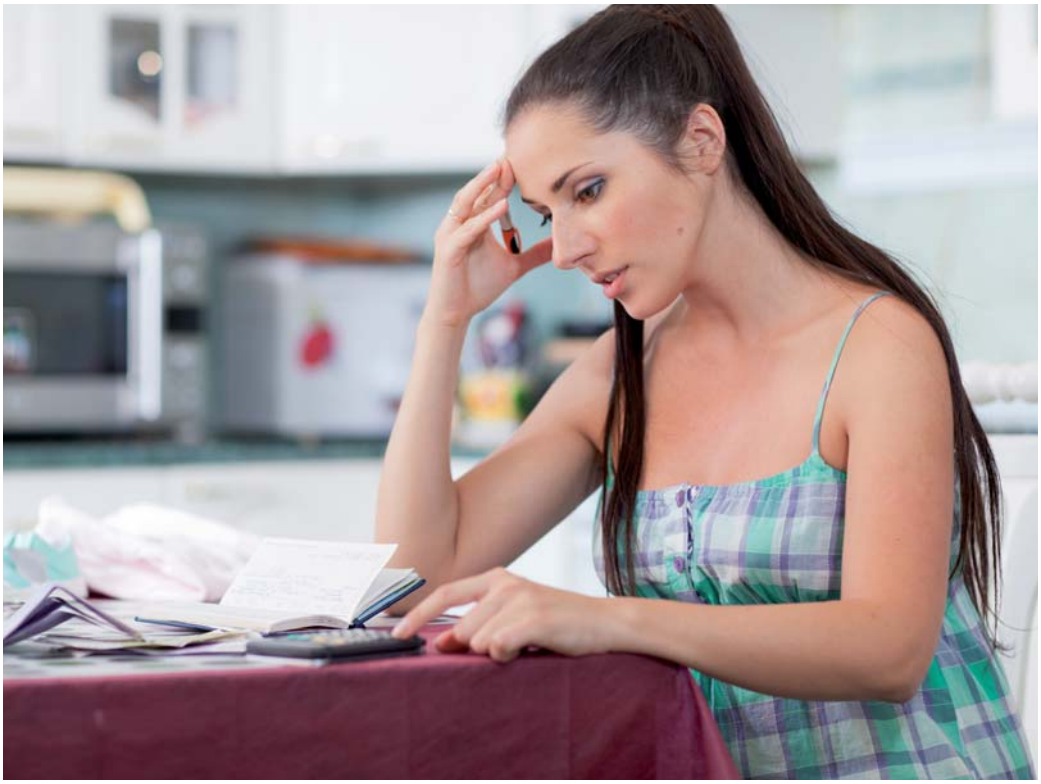
Odesláním formulářů sčítání neskončilo, ba právě naopak

Sebráním vyplněných sčítacích formulářů **skončila nejviditelnější část sčítání**. Další etapy sčítání již nebyly veřejností tolik sledovány, přestože byly stejně důležité. Co se tedy dál odehrávalo na poli sčítání?

Okamžitě po sběru se rozběhla první fáze zpracování výsledků – vytváření vstupních datových souborů. Zatímco údaje z elektronických formulářů byly celkem jednoduše přeneseny do databáze, ty papírové se musely napřed naskenovat (do loňského července jich bylo 12, 9 milionů) a validovat.



Na rozdíl od sčítání v roce 2001 bude potřeba **přiřadit jednotlivé osoby do příslušných domácností** a bytů a byty do domů.

Proč validovat?

Přestože skenery byly výkonné a většinu znaků na formulářích bezpečně rozpoznaly, zůstalo i tak velké množství případů, kdy přístroj nerozpoznal číslici nebo písmeno. V těchto případech bylo nutné, aby nastoupili validátoři a posoudili, co bylo třeba. Nebyla to jednoduchá práce. Nejenom, že probíhala od poloviny dubna až do začátku srpna, ale ve špičce se na ní pracovalo na tři směny.

Pořízená data z elektronických i naskenovaných a zvalidovaných papírových formulářů dále procházela procesem kódování, který automaticky převáděl jak slovní odpovědi, tak i odpovědi zaznamenané křížkem do odpovídajících kódů. Například u adres, oborů vzdělání nebo zaměstnání byla rozmanitost zápisů natolik široká, že k automatickému zakódování nestačily ani rozsáhlé slovníky. Proto museli vyškolení operátoři ručně kódovat značnou část záznamů (5 milionů z 30 milionů). „Někdy to bylo hodně náročné," říká Radka Topinková, jedna z operátorek, která na sčítání pracovala po celou dobu kódování, tj. od konce června do 4. října.

Po skončení kódování proběhla etapa kontrol přípustnosti. V databázi se testovalo, zda se tam vyskytují pouze hodnoty, které jsou přípustné a reálné – např. pouze jeden záznam o pohlaví osoby, měsíc narození osoby pouze v intervalu 1–12 apod.

Dalším krokem zpracování – ve druhé polovině října – bylo porovnání všech vět v databázi a identifikace případných duplicitních záznamů. „Předpokládali jsme, že stejně jako v minulých sčítáních i při sčítání 2011 se někdo sečetl nebo byl sečten dvakrát. Například vysokoškolský student vyplnil sčítací formulář na koleji, kde bydlí, ale jeho rodiče za něj vyplnili formulář i doma, kde je přihlášen k trvalému pobytu," vysvětluje Josef Škrabal, ředitel odboru statis-

Předběžné výsledky

Pro přípravu předběžných výsledků proběhlo zpracování omezeného **množství logických kontrol**. Ty definují logické vazby mezi statistickými proměnnými a obsahují algoritmus pro opravu chybných vazeb.

Na konci listopadu 2011 proběhl podobně omezený proces odvozování ukazatelů, který spočíval v plně automatizovaném výpočtu mikrodat jednotlivých entit pomocí aplikačního vybavení. Po následném výpočtu agregací byla data přenesena do veřejné databáze ČSÚ a **připravena pro publikování**.

tiky obyvatelstva v Českém statistickém úřadu.

Při nalezení duplicitních záznamů se také vyhodnocovala úroveň vyplnění jednotlivých otázek, aby zneplatněním duplicitních záznamů nedocházelo ke ztrátě informací. „V této fázi jsme rovněž již využívali evidenci obyvatel, ze které jsme k existujícím záznamům přebírali některé další údaje, které jsme nezjišťovali dotazem na formuláři, protože jsou v evidenci k dispozici. Zejména adresu trvalého bydliště, ale i rok příjezdu do země u cizinců a podobně," dodává odborník.

Definitivní výsledky

Budou publikovány ve druhém pololetí 2012. Procházejí a budou procházet ještě složitějším procesem zpracování. Vzhledem k tomu, že budou podle požadavku Eurostatu publikovány podle obvyklého pobytu respondentů, je nutné provést podle speciálního algoritmu odvození tohoto ukazatele – adresy obvyklého pobytu. Potom bude následovat jedna z nejsložitějších etap zpracování definitivních výsledků – tzv. datová burza.

Na rozdíl od sčítání v roce 2001 bude potřeba přiřadit jednotlivé osoby do příslušných domácností a bytů a byty do domů. V předchozích sčítáních, kdy soubory formulářů kompletovali sčítací komisaři za svůj sčítací obvod, byly vyplněné formuláře za domácnosti a byty již před pořízením dat pěkně pohromadě.

> *Ve špičce se pracovalo na validaci sčítacích formulářů na tři směny.*

Variabilita způsobů návratnosti formulářů při sčítání 2011 způsobila, že každý člen domácnosti mohl svůj formulář odevzdat jiným způsobem. Kompletace dat za jednu domácnost, jeden byt nebo jeden dům probíhá proto až při zpracování. I tato etapa má část automatickou a manuální.

Součástí zpracování definitivních výsledků je rovněž etapa logických kontrol (kontrola logických vazeb mezi ukazateli) a následně fáze anonymizace. Při té se odstraní z datových souborů osob atributy den a měsíc narození, rodné číslo, jméno a příjmení. V dalším zpracování se budou používat pouze anonymizované údaje, kde už nebude možné konkrétní osobu dohledat.

Štěpánka Morávková
oddělení metodiky, analýz a diseminace sčítání

Naplnění databáze

Zatímco předchozí činnosti probíhaly na zabezpečeném režimovém pracovišti ČSÚ, závěrečná etapa zpracování definitivních výsledků ze sčítání už bude probíhat přímo v ústředí. Jejím základem bude odvození široké škály ukazatelů – tedy naplnění databáze mikrodaty za domy, byty a osoby, ale rovněž výpočet zcela nových údajů, zejména odvozování domácností a jejich charakteristik, výpočet dojížďkových proudů a dalších ukazatelů. Poté budou postupně zveřejňovány definitivní výsledky.

Obyvatelé Prahy dali přednost internetu

Lidé mohli poprvé při loňském sčítání využít **vyplnění elektronických formulářů** prostřednictvím internetu. ČSÚ tak získal 4,33 milionů sčítacích formulářů (25,5 %).

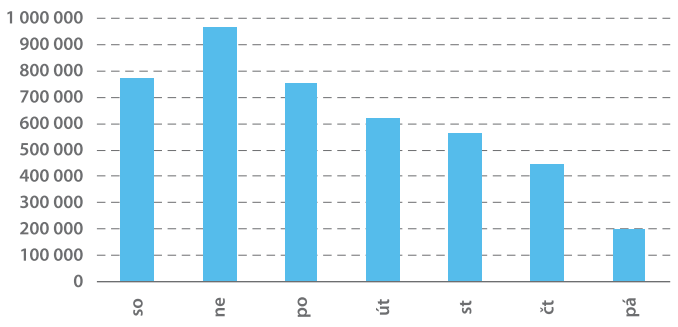
Elektronicky se sčítaly především osoby mladší (30–39 let) se středním nebo úplným středním odborným vzděláním, ekonomicky aktivní, které bydlely v domácnosti vybavené počítačem s připojením k internetu. Vyplnění jednoho formuláře trvalo průměrně 10 minut. Nejvíce formulářů bylo přijato ve víkendové dny (zhruba 40 % všech elektronických formulářů), preferovanou dobou odesílání byly jednoznačně večerní hodiny. Mezi 18. a 22. hodinou bylo doručeno téměř 40 % všech elektronických formulářů. Zajímavé je, že elektronické formuláře byly vyplněny i za nepracující důchodce (11 %) a nezaměstnané (4 %).

Děti a senioři

Soubor formulářů zahrnoval i sčítací listy osoby za děti, které se narodily těsně před rozhodným okamžikem sčítání, tedy před půlnocí z 25. na 26. března 2011. Takových dětí bylo celkem 62 a v rozhodný okamžik jim byl jeden jediný den. Naopak nejstarší elektronicky sečtené osobě bylo 106 let.

Možnost vyplnit sčítací formulář nejčastěji využívali obyvatelé Prahy (32,5 %), poté následovaly tři moravské kraje: Jihomoravský (29,5 %), Zlínský (29,2 %) a Moravskoslezský (27,1 %). Nejméně tuto formu preferovali obyvatelé dvou východočeských krajů – Královéhradeckého (19,8 %) a Pardubického (21,2 %).

Počet odeslaných elektronických formulářů během jednotlivých dní v týdnu



Zdroj: ČSÚ

Alena Géblová
odbor vnější komunikace