

LEHD Restricted-use Data Products: Structure, Access

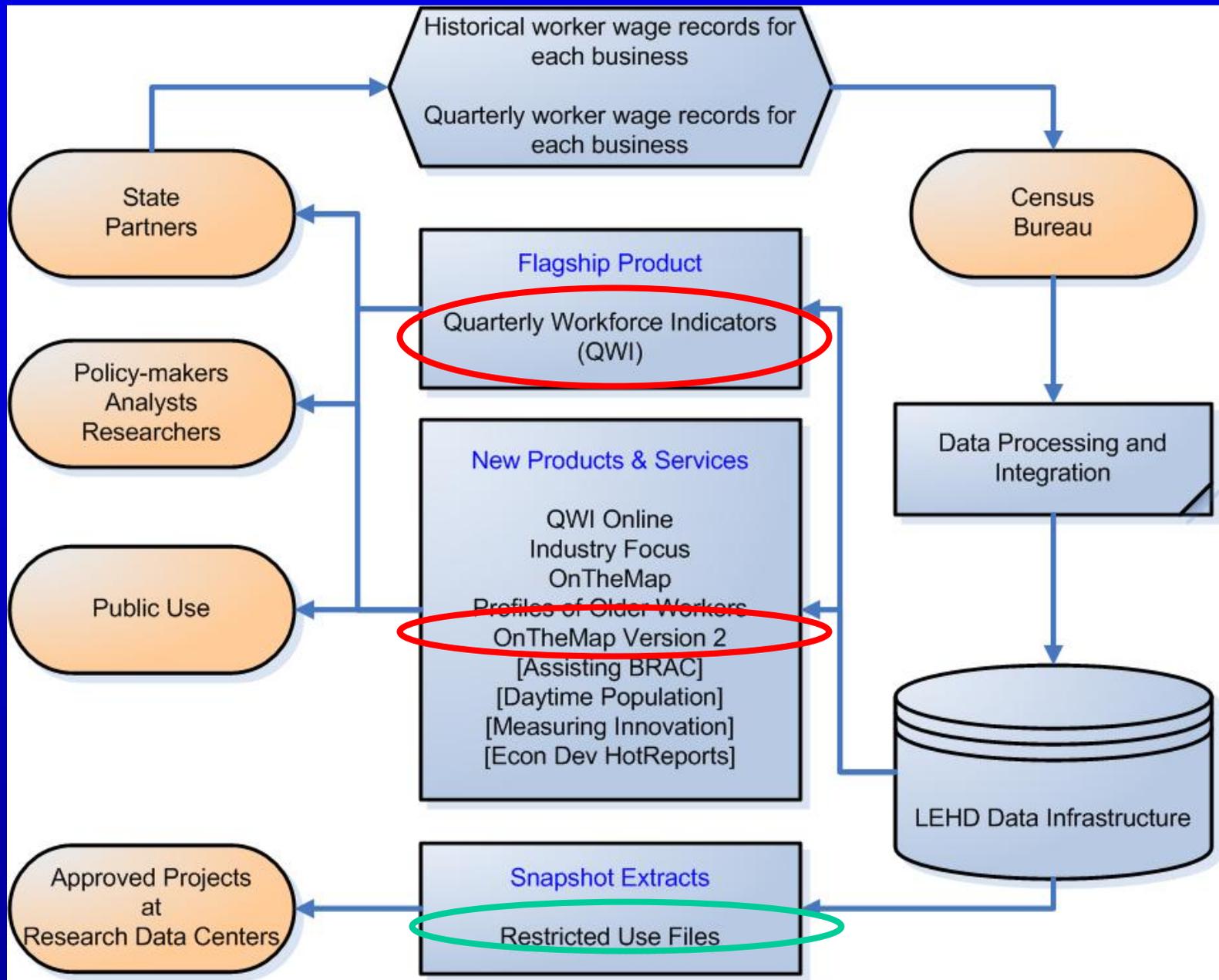
Lars Vilhuber

Cornell University and U.S. Census Bureau

With contributions from Kevin McKinney.

Overview

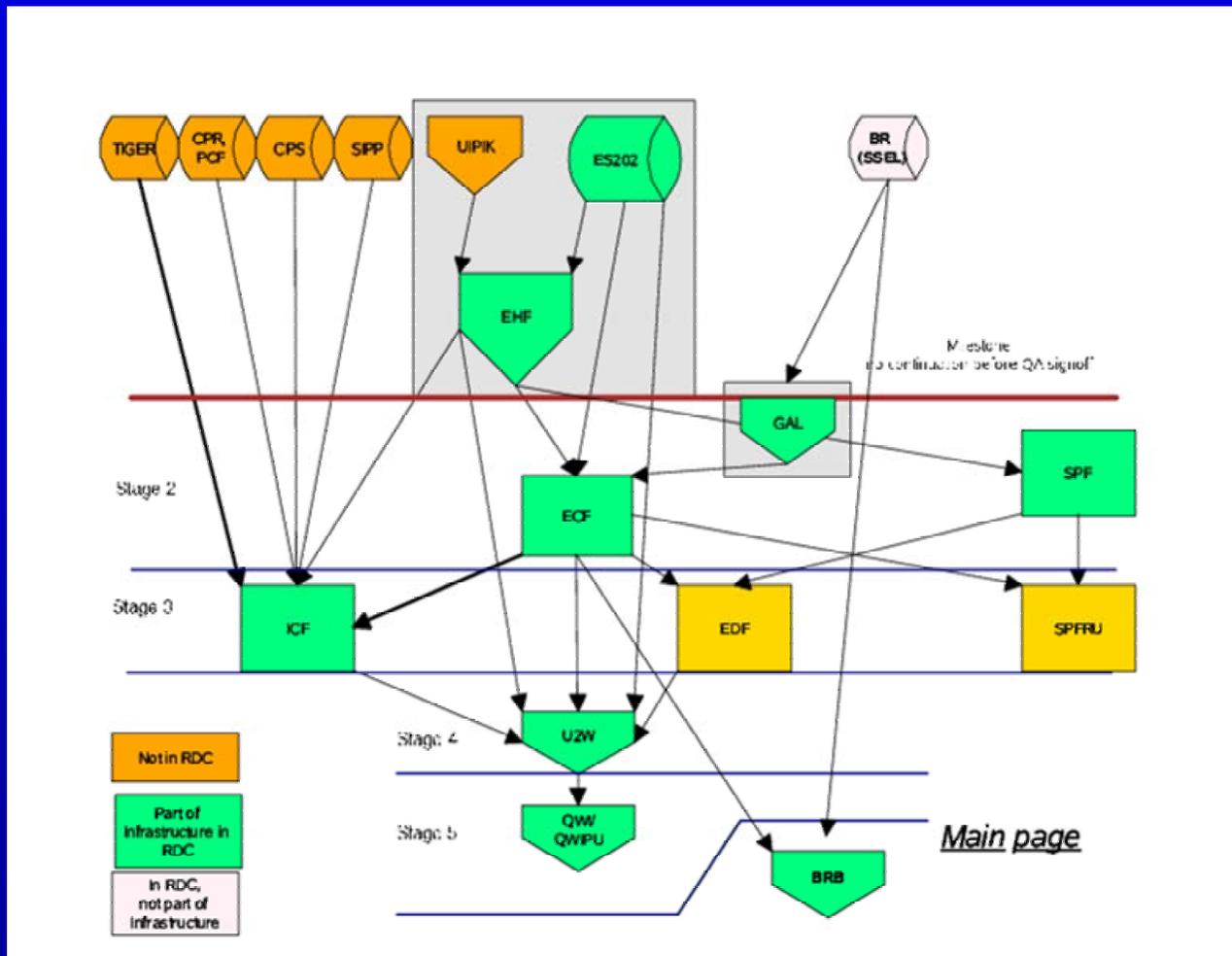
1. Data products
2. Protecting confidentiality
3. QWI: Protection and Analytical Validity
 1. Methods and degree of protection
 2. Analytical validity
4. OTM: Protection and Analytical Validity
 1. Methods
 2. Protection
 3. Analytical validity



The Quarterly Workforce Indicators

- The QWI : a set of 30 time series describing local labor market conditions
- Based on the Longitudinal Employer-Household Dynamics infrastructure file system that integrates
 - Unemployment Insurance wage records,
 - Quarterly Census of Employment and Wages,
 - Census Bureau demographic data
 - Census Bureau economic (business) data

Data flow view of LEHD Infrastructure



Census Research Data Centers

9 Research Data Centers:

- Berkeley, CA
- Los Angeles, CA
- Chicago, IL
- Washington, DC
- New York, NY
- Ithaca, NY
- Durham, NC

Access to Census data upon project approval

- Title 13 benefit to Census Bureau necessary
- Tax data (Title 26) requires additional review by IRS

<http://www.ces.census.gov>

LEHD Infrastructure in the Census Research Data Centers

- Big-picture overview
<http://lehd.dsd.census.gov/led/library/techpapers/tp-2006-01.pdf>.
- no public-use data
- no information related to the disclosure-avoidance measures used in QWI and OTM
- Updated every 2-3 years as a coherent snapshot of the LEHD databases

TREATMENT OF FEDERAL TAX INFORMATION

- Some components have Title-26 protected variables
- Such T26 components need to be requested separately, will trigger additional proposal review

Name	CES abbrev., if different	T26 component	CES abbrev., if different
Business Register Bridge (BRB)		(all)	
Employer Characteristics File (ECF)		ECFT26	ECT
Employment History File (EHF)		none	
ES-202 (QCEW)	ES2	ECFT26	ECT
Individual Characteristics File (ICF)		ICFT26	ICT
Geocoded Address List (GAL)		GALT26	GAT
Quarterly Workforce Indicators (QWI, establishment level files)			
Successor-Predecessor Files (SPF)			
Unit-to-Worker Impute (U2W)			

IDENTIFIERS

- In general, linkages between the different files occur using deterministic match-merge techniques
- Person, firm, and establishment identifiers allow to link all LEHD Infrastructure files among themselves.
- External linkages are generally probabilistic:
 - Linkages to BR (many-to-many match)
 - Linkages to external files by firm name
 - Linkages to external files by establishment location

Individual identifier system (PIK)

- All Social Security Numbers (SSN) have been replaced by Protected Identity Key (PIK)
 - no SSN are available anywhere in this data.
- S2004: 201,317,264 individuals
- ICF contains identifiers linking to other person-level data
 - Current Population Survey (CPS),
 - Survey of Income and Program Participation (SIPP)
 - ACS also feasible using appropriate ACS files with PIK
 - Note that these are generally the Census-internal identifiers and may not have a direct correspondence to external identifiers.

Firm/establishment identifiers

- Firm identifiers are called *State employer identification number* (SEIN) and generally reflect an entity reporting Unemployment insurance data (UI) taxes to state authorities.
 - S2004: 9,869,917 unique firms
- “Establishments” (more precisely: reporting units) are identified by a combination of SEIN and reporting unit (SEINUNIT) (also referred to as SESA)
 - S2004: 11,893,084 unique establishments
- The firm and establishment identifiers are state-specific - within the LEHD Infrastructure, there is no method of linking units of a nation-wide firm across state borders.

Other firm and establishment identifiers

- Federal EIN is available
 - on ECF for most states,
 - on ECFT26 for California
- CFN, Alpha (from Business Register/LDB)
 - accessed using the Business Register Bridge (BRB)

Addresses: GAL

- GAL is a list of all unduplicated addresses from
 - Business Register
 - ACS
 - AHS
 - ES-202
 - Census Master Address File (MAF)
- All addresses from LEHD Infrastructure have been geocoded to the most accurate level possible
 - S2004: 164,928,984 unique addresses
- Cross-walks to all input data sets are available
 - Crosswalk to BR is T26 – separate request necessary

Internal consistency of LEHD Infrastructure

- LEHD Infrastructure is constructed to be internally consistent
 - Firms on EHF = Firms on ECF
(superset of UI and ES202)
 - Individuals on EHF = Individuals on ICF
- ... and complete: all missing data is imputed or edited
 - Addresses (generally at least to block levels, at least to county level)
 - Periodically missing information on firms
 - (research) periodically missing information on individuals/jobs
- Some exceptions to this rule are noted elsewhere

Working with files

- LEHD Infrastructure files are huge when compared to regular research files. In the current version, in all available states and years combined, there are
 - 6,100,912,201 wage records from 201,317,264 individuals in 754,775,697 unique jobs
 - 226,639,116 quarterly observations on 11,893,084 establishments from 9,869,917 firms.
 - Total size of all datasets is about 1.5TB
- Careful planning is required to ensure that adequate resources are available.

Facilitating researcher access

- The LEHD/QWI production system always processes all records – no use for subsetting variables
- The research versions of the LEHD Infrastructure files in the RDC environment have additional random variables that allow for the selection of uniform random subsamples of
 - firms (SEIN),
 - establishments (SEINUNIT), and
 - individuals (PIK).
- No such random variable is available on the EHF, since there is no single good strategy for selecting jobs.

SPECIAL DISCLOSURE RULES

- Confidentiality protection for the QWI uses noise infusion of the micro-data.
- Disclosure Review Board (DRB) does not allow the release of tabulations for sub-state geography that do not use the QWI noise infusion process.
 - In addition, the required noise factors have not been placed on the RDC snapshot files as part of the DRB's normal rules limiting access to the specific parameters of its approved disclosure limitation methods.
 - Consequence: Sub-state geography tables will not be approved.
 - National or multi-state tables may be approved (by DRB) provided they do not compromise the protection system.
- Model-based output is normally allowed.
- The chief disclosure officer for the RDC network will coordinate the reviews.

SPECIAL DATA USE RULES

- The underlying micro-data in the LEHD infrastructure file system were provided to the Census Bureau by state Labor Market Information (LMI) offices under Memoranda of Understanding
 - Analysis of a single state's data in identifiable form requires the permission of the state's LMI officer.
 - Current members of the LED partnership are shown on the LEHD main web page.
 - When reporting results from studies that include multiple states, the results should be pooled across the states.
- The identity of the LED member states is obviously not confidential. You may say which states were used in your analysis.
- The chief disclosure officer for the RDC network will review compliance with this requirement in consultation with the Assistant Division Chief for LEHD

Contact LEHD

Quarterly Workforce Indicators:

<http://lehd.dsd.census.gov/led/datatools/qwiapp.html>

OnTheMap

<http://lehdmap.dsd.census.gov>

LEHD Program/Technical Documentation:

<http://lehd.dsd.census.gov>

Assistant Division Chief, LEHD:

Jeremy.S.Wu@census.gov