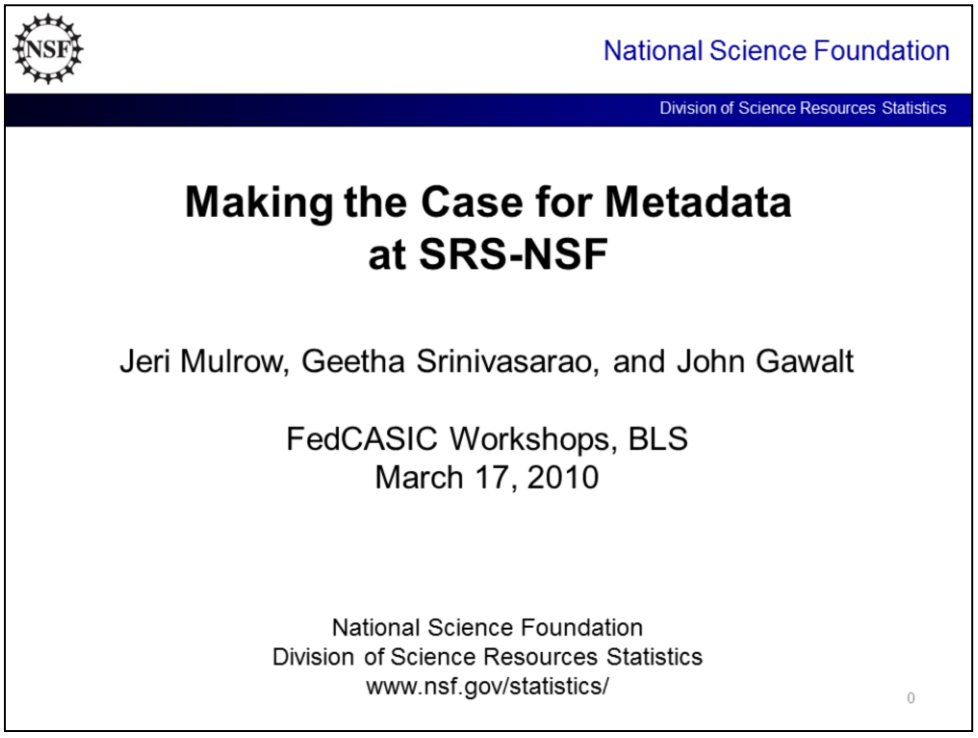
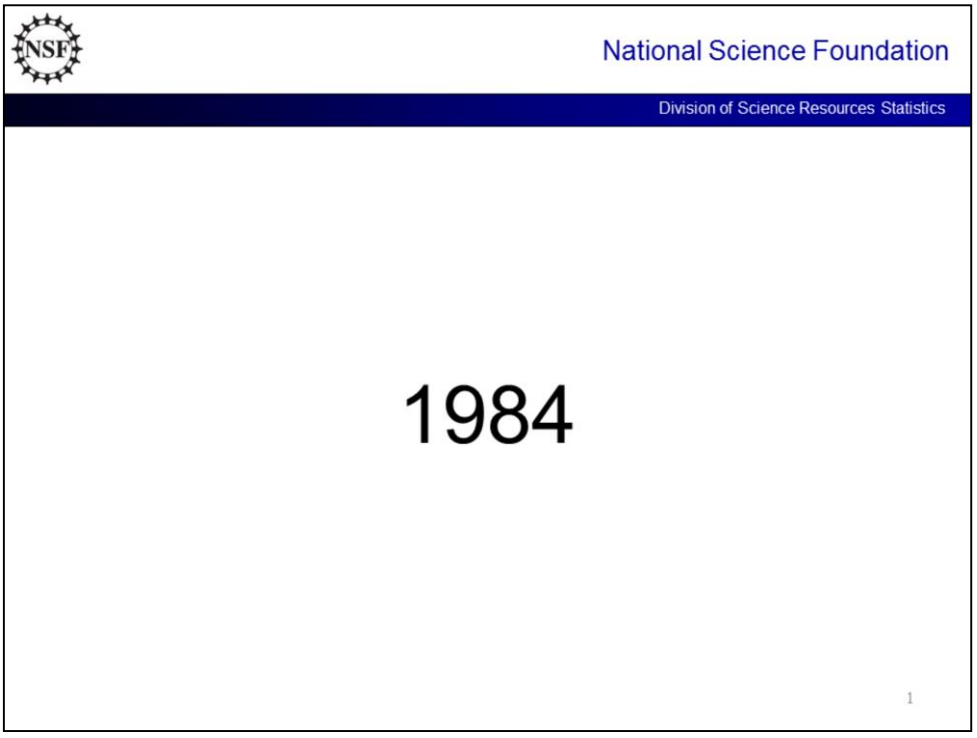


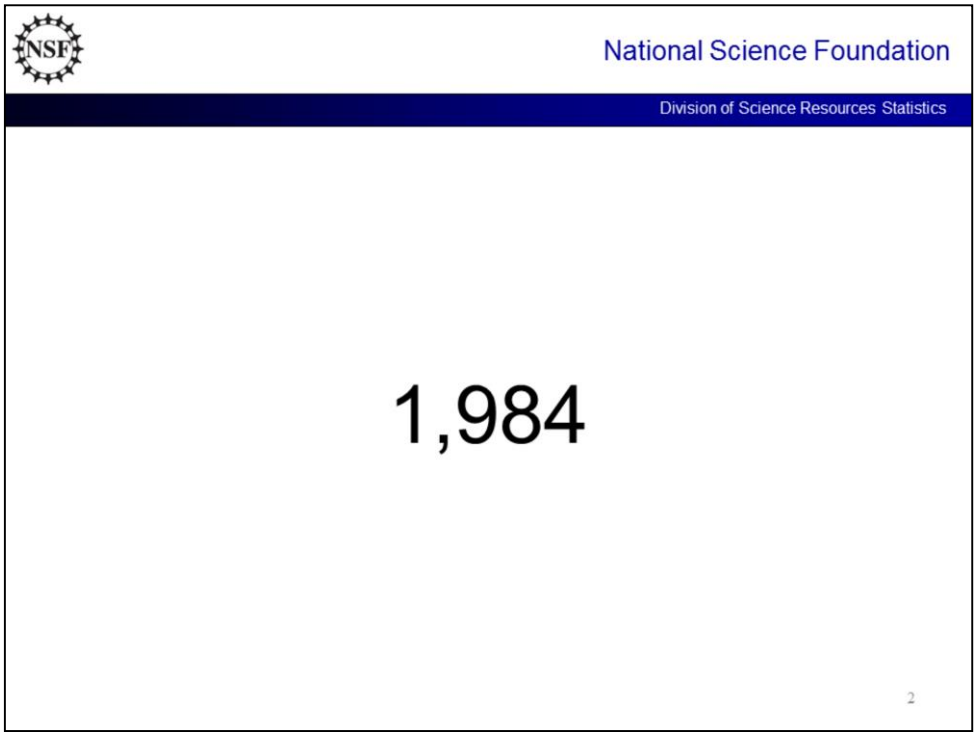
Good Afternoon. My name is Jeri Mulrow, I am a mathematical statistician and I am a heavy user of metadata. You could almost say I am addicted to metadata, but maybe we shouldn't go there.

But seriously before I get started, I would like to thank Dan Gillman for putting this session together and for inviting me to speak.

I would also like to acknowledge my co-authors and colleagues, Geetha and John who have really been the ones to do the bulk of the work on metadata at SRS.



1984. What is this? A year? A book?

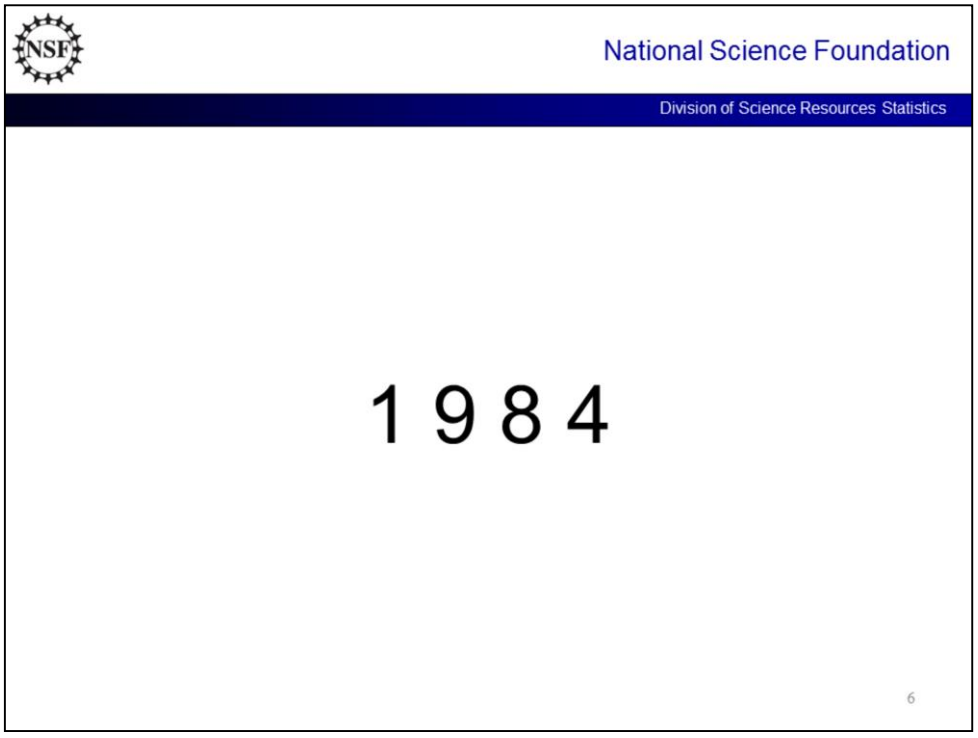


A number 1,984

1

1 9

198



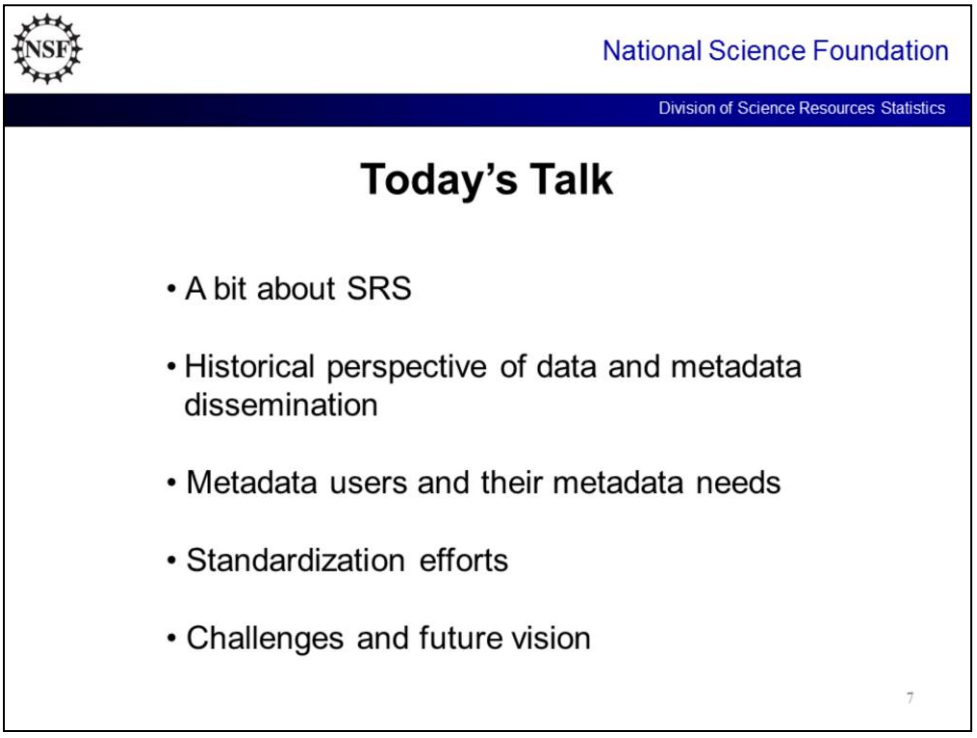
A sequence of numbers: 1 9 8 4

Without any context we don't know what it is. We aren't sure how to interpret it.

Without metadata, without context, we don't know what it means.

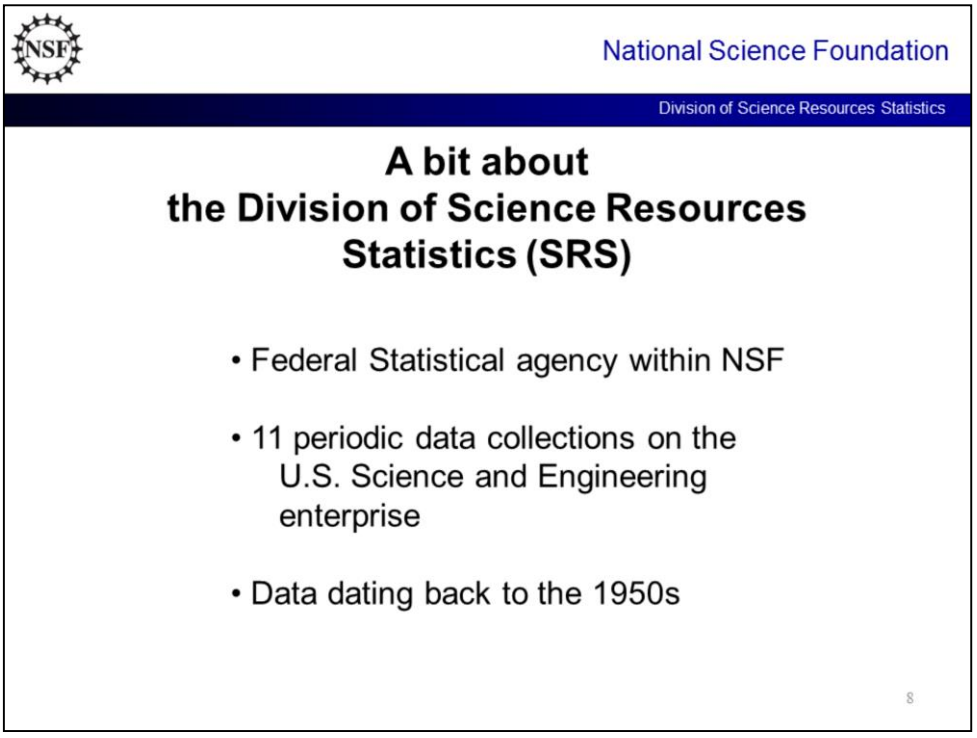
Metadata is important. Done.

But what do I want to talk about today?



To get things going, I'll tell you a little bit about the Division of Science Resources Statistics, in case you don't know.

I'll give a historical perspective of data and metadata dissemination at SRS. Then talk about metadata users (like myself) and our metadata needs. I'll mention some of the standardization efforts that have gone on, but I won't go into those in detail and then tell you about some of the challenges we have faced at SRS in pulling the metadata together and our vision for the near future.



The Division of Science Resources Statistics is the Federal Statistical Agency within the NSF. We are one of the smallest statistical agencies. As such, SRS is responsible for national data collections on the U.S. science and engineering enterprise.

For example, we collect data on research & development (R&D) – who is doing it, where are they doing it, how much are they spending on it?

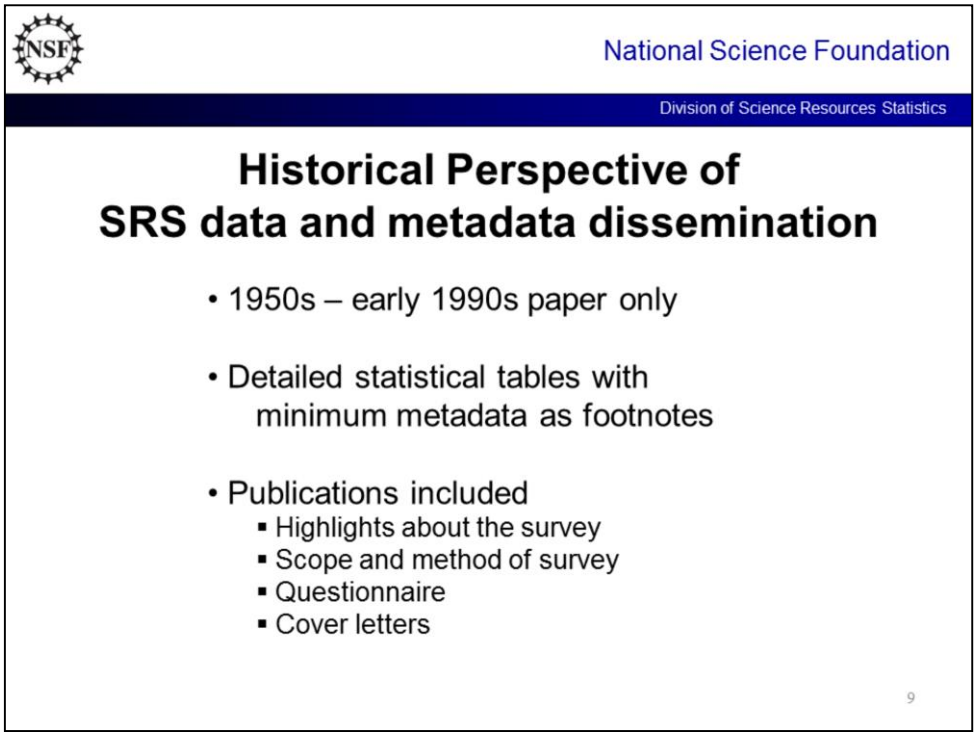
Because people do R&D, we collect data on Scientist and engineers – who are they, what education do they have, where are they employed?

To get this information, we run 11 periodic data collections. We have establishment surveys that go to businesses and academic institutions. We have demographic surveys. Some of our collections are intended to be censuses, others are sample surveys.

We have data going back to the 1950s just after NSF was established.

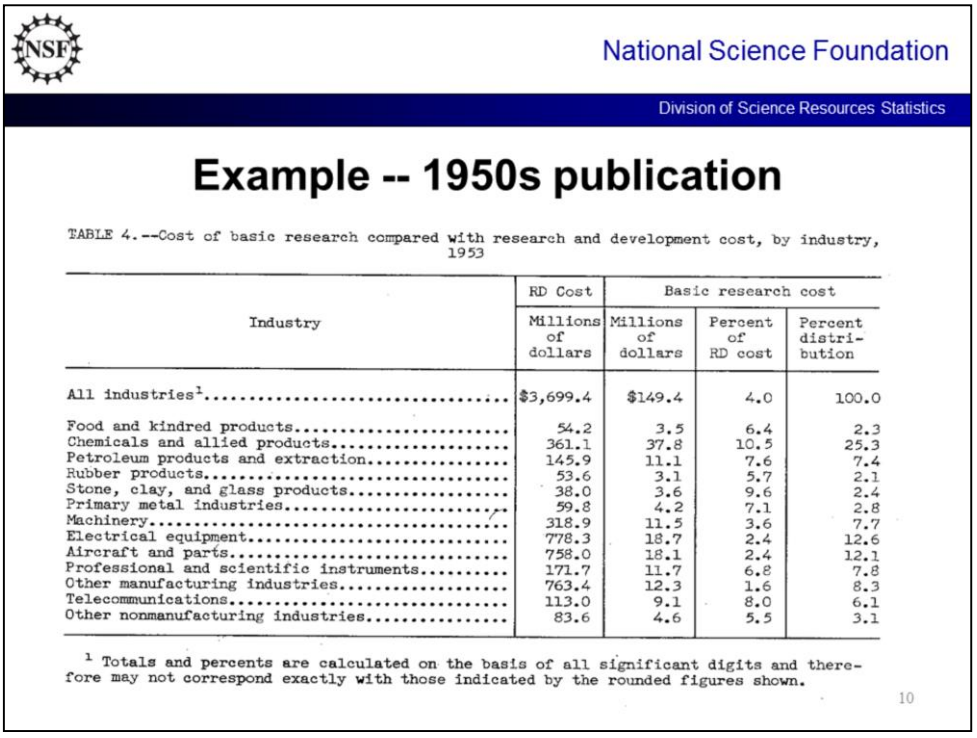
For a small agency, we run the gamet.

We have data dating back to the early 1950s shortly after NSF was founded.



Let me give you a bit of a historical perspective on SRS data and metadata dissemination. Many of you may recognize your own agencies, with some variation of course, in this perspective.

From the 1950s to the early 1990s we disseminated all of our data via paper mostly in for form of detailed statistical tables, which included a minimal amount of metadata using in the form of footnotes. We also provided additional context or information about the data in a separate publication which provided highlights of the survey, covered the scope and methods, and included a copy of the questionnaire and accompanying cover letters.



National Science Foundation

Division of Science Resources Statistics

Example -- 1950s publication

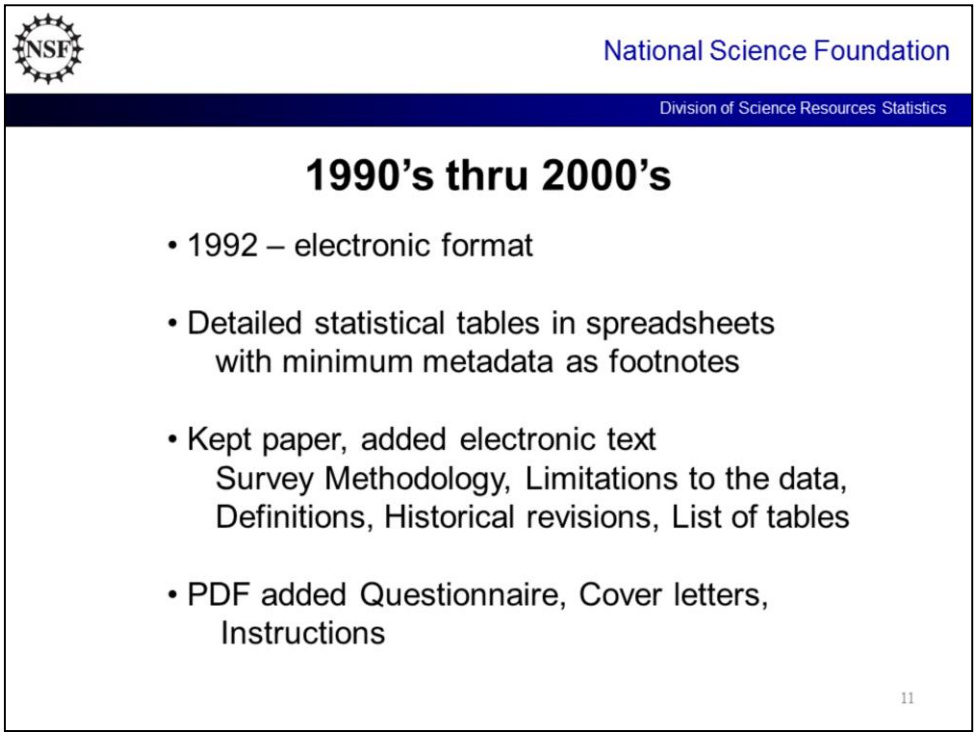
TABLE 4.--Cost of basic research compared with research and development cost, by industry, 1953

Industry	RD Cost	Basic research cost		
	Millions of dollars	Millions of dollars	Percent of RD cost	Percent distribution
All industries[1]	$3,699.4	$149.4	4.0	100.0
Food and kindred products	54.2	3.5	6.4	2.3
Chemicals and allied products	361.1	37.8	10.5	25.3
Petroleum products and extraction	145.9	11.1	7.6	7.4
Rubber products	53.6	3.1	5.7	2.1
Stone, clay, and glass products	38.0	3.6	9.6	2.4
Primary metal industries	59.8	4.2	7.1	2.8
Machinery	318.9	11.5	3.6	7.7
Electrical equipment	778.3	18.7	2.4	12.6
Aircraft and parts	758.0	18.1	2.4	12.1
Professional and scientific instruments	171.7	11.7	6.8	7.8
Other manufacturing industries	763.4	12.3	1.6	8.3
Telecommunications	113.0	9.1	8.0	6.1
Other nonmanufacturing industries	83.6	4.6	5.5	3.1

[1] Totals and percents are calculated on the basis of all significant digits and therefore may not correspond exactly with those indicated by the rounded figures shown.

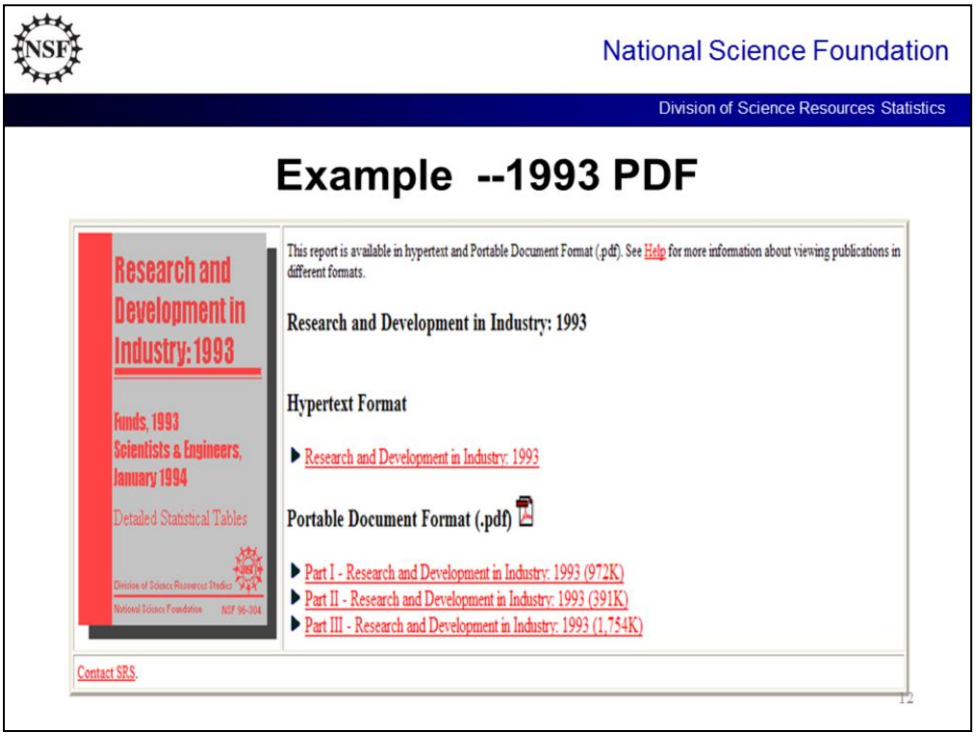
10

Here is an example of a very early publication. We see the data and a bit of metadata about what the values mean – for example, millions of dollars in R&D costs.

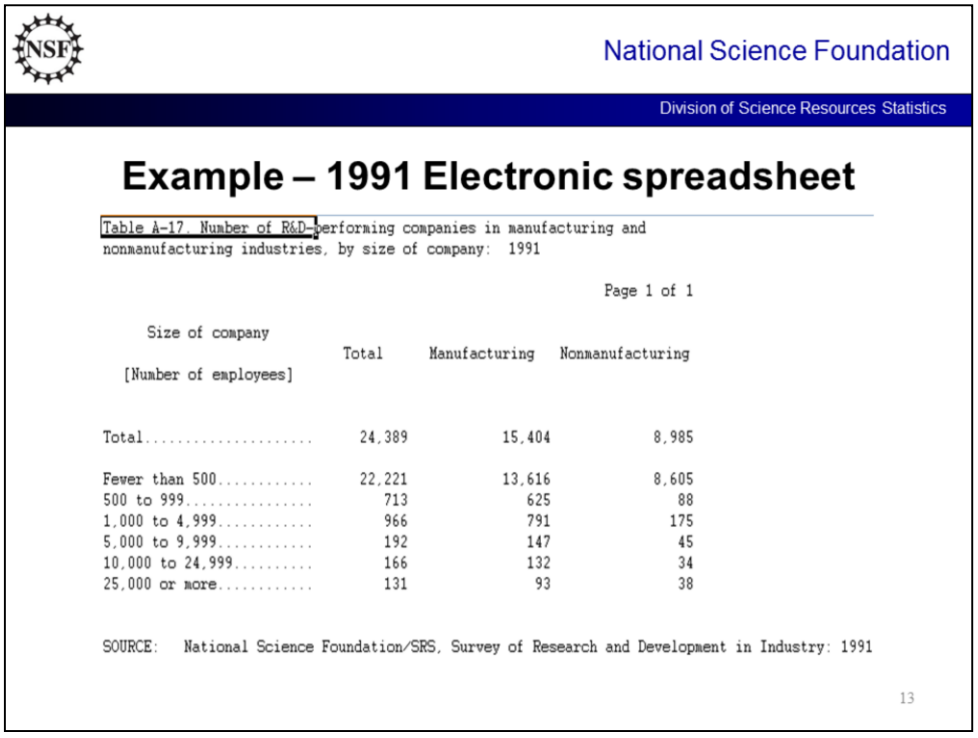


Then in the 1990s we entered the electronic age. We kept the paper format and ADDED electronic spreadsheets for the DSTs. The metadata remained pretty stable over this time. We kept the paper format and ADDED the information in electronic text, covering about the same things as we did before, adding some new information, such as limitations of the data and a historical perspective to the data. We ADDED pdf's of the questionnaire, cover letters and instructions.

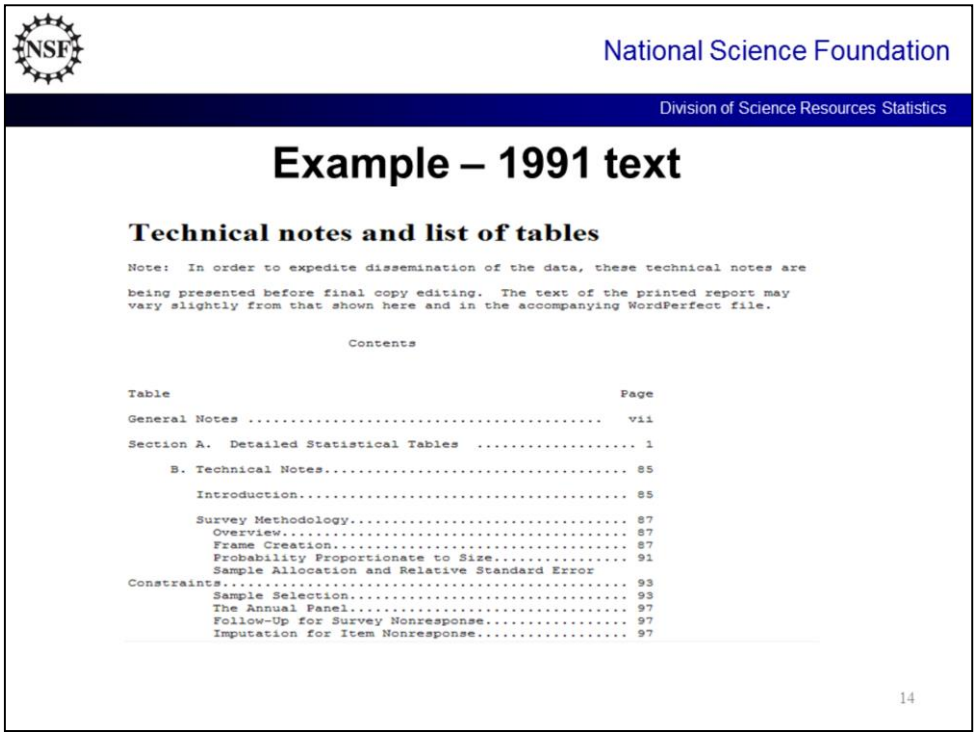
Ah, everything is now available on the web! Let's take a look at it.



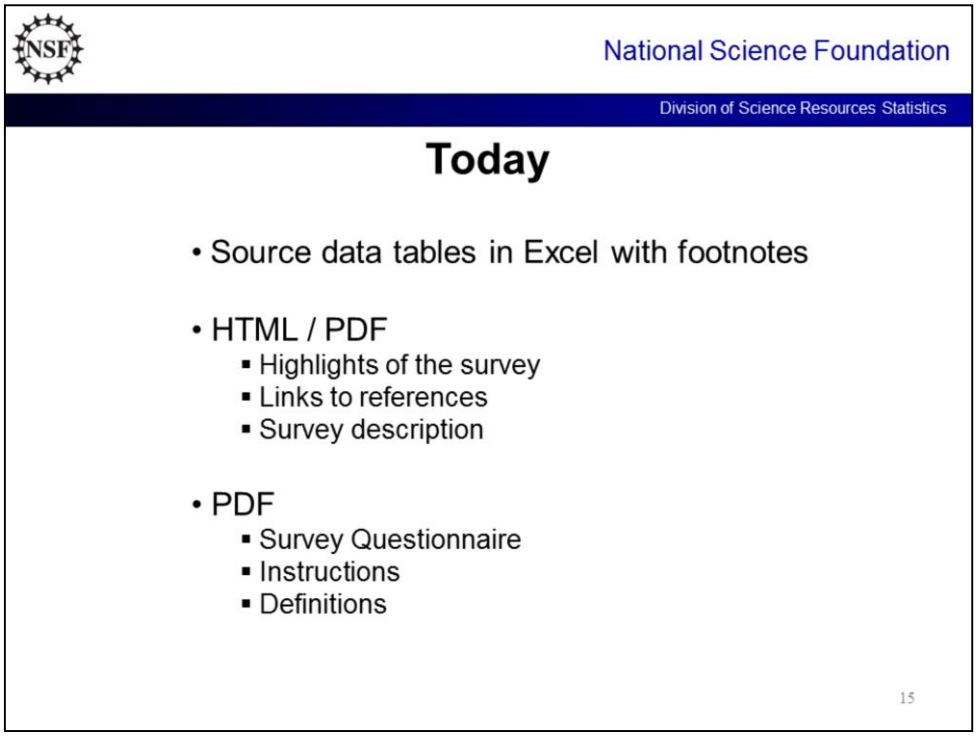
Here is an example of an early electronic publication. Basically we moved our paper to the web. We made it look like the paper.



Here is an example of an early electronic spreadsheet in Excel. We kept the formatting of the paper and moved it to the web.



Accompanying text with the tables was put on the web, in a paper-like version.



And today. Where are we today? For our detailed statistical tables, we have the source data in Excel, still with footnotes.

We have html and pdf formats for our publications and metadata in pdf format for the survey questionnaire, instructions and definitions.

We still have paper but are trying to move away from it. There are still vocal users of paper though.

Example – 2007 Excel spreadsheet

TABLE 1. Funds expended for industrial R&D performance, by source of funds, size of company, and net sales: 2006 and 2007

Selected characteristic	2006	2007	2006	2007
	Current $millions		2000 constant $millions	
Total industrial R&D performance	247,669	269,267	212,271	224,732
Source of funds				
Company and other nonfederal	223,365	242,682	191,440	202,544
Federal	24,304	26,585	20,830	22,188
Size of company (number of employees)				
5–24	7,207	10,854	6,177	9,059
25–49	D	7,884	D	6,577
50–99	9,064	10,068	7,769	8,403
100–249	13,306	13,354	11,404	11,145
250–499	D	8,258	D	6,889
500–999	13,360	14,279	11,451	11,917
1,000–4,999	37,866	41,103	32,454	34,305
5,000–9,999	20,434	22,673	17,513	18,923
10,000–24,999	37,865	45,946	32,453	38,347
25,000 or more	92,925	94,848	79,644	79,161
Net sales*	6,642,500	7,027,049	5,693,116	5,864,818

D = suppressed to avoid disclosure of confidential information

*Dollar values for goods sold or services rendered by companies that perform R&D in the United States to customers outside the company, including the federal government, less such items as returns, allowances, freight charges, and excise taxes. Excludes intracompany transfers and sales by foreign subsidiaries but includes transfers to foreign subsidiaries and export sales to foreign companies.

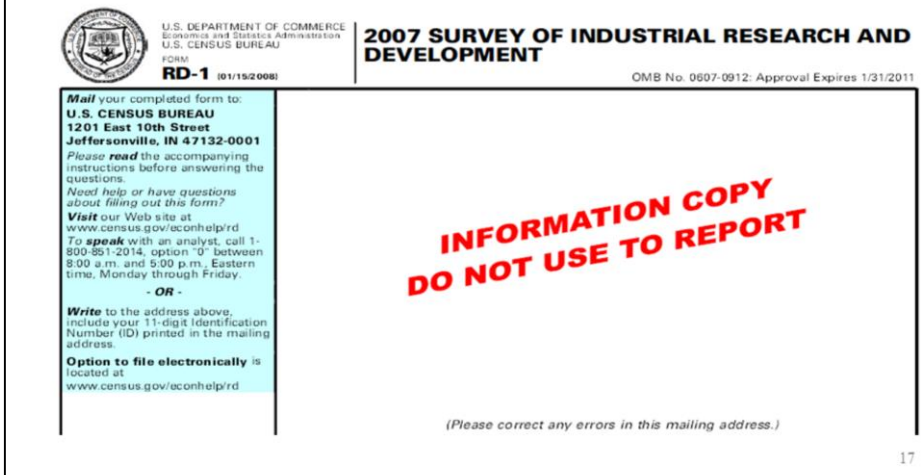
NOTES: Detail may not add to total because of rounding. Excludes data for federally funded research and development centers. 2000 gross domestic product implicit price deflators were used to convert current to constant dollars.

SOURCE: National Science Foundation/Division of Science Resources Statistics, Survey of Industrial Research and

16

Here is an example of a recent Excel spreadsheet. It is still very formatted and made to look like paper. It is not the easiest to use for an analyst without reformatting but it reads well to those who don't want to do any data manipulations.

Example -- 2007 SIRD1



U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU
FORM
RD-1 (01/15/2008)

2007 SURVEY OF INDUSTRIAL RESEARCH AND DEVELOPMENT

OMB No. 0607-0912: Approval Expires 1/31/2011

Mail your completed form to:
U.S. CENSUS BUREAU
1201 East 10th Street
Jeffersonville, IN 47132-0001

Please ***read*** the accompanying instructions before answering the questions.

Need help or have questions about filling out this form?

Visit our Web site at www.census.gov/econhelp/rd

To ***speak*** with an analyst, call 1-800-851-2014, option "0" between 8:00 a.m. and 5:00 p.m., Eastern time, Monday through Friday.

- OR -

Write to the address above, include your 11-digit Identification Number (ID) printed in the mailing address.

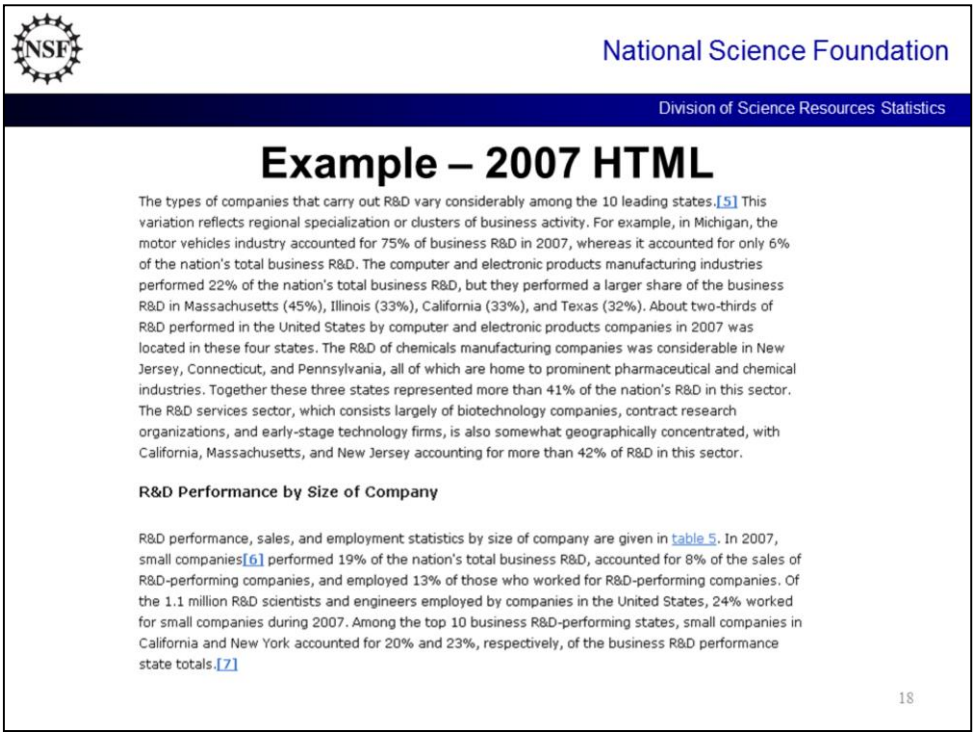
Option to file electronically is located at www.census.gov/econhelp/rd

INFORMATION COPY
DO NOT USE TO REPORT

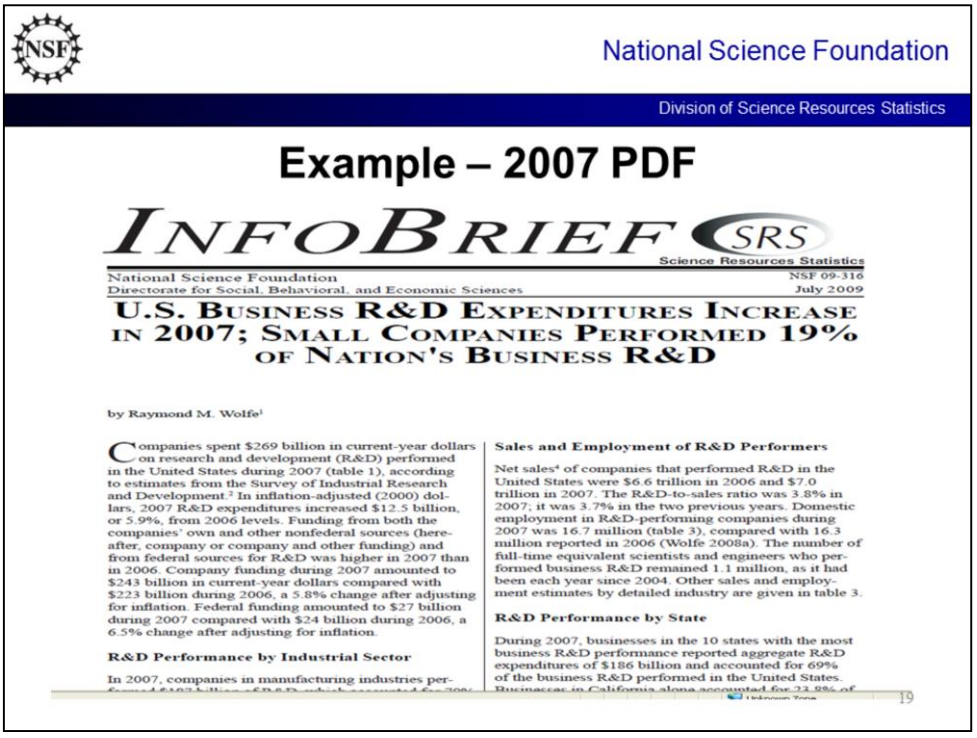
(Please correct any errors in this mailing address.)

17

We continue to have pdfs of our questionnaires.

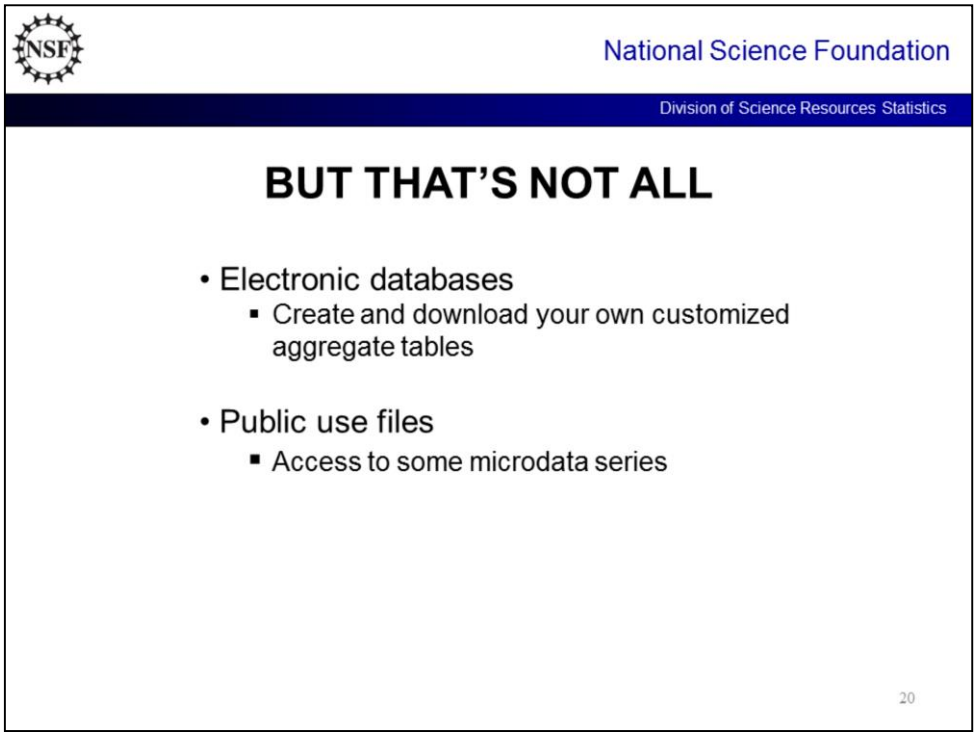


Our text is now in HTML format with embedded links. Ah progress!



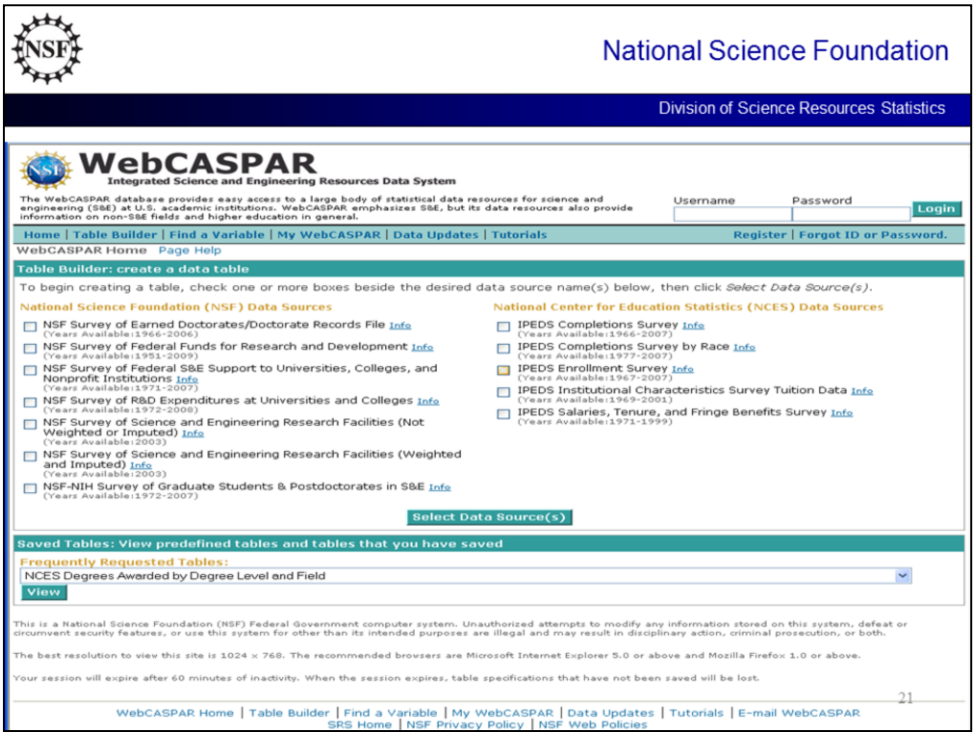
Here is an example of one of our on-line publications. Looks like paper.

That covers an historical perspective of our detailed statistical tables, but wait, there is more.



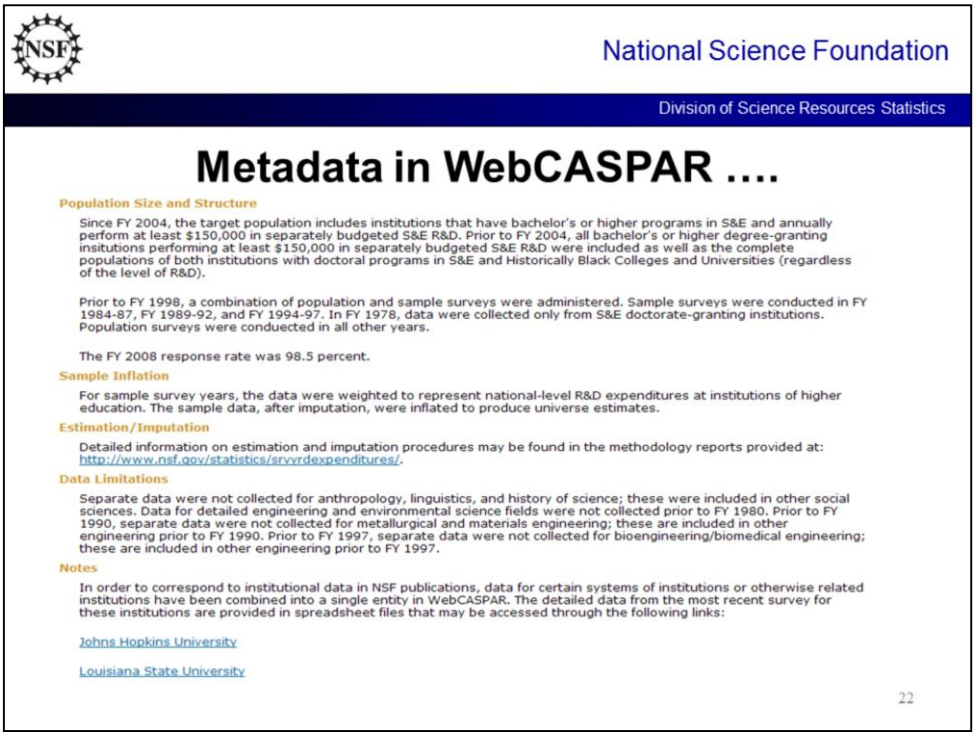
We wanted to be able to provide more access to the microdata so we developed two different electronic databases where a user can create and download their own customized tables. And we have a few public use files. Both of these formats are on the web and accessible to the general public.

We also have data licenses which allow for microdata access, but I am not going to talk about them specifically here as they have the same metadata association with them as the electronic databases.



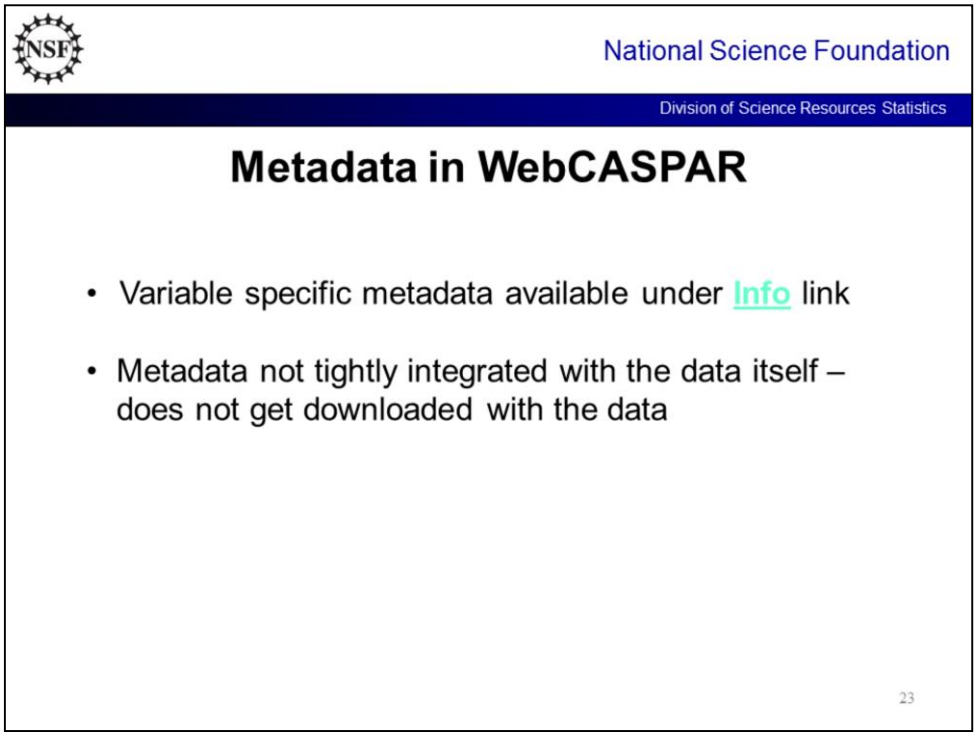
Here is a screenshot of one of our electronic databases. This one is called WebCASPAR – it is an Integrated Science and Engineering Resources Data System for the web and provides access to several of our data series along with data from NCES in an integrated format.

Metadata is available under the Info links.



So if I click on the Info button, we get this type of screen popping up, which gives us information about the population, the sample, data limitations and so on.

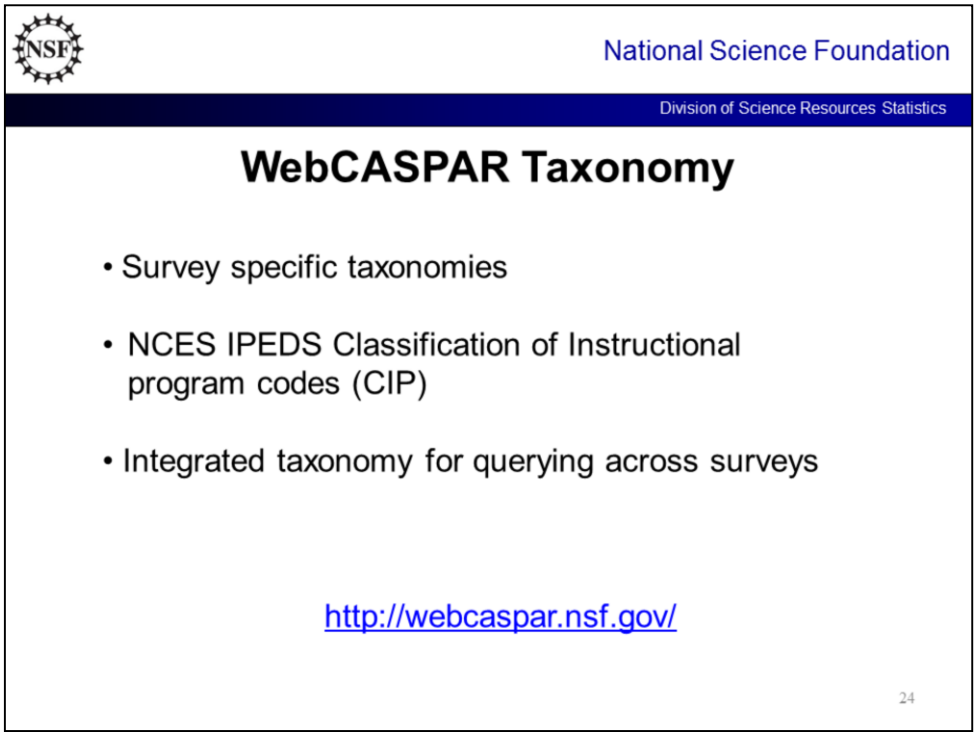
As you go further into the database, there are Info links which lead you to other metadata.



In WebCASPAR, we have included variable specific metadata, available under the Info link. And there is relatively easy access to other metadata about the survey.

Access to the questionnaires and cover letters is outside of the database.

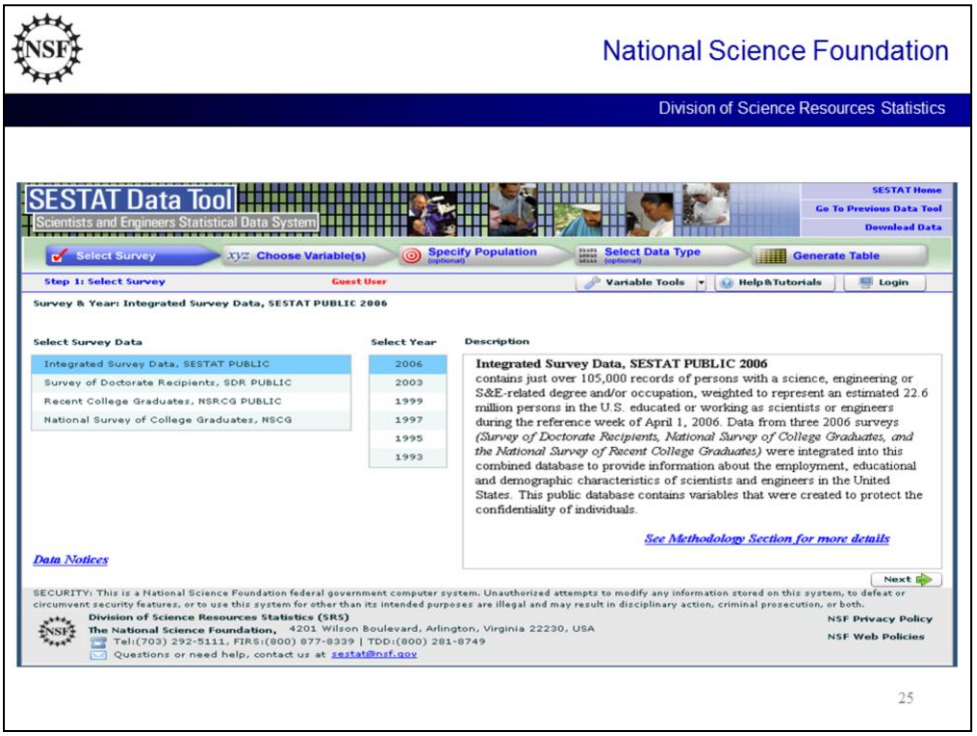
Unfortunately, the metadata are not tightly integrated with the data itself, so when a user downloads their customized table, the metadata does not come automatically with it. The user has to get the metadata in a separate step.



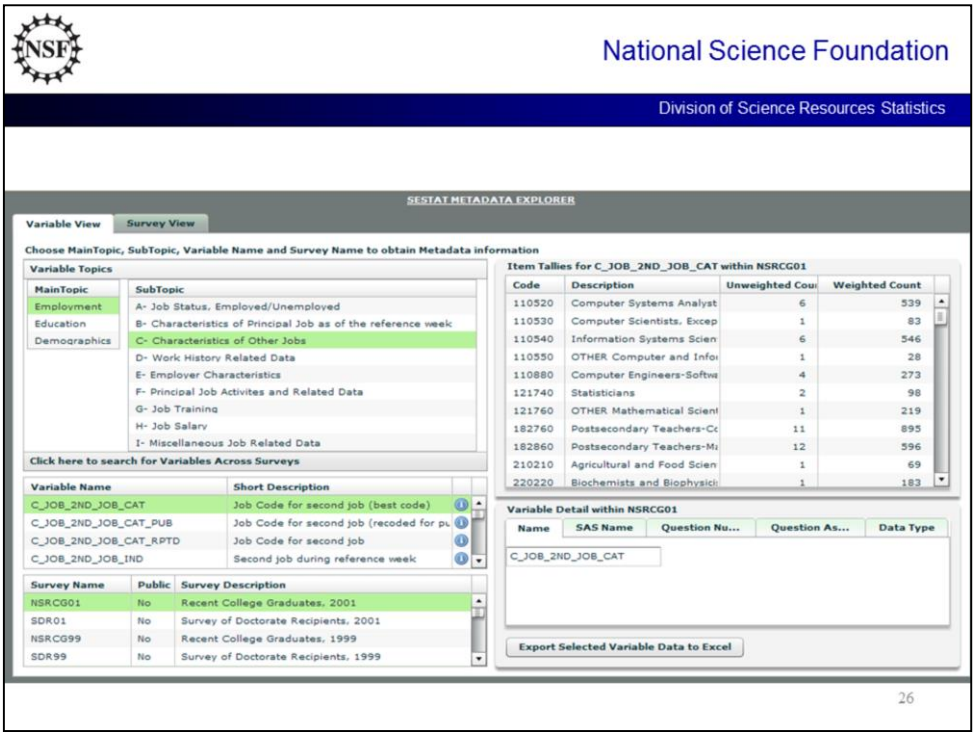
And something that is hidden, but very important, is the set of taxonomies in WebCASPAR. Because this database combines data across several surveys, the system has several different taxonomies for our fields of science built into each. It includes the survey specific taxonomies, which are all not the same, the NCES IPEDS classification, AND an integrated taxonomy that allows for querying across surveys.

This taxonomy information does not come along with the data.

It adds a level of complexity to the data, metadata, and to the system.

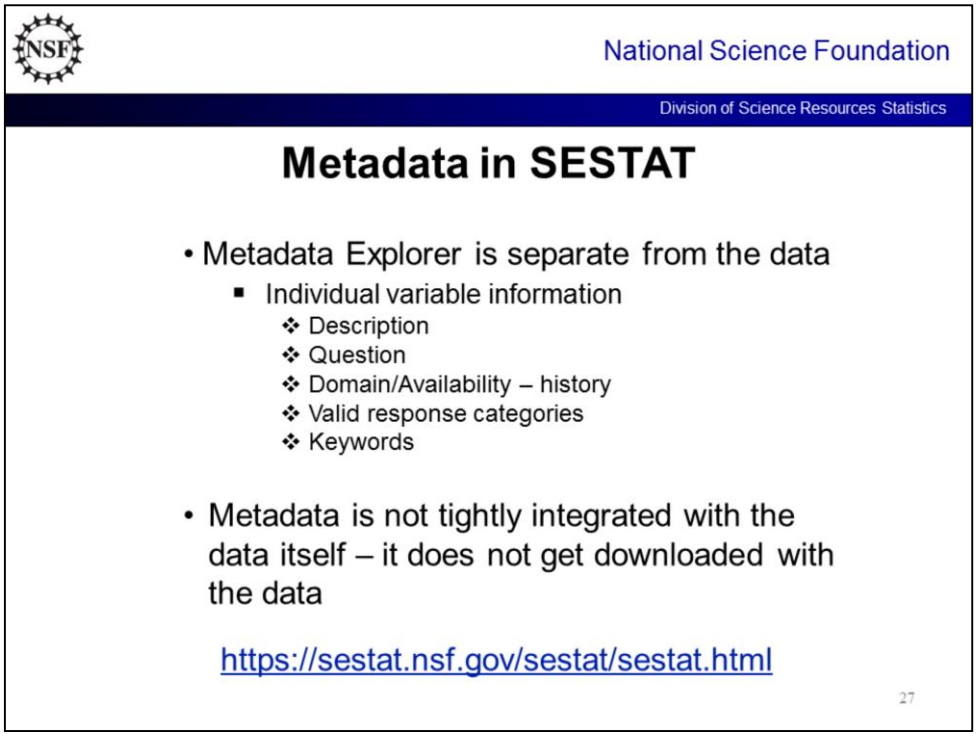


Here is a screenshot of our other electronic database, called SESTAT – the Scientists and Engineers Statistical Data System. This database covers three of our demographic surveys, plus an integrated set.



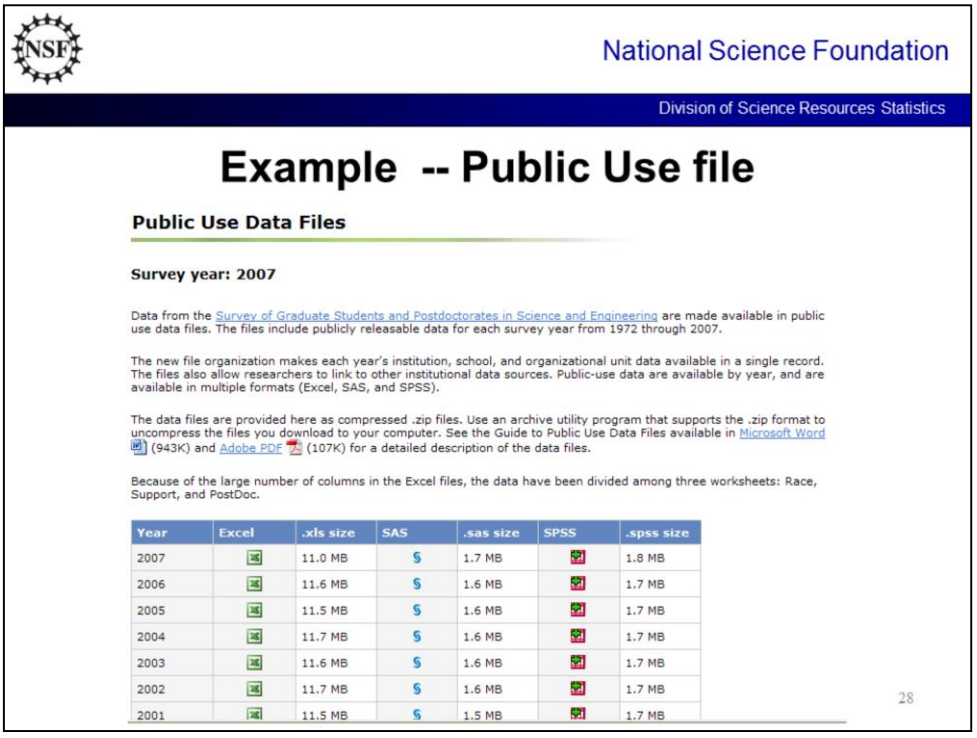
We have built even more metadata into this database. It has a separate Metadata Explorer, shown in this screenshoot, which provides quite a bit of variable specific information, including the variable response categories, the number of unweighted cases and the number of weighted cases.

It is not bad. A bit confusing to use at first, but a lot of information is there in the system.

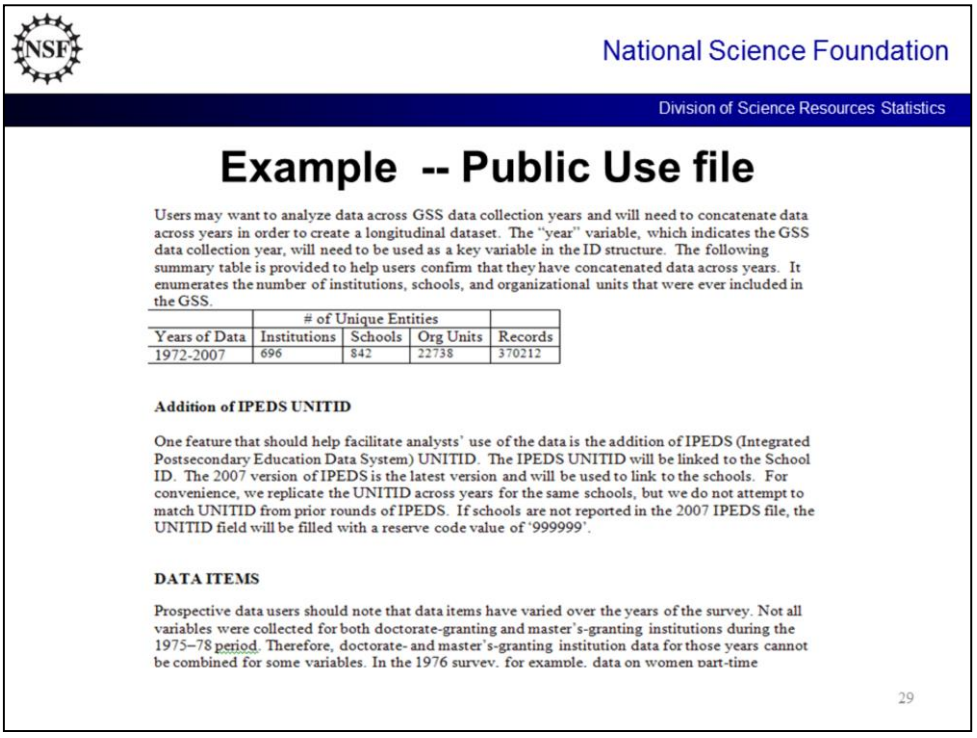


So in SESTAT, the metadata is a bit more sophisticated. (It has its own metadata explorer within the database, but separate from the data. Similar to a codebook.)

But like the other system, the metadata are not tightly integrated with the data itself and does not get downloaded with the data.



Just briefly, I mentioned with have a few public use files.



National Science Foundation

Division of Science Resources Statistics

Example -- Public Use file

Users may want to analyze data across GSS data collection years and will need to concatenate data across years in order to create a longitudinal dataset. The "year" variable, which indicates the GSS data collection year, will need to be used as a key variable in the ID structure. The following summary table is provided to help users confirm that they have concatenated data across years. It enumerates the number of institutions, schools, and organizational units that were ever included in the GSS.

	# of Unique Entities			
Years of Data	Institutions	Schools	Org Units	Records
1972-2007	696	842	22738	370212

Addition of IPEDS UNITID

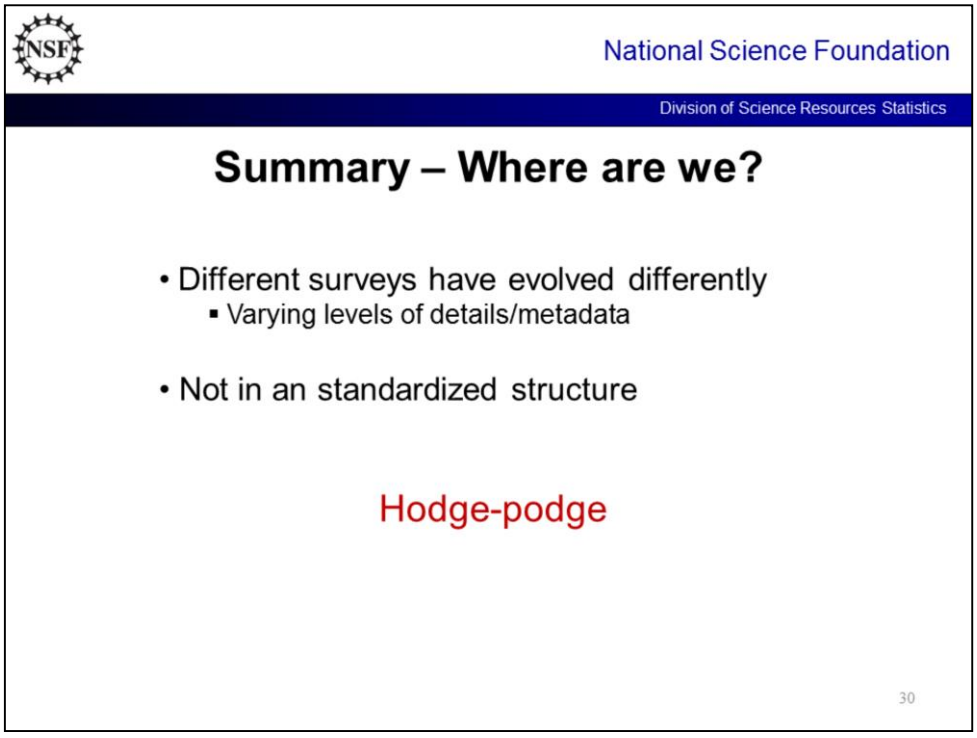
One feature that should help facilitate analysts' use of the data is the addition of IPEDS (Integrated Postsecondary Education Data System) UNITID. The IPEDS UNITID will be linked to the School ID. The 2007 version of IPEDS is the latest version and will be used to link to the schools. For convenience, we replicate the UNITID across years for the same schools, but we do not attempt to match UNITID from prior rounds of IPEDS. If schools are not reported in the 2007 IPEDS file, the UNITID field will be filled with a reserve code value of '999999'.

DATA ITEMS

Prospective data users should note that data items have varied over the years of the survey. Not all variables were collected for both doctorate-granting and master's-granting institutions during the 1975–78 period. Therefore, doctorate- and master's-granting institution data for those years cannot be combined for some variables. In the 1976 survey, for example, data on women part-time

29

This is an example of the documentation that goes with the public use files. It is available as PDF and Word document. It includes, the data file formats, data dictionary, possible values for categorical variables, historical changes to the data items, changes to survey instrument and definition of terms provided on the survey instrument.



So where are we?

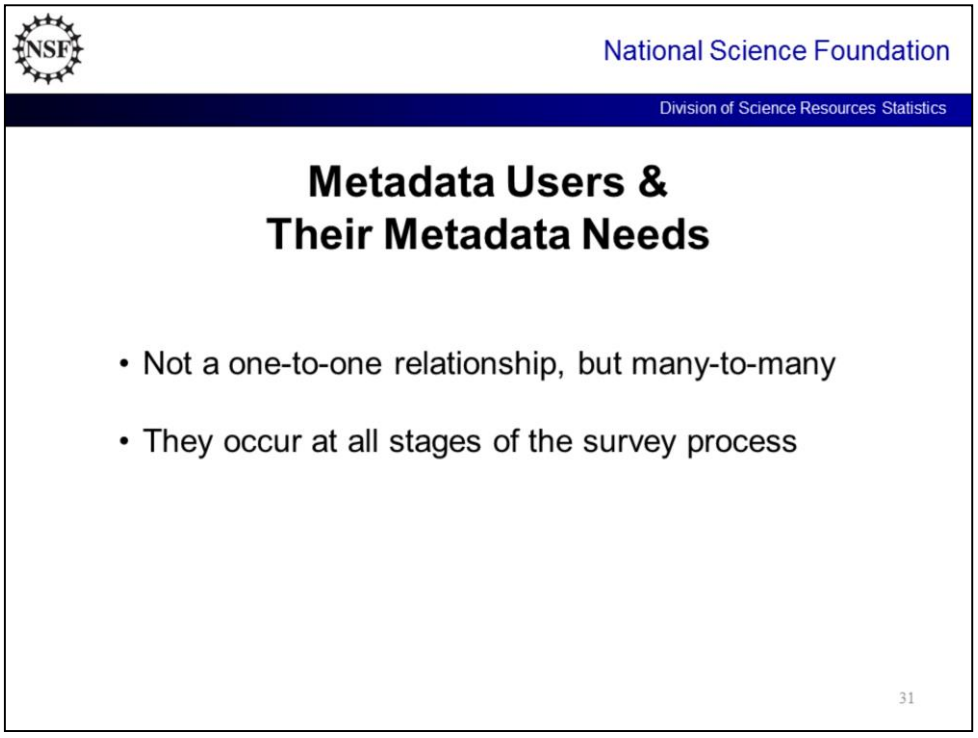
Over time, different surveys within the division have evolved differently with varying levels of metadata and details available on-line. In addition, we have a variety of data and information in a variety of formats.

I was going to say things are not in an organized structure, but they are organized but by different types of publications.

What it is not in is a standardized format. We are starting to see the need for standardization just from the dissemination side.

This is not enough though. This is not enough to make the case for metadata in SRS. Looking at it from this perspective still did not make the case for many in SRS who have to supply the metadata.

So let me look at metadata from a different point of view now.

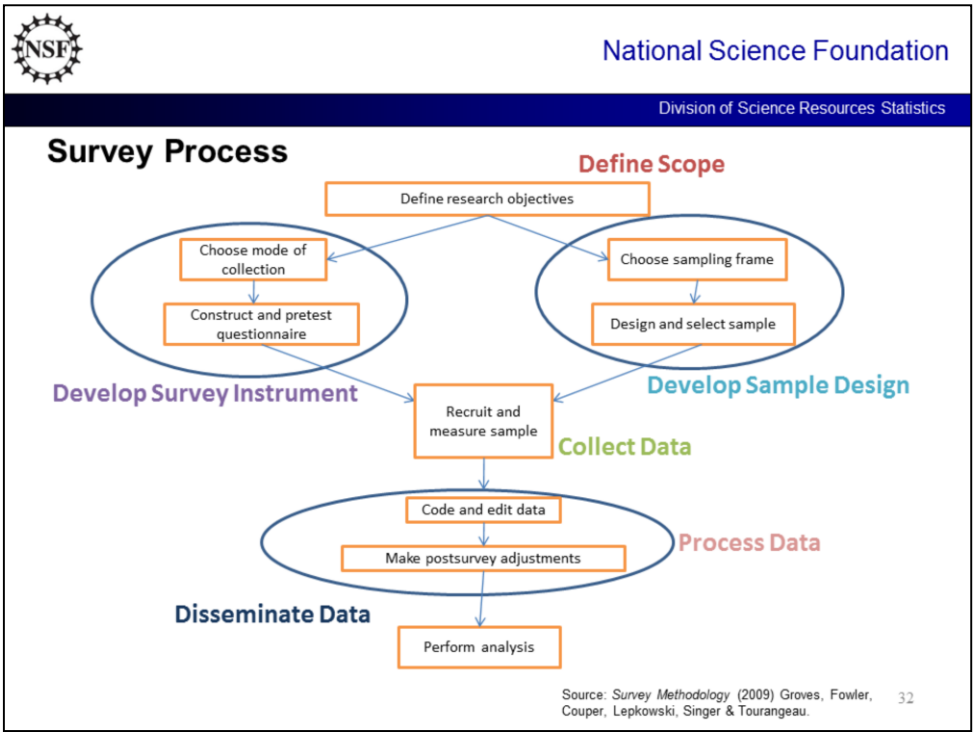


Let's take a look at the different metadata users, of which I am one, and their different needs.

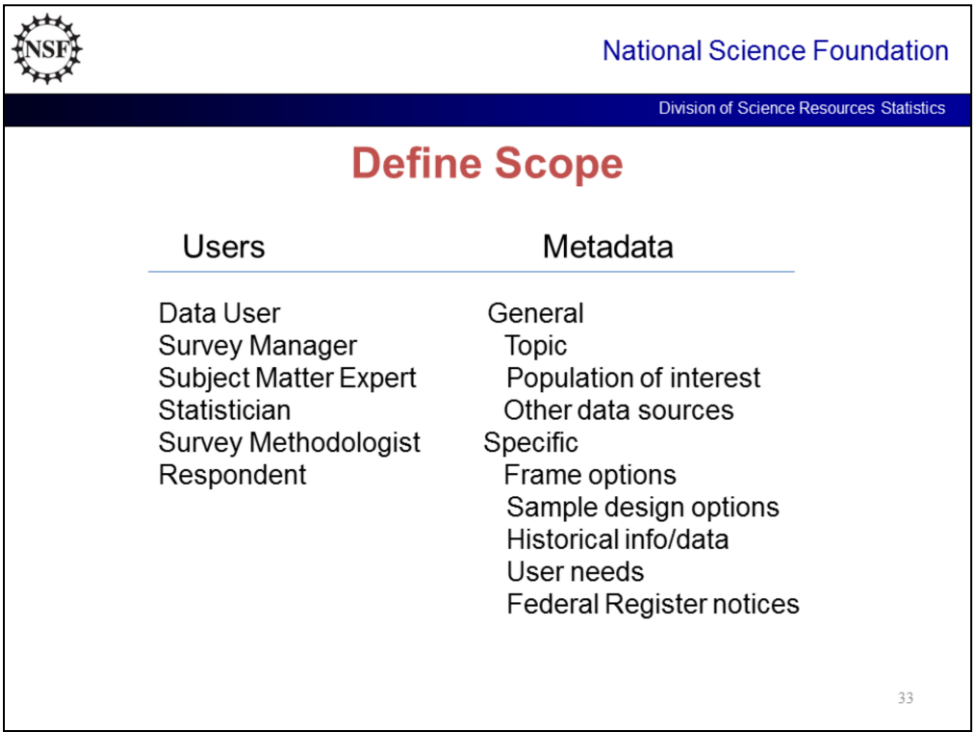
When we started mapping the various metadata users to their various metadata needs, we found that is was not a one to one relationship but a many to many relationship. That is many different users may use the same metadata but in different ways.

We also found that metadata uses and users occur at all stages of the survey process.

I think one of the issues that we have had in SRS is the focus on just the final data user and their need for metadata. But that is not enough of an appeal to internal resources who are often detached from the final data user. I think a realization that internally we use the metadata in many ways can help. So I want to explicitly point out some of the metadata users and their metadata needs.



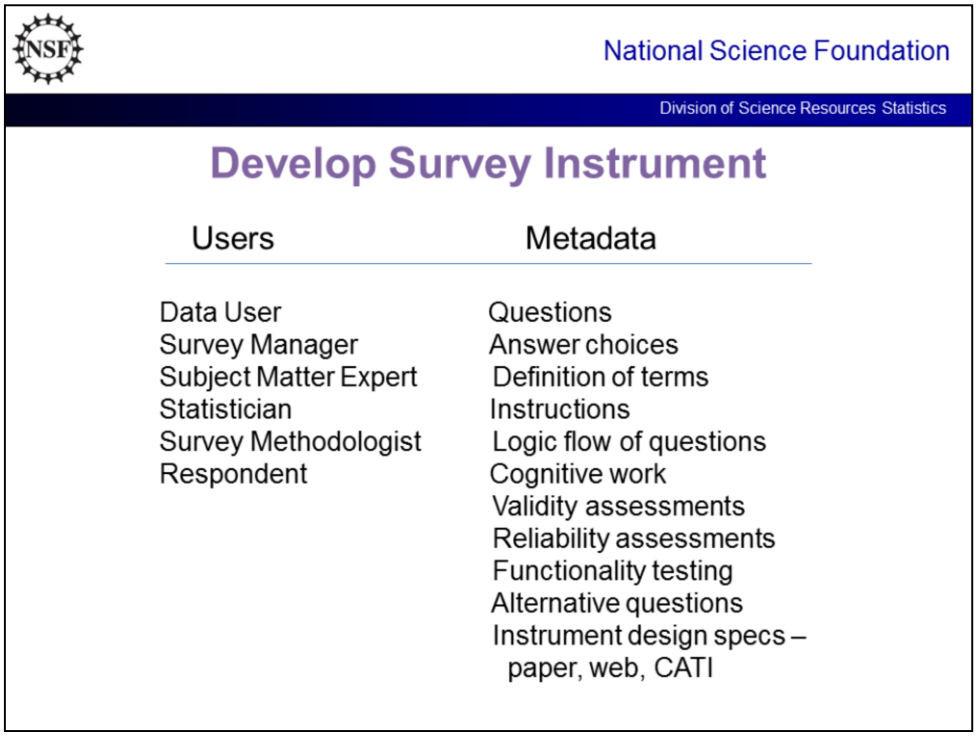
Many of you are familiar with the Groves, et al survey process. Metadata is generated at all stages of this process and Metadata users occur at all stages. We have grouped some of the stages together for simplicity and I will talk a bit about the metadata users and the metadata they might use at each of these different phases.



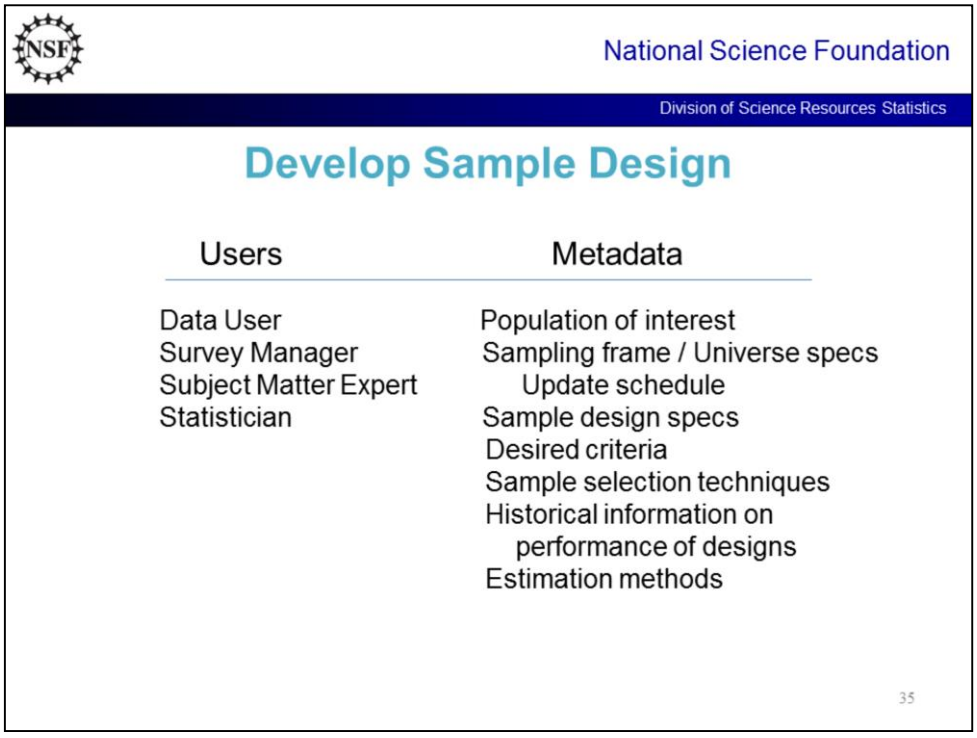
At the beginning stage of defining the scope of the survey or data collection, there are a variety of metadata and metadata users.

Some of the metadata users are: final data user, survey manager, …. And even the respondent might be using some of the metadata generated at this phase.

The same metadata is used differently by different users. For example, take metadata about the different frame options. One user of this information is the statistician who is looking at it in terms of what is the coverage, how up-to-date is it, how easy will it be to gain access to it for as a sampling frame. Another user is the survey manager who is looking at it in terms of cost to access. The respondent may not care about this, only that they were selected to be part of the survey. They may wonder though why they were selected, what was it about them that made it so they have to answer this blasted survey.

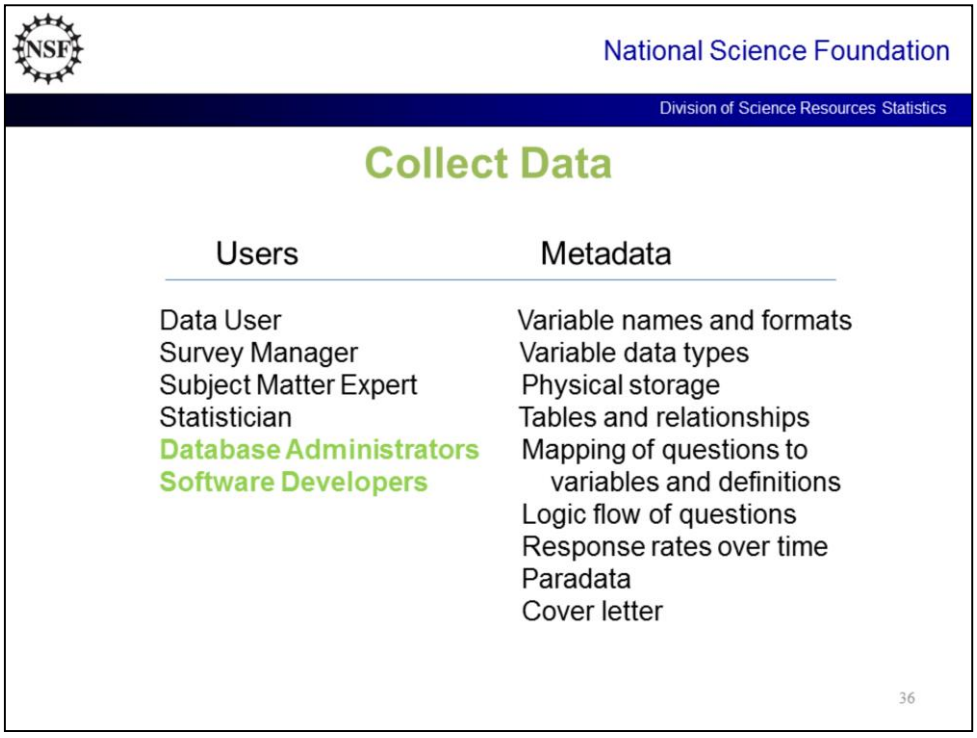


At the Develop Survey Instrument phase, we have the same metadata users, but different types of metadata being generated and used. Again, there may be different users using the same metadata in different ways. The final data user is using some of this information to interpret the data, such as definition of terms and answer choices. The respondent is using the same information, but in trying to figure out how to answer and provide data. The Survey Methodologist is using this information in yet another way.



We might have slightly fewer metadata users at this stage, but it isn't entirely clear to me that that is the case. That is, we did not include the survey methodologies at this stage, but I am not sure that is correct. We also didn't include the respondent. I'll let you think about it.

Anyway, we definitely have a variety of metadata being generated and used. The metadata users would like to have access to this information easily and quickly. Right now, it is stored in a variety of areas in a variety of formats. Much of it is not linked to the data itself and sometimes it 'gets lost', especially over time.

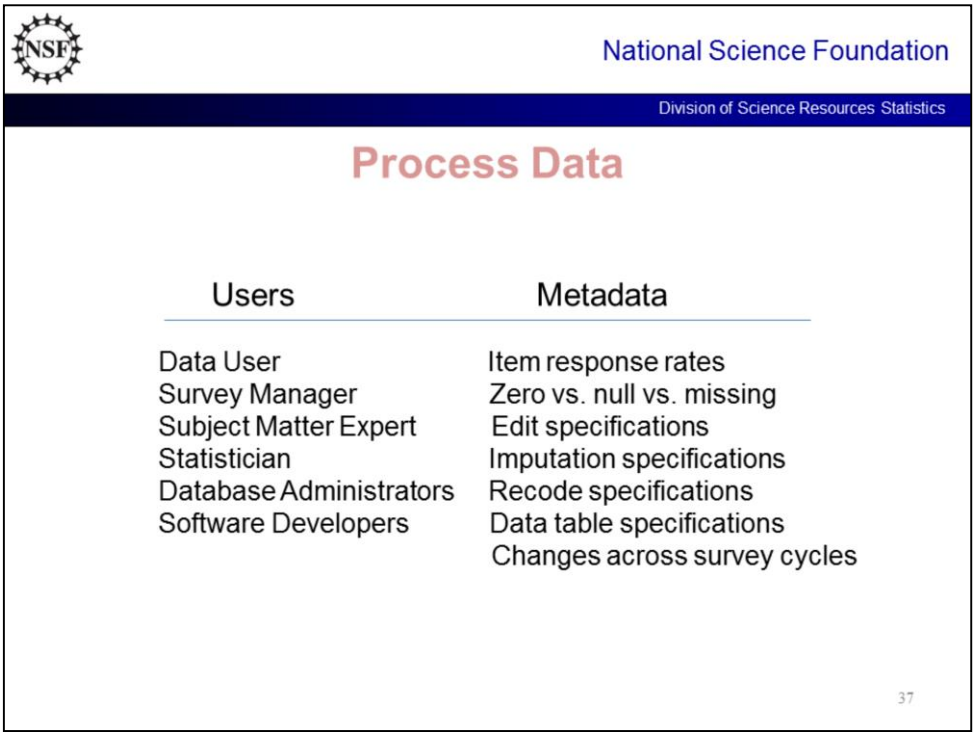


At this phase, we have a couple of new metadata users entering the scene – they are database administrators and software developers. Before I started working with John and Geetha, I was not really thinking about these types of users, however, they are extremely important to the survey process at this stage. They use much of the same, familiar metadata but in different ways than data users.

For example, a database administrator needs to know how the data are structured, what are the relationships among the variables, how do they map to the survey questions and so on. This type of metadata has actually been a bit more difficult for us to obtain, because it is not traditionally thought of as part of the metadata by data collectors.

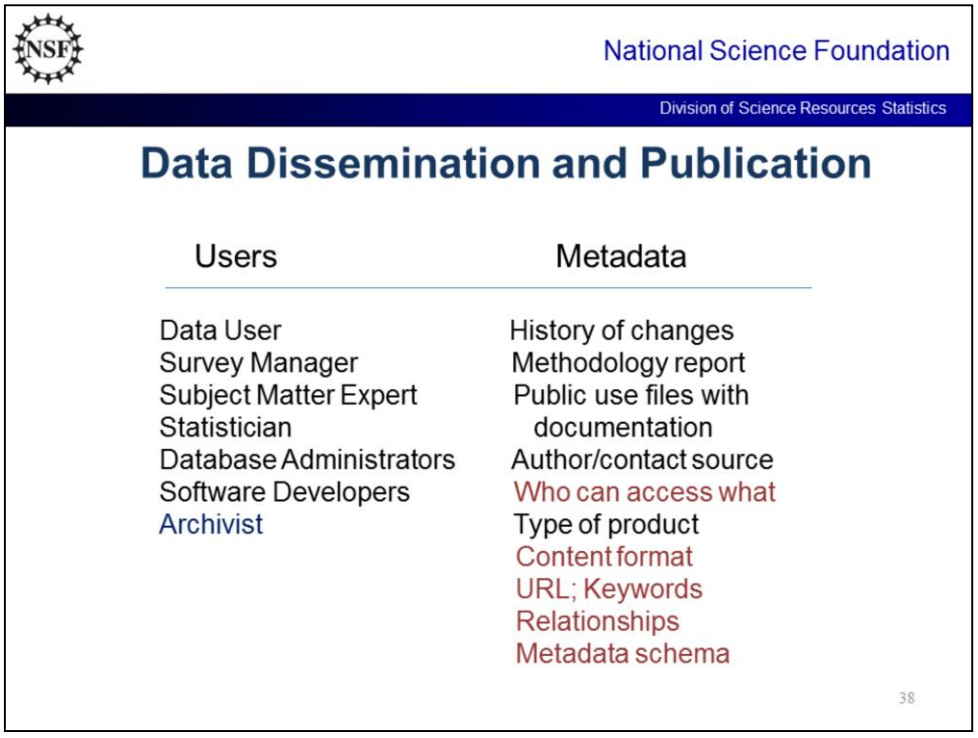
We may be a bit more familiar with the types of metadata used by software developers, that is, the web survey designers (or CAPI or CATI designers), such as logic flow of questions.

Some Paradata is entering the stream here (others comes with the frame in an earlier stage).



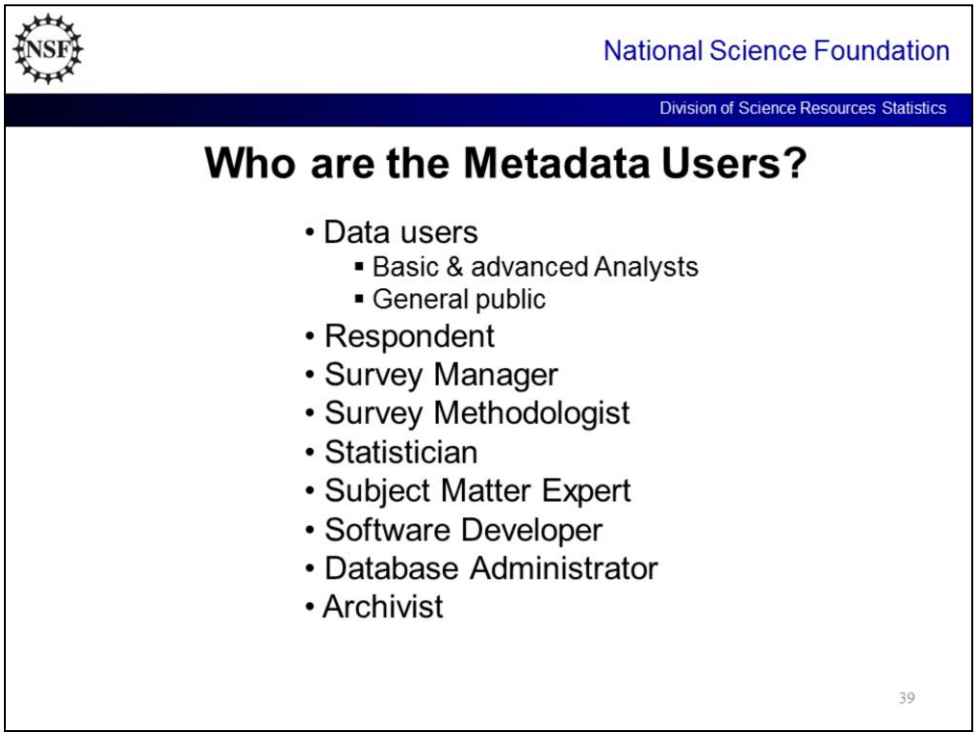
Moving along to data processing, we have the same types of metadata users and some familiar metadata.

We are seeing the added dimension of time here more, with changes across survey cycles. Keeping track over time is an added challenge to all of this.



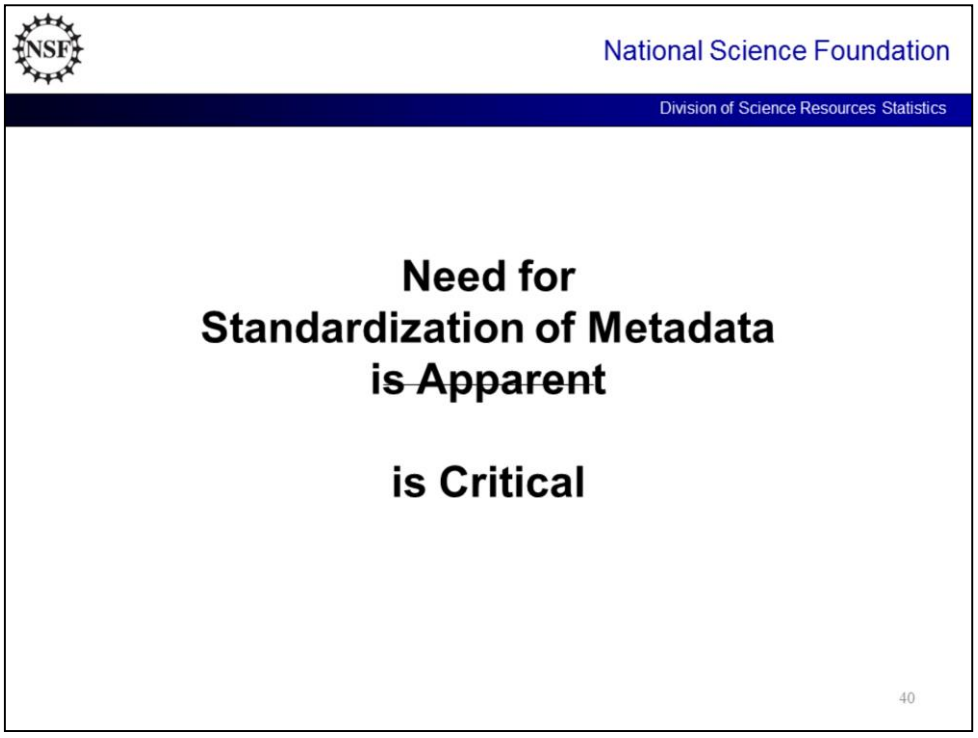
At this stage, we have the archivist as a new user.

The metadata in red is newer to us, and we have not disseminated some of this information until recently.



Let's recap, who are the different metadata users?

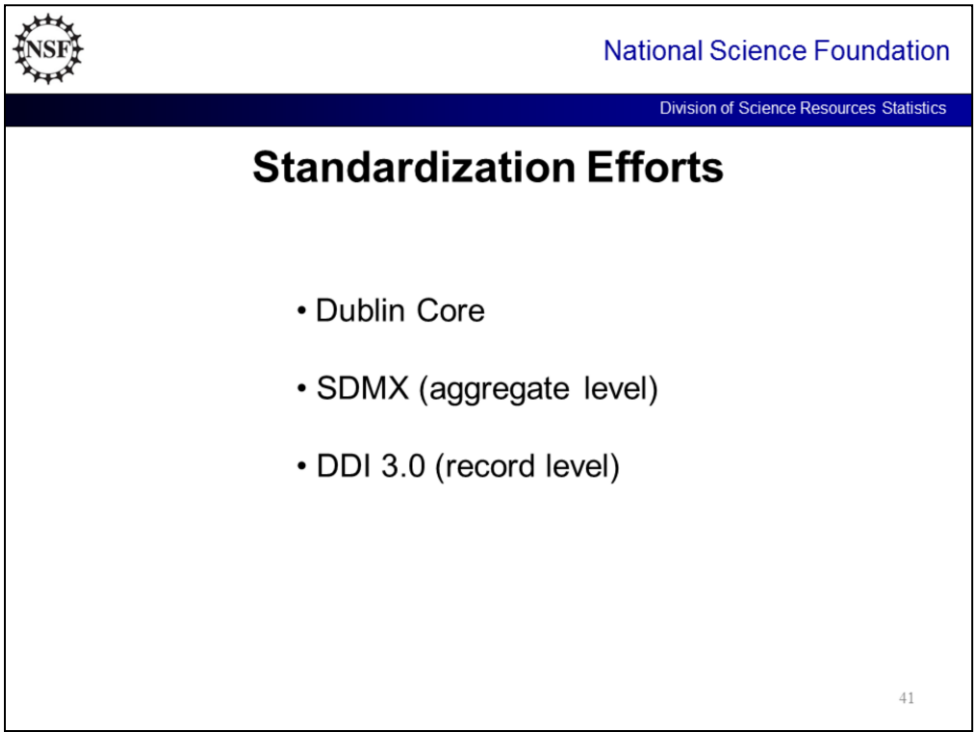
I have listed 9 but there are likely more. They are using a variety of metadata in a variety of ways, sometimes the same metadata in different ways. Which lead us to….



We were starting to see the need for standardization of metadata when we look at this from a historical perspective.

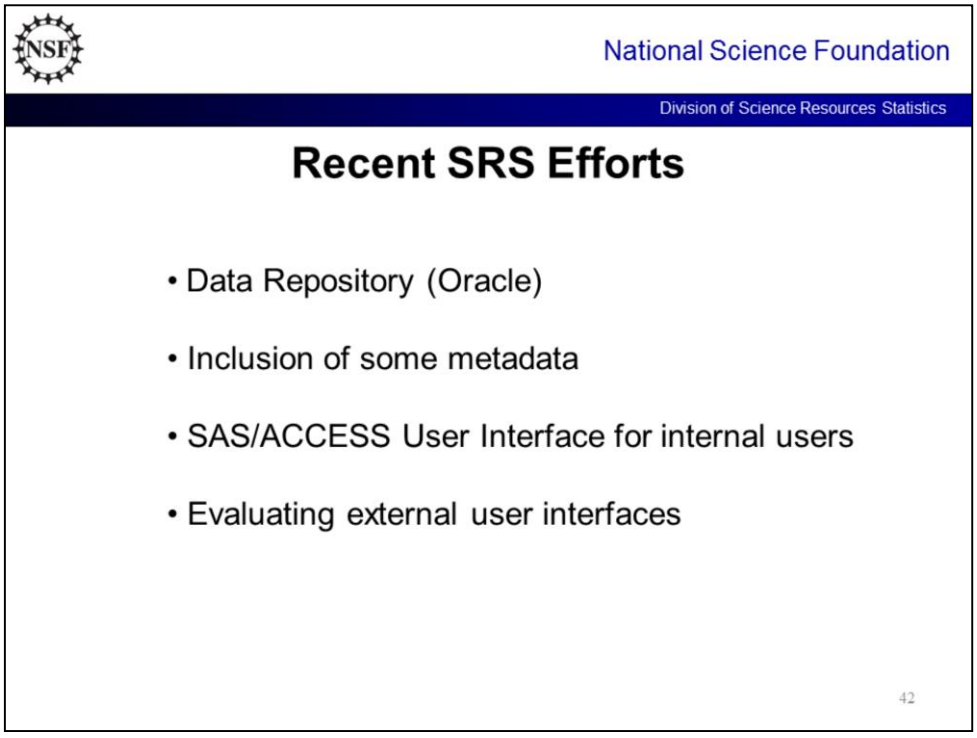
But now, when we look at the different users and their different metadata needs, it is not only is apparent that we need standardization, but it critical.

There is a lot of different information being used in a lot of different ways by a lot of different people.



There have been some standardization efforts. Unfortunately they are not as widely used as they might be, especially in the survey world.

This could be due to the complexity of the problem, the lack of adequate tools, missing metadata or some combination of all of these plus more things. At any rate, it should be clear that we need to do something.



We have recognized this need and we are working on some efforts.

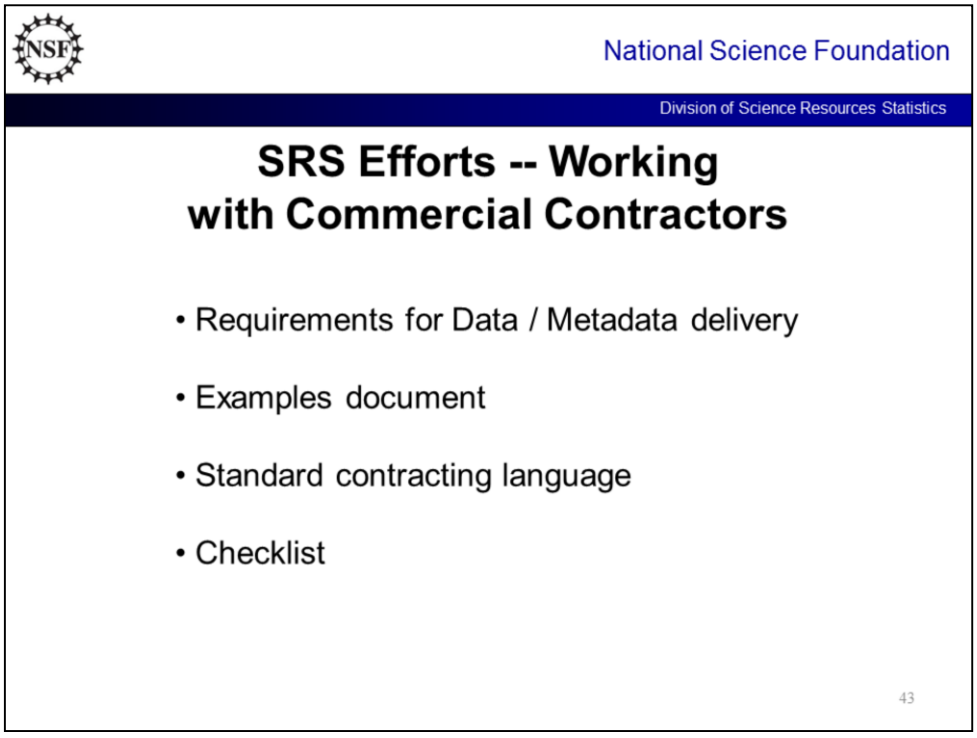
We have been are working to put our data and metadata into a data repository. The repository is in the form of an Oracle database, which provides us flexibility for the future.

We are working to include as much metadata as we can, but it has actually been a challenge to gather some of this. Not only are we dealing with a variety of internal sources, we contract out all of our surveys.

Trying to get the metadata has been a challenge, let alone getting it in any type of standardized format.

On top of the repository we have a SAS/ACCESS user interface for internal SRS users.

We have done some proof of concept and evaluations of different external user interfaces but have not made any final decisions on this.



We found we needed to develop requirements for the data and metadata delivery for our data repository and are currently working with individual survey managers and their contractors to enhance their understanding of these requirements. As part of this, Geetha put together an Examples document to show them what we mean when we say different things.

One of the things that has come out of this process for me is that we all speak a different language. That is, statisticians and database managers don't call the same things by the same names, and survey managers may call the thing something entirely different than either the statistician or the database manager. A lot of the effort has been about communication. We are slowly getting on the same page with things.

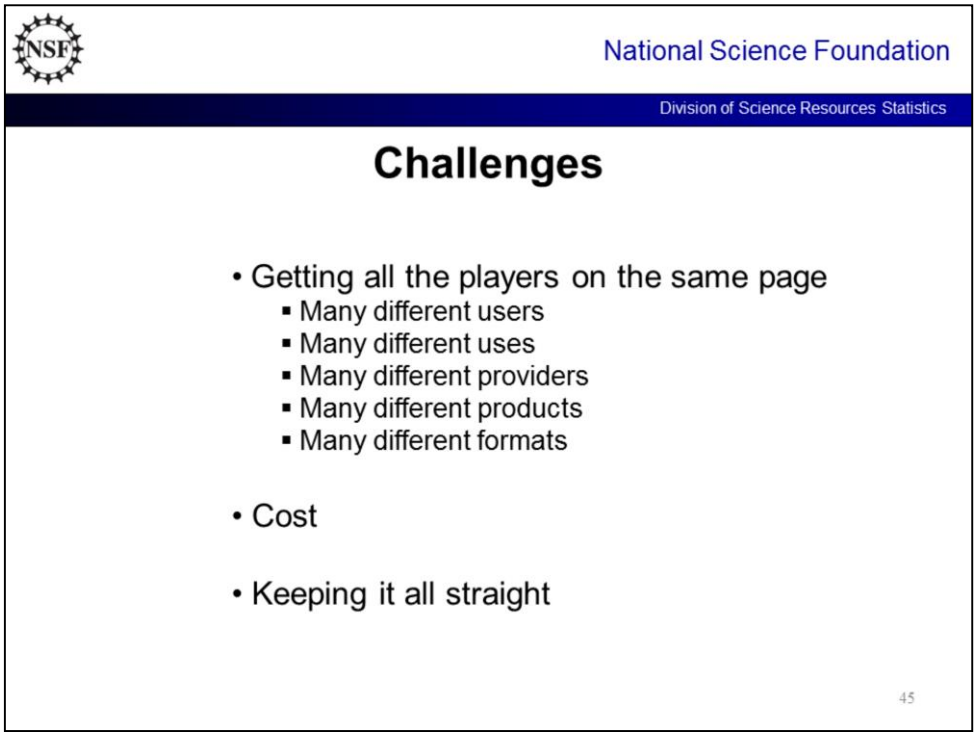
We developed standard contracting language this is to be included in all of our data collection contracts. If you have worked with contracts, you know that if it isn't in there, you don't get it.

To help us and the contractors we developed a checklist. To be honest, this has not been as useful as we had hoped.

We have adopted some basic operating principles to guide us. We are using Oracle to store the data and microdata.

Right now, we are trying to get as much metadata as we can in whatever format it currently exists. It is a start.

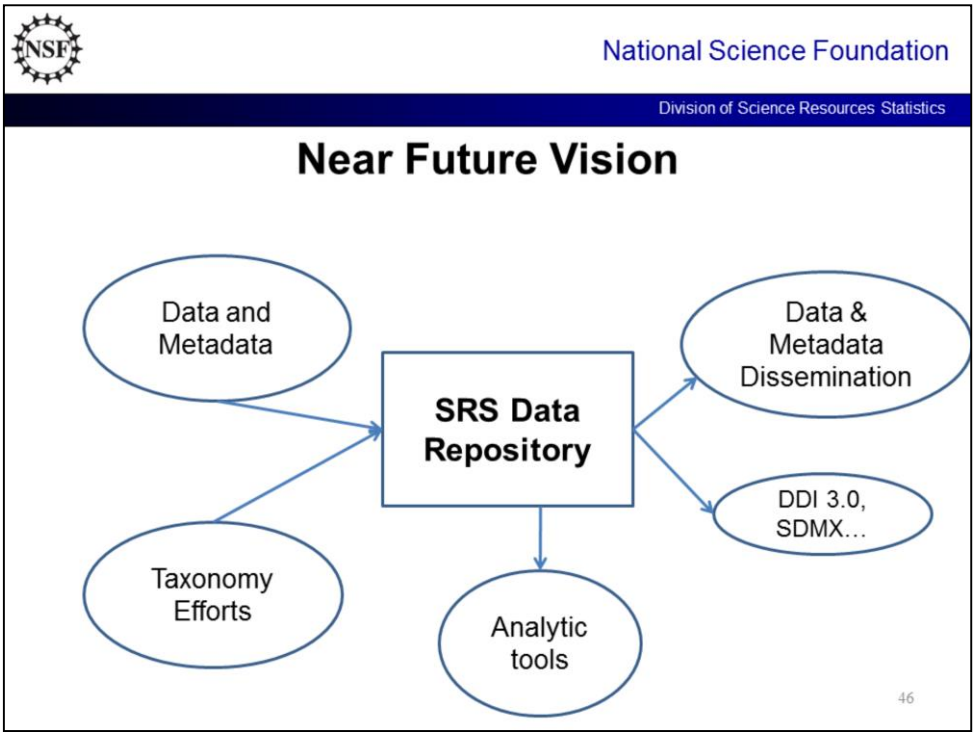
And the key is keeping it all organized. Geetha is handling that!



As I said before, the biggest challenge has been communicating the need for the data and metadata so that we all have the same understanding.

So far, cost has not been a huge challenge, but we can't ignore it.

And it takes some amount of planning to keep it all straight.

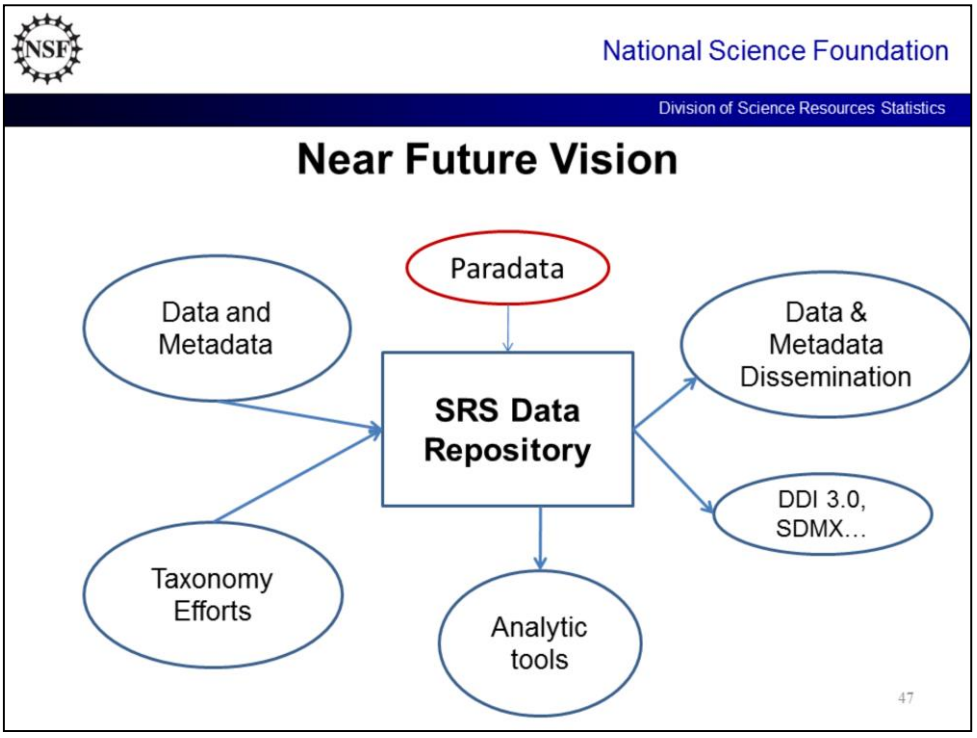


For now our visions looks like this.

Get the data and metadata into the repository. Work on taxonomy efforts to organize and standardize our classifications across surveys.

Work on getting interfaces that allow us to disseminate the data and metadata. Work on standardizing the metadata and data, possibly using DDI 3.0.

Work on getting analytic tools that can be used directly with the data in the repository.

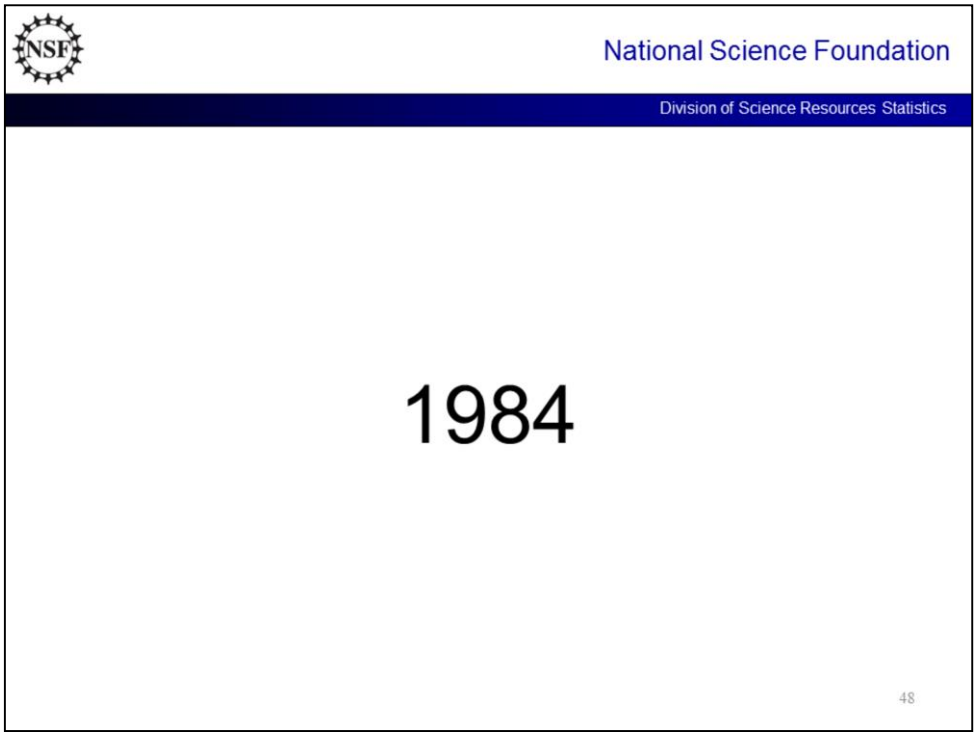


Near future – add paradata into the mix.

Get the data and metadata into the repository. Work on taxonomy efforts to organize and standardize our classifications across surveys.

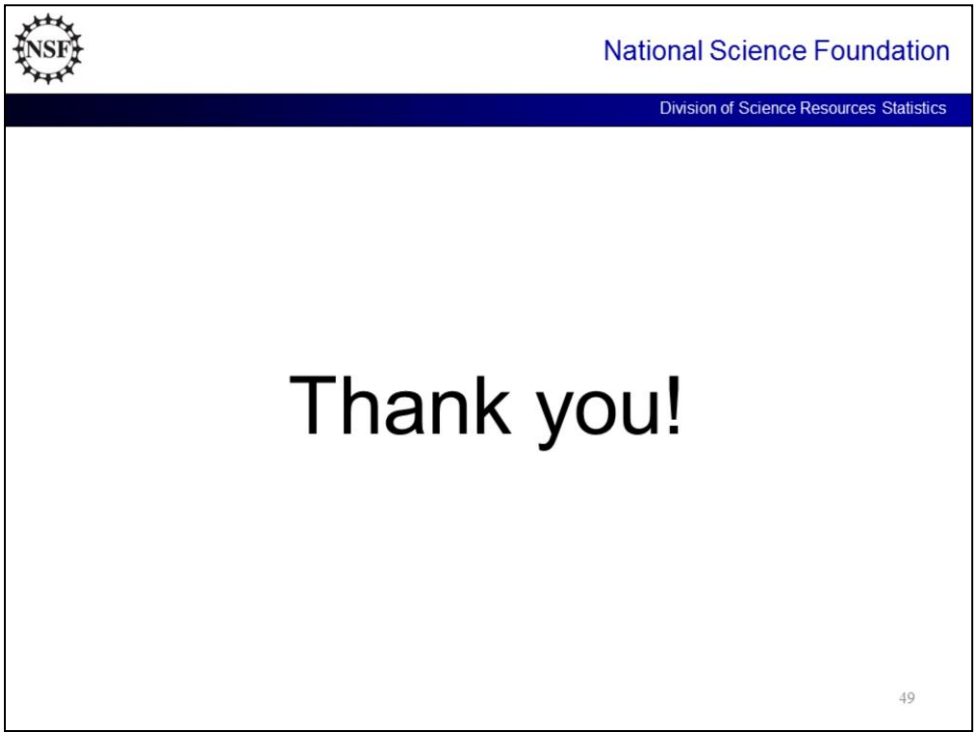
Work on getting interfaces that allow us to disseminate the data and metadata. Work on standardizing the metadata and data, possibly using DDI 3.0.

Work on getting analytic tools that can be used directly with the data in the repository.



So back to the beginning. 1984. What is it?

1984 – For me, that is the year I got married!



So back to the beginning. 1984. What is it?

1984 – For me, that is the year I got married!