Software Development

Steven Ruggles

Minnesota Population Center
MPC

Funded by the National Science Foundation

Why focus on software?

- Important component of project
- NSF expressed concern about this aspect in particular
  - 30K rebudget for expert to evaluate our software development
  - Main goal is to figure out how we should proceed

Background
Software Development at MPC, 1990-1997

- Main projects: IPUMS-USA and data collection projects (samples of 1850-1880, 1920)
  - Data entry software (multiple systems)
  - Data cleaning software (Fortran)
  - Recoding and reformatting (Fortran)
  - Editing and allocation (Fortran)
  - Data Access (Perl, Fortran)
- Programming staff: Ruggles and Todd Gardner, plus the occasional student
- Todd left for Census in 1998.

Software Development at MPC, 1998-2004

– Many new projects: NHGIS, NAPP, complete-count data and high-density samples (1880, 1900, 1930), IPUMS-International, IHIS, Record-Linkage project, Data Sharing Initiative
  - New data sources: International files, data produced by genealogists, and aggregate data files as well as MPC-produced historical data
  - Software increasingly driven my machine-understandable metadata
  - Development of tools to automate many tasks once carried out by hand to accommodate increased scale of work
  - Development of tools to create and verify metadata
  - Much more complex web applications
  - Major long-term projects: replace all Fortran and Perl, develop aggregate data access and mapping tool, develop new generation of data access tools for microdata

Growth of Staff:
MPC Software Development Staff, 2004

Director
William C. Block, Ph.D.

Full-time developers:
- Colin Davis
- Duy Duong
- Cuong Nguyen
- Benjamin Ortega
- Marcus Peterson
- James Shelburne
- Robert Wozniak

Graphic/Web Designer:
Charlot Meyer

Student programmers:
- Tony Collen
- Peter Giencke
- Tony Jiang
- Vladimir Vladykin

Programming pool concept: staff assigned to projects as needed depending on their specific skills (e.g. web design, XML, database management, etc.)—most programmers work on multiple projects over the course of a year

Redesign of Software Development Process

- In August 2003, internal review found that that although routine applications programming was proceeding smoothly, the major long-term development projects were going too slowly
- We decided to bring in an outside expert consultant to evaluate our procedures and develop recommendations for change
- Fortunately, I knew just the person:

Catherine Ruggles, Chief Technology Officer

A seasoned technical executive with extensive experience in the software industry, Ms. Ruggles is responsible for setting the technical vision, strategy, and product direction at WebPutty.

Ms. Ruggles joined WebPutty from Enfish Technology, an Internet startup company focused on delivering personal information management products and services. During her tenure at Enfish, Ms. Ruggles held positions as Chief Technical Officer and Senior Vice President of Software Development. Ms. Ruggles's primary responsibilities included initiating and implementing the standard software development and lifecycle process, and building a strong development team. Prior to joining Enfish, Ms. Ruggles held senior management positions at Symantec, Peter Norton Product Group, Prospect Research Corporation, Sofistry, Transaction Technology, and Bradford National Corporation.

Jo’s Mission:

Evaluate all aspects of MPC software development to improve productivity and software quality, especially project management and software design process

Innovations of the past 18 months

1. Process and policies
   - Evaluation of architecture of all major systems, particular attention to scalability and maintainability
   - Shortened development cycles
   - Achievable and trackable work scoping and milestones
   - Systematic code review
   - Standardization of languages (leaving behind legacy Perl, PHP, Fortran; ultimate goal of standardizing on Java)

Innovations of the past 18 months

2. Tools

- Integrated development environment (IntelliJ)
- Debugger (IntelliJ)
- Unit Tests (JUnit)
- Clover (code coverage tool for testing amount of code that is actually used; a test for Unit Tests)
- Automated builds (ANT)
- CVS (source control)
- Request Tracker (RT), preventing maintenance tasks from interfering with new development

Results

- Unprecedented progress on all long-term software development projects
- IPUMS-International had two major accomplishments
  - Java-based front end for data extracts is in alpha; replaces PHP stopgap system developed in 2002 (as well as the 1996 Perl script still in use for IPUMS-USA)
  - Data editing and missing data allocation software is rewritten, now driven by scripts that can be tuned by analyst

New Senior Search Underway

Responsibilities:

Strategic planning, designing, architecting, analyzing, mentoring and supervising others in writing, testing, and improving computer programs for data extraction and data management.

Researching, testing, and recommending appropriate languages, libraries, and technologies for MPC applications, and ensuring documentation of all code development.

Preferred Qualifications:

- Proven ability and experience with a large-scale software project in a team setting, using source control, bug tracking, automated build, and project management tools.
- Experience architecting web solutions requiring scalability and reliability. Proven experience and skills to create work plans, manage deadlines, and prioritize duties on multiple simultaneous projects.

- No specific salary cap
- 100+ applications in past 10 days

Software development goals

1. Complete tasks envisioned in our original ten-year, $10 million, plan for IPUMS-International
   - Advanced data extract features
     - attaching characteristics of householder, family”head,” own spouse, own mother, own father
     - aggregating across households, families, and sibling sets
   - Develop dynamic documentation browser
   - Develop tools needed to accommodate more censuses and speed processing

Software development goals

2. Add capabilities introduced in current proposal

- Move to new XML metadata standards (DDI)
- Add on-line data analysis
- Speed up data extraction, either by moving to an inverted matrix data format or through other means
- Add capability of attaching contextual characteristics to microdata
- Improve tools for navigating documentation
- Additional data extract features
  - Add capability to replicate past extracts used in publications
  - On-the-fly recoding to obtain common universe
  - excluding imputed values

New Metadata is Key

- Current IPUMS metadata is thousands of static html pages, plus hundreds of tables that drive data conversion and data access software
- Unwieldy and difficult to maintain; even minor changes to variables sometimes require changes to many files
- We are shifting to metadata loosely based on the Data Documentation Initiative XML metadata standard
- Goals are easier maintenance and more dynamic documentation customized to a particular users needs

What we would do with more IT resources

(Dan Newlon asked me to address this question)

1. Improve Online Analysis
   - SDA system: limitations
   - SDA would collaborate on a better-suited system
   - Large potential audience

2. Metadata comparability standards
   - DDI dies not address cross database comparability
   - Can it be modified to address this important need?
   - Do we need to develop a new metadata framework?

End.