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

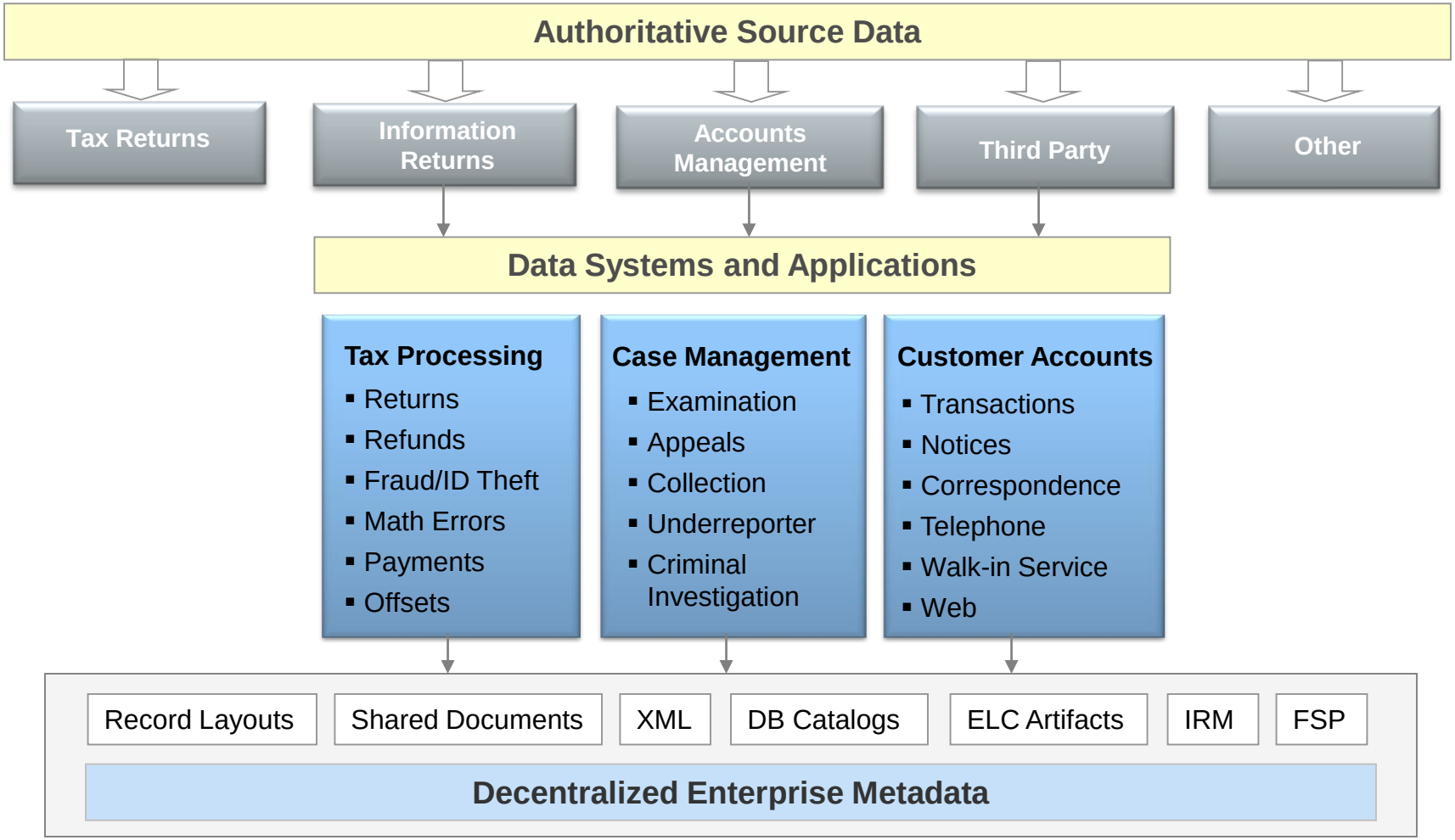
- Overview of the IRS Data Environment
  - IRS enterprise metadata
- Research Data Environment
  - Data and metadata
- Metadata Conceptual Framework and Logical Model
  - Contextual, system, and application properties
  - Controlled vocabulary
  - Examples
- Application Layer
  - Search, Reviews, and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

IRS Enterprise Data Environment

- **High volumes** of taxpayer transactions
  - Over 237 milion 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

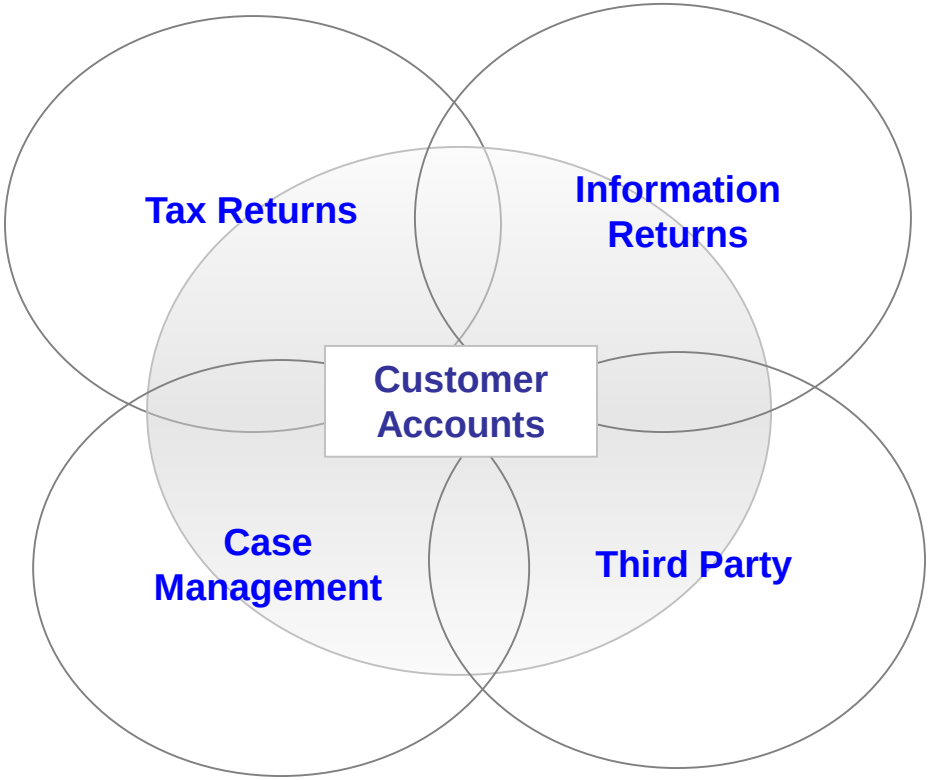
IRS Enterprise Metadata

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

Types of Data

- Forms
- Schedules
- Worksheets
- Attachments
- Images
- Correspondence
- Transactions
- Phone Calls
- Notices
- Transcripts

- Structured
- Sem-structured
- Unstructured



Sources of Data

- Employers
- Banks
- Brokers
- Non-Profits
- Interagency
- Fed/State
- Treaty Partners
- Preparers
- Intermediaries

Metadata Reference Model for IRS Research

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

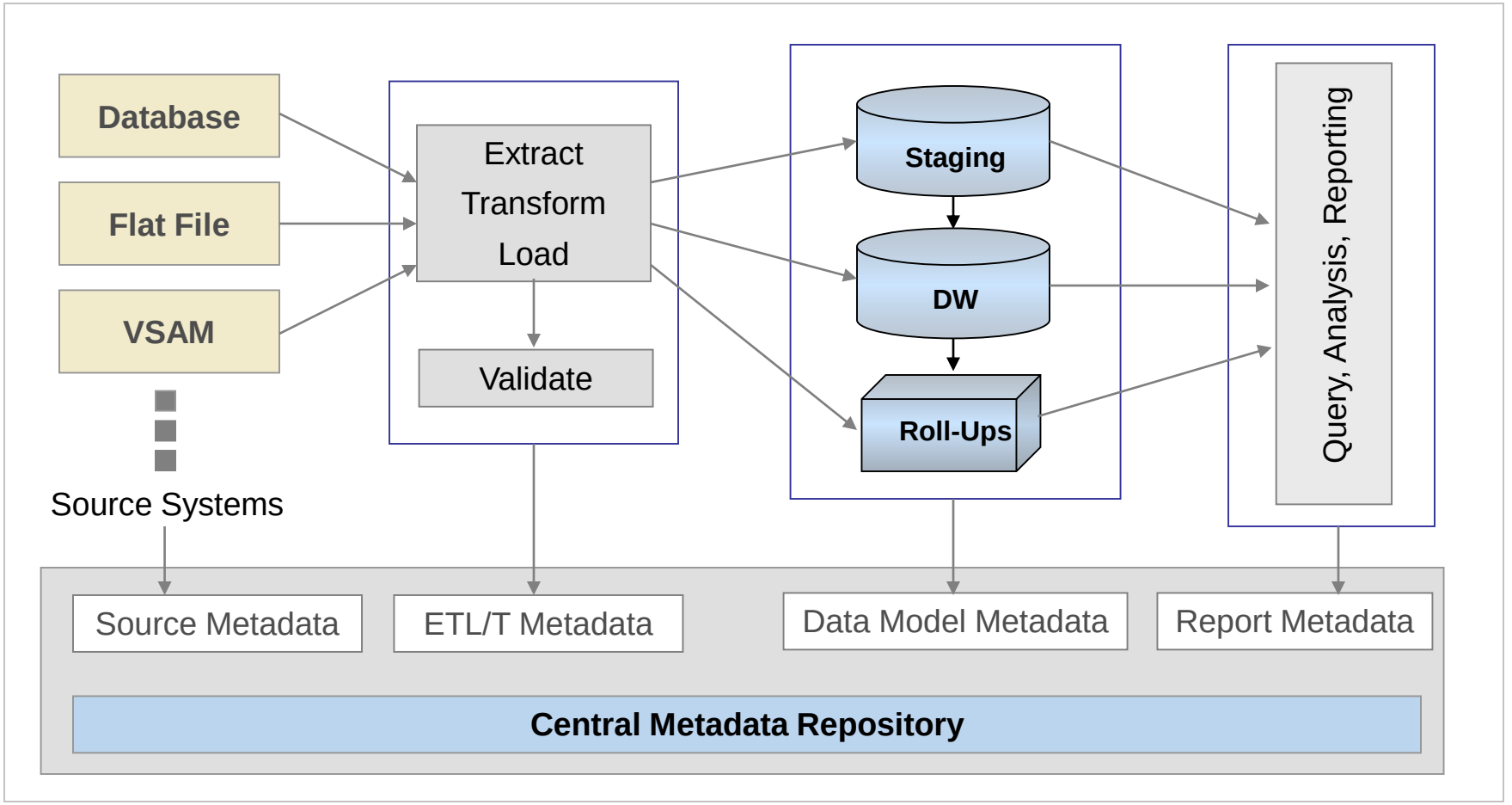
Combining Metadata, Search, and Data Profiling

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 consisteny 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

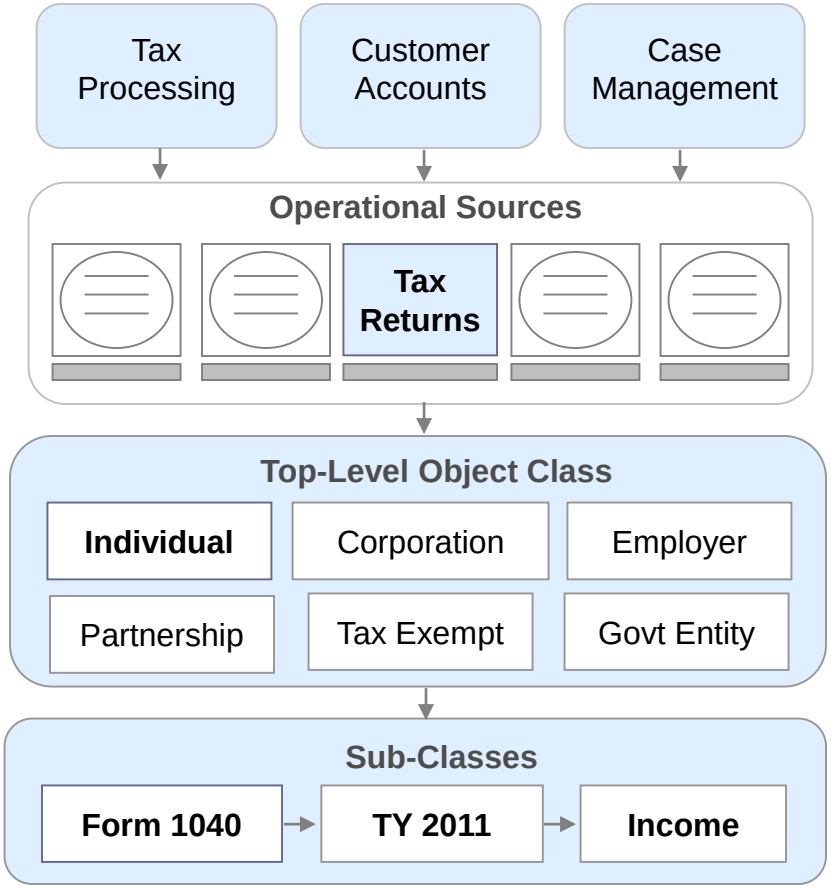
- Simple hierarchical model
- Web accessible
- Fast search times for most terms and phrases
- Easy to maintain

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 chage
- General monotinicity of classes and properties
- Properties are context independent

Metadata Reference Model for IRS Research

Conceptual Framework

System Metadata
Physical properties, data movement, ETL/T, and workflow artifacts

Source System Characteristics
- System properties
- File or table names
- Data element names and definitons
- Data types
- Transformation rules
- Cross-references

Target System Properties
- Table names
- Column names
- Data types
- Indexes
- Partitions or table spaces

Contextual Metadata
Attributes, references, and other searchable content

Data Attributes
- Authoritative system
- Data element name and definiton
- Availability
- Data type
- Join paths
- Legacy source reference
- User reviews
- Links to context-dependent data

Publishing Standards
- Web-based
- Standard format
- Hierarchical and free-form search

Application Metadata
Context dependent logic, conditional rules, and dynamic processing

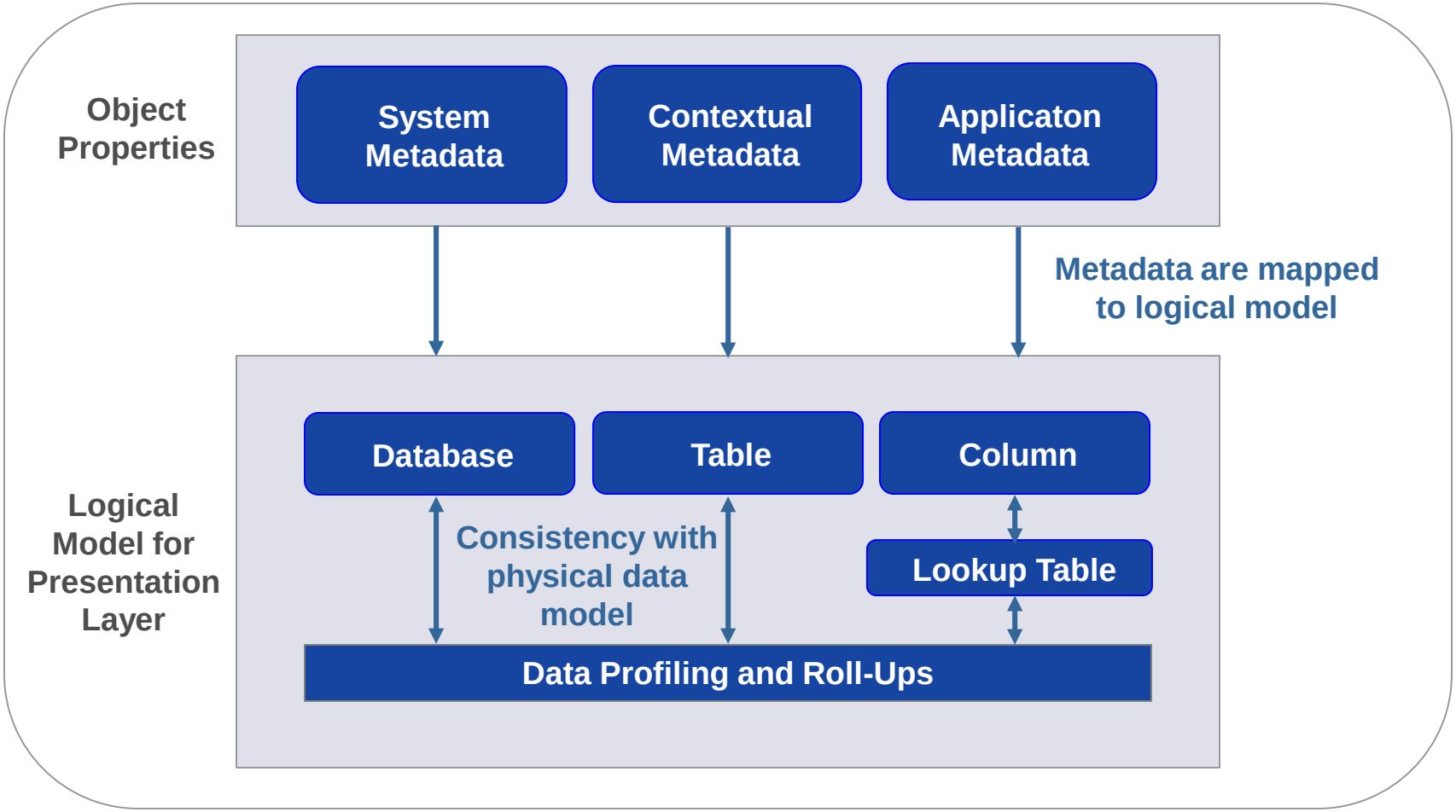
Web-Based Logic
- Reports and roll-ups
- Lookup tables
- URLs and other links
- External communication

Profiling
- Frequencies
- Statistical distributions
- Trend analysis
- Geographic maps

Reviews
- User ID
- Table/column reference
- Feedback

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

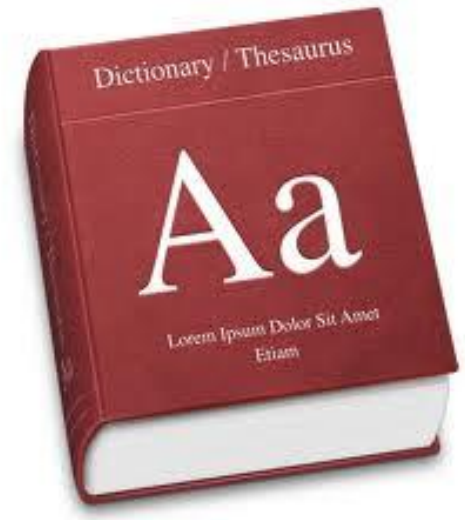
Category	Property	Definition
Contextual	Column Name	Name of the database table
	Column Long Name	Legacy name or display name
	Column Definition	Definition of the column
	First Year	First year (YYYY) in which the column is available
	Last Year	Last year (YYYY) in which the column is available
	Last Updated	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 exsts
	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 defintion 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 defintion
- 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.

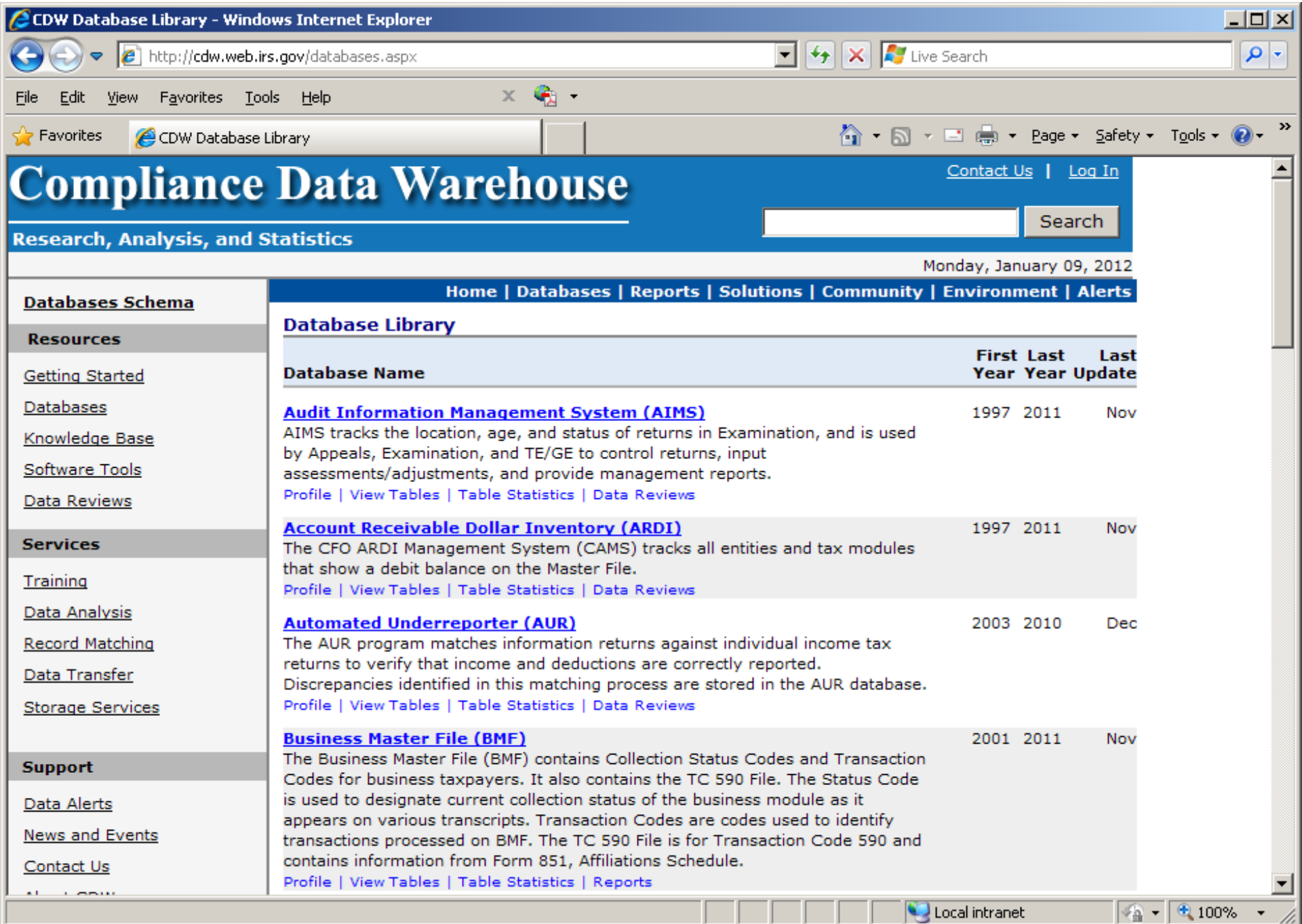
TC973_MHTNDEL: TC 973 Months Delinquent

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.

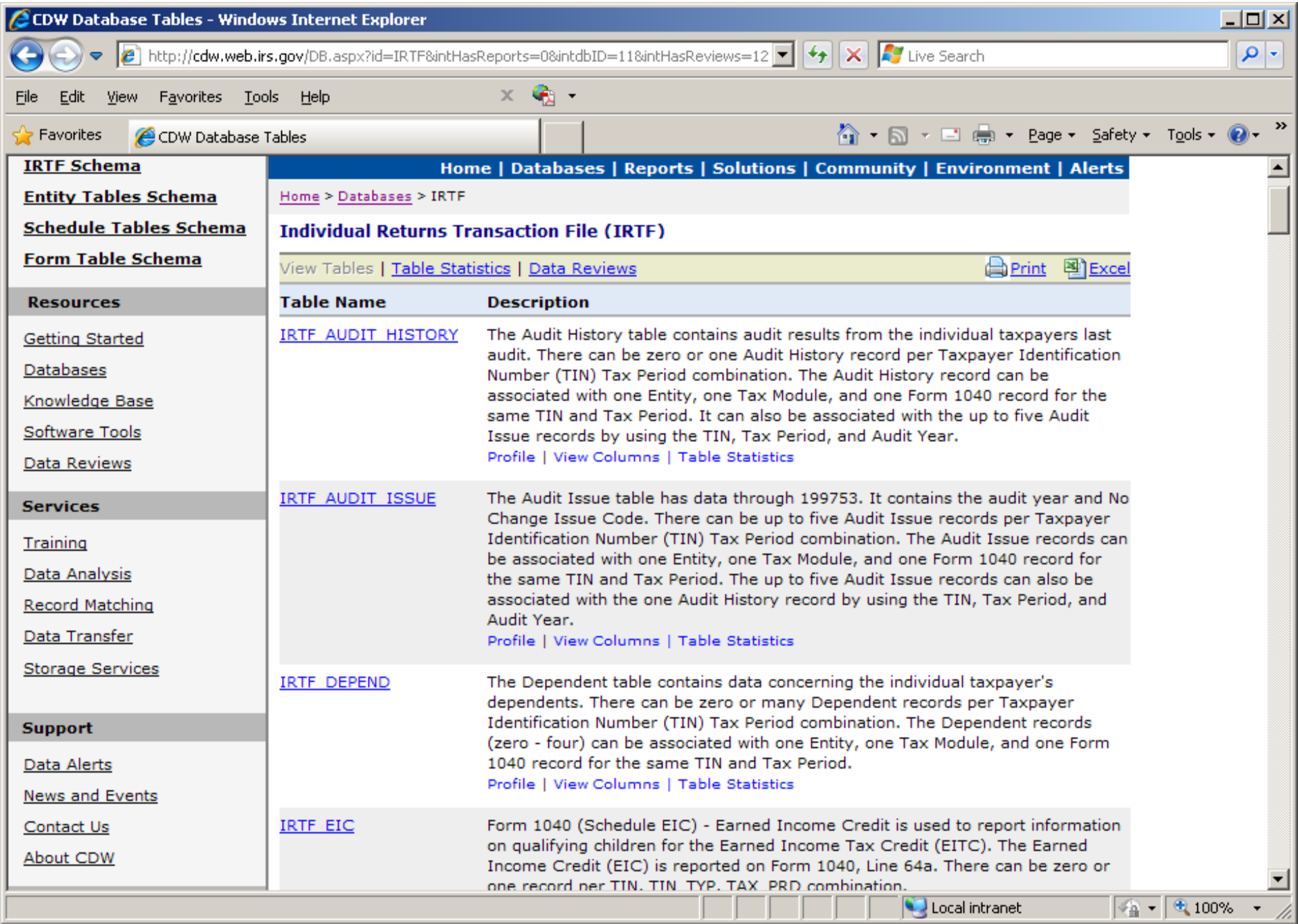
WO_IND: Write-Off Indicator

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.

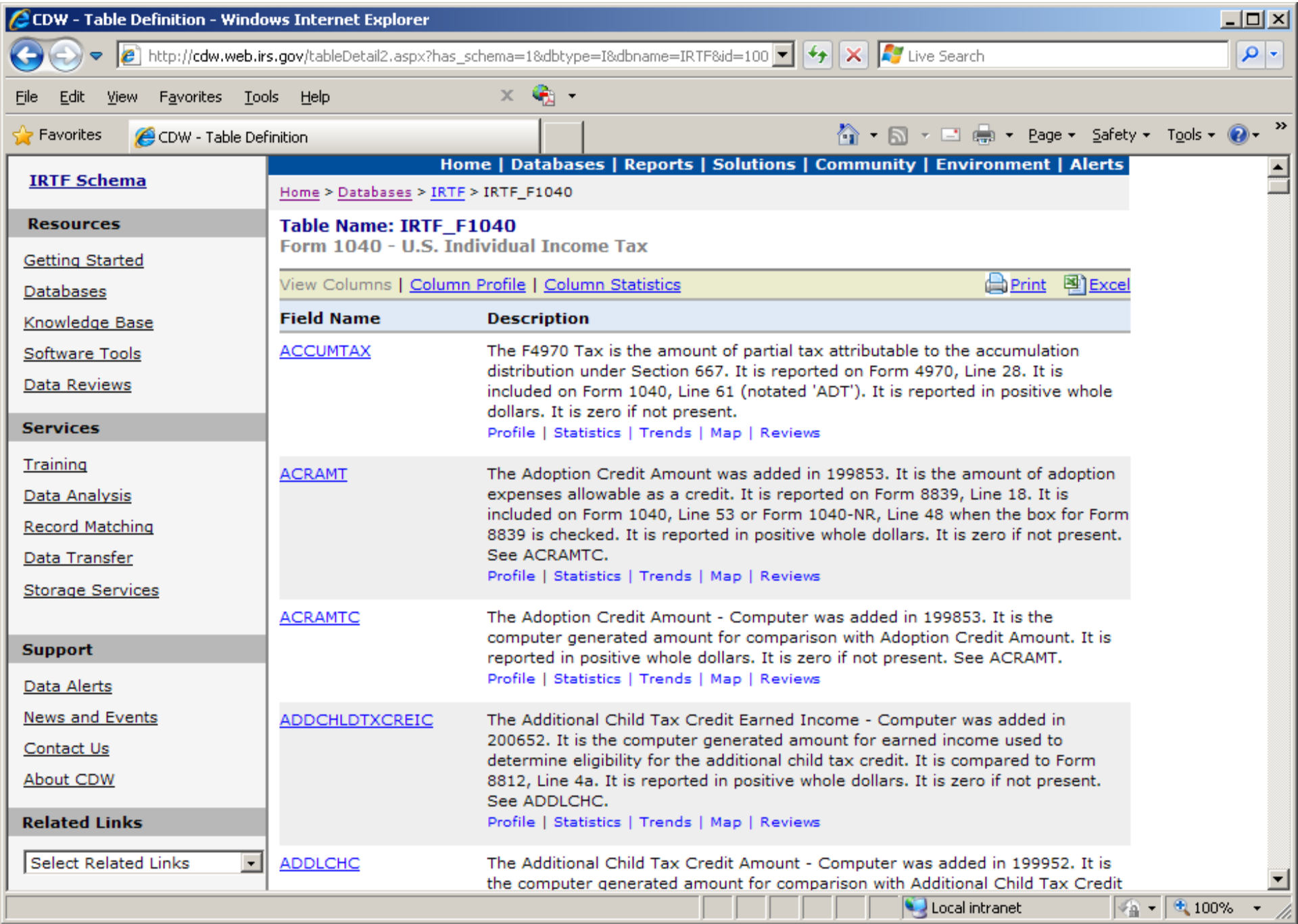
CDW Metadata – Database Level



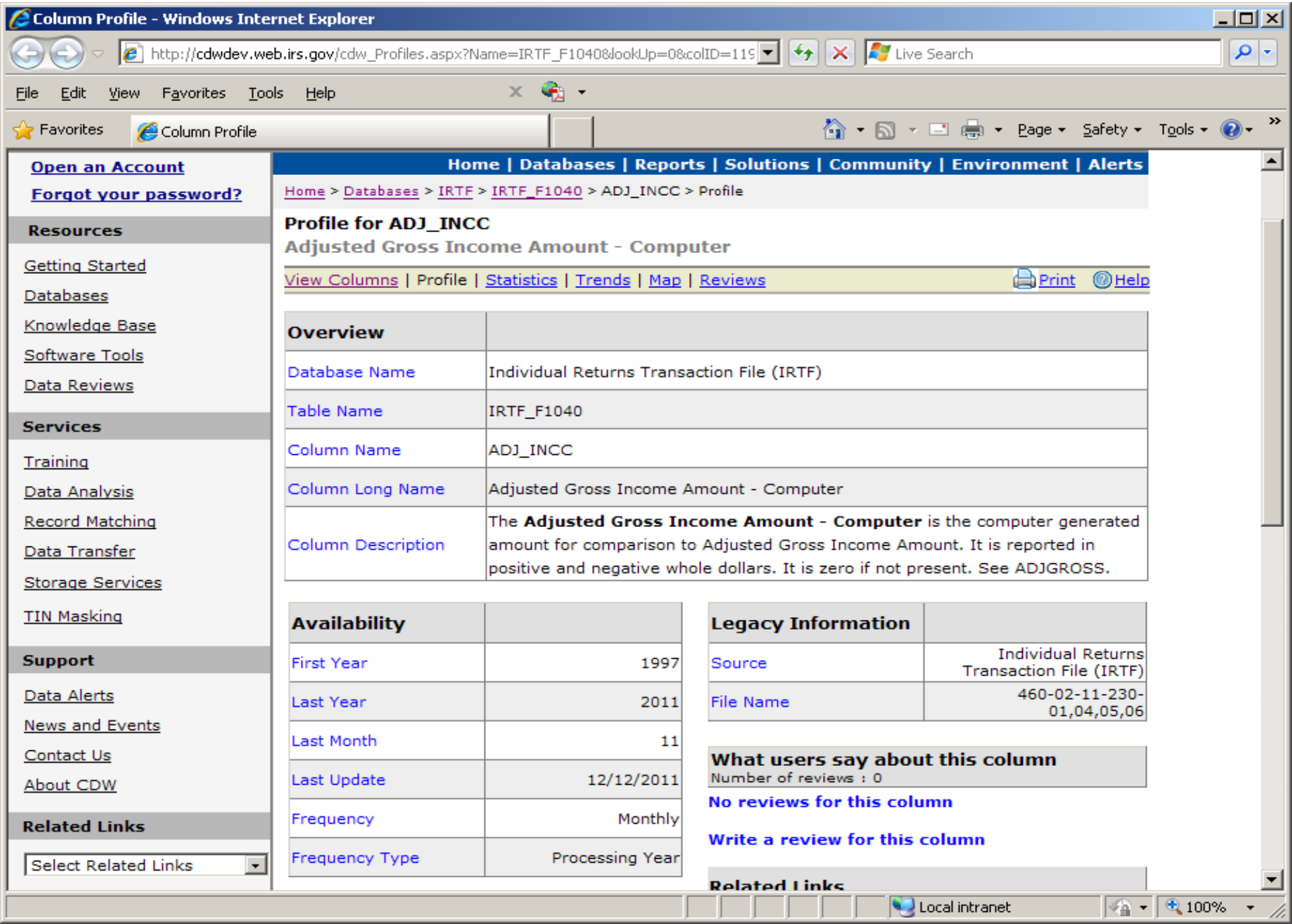
CDW Metadata – Table Level



CDW Metadata – Column Level



CDW Metadata – Column Level

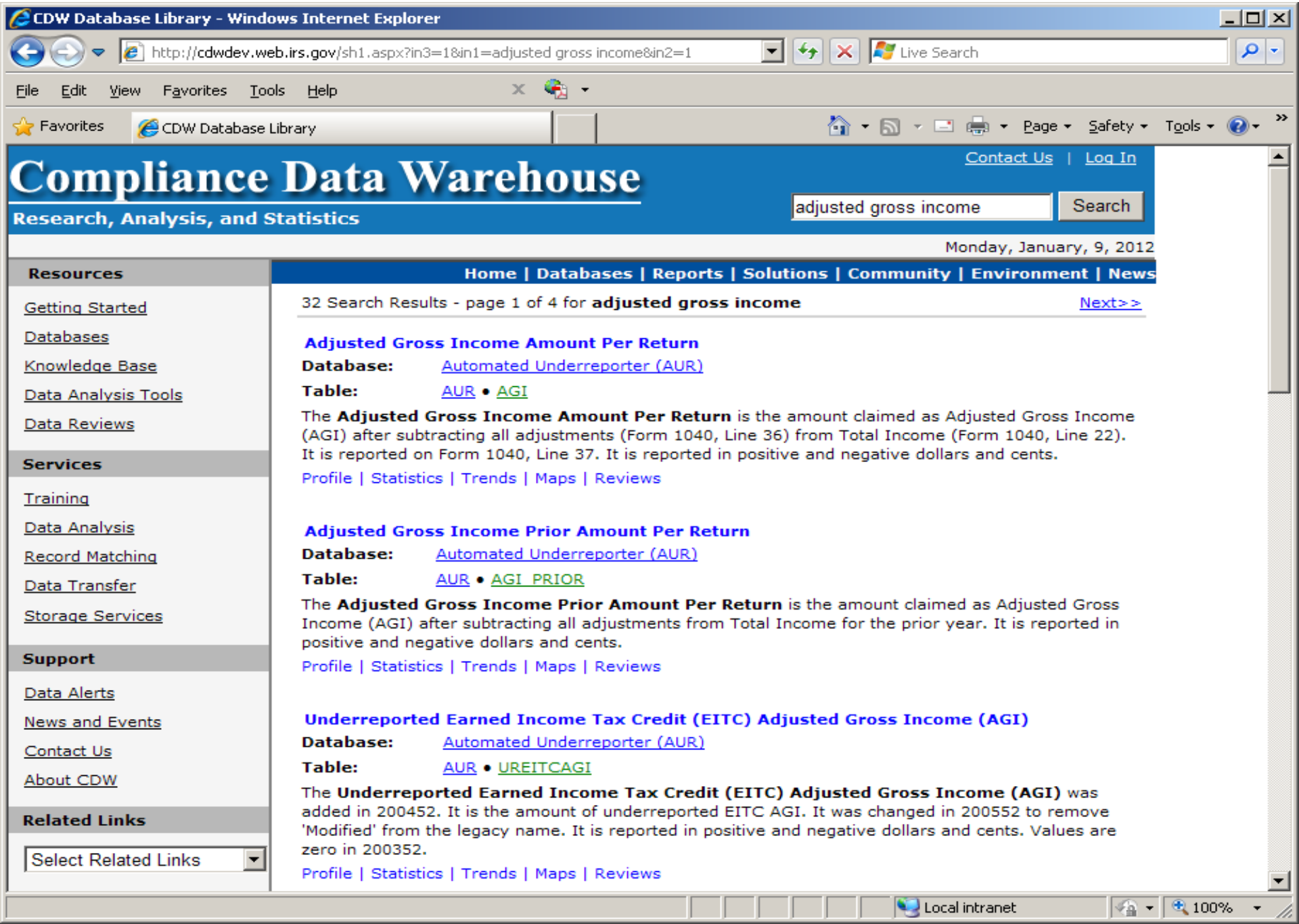


Combining Metadata, Search, and Data Profiling

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

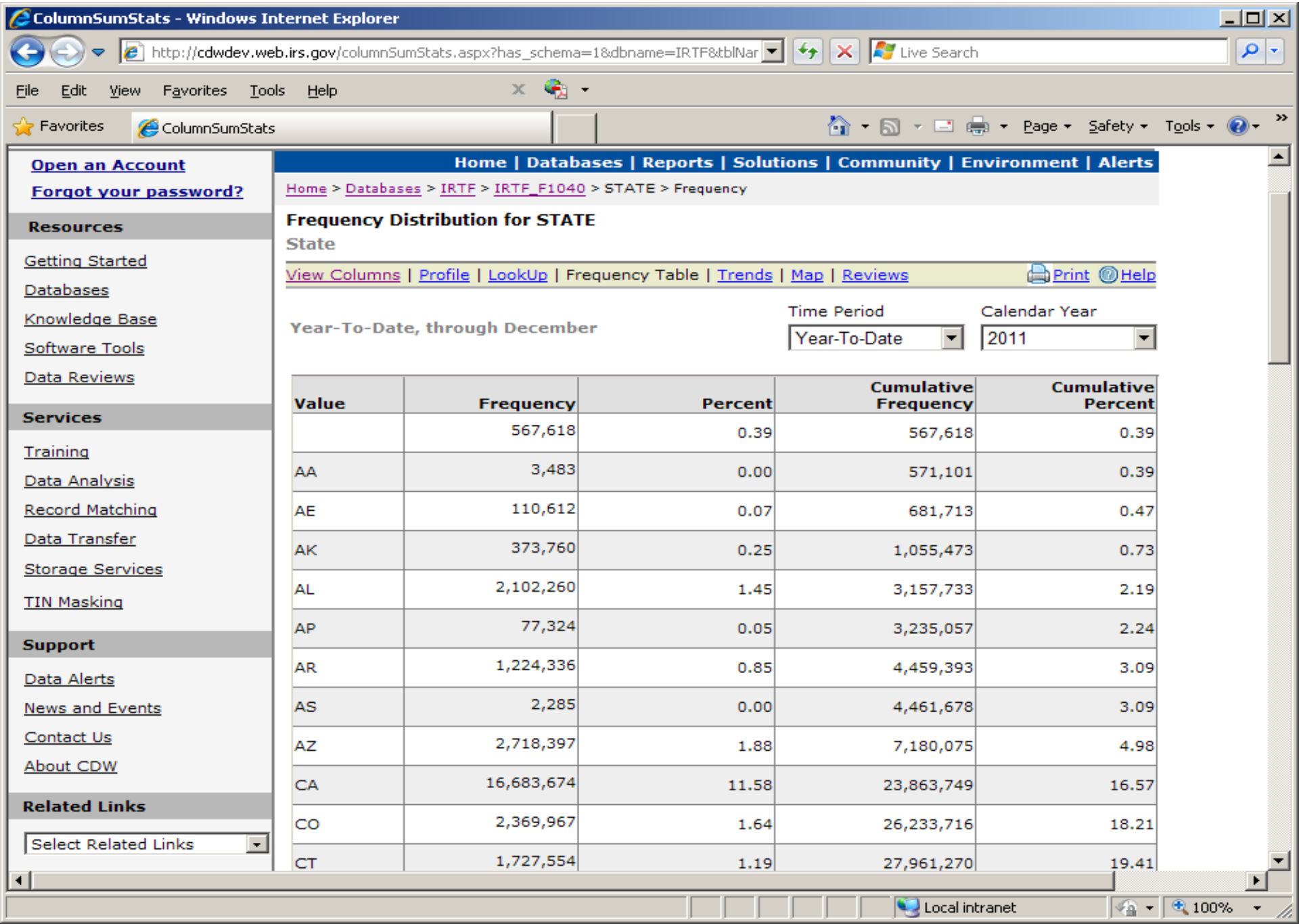
CDW Search Results



Data Profiling

- **Data profiling** is the use of a set of standard functions or rules to expore 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



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

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

Favorites | Column Statistics | Page | Safety | Tools

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 > Statistics

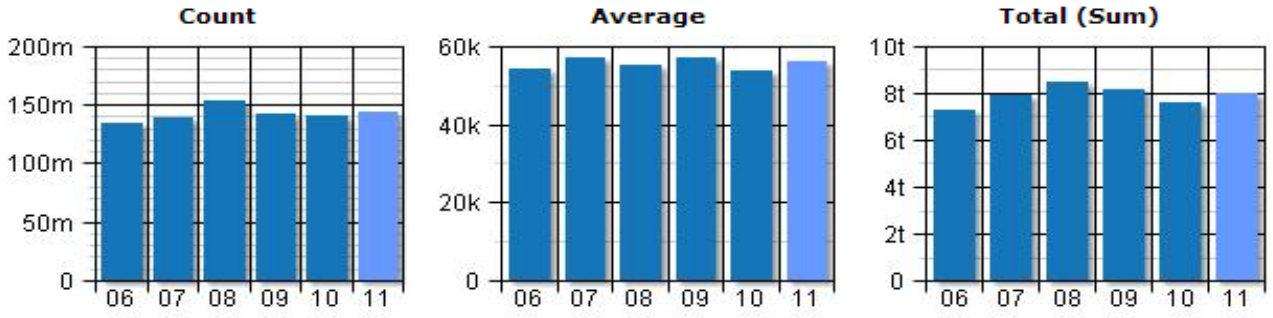
Statistics for ADJ_INCC

Adjusted Gross Income Amount - Computer

View Columns | Profile | Statistics | Trends | Map | Reviews | Help

By Processing Year, Year-To-Date through November

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

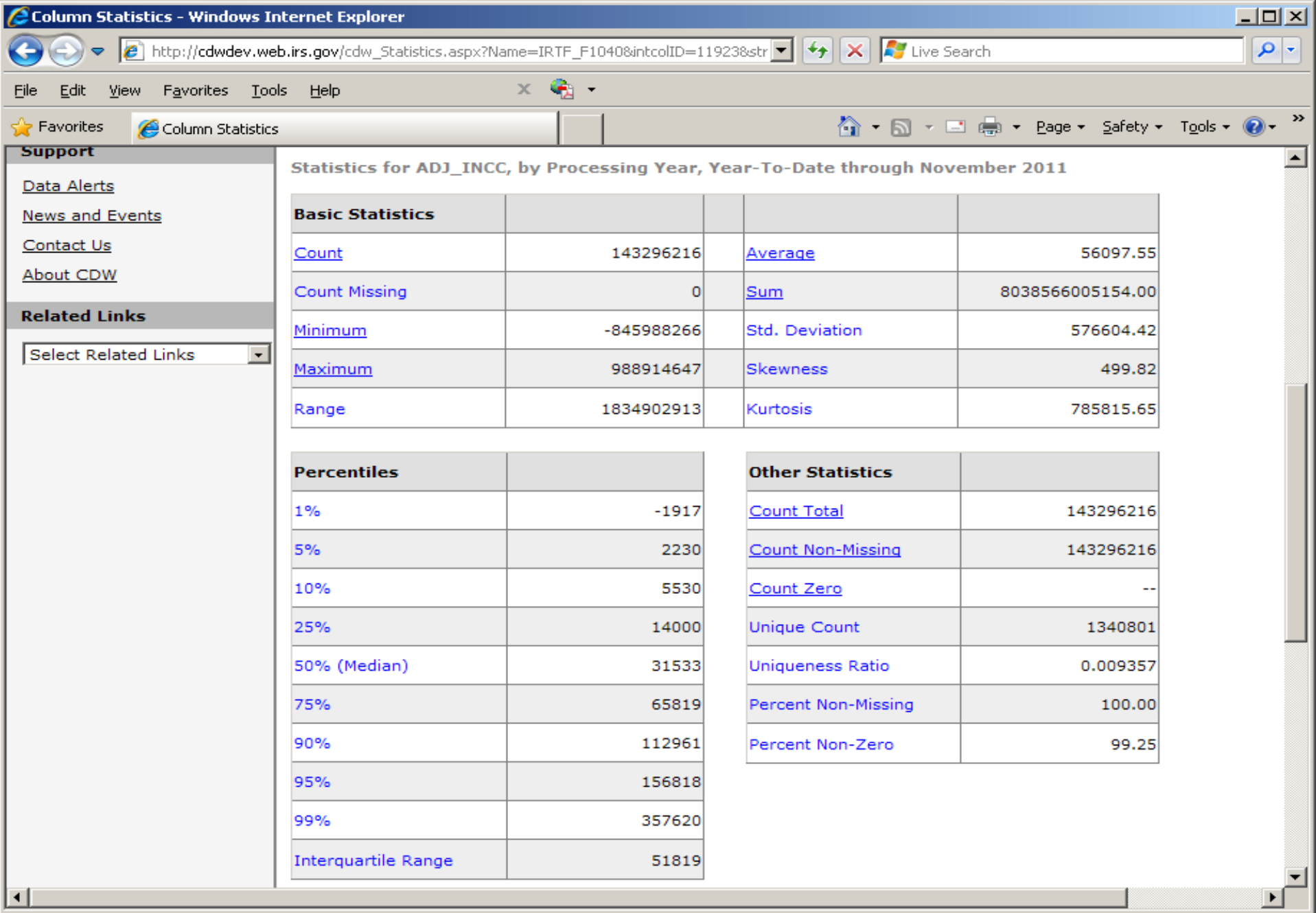


Statistics for ADJ_INCC, by Processing Year, Year-To-Date through November 2011

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

Local intranet | 100%

CDW Data Profiling – Statistical Distributions



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

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

Favorites | Column Trends

Page Safety Tools

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

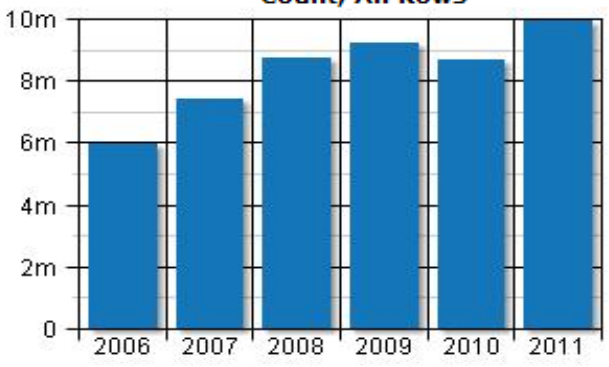
Schedule B Dividends

View Columns | Profile | Statistics | Trends | Map | Reviews Help

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

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

Count, All Rows



Year/Year Change, All Rows

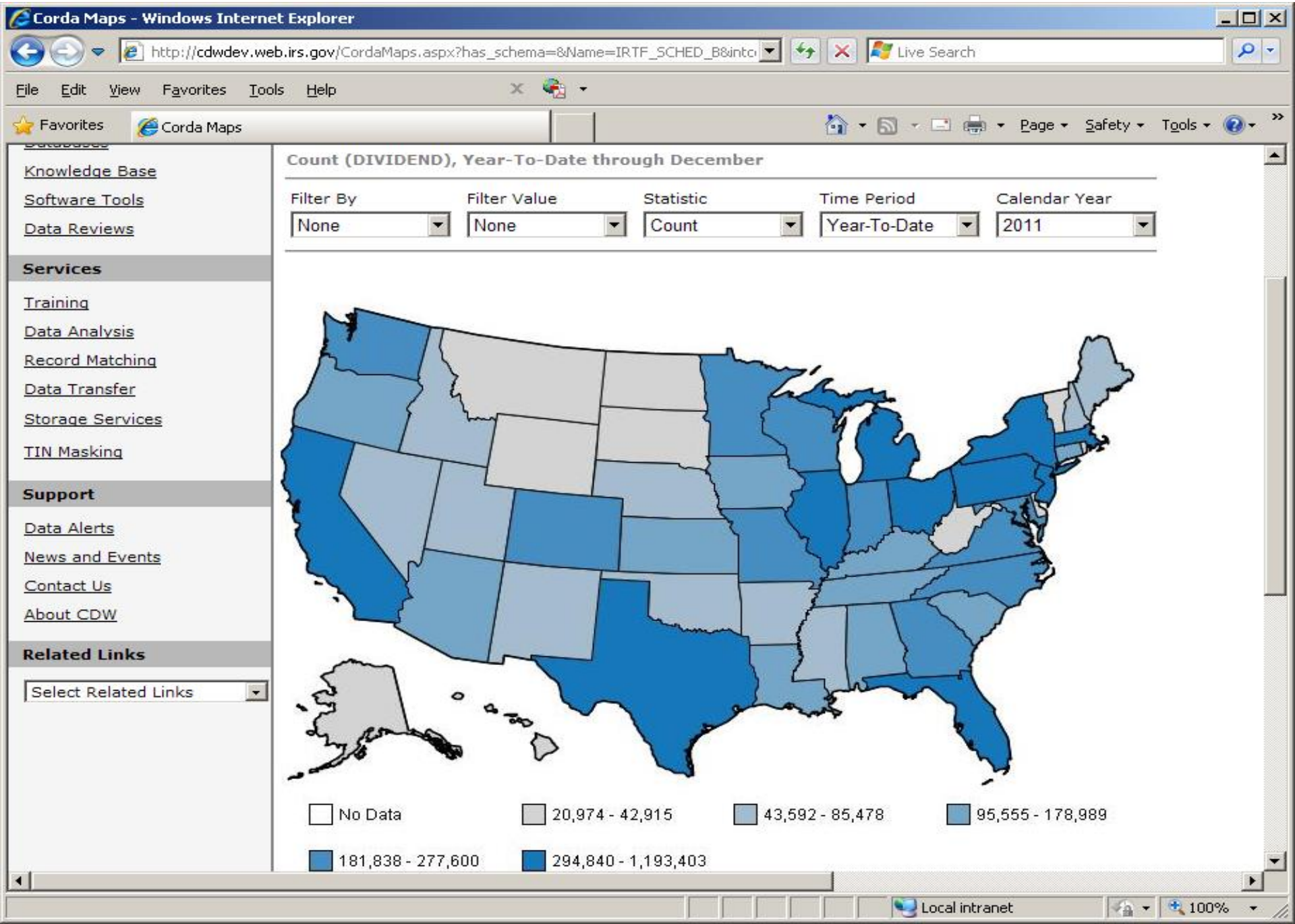
	Count	Change	%Chg
2007	7,428,811	--	--
2008	8,717,895	1,289,084	17.3
2009	9,238,865	520,970	5.9
2010	8,648,427	-590,438	-6.3
2011	9,961,005	1,312,578	15.1

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

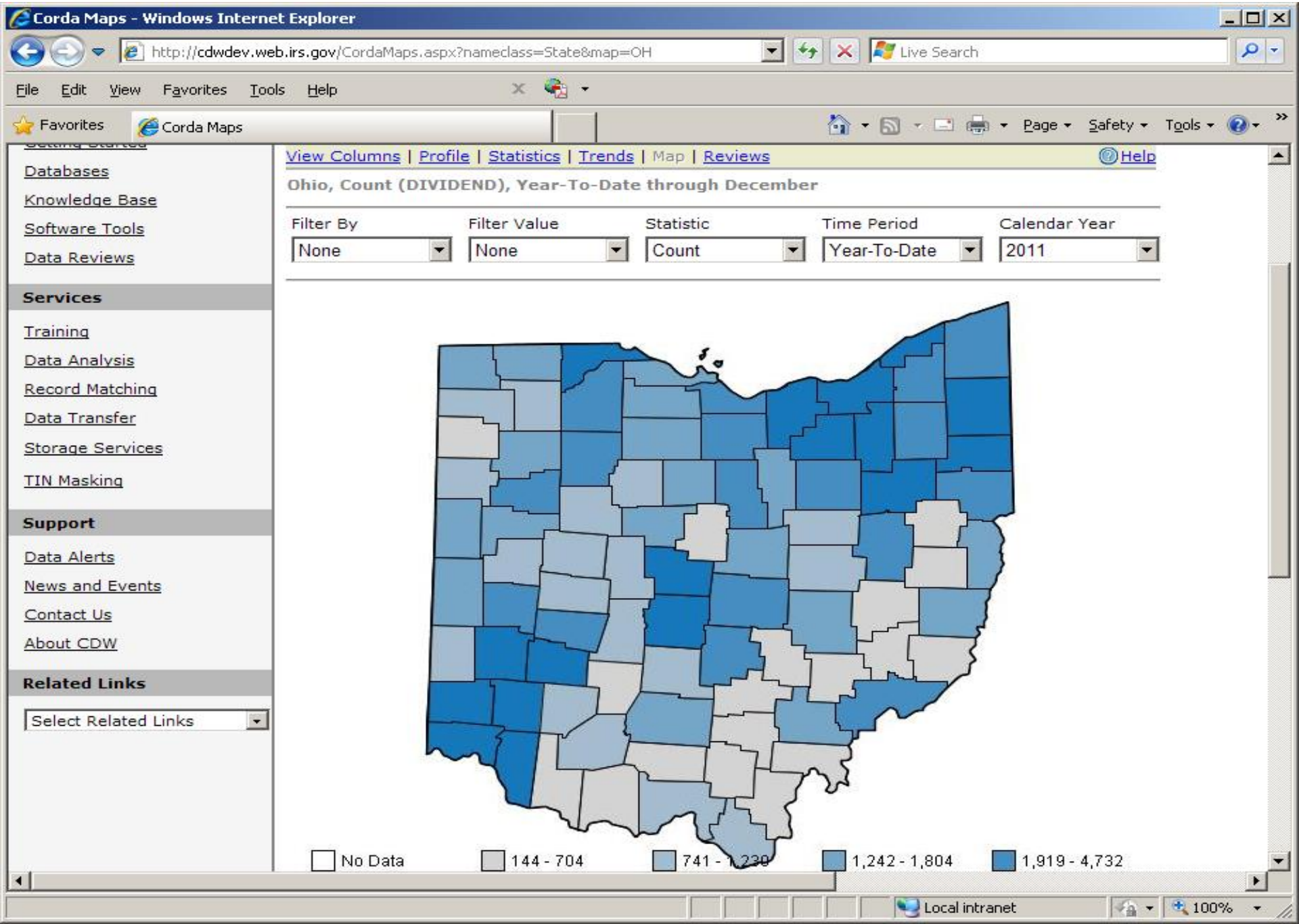
STATE	2009	2010	2011
	11,122	12,915	22,546
AA	294	254	278
AE	5,467	4,644	4,751
AK	29,667	25,164	28,250

Local intranet 100%

CDW Data Profiling – Maps



CDW Data Profiling – Maps



Combining Metadata, Search, and Data Profiling

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” threshhold)
  - 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