The Development of a Generalized Resource Tool for Aggregate Data (GRETA) at the University of Minnesota

A Joint Project of the Machine Readable Data Center and Social Science Research Facility at the University of Minnesota

Wendy L. Treadwell,William C. Block
and John T. Easton

IASSIST - 8 June 2000

Proposal

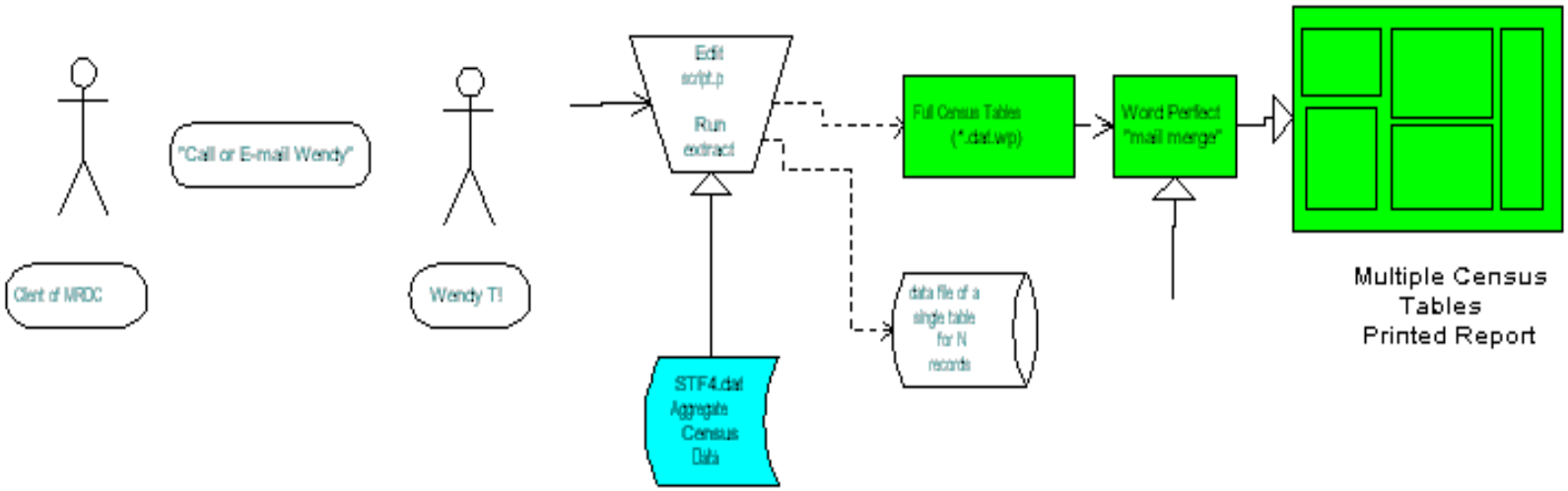
This is a pilot project to create a web-based access tool for the purpose of displaying and extracting data from aggregate data files. The extraction engine will be driven by XML tagged, DDI compliant metadata.

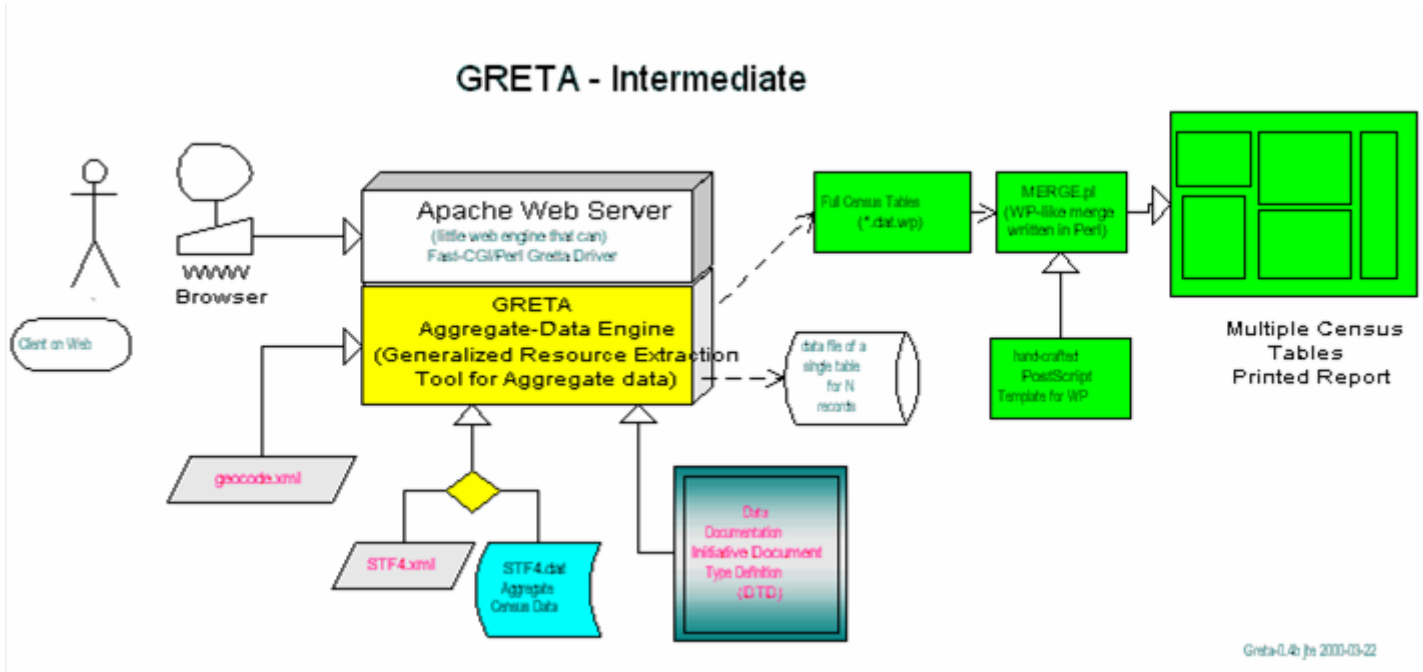
This project is a University of Minnesota Digital Libraries Project funded by the Minnesota Legislature.

Access tools that require specialized programming for each data file or each product mean that:

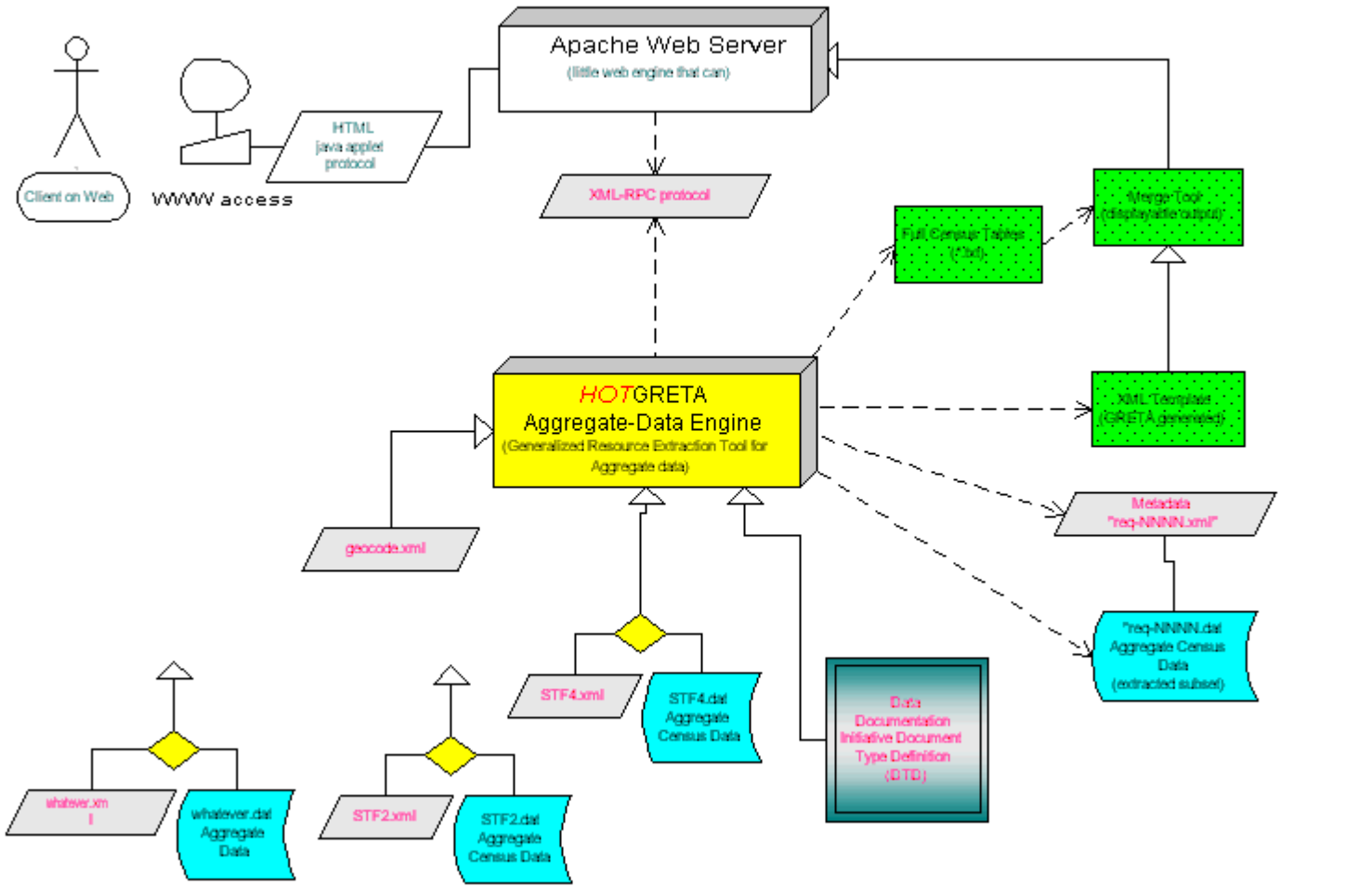
- users get access only to heavily used data sets
- past use drives decisions
- difficult to use or obscure files become lost
- extract is easy but display is harder and often ignored
- each new file requires preparatory work to put it in the required format

Current Method




GRETA - Intermediate
Client on Web
WWW
Browser
Apache Web Server
(little web engine that can)
Fast-CGI/Perl Gretta Driver
GRETA
Aggregate-Data Engine
(Generalized Resource Extraction
Tool for Aggregate data)
geocode.xml
STF4.xml
STF4.dat
Aggregate
Census Data
Data
Documentation
Initiative Document
Type Definition
(DTD)
data file of a
single table
for N
records
Full Census Tables
(*.dat.wp)
MERGE.pl
(WP-like merge
written in Perl)
hand-crafted
PostScript
Template for WP
Multiple Census
Tables
Printed Report
Greta-0.4b jhs 2000-03-22

GRETA - Final System



"Your general appearance is not distasteful." Greta Garbo as Ninotchka in NINOTCHKA (1939)

Generalization

- saves on specialized template development
- runs on archive format
- DDI compliant metadata is the sole requirement
- product-specific search engines are not required
- makes it cost-effective to mount little used but valuable data

What is aggregate data?

- ‘group response’ aggregated for an area or response group
- has meaning in and of itself -- ‘look-up’ data
- frequently forms a matrix or table with related variables

Why additional elements are needed to describe aggregate data

Using current variable and variable group element sets means:

- repetitive coding to capture variable definitions
- loss of relationship between variables
- difficulty in reassembling all or parts of the matrix

"Everybody quarrels. Keeps people from getting bored."
Greta Garbo as Marguerite Gautier in CAMILLE (1937).

Table vs. Matrix

Table:

- graphical representation
- two dimensional - rows and columns
- static

Matrix:

- relational representation
- 1 to n dimensions - vectors and cell coordinates
- can be reorganized without losing relationships

varMtx descriptive levels: MATRIX

151. VACANT-FOR-RENT UNITS BY ASKING RENT AND NUMBER OF BEDROOMS
(Universe: Year-round Vacant-for-rent Units for Which Rent is Tabulated)
Monthly Contract Rent (asking rent) (7) by Number of Bedrooms (5)

Asking less than $40:		
0 bedroom	300	(1,1)
1 bedroom	309	(1,2)
2 bedrooms	318	(1,3)
3 bedrooms	327	(1,4)
4 bedrooms or more	336	(1,5)
$40-$59:		
Repeat Number o f Bedrooms (5)	345	(2,1)
$60-$79:		
Repeat Number o f Bedrooms (5)	390	(3,1)
$80-$99:		
Repeat Number o f Bedrooms (5)	435	(4,1)
$100-$149:		
Repeat Number o f Bedrooms (5)	480	(5,1)
$150-$199:		
Repeat Number o f Bedrooms (5)	525	(6,1)
$200 or more:		
Repeat Number o f Bedrooms (5)	570	(7,1)

```
varMtx*
|--  labl*
|--  txt*
|--  dmnsQnty?
|--  cellQnty?
|--  universe?
|--  imputation?
|--  security?
|--  embargo?
|--  respUnit?
|--  anlysUnit?
|--  verStmt* (with
|       subelements)
|--  concept*
```

varMtx descriptive levels: VECTOR

151. VACANT-FOR-RENT UNITS BY ASKING RENT AND NUMBER OF BEDROOMS
(Universe: Year-round Vacant-for-rent Units for Which Rent is Tabulated)
Monthly Contract Rent (asking rent) (7) by Number of Bedrooms (5)

Asking less than $40:		
0 bedroom	300	(1,1)
1 bedroom	309	(1,2)
2 bedrooms	318	(1,3)
3 bedrooms	327	(1,4)
4 bedrooms or more	336	(1,5)
$40-$59:		
Repeat Number o f Bedrooms (5)	345	(2,1)
$60-$79:		
Repeat Number o f Bedrooms (5)	390	(3,1)
$80-$99:		
Repeat Number o f Bedrooms (5)	435	(4,1)
$100-$149:		
Repeat Number o f Bedrooms (5)	480	(5,1)
$150-$199:		
Repeat Number o f Bedrooms (5)	525	(6,1)
$200 or more:		
Repeat Number o f Bedrooms (5)	570	(7,1)

```
|--  mtxdmns*
|    |--  coord?
|    |--  labl*
|    |--  txt*
|    |--  defntn?
|    |--  cohQnty?
|    |--  qstn*
          (with subelements)
|    |--  derivation?
|    |    |--  drvdesc?
|    |    +--  drvcmd?
```

varMtx descriptive levels: COHORTS

151. VACANT-FOR-RENT UNITS BY ASKING RENT AND NUMBER OF BEDROOMS
(Universe: Year-round Vacant-for-rent Units for Which Rent is Tabulated)
Monthly Contract Rent (asking rent) (7) by Number of Bedrooms (5)

Asking less than $40:		
0 bedroom	300	(1,1)
1 bedroom	309	(1,2)
2 bedrooms	318	(1,3)
3 bedrooms	327	(1,4)
4 bedrooms or more	336	(1,5)
$40-$59:		
Repeat Number o f Bedrooms (5)	345	(2,1)
$60-$79:		
Repeat Number o f Bedrooms (5)	390	(3,1)
$80-$99:		
Repeat Number o f Bedrooms (5)	435	(4,1)
$100-$149:		
Repeat Number o f Bedrooms (5)	480	(5,1)
$150-$199:		
Repeat Number o f Bedrooms (5)	525	(6,1)
$200 or more:		
Repeat Number o f Bedrooms (5)	570	(7,1)

```
|    |--  stdCohort*
|    |    |--  construct?
|    |--  cohortGrp*
|    |    |--  labl*
|    |    +--  txt*
|    +--  cohort*
|         |--  cohValu?
|         |--  labl*
|         +--  txt*
```

varMtx descriptive levels: CELL COORDINATES

151. VACANT-FOR-RENT UNITS BY ASKING RENT AND NUMBER OF BEDROOMS
(Universe: Year-round Vacant-for-rent Units for Which Rent is Tabulated)
Monthly Contract Rent (asking rent) (7) by Number of Bedrooms (5)

Asking less than $40:		
0 bedroom	300	(1,1)
1 bedroom	309	(1,2)
2 bedrooms	318	(1,3)
3 bedrooms	327	(1,4)
4 bedrooms or more	336	(1,5)
$40-$59:		
Repeat Number o f Bedrooms (5)	345	(2,1)
$60-$79:		
Repeat Number o f Bedrooms (5)	390	(3,1)
$80-$99:		
Repeat Number o f Bedrooms (5)	435	(4,1)
$100-$149:		
Repeat Number o f Bedrooms (5)	480	(5,1)
$150-$199:		
Repeat Number o f Bedrooms (5)	525	(6,1)
$200 or more:		
Repeat Number o f Bedrooms (5)	570	(7,1)

```
var* (Add ATTRIBUTE
   varMtx)
|
|--  location*
|--  coordVal*
     (ATT == ID,
      xml:lang,
      source,
      coord)
|--  labl*
|--  (rest of var
     subelements)
```

Implications for Data Access

- Preserve the ‘look-up’ functionality
- Minimize cost of adding new data sets
- Allow for creation of a table template ‘on-the-fly’
- Allow for customization through:
  - rotating the matrix
  - selecting a specific slice of the matrix
  - collapsing cohorts within a vector

Where is GRETA going?

- Possible expansion of IPUMS data collection to include aggregate data using GRETA as the search engine
- Possible integration with other search systems to expand capability to address aggregate data
- Possible project to preserve functional access to U.S. federal depository data

Gif me a visky, ginger ale on the side. Greta Garbo as Anna Christie in ANNA CHRISTIE (1930).

Contacts

http://www.socsci.umn.edu/PDAS/

Wendy Treadwell <wendy@mrdc.lib.umn.edu>

Bill Block <block@socsci.umn.edu>

John Easton <easton@socsci.umn.edu>

photos from:

http://users.deltanet.com/users/dstickne/garbo.htm