

A Metadata Reference Model for IRS Research Data

2012 Federal Computer Assisted Survey Information Collection
(FedCASIC) Conference

March 27-29, 2012
Bureau of Labor Statistics Conference Center
Washington, DC 20212

Internal Revenue Service
Research, Analysis, and Statistics

Presentation Agenda

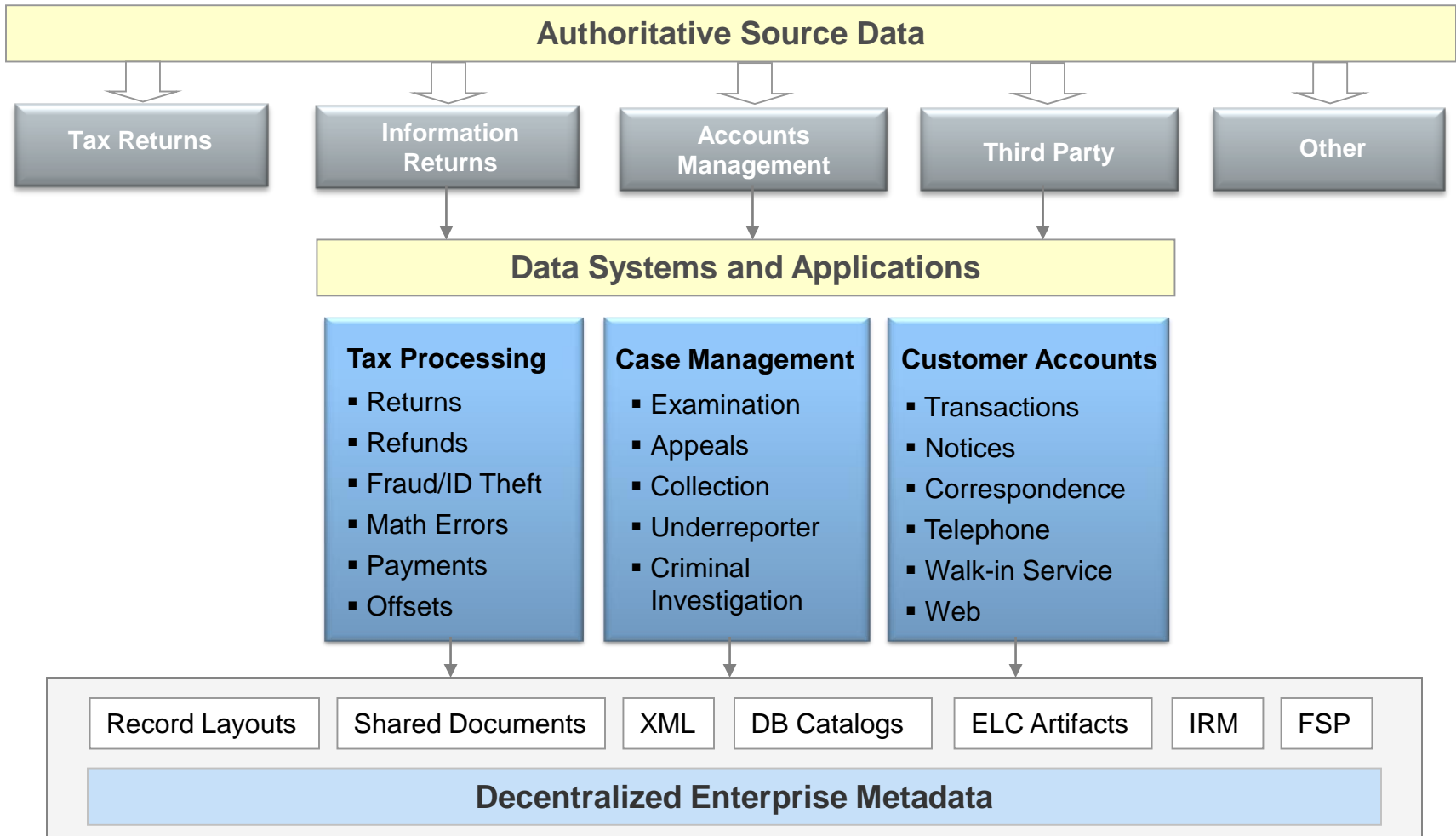
- Overview of the IRS Data Environment
 - IRS enterprise metadata
- Research Data Environment
 - Data and metadata
- Metadata Conceptual Framework and Logical Model
 - Contextual, system, and application properties
 - Controlled vocabulary
 - Examples
- Application Layer
 - Search, Reviews, and Data Profiling
- Wrap-up

IRS Enterprise Data Environment

- **High volumes** of taxpayer transactions
 - Over 237 million tax returns filed
 - \$2.4 trillion in gross receipts
 - 122 million refunds processed, totalling \$415 billion
 - 2.6 billion third-party information returns
 - 305 million visits to IRS web site
 - Nearly 80 million toll-free telephone calls
 - More than 154 million notices issued
- Data are stored in a **variety of formats** across multiple platforms
 - Structured, semi-structured, and unstructured data
 - Mainframe, Unix derivatives, Windows platforms
 - Flat files, VSAM, DB2, Oracle, SQL Server
- Data **access costs** are high
 - Disparate formats, authorization policies, access channels
 - Systems are often designed for a single, operational purpose

Metadata Reference Model for IRS Research

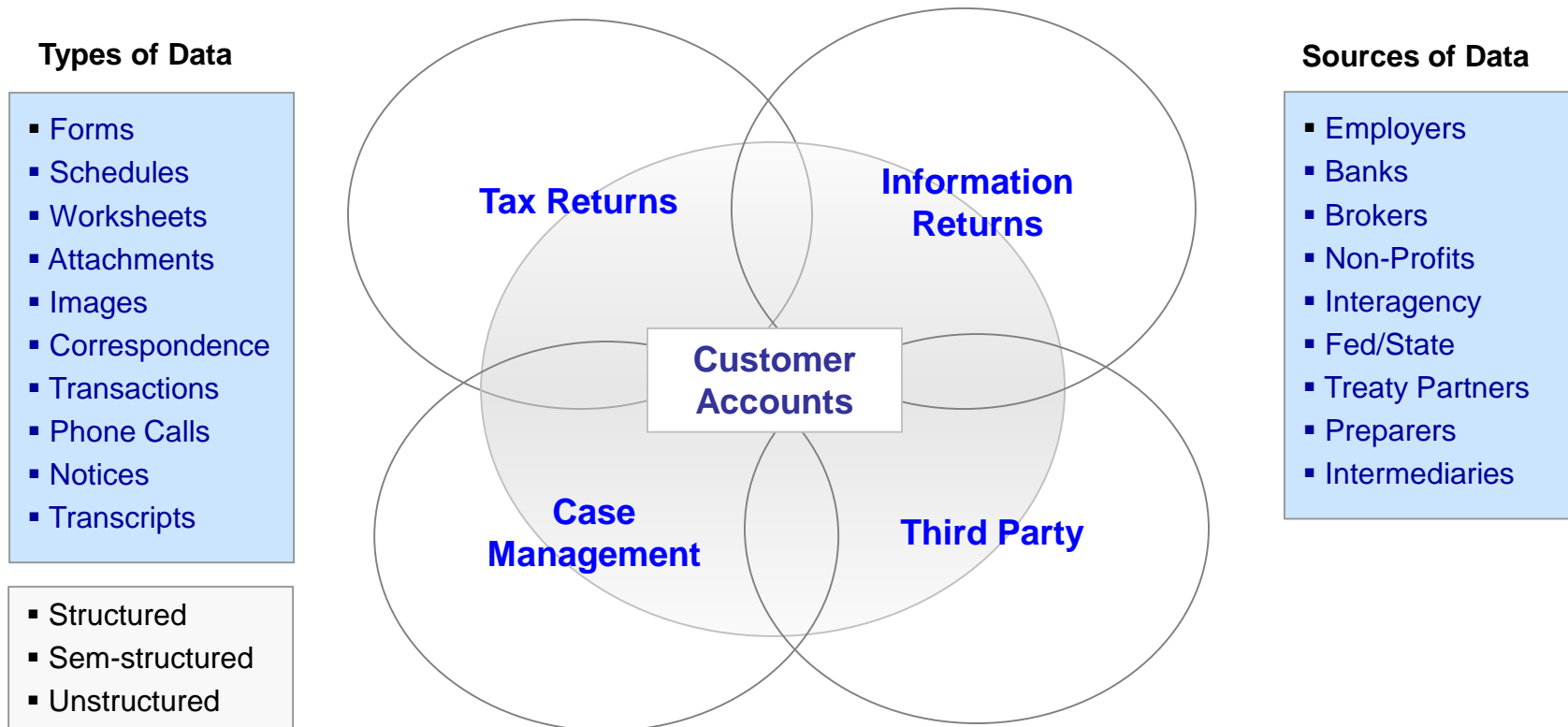
IRS Enterprise Data Environment



Metadata Reference Model for IRS Research

IRS Enterprise Metadata

- **Good news:** Small number of top-level domains with fairly good semantic interoperability



IRS Enterprise Metadata

- Over 425 separate systems or applications in the IRS, each with its own metadata
- Inconsistent tools and standards across legacy systems
- Access to reference material is sometimes limited by role or requires specialized software
- Formats include Excel spreadsheets, PDFs, Word documents, XML, database catalogs
- As IRS architects new enterprise data solutions, they are not always being defined with consistent metadata standards
- No strategies for Master Data Management, Householding, Entity Resolution, and other information quality standards

Metadata Reference Model for IRS Research

IRS Research Data Environment



- IRS Research organization manages and administers its own IT systems
- Centralized database with roughly 30 legacy sources
- Largest database in the IRS
- Standardization of key dimensions, e.g., taxpayers, time, geography
- **Over 40,000 unique data elements**
- Web-based metadata and dynamic data profiling
- Database is accessible via third-party tools, e.g., SAS, SQL, R, Stata, Hyperion, ArcGIS, or any ODBC- or JDBC-compliant application
- Over 900 users across the IRS, Treasury, GAO, and other

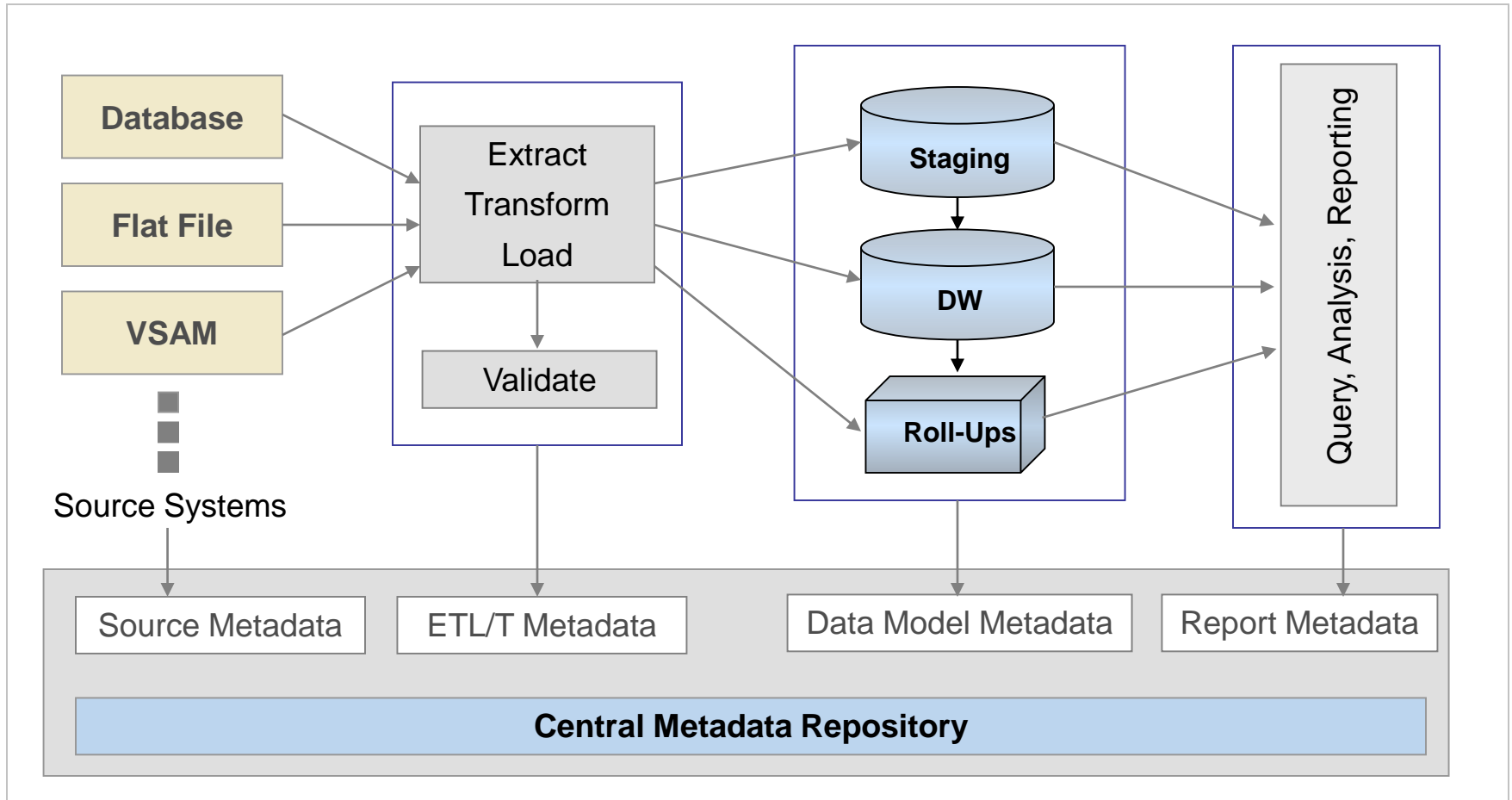
Combining Metadata, Search, and Data Profiling

IRS Research Data Environment

▪ Number of key data sources	28
▪ Number of database tables	1,180
▪ Number of columns.....	41,050
▪ Number of columns with metadata	25,250
▪ Number of metadata-column attributes.....	550,000
▪ Total database storage	420TB
▪ Total disk storage	1.1PB
▪ Number of user accounts	900
▪ Average daily concurrent connections	120
▪ Average daily database queries	3,400
▪ Average daily database queries from the website	1,200

Metadata Reference Model for IRS Research

IRS Research Data Environment



Metadata Reference Model for IRS Research

IRS Research Metadata



- Simple **reference model** is used to guide consistency of data definitions and other searchable artifacts
- Combination of **system, contextual, and application** attributes
- **Controlled vocabulary** for key descriptive elements
- Strategy favors **basic discoverability** rather than developing formal, systematized collections

Strengths

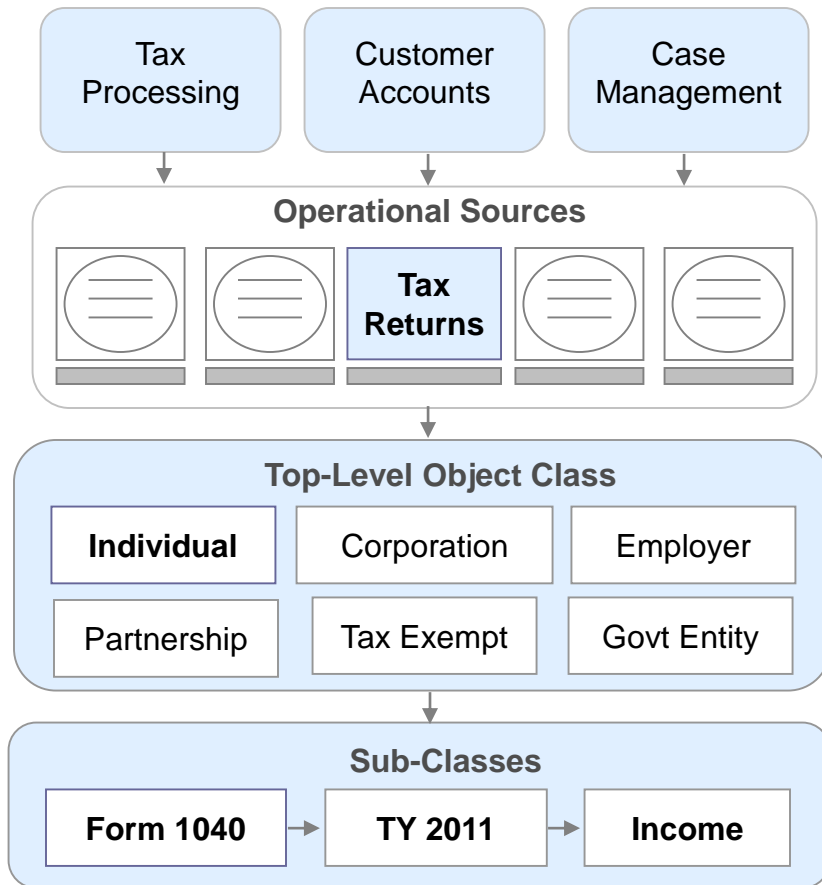
- Simple hierarchical model
- Web accessible
- Fast search times for most terms and phrases
- Easy to maintain

Weaknesses

- No markup language
- Lack of explicit structured connections
- Poor homographic separation

Metadata Reference Model for IRS Research

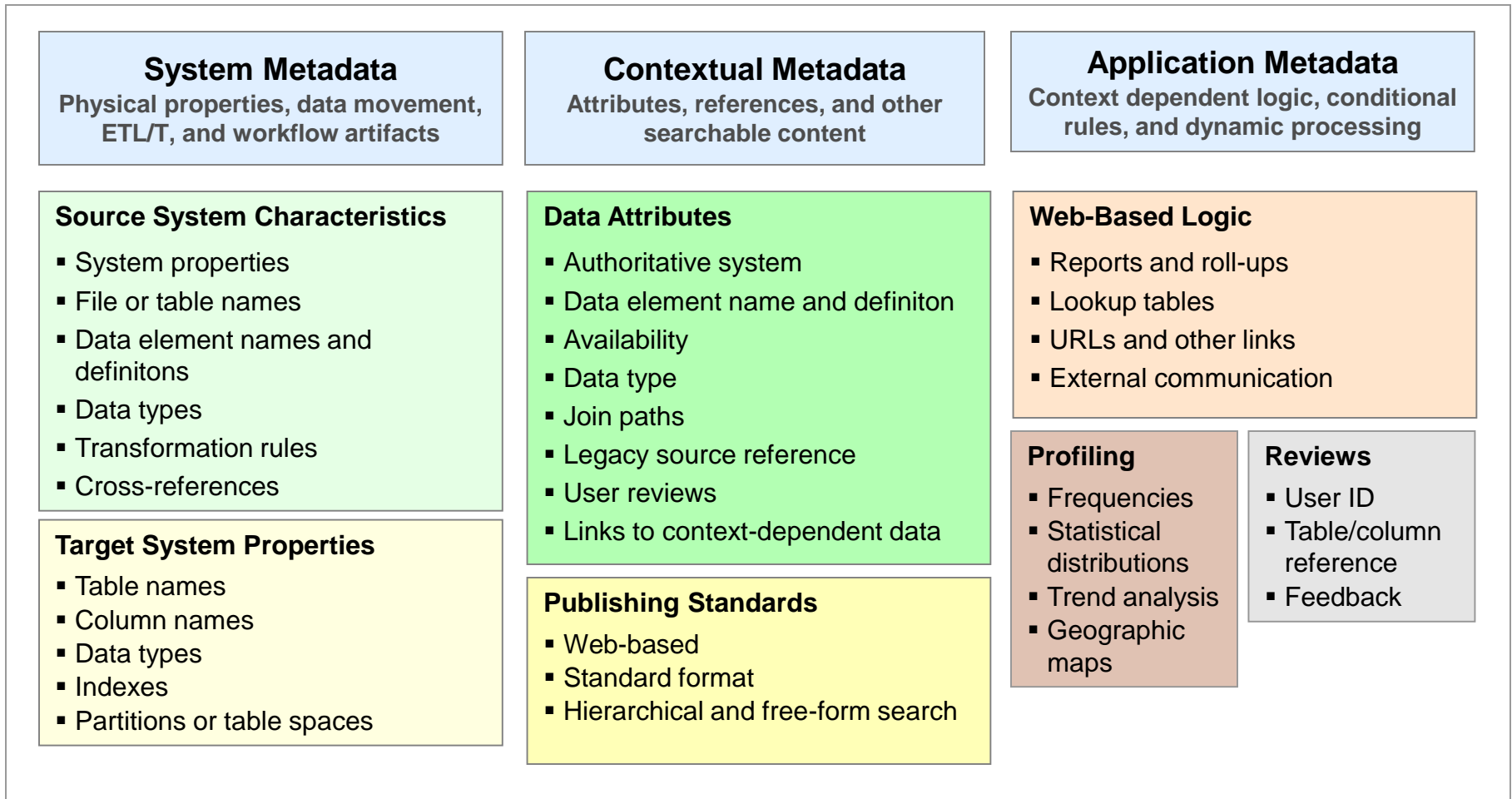
Conceptual Framework



- Small number of top-level domains
- Well-defined object classes within each domain, e.g., Individual Tax Returns is a member of Tax Processing domain
- Sub-classes exist within an object class, e.g., Schedule D Worksheet within the Schedule D
- Object classes can be added over time as source data change
- General monotonicity of classes and properties
- Properties are context independent

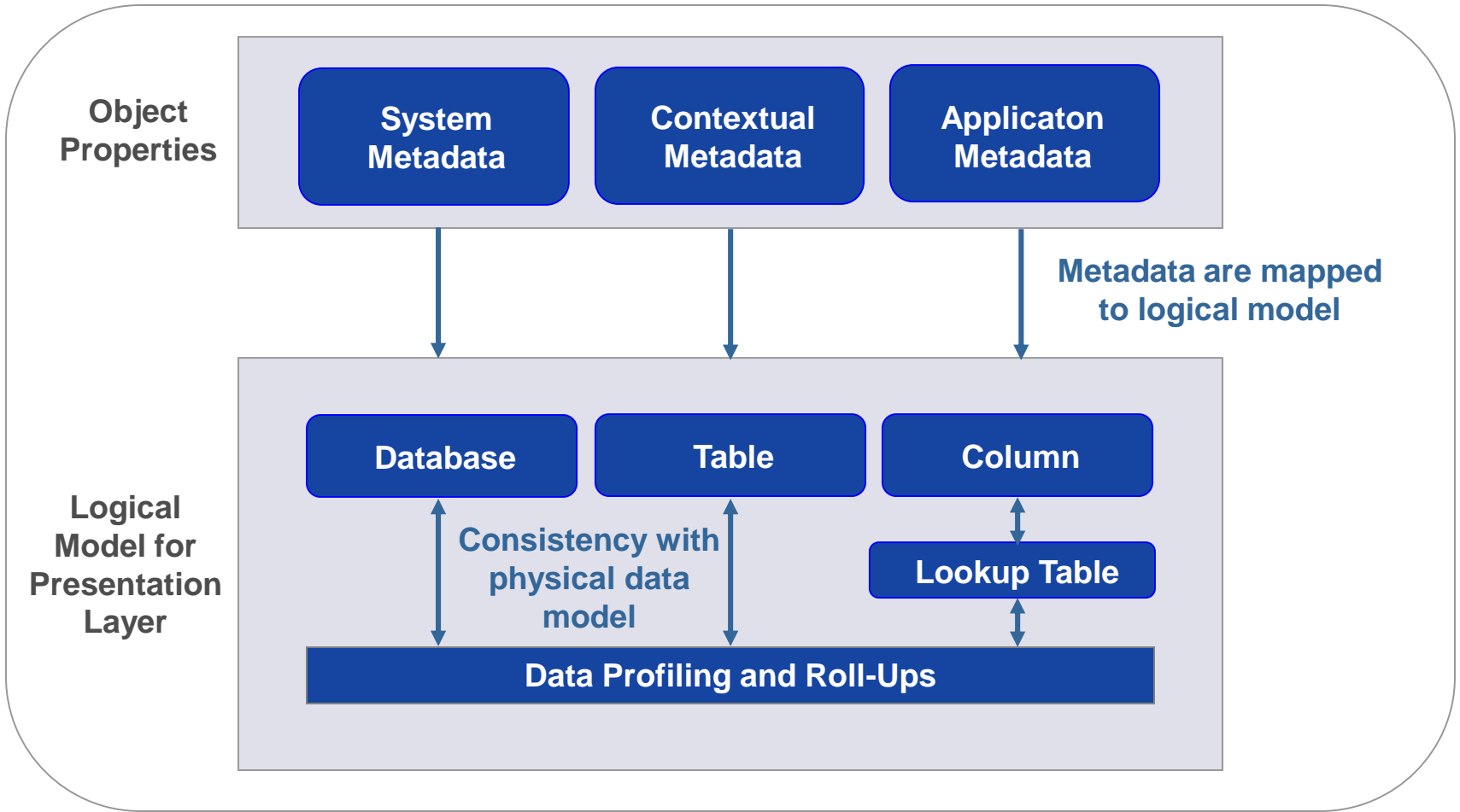
Metadata Reference Model for IRS Research

Conceptual Framework



Metadata Reference Model for IRS Research

Logical Model



Metadata Reference Model for IRS Research

Database-level Metadata

Category	Property	Definition
Contextual	Data Source Name	Name of the legacy or authoritative data source (if applicable)
	Data Source Definition	Definition of the data source
	First Year	First year in which any table is available
	Last Year	Last year in which any table is available
	Last Updated	Last date (YYYYMMDD) that data were updated in the database
	Frequency	Frequency of the data release (Annual, Quarterly, Monthly, Weekly)
	Frequency Type	Calendar Year, Fiscal Year, Tax Year
System	Database ID	Unique integer value of the database (data source)
Application	Has Reports	Boolean indicator for the presence of roll-ups (summary tabulations) for the database
	Num Tables	Number of tables in the database
	Has Schema	Boolean indicator for the presence of a visual schema
	Has Table Statistics	Boolean indicator for the presence of row counts by State, County, and ZIP code

Metadata Reference Model for IRS Research

Table-level Metadata

Category	Property	Definition
Contextual	Table Name	Name of the database table
	Table Definition	Definition of the table
	First Year	First year in which the table is available
	Last Year	Last year in which the table is available
	Last Updated	Last date (YYYYMMDD) that data were updated in the table
	Last Month	Last month (January – December) that data were updated
System	Table ID	Unique integer value of the table
	Database ID	Integer value of the associated database
	Year Type	Column name associated with Frequency_Type
Application	Num Columns	Number of columns in the table
	Has Table Statistics	Boolean indicator for the presence of roll-up tables
	Has TLine	Boolean indicator for the presence of PDF files
	Has Partition	Boolean indicator for the presence of partitioned database tables
	Display Name	Short description of the table for page display
	Table Label	Calendar-specific label for roll-ups

Metadata Reference Model for IRS Research

Column-level Metadata

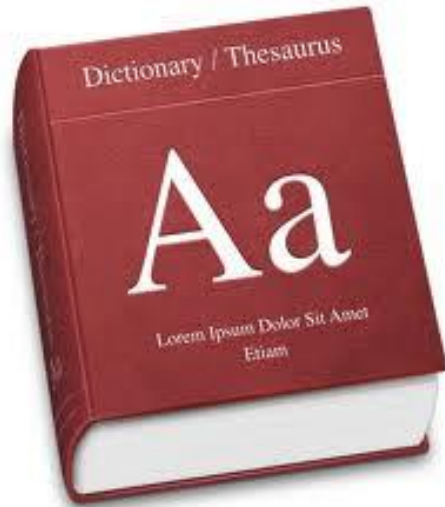
Category	Property	Definition
Contextual	Column Name	Name of the database table
	Column Long Name	Legacy name or display name
	Column Definition	Definition of the column
	First Year	First year (YYYY) in which the column is available
	Last Year	Last year (YYYY) in which the column is available
	Last Updated	Last date (YYYYMMDD) that column was updated
	Last Month	Last month (January – December) that column was updated
	Data Type (General)	Numeric or Character
	Data Type	Database data type
	Is Primary Key	Boolean value identifying if the column is part of a primary key
	Nulls Allowed	Boolean value identifying if nulls are present in the column
	Distribution Type	A value of Discrete or Continuous
	Range Type	A value of Positive, Negative, Positive and Negative, or Null
	Legacy File Source	Specific file section, if exists, of the legacy source

Metadata Reference Model for IRS Research

Column-level Metadata

Category	Property	Definition
Contextual	URL	HTTP address of any links in the column definition
	Min Length	Minimum number of positions in the column value
	Max Length	Maximum number of positions in the column value
System	Column ID	Unique integer value of the column
	Table ID	Unique integer value of the table
	Refresh Date	Last date (YYYYMMDD) that any column attributes were updated
Application	Has Lookup	Boolean value indicating the presence of a lookup table
	Lookup Table Name	Name of the lookup table, if exists
	Has Frequency	Boolean value for frequency table (discrete Distribution Type only)
	Has Statistics	Boolean value for summary statistics (continuous Distribution Type only)
	Has Trends	Boolean value for trend analysis (any Distribution Type)
	Has Maps	Boolean value for frequency table (any Distribution Type)

Column Definition Standards



Controlled Vocabulary

- A standard format is used for column definitions to facilitate consistency across domains and classes
- Column definition includes terms with some ordering implied
- Terms are easily differentiated from each other to avoid overlap
- Mutual exclusivity of terms helps to speed search and retrieval

Other Standards

- Capitalize first letters in legacy names
- Related terms (columns) should appear at the end of the definition
- Reserve all capitals for acronyms

Column Definition Standards

Key Properties

- Legacy name or equivalent
- Beginning cycle
- Short description
- Line item number if tax form or schedule
- Is-a relation
- Format
- Range of values
- Valid values
- Related columns

Standard Template for Column Definitions

The <legacy name or equivalent>

Choose all that apply:

[was added in <Cycle>]. [It has data through <Cycle>]. [It is <short description>]. [It is reported on <Form Name, Line Number>.] [The format is <number of characters> or <numeric>.] [It is reported in <positive, negative, or both positive and negative whole dollars or dollars and cents>.] [Valid values are <enumerated list or range>.] [It is <zero, blank, null, or other condition> if not present or not applicable]. [See <related columns>.]

Column Definition Standards - Examples

THFTLSSC: Total Casualty Theft Loss - Computer

The Total Casualty Theft Loss - Computer is the computer generated amount for comparison with amount of Casualty or Theft losses (from Form 4684, Line 18) reported on Form 1040, Schedule A, Line 20 or Form 1040-NR, Schedule A, Line 8. It is reported in positive whole dollars. It is zero if not present.

TC973_MHTNDEL: TC 973 Months Delinquent

The TC 973 Months Delinquent is the number of months generated by TC 973 for which delinquency penalty accessed. The format is numeric. It is null if not applicable.

WO_IND: Write-Off Indicator

The Write-Off Indicator is used to indicate an unreversed Transaction Code (TC) 530 - Currently not Collectible Account with very little or no chance of collection. The format is numeric. Valid values are 1, 7 to 9, 12, and 13 for IMF and 2 to 7, 10, 13, 15 and 16 for BMF. It is zero or null (prior to 200039) if not present.

CDW Metadata – Database Level

CDW Database Library - Windows Internet Explorer

http://cdw.web.irs.gov/databases.aspx

File Edit View Favorites Tools Help

CDW Database Library

Compliance Data Warehouse

Research, Analysis, and Statistics

Monday, January 09, 2012

[Contact Us](#) | [Log In](#)

[Home](#) | [Databases](#) | [Reports](#) | [Solutions](#) | [Community](#) | [Environment](#) | [Alerts](#)

Database Library

Database Name	First Year	Last Year	Last Update
Audit Information Management System (AIMS) AIMS tracks the location, age, and status of returns in Examination, and is used by Appeals, Examination, and TE/GE to control returns, input assessments/adjustments, and provide management reports. Profile View Tables Table Statistics Data Reviews	1997	2011	Nov
Account Receivable Dollar Inventory (ARDI) The CFO ARDI Management System (CAMS) tracks all entities and tax modules that show a debit balance on the Master File. Profile View Tables Table Statistics Data Reviews	1997	2011	Nov
Automated Underreporter (AUR) The AUR program matches information returns against individual income tax returns to verify that income and deductions are correctly reported. Discrepancies identified in this matching process are stored in the AUR database. Profile View Tables Table Statistics Data Reviews	2003	2010	Dec
Business Master File (BMF) The Business Master File (BMF) contains Collection Status Codes and Transaction Codes for business taxpayers. It also contains the TC 590 File. The Status Code is used to designate current collection status of the business module as it appears on various transcripts. Transaction Codes are codes used to identify transactions processed on BMF. The TC 590 File is for Transaction Code 590 and contains information from Form 851, Affiliations Schedule. Profile View Tables Table Statistics Reports	2001	2011	Nov

Local intranet 100%

CDW Metadata – Table Level

CDW Database Tables - Windows Internet Explorer

http://cdw.web.irs.gov/DB.aspx?id=IRTF&intHasReports=0&intdbID=11&intHasReviews=12

File Edit View Favorites Tools Help

Home | Databases | Reports | Solutions | Community | Environment | Alerts

Home > Databases > IRTF

Individual Returns Transaction File (IRTF)

View Tables | [Table Statistics](#) | [Data Reviews](#) [Print](#) [Excel](#)

Table Name	Description
IRTF_AUDIT_HISTORY	The Audit History table contains audit results from the individual taxpayers last audit. There can be zero or one Audit History record per Taxpayer Identification Number (TIN) Tax Period combination. The Audit History record can be associated with one Entity, one Tax Module, and one Form 1040 record for the same TIN and Tax Period. It can also be associated with the up to five Audit Issue records by using the TIN, Tax Period, and Audit Year. Profile View Columns Table Statistics
IRTF_AUDIT_ISSUE	The Audit Issue table has data through 199753. It contains the audit year and No Change Issue Code. There can be up to five Audit Issue records per Taxpayer Identification Number (TIN) Tax Period combination. The Audit Issue records can be associated with one Entity, one Tax Module, and one Form 1040 record for the same TIN and Tax Period. The up to five Audit Issue records can also be associated with the one Audit History record by using the TIN, Tax Period, and Audit Year. Profile View Columns Table Statistics
IRTF_DEPEND	The Dependent table contains data concerning the individual taxpayer's dependents. There can be zero or many Dependent records per Taxpayer Identification Number (TIN) Tax Period combination. The Dependent records (zero - four) can be associated with one Entity, one Tax Module, and one Form 1040 record for the same TIN and Tax Period. Profile View Columns Table Statistics
IRTF_EIC	Form 1040 (Schedule EIC) - Earned Income Credit is used to report information on qualifying children for the Earned Income Tax Credit (EITC). The Earned Income Credit (EIC) is reported on Form 1040, Line 64a. There can be zero or one record per TIN. TIN TYP. TAX PRD combination.

Local intranet 100%

CDW Metadata – Column Level

CDW - Table Definition - Windows Internet Explorer

http://cdw.web.irs.gov/tableDetail2.aspx?has_schema=1&dbtype=I&dbname=IRTF&id=100

File Edit View Favorites Tools Help

Home | Databases | Reports | Solutions | Community | Environment | Alerts

Home > Databases > IRTF > IRTF_F1040

Table Name: IRTF_F1040
Form 1040 - U.S. Individual Income Tax

[View Columns](#) | [Column Profile](#) | [Column Statistics](#) [Print](#) [Excel](#)

Field Name	Description
ACCUMTAX	The F4970 Tax is the amount of partial tax attributable to the accumulation distribution under Section 667. It is reported on Form 4970, Line 28. It is included on Form 1040, Line 61 (notated 'ADT'). It is reported in positive whole dollars. It is zero if not present. Profile Statistics Trends Map Reviews
ACRAMT	The Adoption Credit Amount was added in 199853. It is the amount of adoption expenses allowable as a credit. It is reported on Form 8839, Line 18. It is included on Form 1040, Line 53 or Form 1040-NR, Line 48 when the box for Form 8839 is checked. It is reported in positive whole dollars. It is zero if not present. See ACRAMTC. Profile Statistics Trends Map Reviews
ACRAMTC	The Adoption Credit Amount - Computer was added in 199853. It is the computer generated amount for comparison with Adoption Credit Amount. It is reported in positive whole dollars. It is zero if not present. See ACRAMT. Profile Statistics Trends Map Reviews
ADDCHLDTXCREIC	The Additional Child Tax Credit Earned Income - Computer was added in 200652. It is the computer generated amount for earned income used to determine eligibility for the additional child tax credit. It is compared to Form 8812, Line 4a. It is reported in positive whole dollars. It is zero if not present. See ADDLCHC. Profile Statistics Trends Map Reviews
ADDLCHC	The Additional Child Tax Credit Amount - Computer was added in 199952. It is the computer generated amount for comparison with Additional Child Tax Credit

Select Related Links

Local intranet 100%

CDW Metadata – Column Level

Column Profile - Windows Internet Explorer

http://cdwdev.web.irs.gov/cdw_Profiles.aspx?Name=IRTF_F1040&lookUp=0&colID=115

File Edit View Favorites Tools Help

Open an Account
Forgot your password?

Resources

- Getting Started
- Databases
- Knowledge Base
- Software Tools
- Data Reviews

Services

- Training
- Data Analysis
- Record Matching
- Data Transfer
- Storage Services
- TIN Masking

Support

- Data Alerts
- News and Events
- Contact Us
- About CDW

Related Links

Select Related Links

Home | Databases | Reports | Solutions | Community | Environment | Alerts

Home > Databases > IRTF > IRTF_F1040 > ADJ_INCC > Profile

Profile for ADJ_INCC

Adjusted Gross Income Amount - Computer

[View Columns](#) | Profile | [Statistics](#) | [Trends](#) | [Map](#) | [Reviews](#) [Print](#) [Help](#)

Overview	
Database Name	Individual Returns Transaction File (IRTF)
Table Name	IRTF_F1040
Column Name	ADJ_INCC
Column Long Name	Adjusted Gross Income Amount - Computer
Column Description	The Adjusted Gross Income Amount - Computer is the computer generated amount for comparison to Adjusted Gross Income Amount. It is reported in positive and negative whole dollars. It is zero if not present. See ADJGROSS.

Availability	
First Year	1997
Last Year	2011
Last Month	11
Last Update	12/12/2011
Frequency	Monthly
Frequency Type	Processing Year

Legacy Information	
Source	Individual Returns Transaction File (IRTF)
File Name	460-02-11-230-01,04,05,06

What users say about this column

Number of reviews : 0

[No reviews for this column](#)

[Write a review for this column](#)

Related Links

Local intranet 100%

Free-Form Search

- Free-form search is available when the location of a column is unknown through a pre-determined drill path (i.e., database-table-column)
- CDW presents standardized search results for consistency
 - Includes database, table, column name, and column description
 - Data profiling functions also available
- Indirect benefit of exposing inconsistencies across databases
 - Same column in two different databases may have a different name or description
 - Often leads to uncovering additional inconsistencies
 - Useful data quality tool

CDW Search Results

CDW Database Library - Windows Internet Explorer

http://cdwdev.web.irs.gov/sh1.aspx?in3=1&in1=adjusted gross income&in2=1

File Edit View Favorites Tools Help

CDW Database Library

Compliance Data Warehouse

Research, Analysis, and Statistics

adjusted gross income Search

Monday, January 9, 2012

Home | Databases | Reports | Solutions | Community | Environment | News

32 Search Results - page 1 of 4 for **adjusted gross income** [Next>>](#)

Adjusted Gross Income Amount Per Return
Database: [Automated Underreporter \(AUR\)](#)
Table: [AUR](#) • [AGI](#)

The **Adjusted Gross Income Amount Per Return** is the amount claimed as Adjusted Gross Income (AGI) after subtracting all adjustments (Form 1040, Line 36) from Total Income (Form 1040, Line 22). It is reported on Form 1040, Line 37. It is reported in positive and negative dollars and cents.

[Profile](#) | [Statistics](#) | [Trends](#) | [Maps](#) | [Reviews](#)

Adjusted Gross Income Prior Amount Per Return
Database: [Automated Underreporter \(AUR\)](#)
Table: [AUR](#) • [AGI_PRIOR](#)

The **Adjusted Gross Income Prior Amount Per Return** is the amount claimed as Adjusted Gross Income (AGI) after subtracting all adjustments from Total Income for the prior year. It is reported in positive and negative dollars and cents.

[Profile](#) | [Statistics](#) | [Trends](#) | [Maps](#) | [Reviews](#)

Underreported Earned Income Tax Credit (EITC) Adjusted Gross Income (AGI)
Database: [Automated Underreporter \(AUR\)](#)
Table: [AUR](#) • [UREITCAGI](#)

The **Underreported Earned Income Tax Credit (EITC) Adjusted Gross Income (AGI)** was added in 200452. It is the amount of underreported EITC AGI. It was changed in 200552 to remove 'Modified' from the legacy name. It is reported in positive and negative dollars and cents. Values are zero in 200352.

[Profile](#) | [Statistics](#) | [Trends](#) | [Maps](#) | [Reviews](#)

Resources

- [Getting Started](#)
- [Databases](#)
- [Knowledge Base](#)
- [Data Analysis Tools](#)
- [Data Reviews](#)

Services

- [Training](#)
- [Data Analysis](#)
- [Record Matching](#)
- [Data Transfer](#)
- [Storage Services](#)

Support

- [Data Alerts](#)
- [News and Events](#)
- [Contact Us](#)
- [About CDW](#)

Related Links

Select Related Links

Local intranet 100%

Data Profiling

- **Data profiling** is the use of a set of standard functions or rules to explore distributional aspects of data to identify data quality problems
- Standard data profiling functions typically include:
 - Frequency counts and outliers
 - Statistical analysis
 - Pattern matching and rule validation
- CDW provides basic exploratory statistics through the website that can be used to support data profiling, including
 - Frequency tables
 - Statistical distributions
 - Trend analysis
 - Geospatial patterns
- Functions are available at both the table and column level
- Type of function presented depends on column distribution, i.e., discrete or continuous

CDW Data Profiling – Frequency Tables

ColumnSumStats - Windows Internet Explorer

http://cdwdev.web.irs.gov/columnSumStats.aspx?has_schema=1&dbname=IRTF&tblNar

File Edit View Favorites Tools Help

Home | Databases | Reports | Solutions | Community | Environment | Alerts

Home > Databases > IRTF > IRTF_F1040 > STATE > Frequency

Frequency Distribution for STATE

State

[View Columns](#) | [Profile](#) | [LookUp](#) | Frequency Table | [Trends](#) | [Map](#) | [Reviews](#) [Print](#) [Help](#)

Time Period: Year-To-Date
Calendar Year: 2011

Year-To-Date, through December

Value	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	567,618	0.39	567,618	0.39
AA	3,483	0.00	571,101	0.39
AE	110,612	0.07	681,713	0.47
AK	373,760	0.25	1,055,473	0.73
AL	2,102,260	1.45	3,157,733	2.19
AP	77,324	0.05	3,235,057	2.24
AR	1,224,336	0.85	4,459,393	3.09
AS	2,285	0.00	4,461,678	3.09
AZ	2,718,397	1.88	7,180,075	4.98
CA	16,683,674	11.58	23,863,749	16.57
CO	2,369,967	1.64	26,233,716	18.21
CT	1,727,554	1.19	27,961,270	19.41

Local intranet 100%

CDW Data Profiling – Statistical Distributions

Column Statistics - Windows Internet Explorer

http://cdwdev.web.irs.gov/cdw_Statistics.aspx?Name=IRTF_F1040&intcolID=119238&str

File Edit View Favorites Tools Help

Column Statistics

Home | Databases | Reports | Solutions | Community | Environment | Alerts

Home > Databases > IRTF > IRTF_F1040 > ADJ_INCC > Statistics

Statistics for ADJ_INCC

Adjusted Gross Income Amount - Computer

[View Columns](#) | [Profile](#) | [Statistics](#) | [Trends](#) | [Map](#) | [Reviews](#) [Help](#)

By Processing Year, Year-To-Date through November

Filter By: None | Filter Value: None | Time Period: Year-To-Date | Calendar Year: 2011

Year	Count
06	130m
07	135m
08	150m
09	140m
10	135m
11	145m

Year	Average
06	50k
07	55k
08	50k
09	55k
10	50k
11	55k

Year	Total (Sum)
06	7.5t
07	8.0t
08	8.5t
09	8.0t
10	7.5t
11	8.0t

Statistics for ADJ_INCC, by Processing Year, Year-To-Date through November 2011

Basic Statistics	
Count	143296216
Count Missing	0
Minimum	-845988266
Maximum	988914647
Average	56097.55
Sum	8038566005154.00
Std. Deviation	576604.42
Skewness	499.82

Local intranet 100%

CDW Data Profiling – Statistical Distributions

Column Statistics - Windows Internet Explorer

http://cdwdev.web.irs.gov/cdw_Statistics.aspx?Name=IRTF_F1040&intcolID=11923&str

File Edit View Favorites Tools Help

Column Statistics

Support

- Data Alerts
- News and Events
- Contact Us
- About CDW

Related Links

Select Related Links

Statistics for ADJ_INCC, by Processing Year, Year-To-Date through November 2011

Basic Statistics			
Count	143296216	Average	56097.55
Count Missing	0	Sum	8038566005154.00
Minimum	-845988266	Std. Deviation	576604.42
Maximum	988914647	Skewness	499.82
Range	1834902913	Kurtosis	785815.65

Percentiles	
1%	-1917
5%	2230
10%	5530
25%	14000
50% (Median)	31533
75%	65819
90%	112961
95%	156818
99%	357620
Interquartile Range	51819

Other Statistics	
Count Total	143296216
Count Non-Missing	143296216
Count Zero	--
Unique Count	1340801
Uniqueness Ratio	0.009357
Percent Non-Missing	100.00
Percent Non-Zero	99.25

Done

Local intranet 100%

CDW Data Profiling – Trend Analysis

Column Trends - Windows Internet Explorer

http://cdwdev.web.irs.gov/cdw_Trends.aspx?Name=IRTF_SCHED_B&intcolID=12634&st

File Edit View Favorites Tools Help

Resources

- Getting Started
- Databases
- Knowledge Base
- Software Tools
- Data Reviews

Services

- Training
- Data Analysis
- Record Matching
- Data Transfer
- Storage Services
- TIN Masking

Support

- Data Alerts
- News and Events
- Contact Us
- About CDW

Related Links

Select Related Links

Trend Analysis for DIVIDEND

Schedule B Dividends

[View Columns](#) | [Profile](#) | [Statistics](#) | Trends | [Map](#) | [Reviews](#) [Help](#)

Count (DIVIDEND), Year-To-Date through December

Group By: STATE | Statistic: Count | Time Period: Year-To-Date | Calendar Year: 3 Years

Count, All Rows

Year	Count
2006	6,000,000
2007	7,428,811
2008	8,717,895
2009	9,238,865
2010	8,648,427
2011	9,961,005

Year/Year Change, All Rows

Year	Count	Change	%Chg
2007	7,428,811	--	--
2008	8,717,895	1,289,084	17.3
2009	9,238,865	520,970	5.9
2010	8,648,427	-590,438	-6.3
2011	9,961,005	1,312,578	15.1

Count, by STATE (STATE), Year-To-Date through December

STATE	2009	2010	2011
	11,122	12,915	22,546
AA	294	254	278
AE	5,467	4,644	4,751
AK	29,667	25,164	28,250

CDW Data Profiling – Maps

Windows Internet Explorer

http://cdwdev.web.irs.gov/CordaMaps.aspx?has_schema=8&Name=IRTF_SCHED_B&into

File Edit View Favorites Tools Help

Knowledge Base
Software Tools
Data Reviews

Services
Training
Data Analysis
Record Matching
Data Transfer
Storage Services
TIN Masking

Support
Data Alerts
News and Events
Contact Us
About CDW

Related Links
Select Related Links

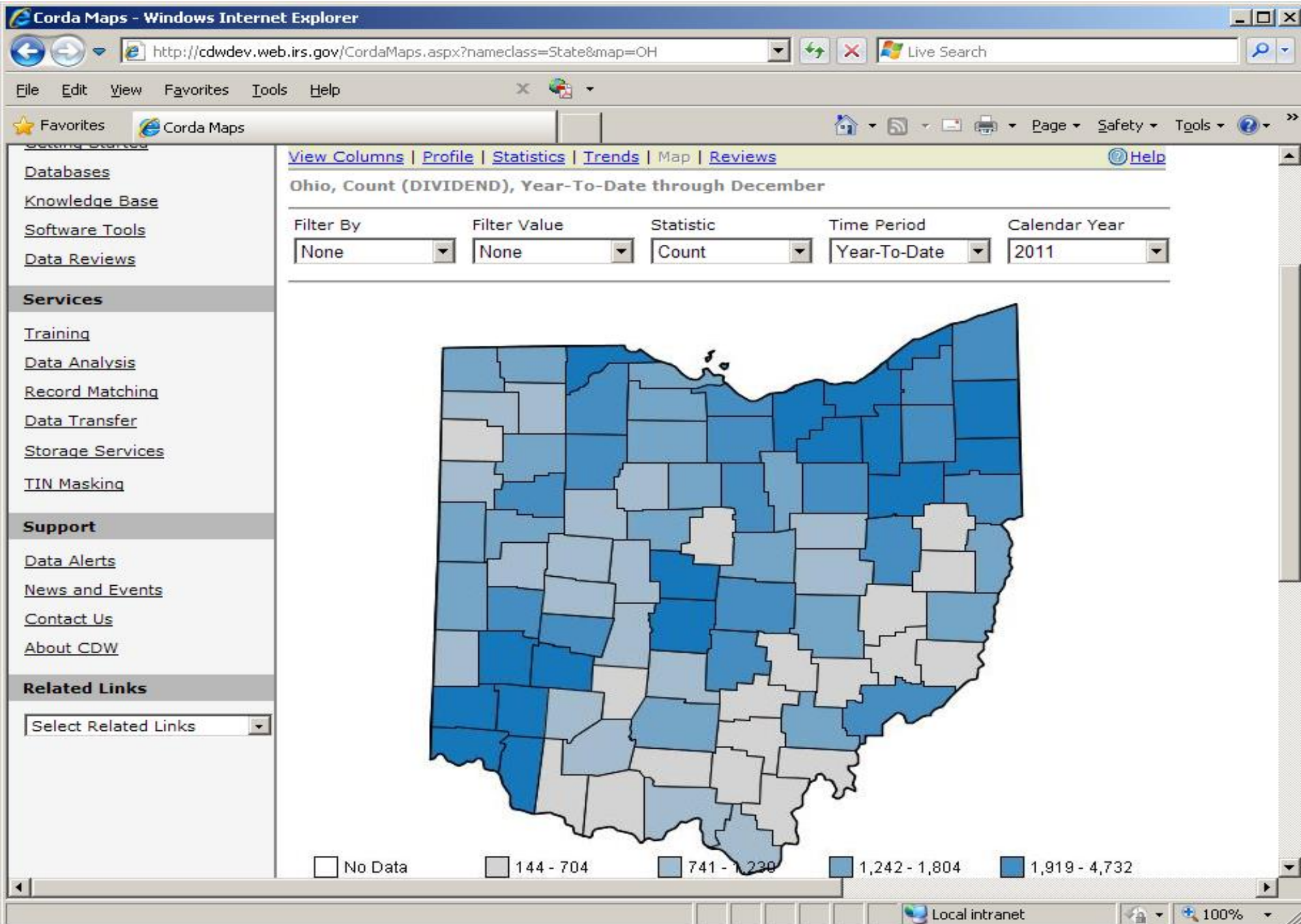
Count (DIVIDEND), Year-To-Date through December

Filter By	Filter Value	Statistic	Time Period	Calendar Year
None	None	Count	Year-To-Date	2011

Color	Range
White	No Data
Lightest Blue	20,974 - 42,915
Light Blue	43,592 - 85,478
Medium Blue	95,555 - 178,989
Dark Blue	181,838 - 277,600
Very Dark Blue	294,840 - 1,193,403

Local intranet 100%

CDW Data Profiling – Maps



Comments and Conclusion

- Metadata and search can be expanded to include data profiling capabilities through a single, web-based experience
- Database-driven queries via the browser are now possible for even the largest databases, e.g., greater than 20TB (current “big data” threshold)
 - Choice of database technology still matters
- CDW provides basic exploratory statistics through the website for data profiling that includes
 - Frequency tables, statistical distributions, trends, maps
- Exposing data via the browser also exposes any problems in the data, but this “crowd sourcing” can actually generate more user feedback about information quality

CDW Data Profiling – Data Reviews

CDW Data Reviews - Windows Internet Explorer

http://cdw.web.irs.gov/Cdw_SearchReviews.aspx

File Edit View Favorites Tools Help

Home | Databases | Reports | Solutions | Community | Environment | Alerts

Home > Data Reviews > All Databases

Recent Data Reviews -- All Databases --

1 2 3 4 5

Tax Table Income - Computer, November 18, 2011
By [McGovern Mary Ellen A](#)
Database: [Individual Returns Transaction File \(IRTF\)](#)
Table: [IRTF_F1040](#) • [TAX_TBLC](#)
There is no Tax_Tbl. What is the name of the variable that the taxpayer captures what the taxpayer entered into Line 41? Only the computer generated amount if present.

Exemption Amount - Computer, November 18, 2011
By [McGovern Mary Ellen A](#)
Database: [Individual Returns Transaction File \(IRTF\)](#)
Table: [IRTF_F1040](#) • [EXEMAMTC](#)
There is no ExemAmt. What is the variable name of the non-computed exemption amount entered on Line 42 of the 2006 Form 1040 Tax Return?

Standard Deduction - Computer, November 18, 2011
By [McGovern Mary Ellen A](#)
Database: [Individual Returns Transaction File \(IRTF\)](#)
Table: [IRTF_F1040](#) • [STDDDED](#)
This does not have a C on the end of it as the explanation implies. There is not a STDDDED at all.

Employment Code, November 08, 2011
By [Ludlum David J](#)
Database: [Business Returns Transaction File \(BRTF\)](#)
Table: [BRTF_ENTITY](#) • [EMPLCODE](#)

Resources

- [Getting Started](#)
- [Databases](#)
- [Knowledge Base](#)
- [Software Tools](#)
- [Data Reviews](#)

Services

- [Training](#)
- [Data Analysis](#)
- [Record Matching](#)
- [Data Transfer](#)
- [Storage Services](#)

Support

- [Data Alerts](#)
- [News and Events](#)
- [Contact Us](#)
- [About CDW](#)

Related Links

Select Related Links

Done Local intranet 100%