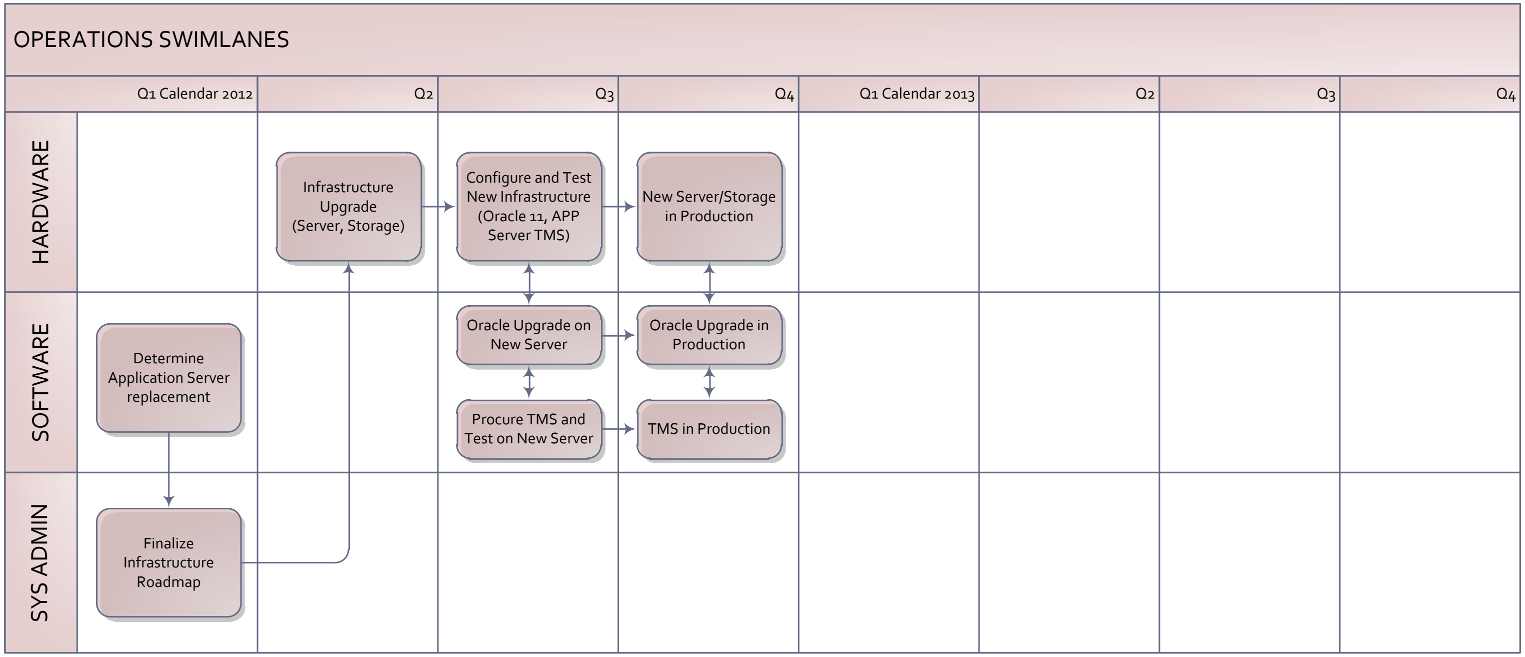
2. Data Access and Tools Timeline cont’d



DATA ACCESS TOOLS SWIMLANES — 2

	Q1 Calendar 2012	Q2	Q3	Q4	Q1 Calendar 2013	Q2	Q3	Q4
SAS Training	JMP Training	SAS Enterprise Guide Training	Deploy Shared Processes (Phase I)	Pilot Cross-Survey Analysis	Deploy Shared Processes (Phase II)			
SAS Infrastructure	Install SAS 9.3, Workspace and Metadata Server	Evaluate SAS Stored Process Server	Purchase SAS Integration Technologies					
BI Tool (Web Interface)			BI Tool Make/ Buy Decision					
Metadata Tools				Metadata Explorer Enhanced to Support Repository Surveys (Phase I)		Metadata Explorer Enhanced to Support Repository Surveys (Phase II)		

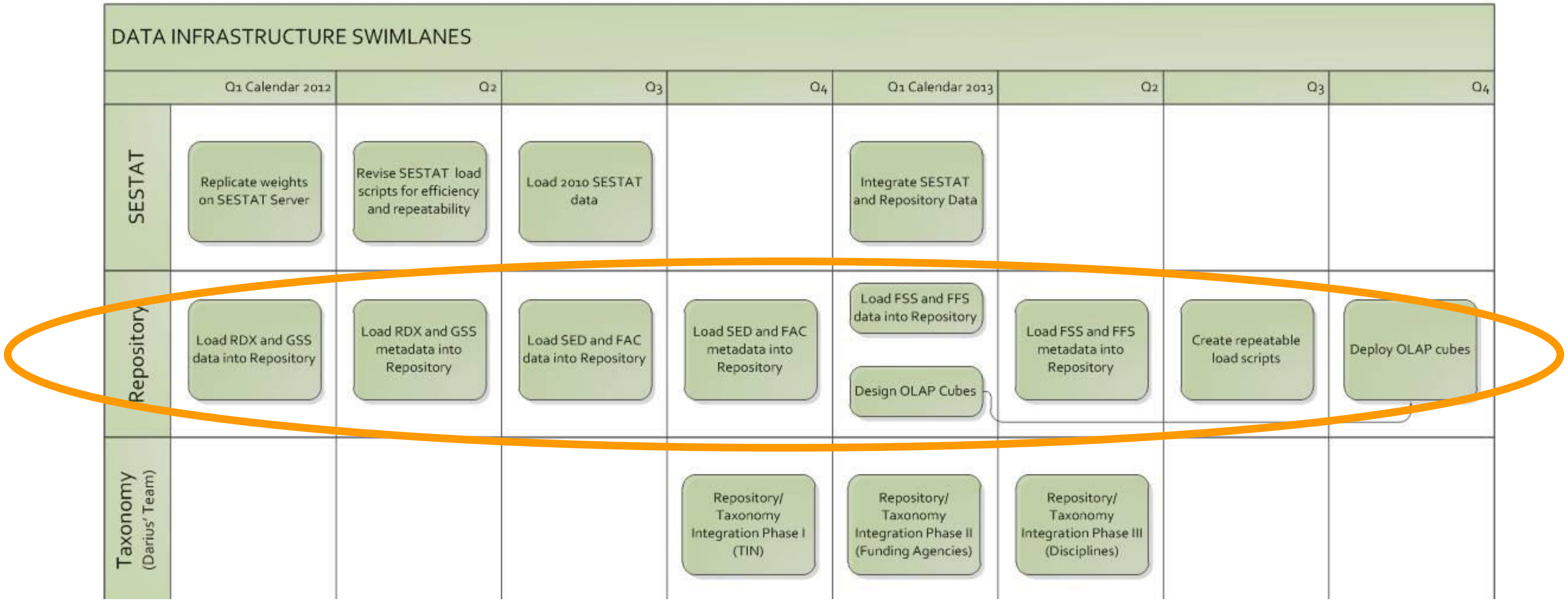
3. Operations Timeline



OPERATIONS SWIMLANES

	Q1 Calendar 2012	Q2	Q3	Q4	Q1 Calendar 2013	Q2	Q3	Q4
HARDWARE		Infrastructure Upgrade (Server, Storage)	Configure and Test New Infrastructure (Oracle 11, APP Server TMS)	New Server/Storage in Production				
SOFTWARE	Determine Application Server replacement		Oracle Upgrade on New Server Procure TMS and Test on New Server	Oracle Upgrade in Production TMS in Production				
SYS ADMIN	Finalize Infrastructure Roadmap							

Data Infrastructure: Repository



DATA INFRASTRUCTURE SWIMLANES

	Q1 Calendar 2012	Q2	Q3	Q4	Q1 Calendar 2013	Q2	Q3	Q4
SESTAT	Replicate weights on SESTAT Server	Revise SESTAT load scripts for efficiency and repeatability	Load 2010 SESTAT data		Integrate SESTAT and Repository Data			
Repository	Load RDX and GSS data into Repository	Load RDX and GSS metadata into Repository	Load SED and FAC data into Repository	Load SED and FAC metadata into Repository	Load FSS and FFS data into Repository Design OLAP Cubes	Load FSS and FFS metadata into Repository	Create repeatable load scripts	Deploy OLAP cubes
Taxonomy (Darius' Team)				Repository/ Taxonomy Integration Phase I (TIN)	Repository/ Taxonomy Integration Phase II (Funding Agencies)	Repository/ Taxonomy Integration Phase III (Disciplines)		

Data infrastructure to store, maintain, and disseminate all NCSES data collections

- Microdata and Metadata
- Define requirements and standards
- Define storage schema

Data Repository Delivery Requirements

Survey contractor:

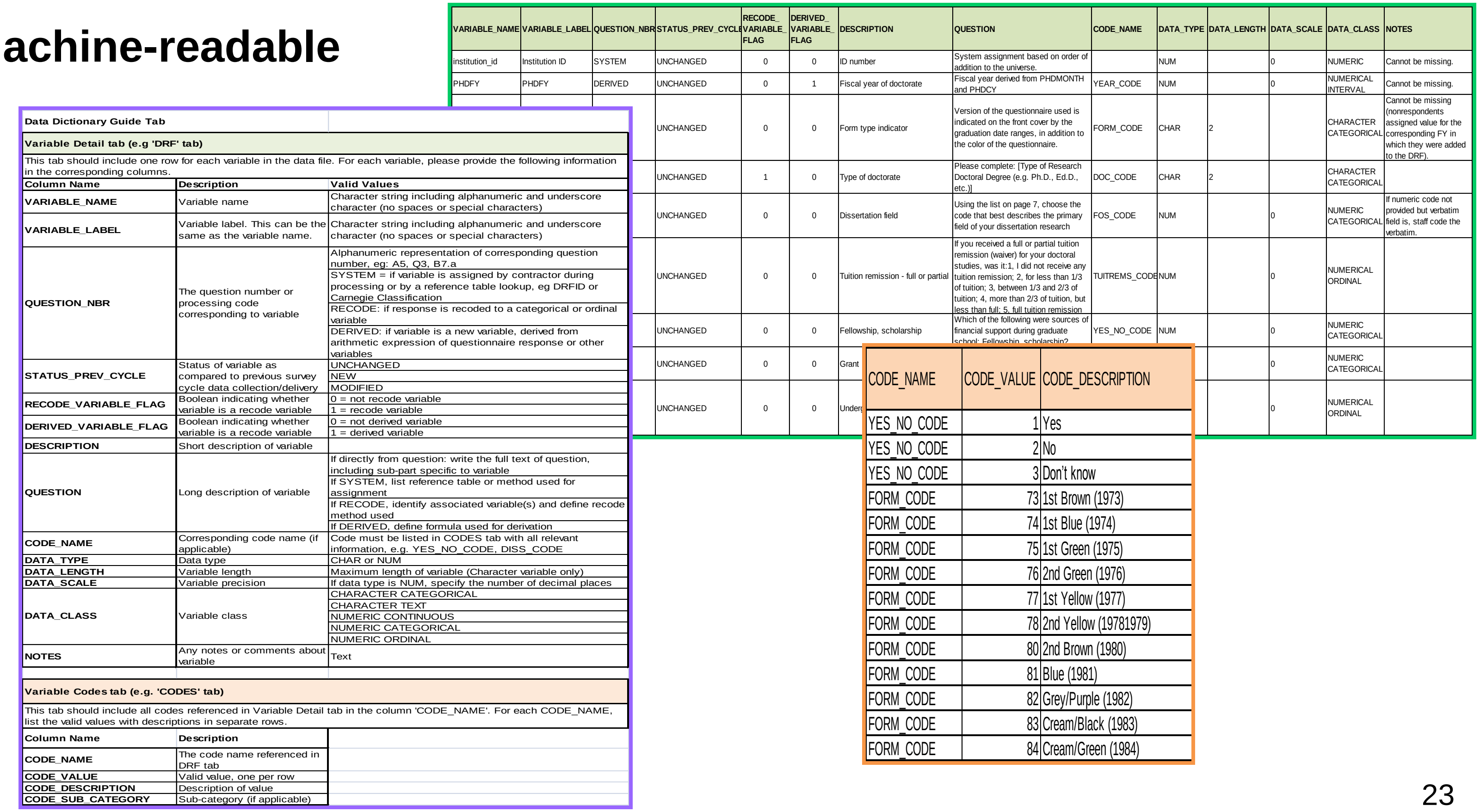
- **Shall deliver the following to the NCSES data repository at the end of each survey cycle:**
  - **Microdata and Associated Files**
    - Microdata
    - Metadata
    - Computer Code
    - Documentation
  - **Product and Report Data and Associated Files**
- **According to specified requirements**

Requirements

Data Repository Delivery Requirements

Metadata requirements

- Comprehensive
- Standard formats
- Machine-readable

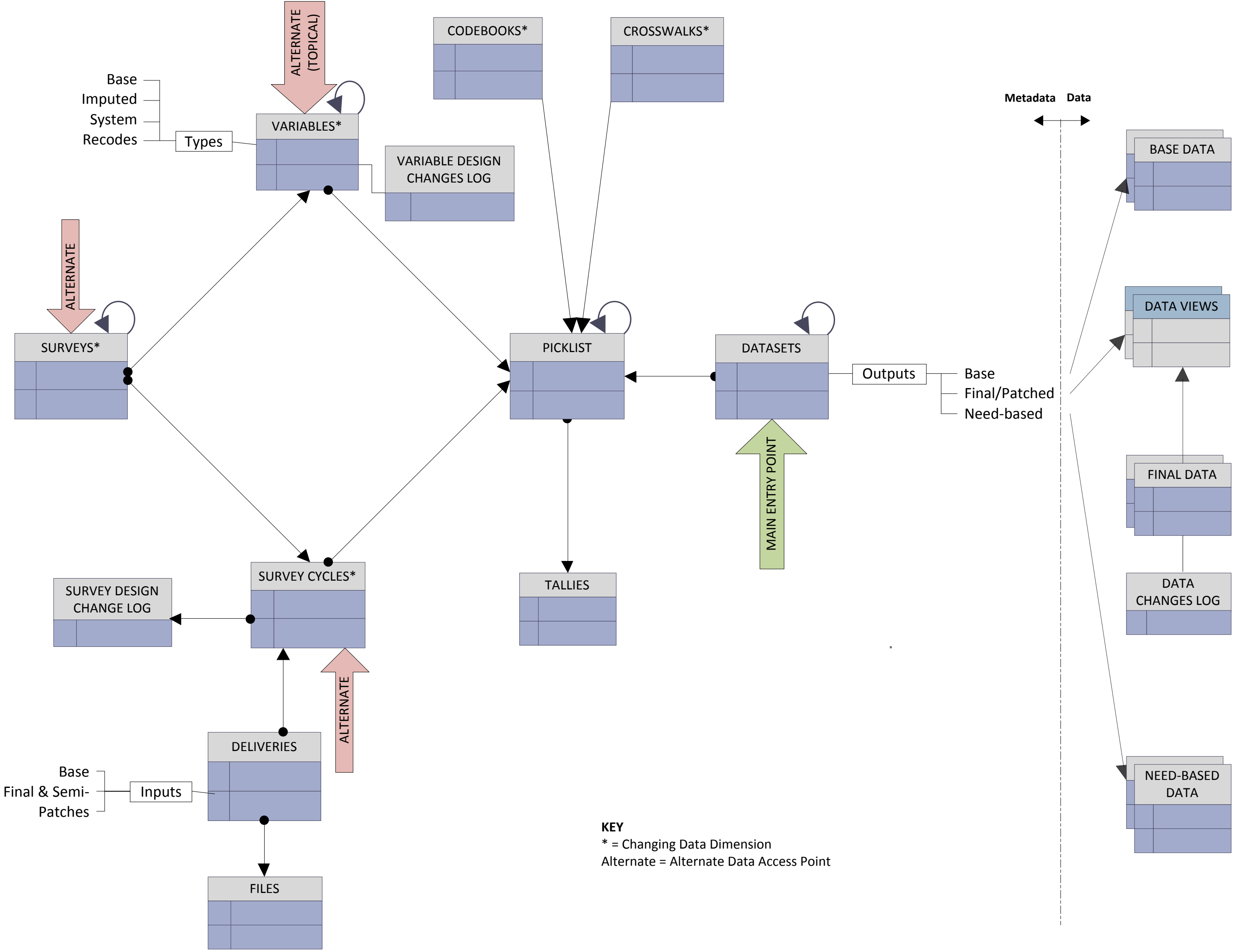


VARIABLE_NAME	VARIABLE_LABEL	QUESTION_NBR	STATUS_PREV_CYCLE	RECODE_ VARIABLE_ FLAG	DERIVED_ VARIABLE_ FLAG	DESCRIPTION	QUESTION	CODE_NAME	DATA_TYPE	DATA_LENGTH	DATA_SCALE	DATA_CLASS	NOTES
institution_id	Institution ID	SYSTEM	UNCHANGED	0	0	ID number	System assignment based on order of addition to the universe.		NUM		0	NUMERIC	Cannot be missing.
PHDFY	PHDFY	DERIVED	UNCHANGED	0	1	Fiscal year of doctorate	Fiscal year derived from PHDMONTH and PHDCY	YEAR_CODE	NUM		0	NUMERICAL INTERVAL	Cannot be missing.
			UNCHANGED	0	0	Form type indicator	Version of the questionnaire used is indicated on the front cover by the graduation date ranges, in addition to the color of the questionnaire.	FORM_CODE	CHAR	2		CHARACTER CATEGORICAL	Cannot be missing (nonrespondents assigned value for the corresponding FY in which they were added to the DRF).
			UNCHANGED	1	0	Type of doctorate	Please complete: [Type of Research Doctoral Degree (e.g. Ph.D., Ed.D., etc.)]	DOC_CODE	CHAR	2		CHARACTER CATEGORICAL	
			UNCHANGED	0	0	Dissertation field	Using the list on page 7, choose the code that best describes the primary field of your dissertation research	FOS_CODE	NUM		0	NUMERIC CATEGORICAL	If numeric code not provided but verbatim field is, staff code the verbatim.
			UNCHANGED	0	0	Tuition remission - full or partial	If you received a full or partial tuition remission (waiver) for your doctoral studies, was it:1, I did not receive any tuition remission; 2, for less than 1/3 of tuition; 3, between 1/3 and 2/3 of tuition; 4, more than 2/3 of tuition, but less than full; 5, full tuition remission	TUITREMS_CODE	NUM		0	NUMERICAL ORDINAL	
			UNCHANGED	0	0	Fellowship, scholarship	Which of the following were sources of financial support during graduate school: Fellowship, scholarship?	YES_NO_CODE	NUM		0	NUMERIC CATEGORICAL	
			UNCHANGED	0	0	Grant					0	NUMERIC CATEGORICAL	
			UNCHANGED	0	0	Underg					0	NUMERICAL ORDINAL	

Data Dictionary Guide Tab		
Variable Detail tab (e.g 'DRF' tab)		
This tab should include one row for each variable in the data file. For each variable, please provide the following information in the corresponding columns.		
Column Name	**Description**	**Valid Values**
VARIABLE_NAME	Variable name	Character string including alphanumeric and underscore character (no spaces or special characters)
VARIABLE_LABEL	Variable label. This can be the same as the variable name.	Character string including alphanumeric and underscore character (no spaces or special characters)
QUESTION_NBR	The question number or processing code corresponding to variable	Alphanumeric representation of corresponding question number, eg: A5, Q3, B7.a SYSTEM = if variable is assigned by contractor during processing or by a reference table lookup, eg DRFID or Carnegie Classification RECODE: if response is recoded to a categorical or ordinal variable DERIVED: if variable is a new variable, derived from arithmetic expression of questionnaire response or other variables
STATUS_PREV_CYCLE	Status of variable as compared to previous survey cycle data collection/delivery	UNCHANGED NEW MODIFIED
RECODE_VARIABLE_FLAG	Boolean indicating whether variable is a recode variable	0 = not recode variable 1 = recode variable
DERIVED_VARIABLE_FLAG	Boolean indicating whether variable is a recode variable	0 = not derived variable 1 = derived variable
DESCRIPTION	Short description of variable	
QUESTION	Long description of variable	If directly from question: write the full text of question, including sub-part specific to variable If SYSTEM, list reference table or method used for assignment If RECODE, identify associated variable(s) and define recode method used If DERIVED, define formula used for derivation
CODE_NAME	Corresponding code name (if applicable)	Code must be listed in CODES tab with all relevant information, e.g. YES_NO_CODE, DISS_CODE
DATA_TYPE	Data type	CHAR or NUM
DATA_LENGTH	Variable length	Maximum length of variable (Character variable only)
DATA_SCALE	Variable precision	If data type is NUM, specify the number of decimal places
DATA_CLASS	Variable class	CHARACTER CATEGORICAL CHARACTER TEXT NUMERIC CONTINUOUS NUMERIC CATEGORICAL NUMERIC ORDINAL
NOTES	Any notes or comments about variable	Text
Variable Codes tab (e.g. 'CODES' tab)		
This tab should include all codes referenced in Variable Detail tab in the column 'CODE_NAME'. For each CODE_NAME, list the valid values with descriptions in separate rows.		
Column Name	**Description**	
CODE_NAME	The code name referenced in DRF tab	
CODE_VALUE	Valid value, one per row	
CODE_DESCRIPTION	Description of value	
CODE_SUB_CATEGORY	Sub-category (if applicable)	

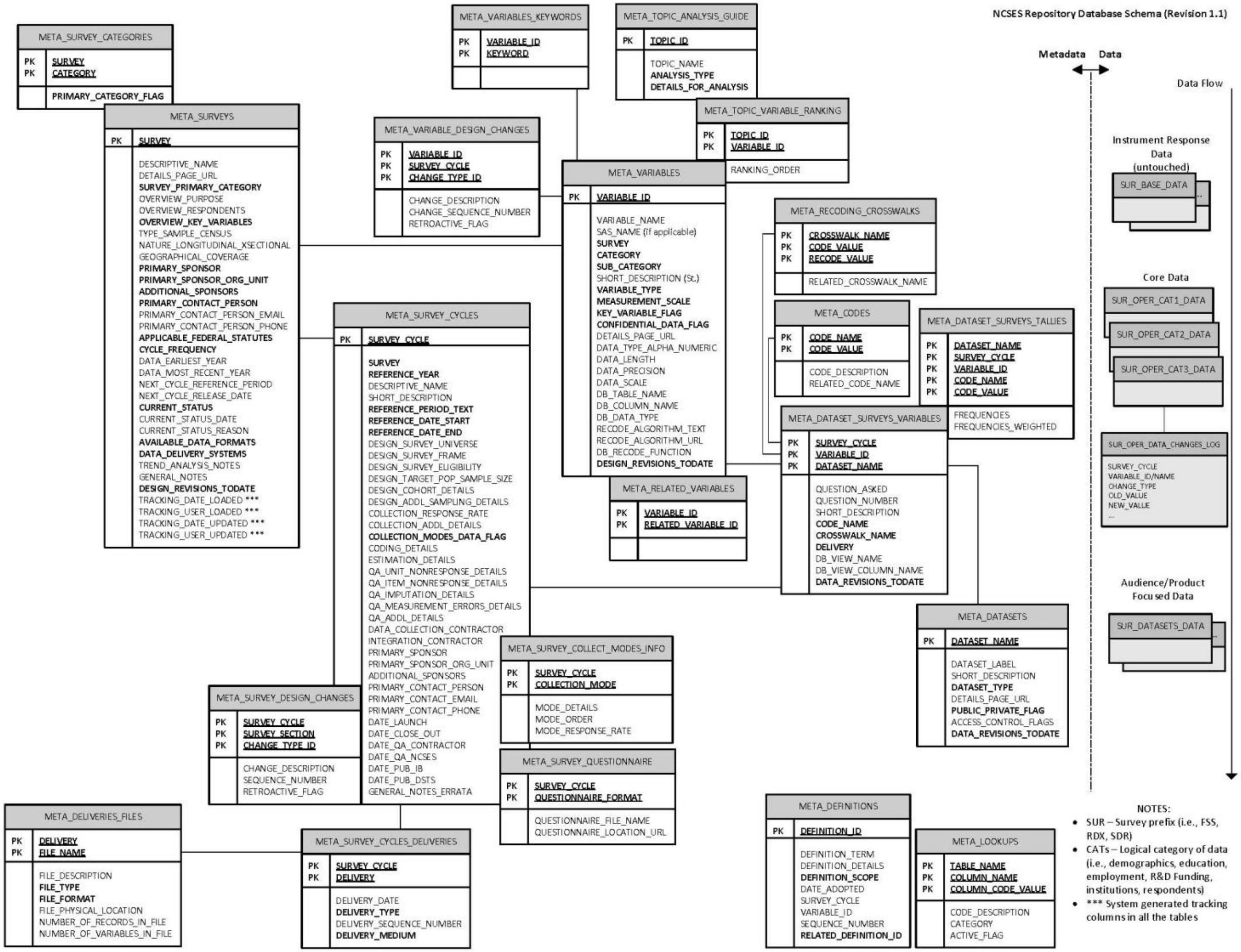
CODE_NAME	CODE_VALUE	CODE_DESCRIPTION
YES_NO_CODE	1	Yes
YES_NO_CODE	2	No
YES_NO_CODE	3	Don't know
FORM_CODE	73	1st Brown (1973)
FORM_CODE	74	1st Blue (1974)
FORM_CODE	75	1st Green (1975)
FORM_CODE	76	2nd Green (1976)
FORM_CODE	77	1st Yellow (1977)
FORM_CODE	78	2nd Yellow (19781979)
FORM_CODE	80	2nd Brown (1980)
FORM_CODE	81	Blue (1981)
FORM_CODE	82	Grey/Purple (1982)
FORM_CODE	83	Cream/Black (1983)
FORM_CODE	84	Cream/Green (1984)

Data Repository Schema


ALTERNATE (TOPICAL)
CODEBOOKS*
CROSSWALKS*
Base
Imputed
System
Recodes
Types
VARIABLES*
VARIABLE DESIGN CHANGES LOG
Metadata
Data
BASE DATA
ALTERNATE
SURVEYS*
PICKLIST
DATASETS
DATA VIEWS
Outputs
Base
Final/Patched
Need-based
MAIN ENTRY POINT
FINAL DATA
SURVEY CYCLES*
SURVEY DESIGN CHANGE LOG
TALLIES
DATA CHANGES LOG
ALTERNATE
DELIVERIES
Base
Final & Semi-Patches
Inputs
NEED-BASED DATA
FILES

KEY
* = Changing Data Dimension
Alternate = Alternate Data Access Point

Data Repository Metadata Design



NCSES Repository Database Schema (Revision 1.1)

	META_SURVEY_CATEGORIES
PK	**SURVEY**
PK	**CATEGORY**
	PRIMARY_CATEGORY_FLAG

	META_SURVEYS
PK	**SURVEY**
	DESCRIPTIVE_NAME
	DETAILS_PAGE_URL
	SURVEY_PRIMARY_CATEGORY
	OVERVIEW_PURPOSE
	OVERVIEW_RESPONDENTS
	OVERVIEW_KEY_VARIABLES
	TYPE_SAMPLE_CENSUS
	NATURE_LONGITUDINAL_XSECTIONAL
	GEOGRAPHICAL_COVERAGE
	PRIMARY_SPONSOR
	PRIMARY_SPONSOR_ORG_UNIT
	ADDITIONAL_SPONSORS
	PRIMARY_CONTACT_PERSON
	PRIMARY_CONTACT_PERSON_EMAIL
	PRIMARY_CONTACT_PERSON_PHONE
	APPLICABLE_FEDERAL_STATUTES
	CYCLE_FREQUENCY
	DATA_EARLIEST_YEAR
	DATA_MOST_RECENT_YEAR
	NEXT_CYCLE_REFERENCE_PERIOD
	NEXT_CYCLE_RELEASE_DATE
	CURRENT_STATUS
	CURRENT_STATUS_DATE
	CURRENT_STATUS_REASON
	AVAILABLE_DATA_FORMATS
	DATA_DELIVERY_SYSTEMS
	TREND_ANALYSIS_NOTES
	GENERAL_NOTES
	DESIGN_REVISIONS_TODATE
	TRACKING_DATE_LOADED ***
	TRACKING_USER_LOADED ***
	TRACKING_DATE_UPDATED ***
	TRACKING_USER_UPDATED ***

	META_VARIABLES_KEYWORDS
PK	**VARIABLE_ID**
PK	**KEYWORD**

	META_VARIABLE_DESIGN_CHANGES
PK	**VARIABLE_ID**
PK	**SURVEY_CYCLE**
PK	**CHANGE_TYPE_ID**
	CHANGE_DESCRIPTION
	CHANGE_SEQUENCE_NUMBER
	RETROACTIVE_FLAG

	META_TOPIC_ANALYSIS_GUIDE
PK	**TOPIC_ID**
	TOPIC_NAME
	ANALYSIS_TYPE
	DETAILS_FOR_ANALYSIS

	META_TOPIC_VARIABLE_RANKING
PK	**TOPIC_ID**
PK	**VARIABLE_ID**
	RANKING_ORDER

	META_VARIABLES
PK	**VARIABLE_ID**
	VARIABLE_NAME
	SAS_NAME (if applicable)
	SURVEY
	CATEGORY
	SUB_CATEGORY
	SHORT_DESCRIPTION (St.)
	VARIABLE_TYPE
	MEASUREMENT_SCALE
	KEY_VARIABLE_FLAG
	CONFIDENTIAL_DATA_FLAG
	DETAILS_PAGE_URL
	DATA_TYPE_ALPHA_NUMERIC
	DATA_LENGTH
	DATA_PRECISION
	DATA_SCALE
	DB_TABLE_NAME
	DB_COLUMN_NAME
	DB_DATA_TYPE
	RECODE_ALGORITHM_TEXT
	RECODE_ALGORITHM_URL
	DB_RECODE_FUNCTION
	DESIGN_REVISIONS_TODATE

	META_RECODING_CROSSWALKS
PK	**CROSSWALK_NAME**
PK	**CODE_VALUE**
PK	**RECODE_VALUE**
	RELATED_CROSSWALK_NAME

	META_CODES
PK	**CODE_NAME**
PK	**CODE_VALUE**
	CODE_DESCRIPTION
	RELATED_CODE_NAME

	META_DATASET_SURVEYS_TALLIES
PK	**DATASET_NAME**
PK	**SURVEY_CYCLE**
PK	**VARIABLE_ID**
PK	**CODE_NAME**
PK	**CODE_VALUE**
	FREQUENCIES
	FREQUENCIES_WEIGHTED

	META_SURVEY_CYCLES
PK	**SURVEY_CYCLE**
	SURVEY
	REFERENCE_YEAR
	DESCRIPTIVE_NAME
	SHORT_DESCRIPTION
	REFERENCE_PERIOD_TEXT
	REFERENCE_DATE_START
	REFERENCE_DATE_END
	DESIGN_SURVEY_UNIVERSE
	DESIGN_SURVEY_FRAME
	DESIGN_SURVEY_ELIGIBILITY
	DESIGN_TARGET_POP_SAMPLE_SIZE
	DESIGN_COHORT_DETAILS
	DESIGN_ADDL_SAMPLING_DETAILS
	COLLECTION_RESPONSE_RATE
	COLLECTION_ADDL_DETAILS
	COLLECTION_MODES_DATA_FLAG
	CODING_DETAILS
	ESTIMATION_DETAILS
	QA_UNIT_NONRESPONSE_DETAILS
	QA_ITEM_NONRESPONSE_DETAILS
	QA_IMPUTATION_DETAILS
	QA_MEASUREMENT_ERRORS_DETAILS
	QA_ADDL_DETAILS
	DATA_COLLECTION_CONTRACTOR
	INTEGRATION_CONTRACTOR
	PRIMARY_SPONSOR
	PRIMARY_SPONSOR_ORG_UNIT
	ADDITIONAL_SPONSORS
	PRIMARY_CONTACT_PERSON
	PRIMARY_CONTACT_EMAIL
	PRIMARY_CONTACT_PHONE
	DATE_LAUNCH
	DATE_CLOSE_OUT
	DATE_QA_CONTRACTOR
	DATE_QA_NCSES
	DATE_PUB_IB
	DATE_PUB_DSTS
	GENERAL_NOTES_ERRATA

	META_RELATED_VARIABLES
PK	**VARIABLE_ID**
PK	**RELATED_VARIABLE_ID**

	META_DATASET_SURVEYS_VARIABLES
PK	**SURVEY_CYCLE**
PK	**VARIABLE_ID**
PK	**DATASET_NAME**
	QUESTION_ASKED
	QUESTION_NUMBER
	SHORT_DESCRIPTION
	CODE_NAME
	CROSSWALK_NAME
	DELIVERY
	DB_VIEW_NAME
	DB_VIEW_COLUMN_NAME
	DATA_REVISIONS_TODATE

	META_DATASETS
PK	**DATASET_NAME**
	DATASET_LABEL
	SHORT_DESCRIPTION
	DATASET_TYPE
	DETAILS_PAGE_URL
	PUBLIC_PRIVATE_FLAG
	ACCESS_CONTROL_FLAGS
	DATA_REVISIONS_TODATE

	META_SURVEY_COLLECT_MODES_INFO
PK	**SURVEY_CYCLE**
PK	**COLLECTION_MODE**
	MODE_DETAILS
	MODE_ORDER
	MODE_RESPONSE_RATE

	META_SURVEY_DESIGN_CHANGES
PK	**SURVEY_CYCLE**
PK	**SURVEY_SECTION**
PK	**CHANGE_TYPE_ID**
	CHANGE_DESCRIPTION
	SEQUENCE_NUMBER
	RETROACTIVE_FLAG

	META_SURVEY_QUESTIONNAIRE
PK	**SURVEY_CYCLE**
PK	**QUESTIONNAIRE_FORMAT**
	QUESTIONNAIRE_FILE_NAME
	QUESTIONNAIRE_LOCATION_URL

	META_DELIVERIES_FILES
PK	**DELIVERY**
PK	**FILE_NAME**
	FILE_DESCRIPTION
	FILE_TYPE
	FILE_FORMAT
	FILE_PHYSICAL_LOCATION
	NUMBER_OF_RECORDS_IN_FILE
	NUMBER_OF_VARIABLES_IN_FILE

	META_SURVEY_CYCLES_DELIVERIES
PK	**SURVEY_CYCLE**
PK	**DELIVERY**
	DELIVERY_DATE
	DELIVERY_TYPE
	DELIVERY_SEQUENCE_NUMBER
	DELIVERY_MEDIUM

	META_DEFINITIONS
PK	**DEFINITION_ID**
	DEFINITION_TERM
	DEFINITION_DETAILS
	DEFINITION_SCOPE
	DATE_ADOPTED
	SURVEY_CYCLE
	VARIABLE_ID
	SEQUENCE_NUMBER
	RELATED_DEFINITION_ID

	META_LOOKUPS
PK	**TABLE_NAME**
PK	**COLUMN_NAME**
PK	**COLUMN_CODE_VALUE**
	CODE_DESCRIPTION
	CATEGORY
	ACTIVE_FLAG

Metadata Data

Data Flow

Instrument Response Data (untouched)

SUR_BASE_DATA ...

Core Data

SUR_OPER_CAT1_DATA

SUR_OPER_CAT2_DATA

SUR_OPER_CAT3_DATA

SUR_OPER_DATA_CHANGES_LOG

SURVEY_CYCLE
VARIABLE_ID/NAME
CHANGE_TYPE
OLD_VALUE
NEW_VALUE
...

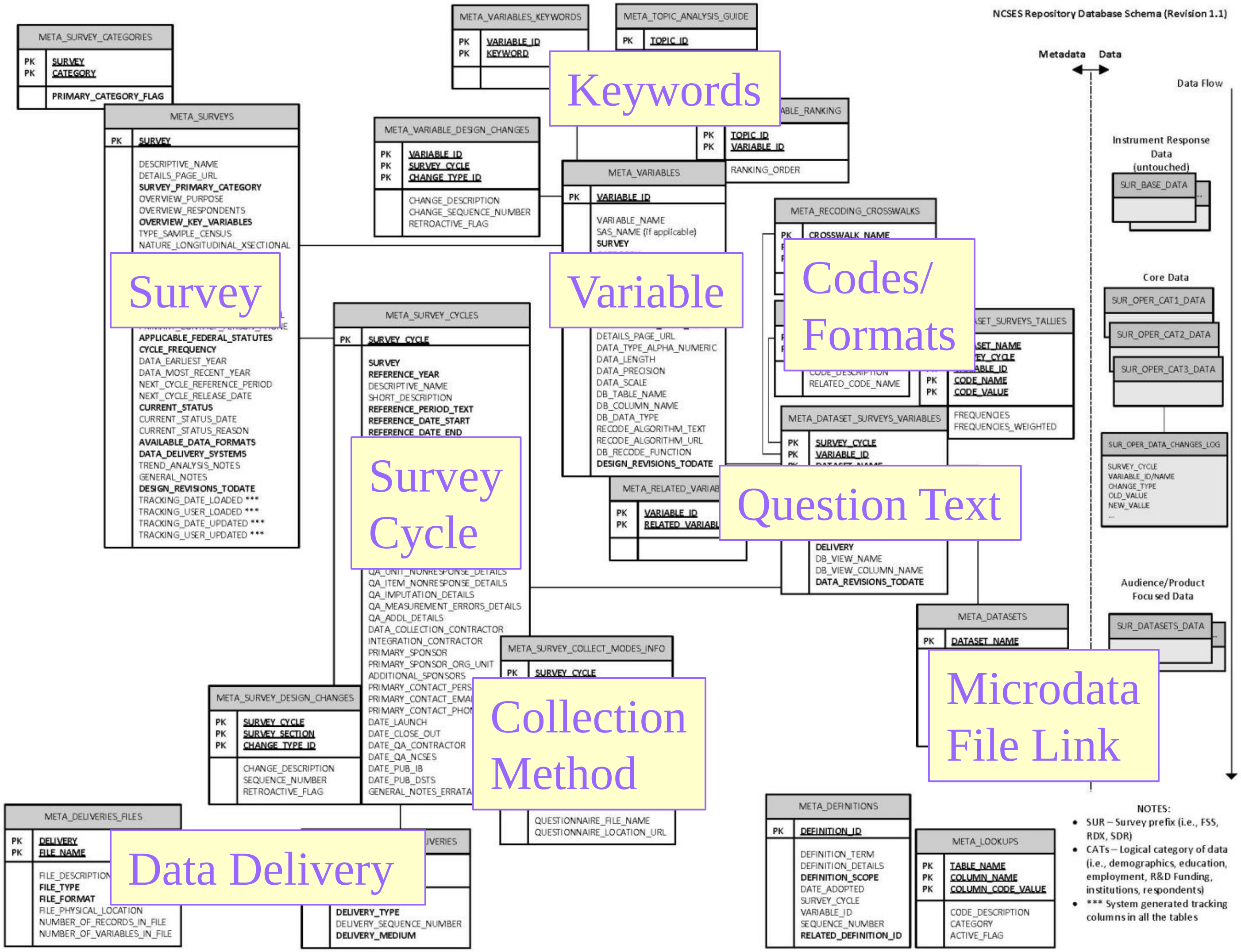
Audience/Product Focused Data

SUR_DATASETS_DATA ...

NOTES:

- SUR – Survey prefix (i.e., FSS, RDX, SDR)
- CATs – Logical category of data (i.e., demographics, education, employment, R&D Funding, institutions, respondents)
- *** System generated tracking columns in all the tables

Data Repository Metadata Design


NCSES Repository Database Schema (Revision 1.1)
Keywords
Survey
Variable
Codes/ Formats
Survey Cycle
Question Text
Collection Method
Microdata File Link
Data Delivery
META_SURVEY_CATEGORIES
PK SURVEY
PK CATEGORY
PRIMARY_CATEGORY_FLAG
META_SURVEYS
PK SURVEY
DESCRIPTIVE_NAME
DETAILS_PAGE_URL
SURVEY_PRIMARY_CATEGORY
OVERVIEW_PURPOSE
OVERVIEW_RESPONDENTS
OVERVIEW_KEY_VARIABLES
TYPE_SAMPLE_CENSUS
NATURE_LONGITUDINAL_XSECTIONAL
APPLICABLE_FEDERAL_STATUTES
CYCLE_FREQUENCY
DATA_EARLIEST_YEAR
DATA_MOST_RECENT_YEAR
NEXT_CYCLE_REFERENCE_PERIOD
NEXT_CYCLE_RELEASE_DATE
CURRENT_STATUS
CURRENT_STATUS_DATE
CURRENT_STATUS_REASON
AVAILABLE_DATA_FORMATS
DATA_DELIVERY_SYSTEMS
TREND_ANALYSIS_NOTES
GENERAL_NOTES
DESIGN_REVISIONS_TODATE
TRACKING_DATE_LOADED ***
TRACKING_USER_LOADED ***
TRACKING_DATE_UPDATED ***
TRACKING_USER_UPDATED ***
META_VARIABLES_KEYWORDS
PK VARIABLE_ID
PK KEYWORD
META_TOPIC_ANALYSIS_GUIDE
PK TOPIC_ID
ABLE_RANKING
PK TOPIC_ID
PK VARIABLE_ID
RANKING_ORDER
META_VARIABLE_DESIGN_CHANGES
PK VARIABLE_ID
PK SURVEY_CYCLE
PK CHANGE_TYPE_ID
CHANGE_DESCRIPTION
CHANGE_SEQUENCE_NUMBER
RETROACTIVE_FLAG
META_VARIABLES
PK VARIABLE_ID
VARIABLE_NAME
SAS_NAME (if applicable)
SURVEY
DETAILS_PAGE_URL
DATA_TYPE_ALPHA_NUMERIC
DATA_LENGTH
DATA_PRECISION
DATA_SCALE
DB_TABLE_NAME
DB_COLUMN_NAME
DB_DATA_TYPE
RECODE_ALGORITHM_TEXT
RECODE_ALGORITHM_URL
DB_RECODE_FUNCTION
DESIGN_REVISIONS_TODATE
META_RECODING_CROSSWALKS
PK CROSSWALK_NAME
CODE_DESCRIPTION
RELATED_CODE_NAME
SET_SURVEYS_TALLIES
SET_NAME
EY_CYCLE
ABLE_ID
PK CODE_NAME
PK CODE_VALUE
FREQUENCIES
FREQUENCIES_WEIGHTED
META_DATASET_SURVEYS_VARIABLES
PK SURVEY_CYCLE
PK VARIABLE_ID
DELIVERY
DB_VIEW_NAME
DB_VIEW_COLUMN_NAME
DATA_REVISIONS_TODATE
META_RELATED_VARIAB
PK VARIABLE_ID
PK RELATED_VARIABL
META_SURVEY_CYCLES
PK SURVEY_CYCLE
SURVEY
REFERENCE_YEAR
DESCRIPTIVE_NAME
SHORT_DESCRIPTION
REFERENCE_PERIOD_TEXT
REFERENCE_DATE_START
REFERENCE_DATE_END
QA_UNIT_NONRESPONSE_DETAILS
QA_ITEM_NONRESPONSE_DETAILS
QA_IMPUTATION_DETAILS
QA_MEASUREMENT_ERRORS_DETAILS
QA_ADDL_DETAILS
DATA_COLLECTION_CONTRACTOR
INTEGRATION_CONTRACTOR
PRIMARY_SPONSOR
PRIMARY_SPONSOR_ORG_UNIT
ADDITIONAL_SPONSORS
PRIMARY_CONTACT_PERS
PRIMARY_CONTACT_EMAI
PRIMARY_CONTACT_PHON
DATE_LAUNCH
DATE_CLOSE_OUT
DATE_QA_CONTRACTOR
DATE_QA_NCSES
DATE_PUB_IB
DATE_PUB_DSTS
GENERAL_NOTES_ERRATA
META_SURVEY_COLLECT_MODES_INFO
PK SURVEY_CYCLE
QUESTIONNAIRE_FILE_NAME
QUESTIONNAIRE_LOCATION_URL
META_SURVEY_DESIGN_CHANGES
PK SURVEY_CYCLE
PK SURVEY_SECTION
PK CHANGE_TYPE_ID
CHANGE_DESCRIPTION
SEQUENCE_NUMBER
RETROACTIVE_FLAG
META_DELIVERIES_FILES
PK DELIVERY
PK FILE_NAME
FILE_DESCRIPTION
FILE_TYPE
FILE_FORMAT
FILE_PHYSICAL_LOCATION
NUMBER_OF_RECORDS_IN_FILE
NUMBER_OF_VARIABLES_IN_FILE
IVERIES
DELIVERY_TYPE
DELIVERY_SEQUENCE_NUMBER
DELIVERY_MEDIUM
META_DATASETS
PK DATASET_NAME
META_DEFINITIONS
PK DEFINITION_ID
DEFINITION_TERM
DEFINITION_DETAILS
DEFINITION_SCOPE
DATE_ADOPTED
SURVEY_CYCLE
VARIABLE_ID
SEQUENCE_NUMBER
RELATED_DEFINITION_ID
META_LOOKUPS
PK TABLE_NAME
PK COLUMN_NAME
PK COLUMN_CODE_VALUE
CODE_DESCRIPTION
CATEGORY
ACTIVE_FLAG
Metadata
Data
Data Flow
Instrument Response Data (untouched)
SUR_BASE_DATA
Core Data
SUR_OPER_CAT1_DATA
SUR_OPER_CAT2_DATA
SUR_OPER_CAT3_DATA
SUR_OPER_DATA_CHANGES_LOG
SURVEY_CYCLE
VARIABLE_ID/NAME
CHANGE_TYPE
OLD_VALUE
NEW_VALUE
Audience/Product Focused Data
SUR_DATASETS_DATA

NOTES:

- SUR – Survey prefix (i.e., FSS, RDX, SDR)
- CATs – Logical category of data (i.e., demographics, education, employment, R&D Funding, institutions, respondents)
- *** System generated tracking columns in all the tables

Next Steps

Data Infrastructure

- **Load microdata**
- **Load metadata**

Data Tools and Access

- **Implement integration technologies**
- **Provide training for staff on existing data tools**
- **Pilot dynamic data tables**

Thank you!

Contact Information
Kimberly Noonan
knoonan@nsf.gov
703-292-5159