GCN
gcn.com
6.5.06
The Technology Publication for Government
PostNewsweek Tech Media
Government Computer News

Fuzzy matching

Even data sharing languages leave room for interpretation

YOU'RE WORKING on a data-sharing project and have established a vocabulary. Great. But even after the terms are agreed upon, agencies must work to make sure all variations in a language are covered.

Take the IRS, for instance. The IRS runs a name-matching service available to all offices within the agency. Given a name and street address, the Name Search Project can quickly return a Social Security Number or an Employer Identification Number, as well as other information, such as birth date and previous tax returns. It's used across multiple departments, for multiple purposes. Any IRS agent who assigns taxpayer identification numbers checks with the system to make sure that person doesn't already have an identifier. Other employees use the system to check addresses or conduct examinations, collections and criminal investigations.

Offering such a data-sharing service, which runs on an IBM 3090 mainframe computer, is a challenge, given that the agency has over 700 million taxpayer identification numbers, said Winnie Wilkinson, lead information technology specialist for the service. But the biggest challenge has been dealing with variations in the spelling of names and street addresses.

BIG NUMBERS: Various spellings for 700 million taxpayer IDs make sharing data a challenge for the IRS. — Winnie Wilkinson, IRS

Before the current service was implemented, IRS employees queried an index file that could only produce exact matches, a trait that would produce maddeningly inconsistent results. "J. Doe" at "123 Main St." would not return "John Doe" at "123 Main Street." Nor would "Ajax Widget Co." bring back any information about "Ajax Widget Company."

To get past this and build some flexibility into the system, the agency adopted software from Identity Systems, a division of Nokia Corp. of Finland. The SSA-Name3 software employs fuzzy-logic searching, which lets users obtain results from nicknames, phonetically similar names and slight misspellings. The FBI also uses the matching software.

Eventually the program will expand its scope beyond user searches to offer an automated query interface that can be deployed by other systems. For instance, other offices have shown an interest in submitting batch jobs. The name-matching service could ingest a list of names and return information for each individual, freeing IRS officers from spending long periods entering names individually. **—Joab Jackson**