

# The Social Science Variables Database at ICPSR

Sanda Ionescu,  
ICPSR

Question Database Workshop  
Paris, France  
January 28, 2011

# Variables Search

- Enables ICPSR users to search variables across datasets
- Assists in data discovery, comparison, harvesting, and analysis
- Useful in question mining for designing new research



# The Social Science Variables Database (SSVD) at ICPSR

- Concept first tested in a pilot project completed by the end of 2004
  - Good functionality
  - Demonstrated benefits of using DDI markup: easy parsing/import; complex, granular searches; user-friendly display
  - Limited number of data sets (69 ICPSR studies included)



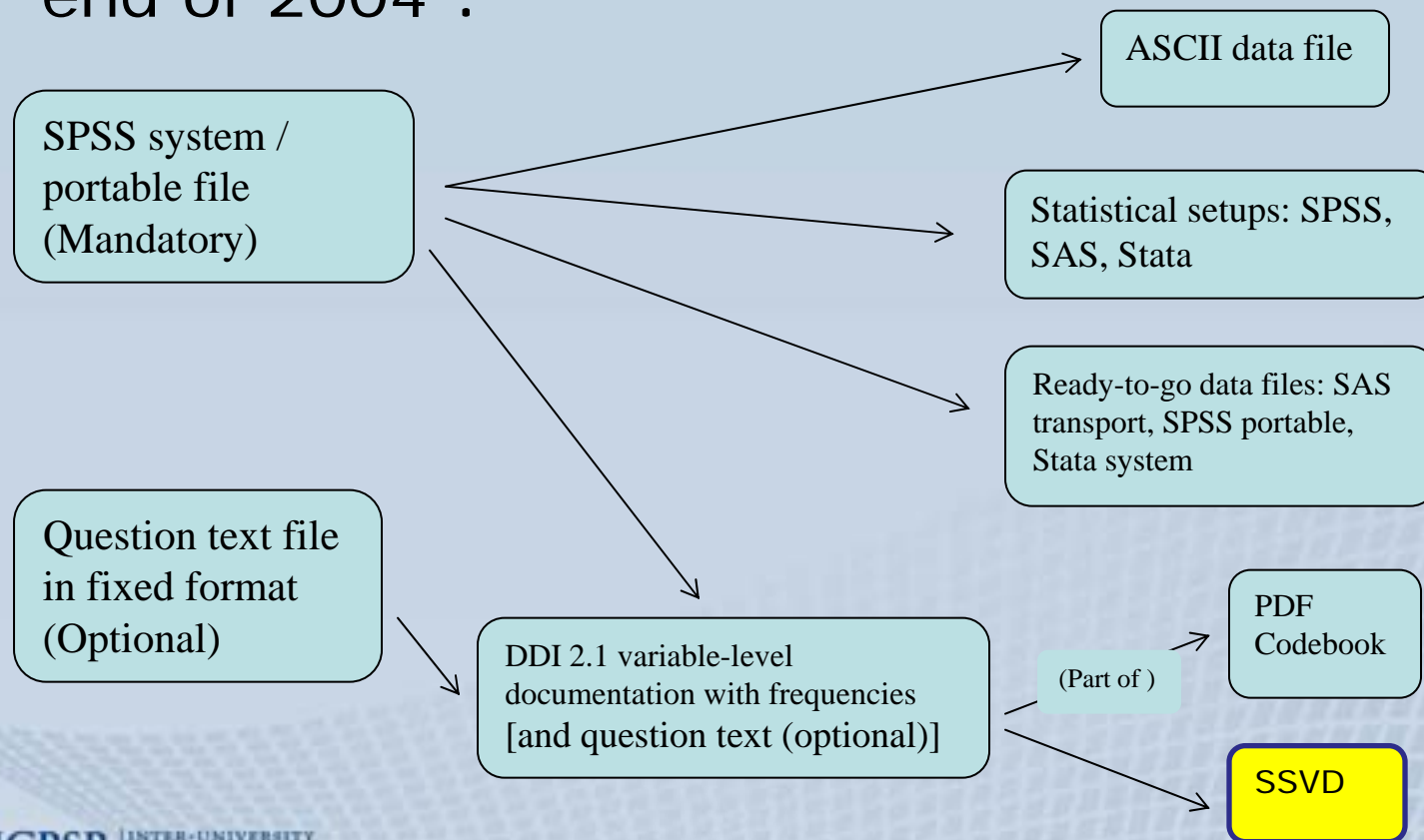
# SSVD at ICPSR

- Expand the project to ultimately include most of ICPSR's holdings
  - Generate variable-level DDI documentation for most ICPSR studies
    - Need for automated production
  - Develop effective technology to store and search the DDI files
    - Handle large number of files
    - Support multiple applications

# SVD at ICPSR

## Creating and Harvesting DDI documentation

- The *Hermes* batch processing system \*, launched end of 2004 :



# SVD at ICPSR

## DDI documentation

- *Hermes:*
  - Consistent, reliable source of variables descriptions in DDI
  - Automated production
    - DDI produced with no additional staff effort
    - Predictable file structure
    - Consistency in fields supported /generated
    - Valid DDI

# SVD at ICPSR

## DDI documentation

- *Hermes DDI:*
  - Quality of field content – less consistent (dependent on available processing resources as well as the individual archives' contractual agreements)
    - Truncated or abbreviated text, particularly in variable/value labels (acceptable in statistical syntax)
    - Question text may be missing although available in the original / deposited documentation

# SVD at ICPSR

## DDI documentation

- Additional quality standards necessary for DDI documentation, to maximize effectiveness of Public Search:
  - Presence of question text, whenever available
  - Increased readability of variable/value labels, especially if question text is not present

# SSVD at ICPSR

## Studies selection

- Not all ICPSR studies qualify for variable-level searches
- Criteria for selecting studies; not included:
  - Some aggregate/statistical data (ex. Census data, Data Books, Roll Call records, etc.)
  - Poor documentation (very rare)
  - Studies with restricted documentation (for restricted data, DDI is released if the documentation is public)

# SSVD at ICPSR

## Harvesting DDI

- Pre-SSVD upload:
  - Review of DDI output from Hermes to apply content quality standards and study selection criteria
  - Additional work to upgrade DDI where necessary (and feasible)
    - Add question text
    - Complete truncated text
    - Improve readability of labels
    - Add frequencies if missing

# SSVD at ICPSR

## Harvesting DDI

- Harvesting and preparing DDI for SSVD
  - Started 2005-2006
  - Goal: have a sizable batch of DDI files ready, to test database and search performance when made available

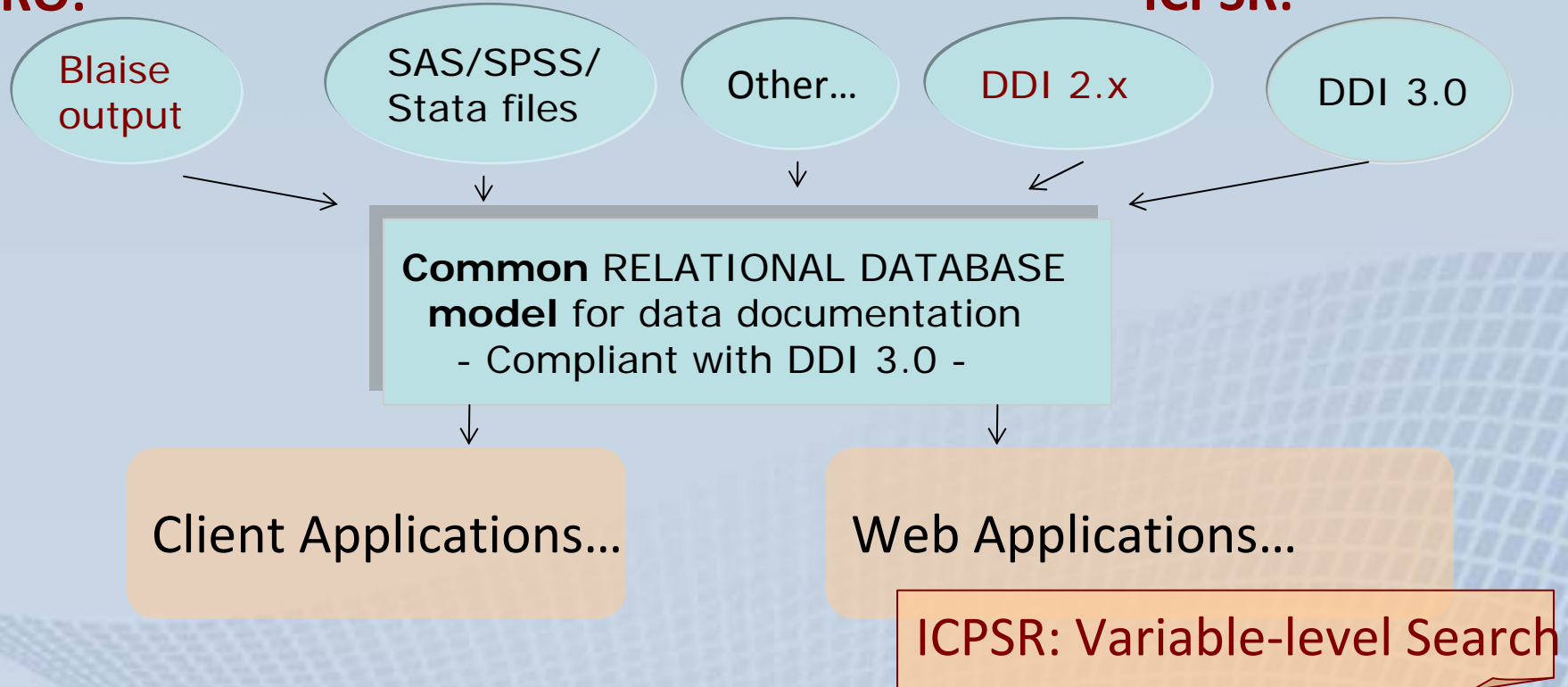
# SSVD at ICPSR

## DDI Storage and Search

- Creating a DDI-compliant Variables Database – an SRO-ICPSR collaboration project

**SRO:**

**ICPSR:**



# SSVD at ICPSR

## DDI Storage and Search

- The ICPSR variables database:
  - Built in Oracle as a separate entity, with links to studies' and series' descriptions (also stored in Oracle)
  - Compatible with both DDI 2 and 3 (input and output)
  - Initial upload programmed from DDI 2.1: variable descriptions complete with (unweighted) frequencies and question text, where available
  - DDI 3 output for Web display

# SSVD at ICPSR

## DDI Storage and Search

- New database finalized late 2008
- Early 2009 first batch of approximately 3,500 DDI files (one file per dataset) uploaded into SSVD. These represented
  - Approx. 1,300 ICPSR studies, or approx. 30 percent of ICPSR holdings with data and setups
  - Over 1,000,000 individual variable descriptions



# SSVD at ICPSR

## Current content

- End of 2010:
  - 1.6 million variables (+ .5 million)
  - 2,300 studies (+ 1,000)
  - 43 percent of ICPSR holdings with data and setups (+ 13 percent)
- Retrofits added as available (small scale projects)



# SSVD at ICPSR

## DDI Storage and Search

- Oracle Text searches used in Beta-testing phase
  - Slow retrieval
  - Limited to 500 results



# SSVD at ICPSR

## DDI Storage and Search

- Autumn 2009 switched to Solr/Lucene:
  - Easy indexing
  - Faster searches, unlimited hits
  - Facets/Filters imported from Study Descriptions (also DDI 2 compatible)
    - Study
    - Series
    - Time Period
    - Geography
  - XML itself is being indexed and searched – no longer uploaded in the database

# SSVD at ICPSR

- 2010 developments:
  - Web presentation: variables search enabled from individual Study and Series pages led to increased visibility and traffic
  - Advanced search introduced to include fielded searches
  - Significant improvements in the quality of DDI content, as all on-going projects and topical archives realize the benefits of having their studies included in the variables search
  - More flexibility in applying standards for content quality and study selection

# SSVD at ICPSR

- 2011:
  - Workflow: DDI files to be reviewed prior to release of archival and distribution information packages, to ensure archiving most updated copy of DDI (same as SSVD copy)
  - Web: Replacing the current study description search (main search) with an “integrated” search that will also include full text of codebooks, bibliographic citations, and DDI variable descriptions

# SSVD at ICPSR

- DDI fields searched:
  - Variable name ^2
  - Variable label ^5
  - Question text sequence ^5
  - Descriptive text ^2
  - Category label ^2
- Variable notes – not indexed /searched, but they are displayed.

# SSVD at ICPSR

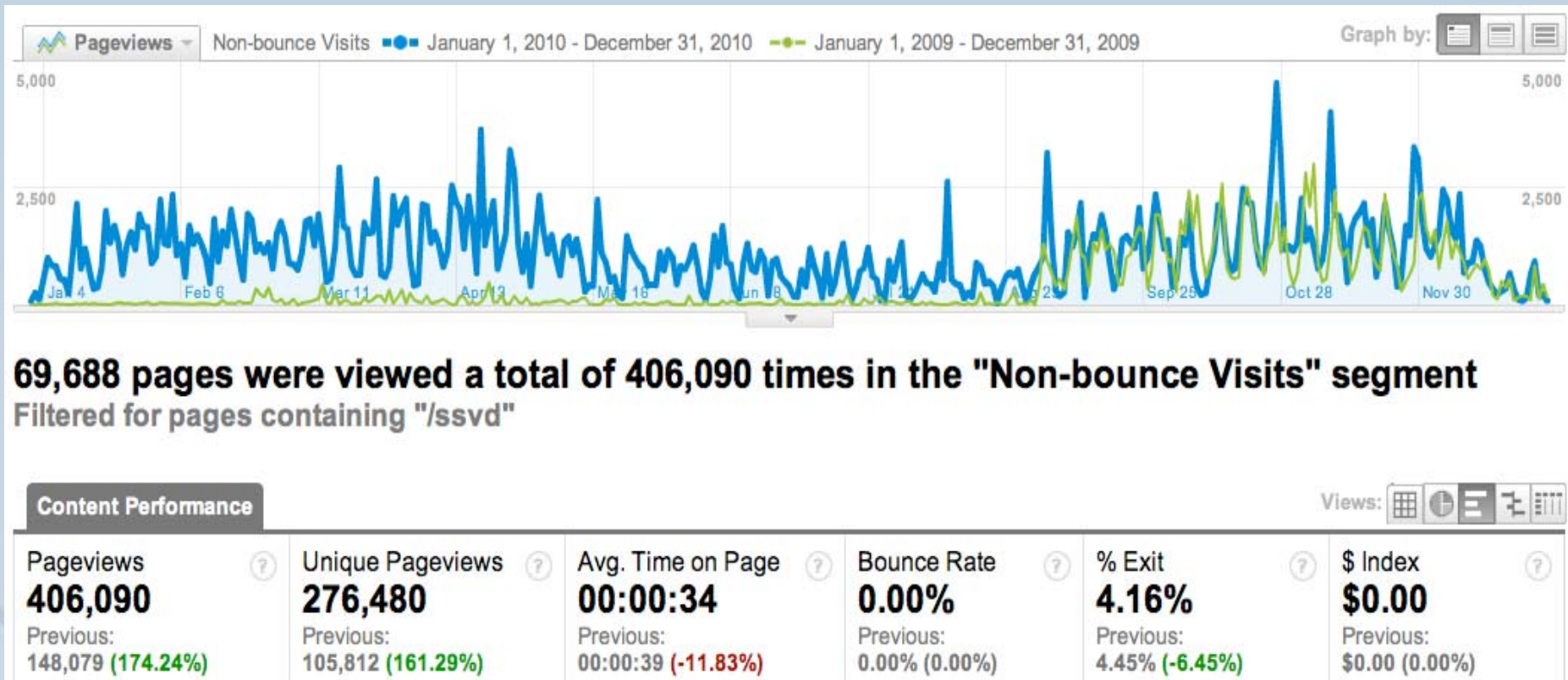
- The Public Search Features:
  - Stemming
  - “Phrase searches”
  - Fielded searches (treated as a default Boolean “and”: Boolean operators “or,” and “not” are ignored)
    - Variable label
    - Question text
    - Value labels

<http://www.icpsr.umich.edu/icpsrweb/ICPSR/>

# SSVD at ICPSR

## Usage data

- "Quantitative":





# SSVD at ICPSR

## Usage data

- “Qualitative”:
  - One of our most popular Web services
  - Preferred to main (study description) search for data discovery
  - Requests to increase the number of studies searched, including by retrofitting the collection
  - Requests for more refined searches

# SSVD at ICPSR

## The Internal Search

- Currently reserved for staff
- Used for internal projects (ex.: production of harmonized datasets)
- Still using Oracle Text
- Provides same page display of variables' descriptions for easy comparison
- Allows narrowing down search results to customized selection and export to spreadsheet format to facilitate production of translation table

<http://www.icpsr.umich.edu/ssvd2/overview.jsp>

# SSVD at ICPSR

## The Internal Search

A	B	C	D	E	F	G	H
Study Name	Study No	Dataset	Variable	Variable Label	Question Text	Category Label	Category Value
Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 2000	3184	0004	V3453	003E06B:DRG USE+,MY FRND	How about using drugs (other than marijuana or alcohol)-- does that cause a student to be looked up to or looked down on? Among my own group of friends, such drug use is . . .	* DOWN LOT:(1) * DOWN SOM:(2) * NEITHER:(3) * UP SOME:(4) * UP A LOT:(5) * Missing	1 2 3 4 5 -9
Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 2001	3425	0004	V3453	013E06B:DRG USE+,MY FRND	How about using drugs (other than marijuana or alcohol)-- does that cause a student to be looked up to or looked down on? Among my own group of friends, such drug use is . . .	* DOWN LOT:(1) * DOWN SOM:(2) * NEITHER:(3) * UP SOME:(4) * UP A LOT:(5) * MISSING	1 2 3 4 5 -9
Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 2002	3753	0004	V3453	023E06B:DRG USE+,MY FRND	How about using drugs (other than marijuana or alcohol)--does that cause a student to be looked up to or looked down on? B: Among my own group of friends, such drug use is . . .	* DOWN LOT:(1) * DOWN SOM:(2) * NEITHER:(3) * UP SOME:(4) * UP A LOT:(5) * MISSING	1 2 3 4 5 -9
Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 2003	4019	0004	V3453	033E06B:DRG USE+,MY FRND	How about using drugs (other than marijuana or alcohol)--does that cause a student to be looked up to or looked down on? B: Among my own group of friends, such drug use is . . .	* DOWN LOT:(1) * DOWN SOM:(2) * NEITHER:(3) * UP SOME:(4) * UP A LOT:(5) * MISSING	1 2 3 4 5 -9
Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 2004	4264	0004	V3453	043E06B:DRG USE+,MY FRND	How about using drugs (other than marijuana or alcohol)--does that cause a student to be looked up to or looked down on? B: Among my own group of friends, such drug use is . . .	* DOWN LOT:(1) * DOWN SOM:(2) * NEITHER:(3) * UP SOME:(4) * UP A LOT:(5) * MISSING	1 2 3 4 5 -9
<b>Search Criteria -</b> <b>Search Type: Internal Search</b> <b>1st Keyword: drug use</b> <b>1st Search Field: all search fields</b> <b>Search Option: and</b> <b>2nd Keyword: marijuana</b> <b>2nd Search Field: all search fields</b>							

# Questions?

