Report o procesech sběru dat a propojení datové základny pro hodnocení dopadu programů TA ČR

1. Sběr dat a jejich zpracování v datové základně

Datová základna DAFOS obsahuje integrovaná data vycházející z více datových zdrojů.

Tabulka 1 Data v datové základně, jejich aktualizace a vlastník

Datový zdroj	Aktualizace dat	Vlastník
ISVAV	ISVAV, API	RVVI[1]
ARES (RES, VREO)	API	MF ČR[2]
PATSTAT	PATSTAT, API	EPO[3]

Zdroj: Stav datové základny k 31.12.2019

Aktualizace dat je zajištěna pomocí API s feature overlaid fixture (zajišťuje zachování upravených názvů a kódů).

2. Dostupnost dat pro hodnocení dopadu

Tato kapitola popisuje, jaká data je možné ihned využít a jaká je nutno ještě zpracovat. Krátce jsou popsána i data, které existují, ale nejsou dostupná a data, které by bylo dobré mít, ale nejsou systematicky evidována.

2.1. Data dostupná a strukturovaná

Tato data je možné získat ve strukturované podobě, tak aby je bylo možné přiřadit podle konkrétního společného klíče. V tomto případě je nejspolehlivější společným klíčem identifikační číslo – IČ společnosti, který je jednotný pro zmíněné zdroje dat v následující tabulce.

Tabulka 2 Dostupná a zpracovaná data

Data	Zdroj dat

[1] Rada pro výzkum, vývoj a inovace
[2] Ministerstvo financí ČR
[3] European patent office

Informace o udělení podpory na účely VaV	IS VAV
Informace o výši dotace	IS VaV
Informace o podaném projektu do Programu ALFA	Interní TA ČR
Finanční data (Obrat, zisk, aktiva, přidaná hodnota, ROA)	MagnusWeb databáze
Informace o kategorii zaměstnanců	RES
Informace o odvětví podniku	RES
Informace o vlastníkovi (veřejný/soukromý český/soukromý zahraniční)	RES
Informace o lokalitě (NUTS 3)	RES
Informace o spolupráci s VO	IS VaV
Vznik/zánik/transformace podniku	TC data, RES, Magnus, individuálně z internetu

2.2. Dostupná data nestrukturovaná

Za data nutná ke zpracování jsou považována ta data, která není možné využít a je třeba je převést do strukturované podoby, tak aby je bylo možné pomocí jednotného klíče propojit s ostatními daty. Mezi nejvýznamnější zdroje dat tohoto charakteru patří data Úřadu pro průmyslové vlastnictví (ÚPV) a data z patentové databáze Patstat. Pro zajištění použitelné struktury dat ÚPV je nutné data stáhnout z online databáze pomocí scriptu[4]. Je reálné, aby tuto aktivitu zajistil tým datové části projektu PROEVAL. Struktura dat v databázi Patstat (kterou TA ČR vlastní) je nekoherentní s identifikátorem podniku IČ. Tento problém byl již úspěšně vyřešen institucí CERGE-EI (IDEA)[5]. Pokud nebude možné vyjednat získání strukturovaných dat Patstat od CERGE-EI, jejich transformace by zabrala přibližně měsíc práce jednoho pracovníka[6]. Další důležitá data se týkají počtu jednotlivých zaměstnanců podniků. Tato data jsou mimo jiné získávána i společností BISNODE ČESKÁ REPUBLIKA, A.S. a je možné je pořídit. Pro podpořené a kontrolní skupinu nebyl údaj dostupný v potřebné míře, a proto nakonec nebyl údaj zahrnut.

2.3. Data existující, ale nedostupná

[4] Sada automatizovaných úkonů, např. stahování textového obsahu html stránky
[5] Zodpovědná osoba Martin Srholec
[6] Tuto práci není možné automatizovat, protože je nutno k jednotlivým podnikům se špatným názvem – které pro stejný podnik nejsou unikátní – přiřadit jednotlivá IČ a správné názvy. Jediný zisk dat nad ty automatizovaně získatelná z ÚPV jsou informace o mezinárodních patentových přihláškách českých podniků.

Pro vyhodnocení dotací výzkumu a vývoje je vhodné disponovat daty o výdajích na výzkum a vývoj (a inovace). Tato data jsou systematicky sbírána prostřednictvím Českého statistického úřadu (ČSÚ) formou dotazníku VTR5-01 a dotazníku o inovacích. Individuální data (potřebná pro takové hodnocení) nejsou TA ČR přístupná.

2.4. Data, která nejsou systematicky sbírána pro velký vzorek podniků

Data nejvyšší kvality a vysoké vypovídající hodnoty na individuální úrovni podniků jsou získávána v mapování inovačních kapacit (INKA). V rámci prvního mapování (2014) bylo zajištěno mnoho systematicky informací pro 452 podniků. Pro dopadové analýzy by tento vzorek musel být větší (alespoň v řádu jednotek tisíců). Primární data z mapování INKA jsou pro podobné analýzy výborná, protože obsahují jak kvantitativní údaje, která nejsou pro TA ČR dostupná (z ČSÚ), ale hlavně také ojedinělé kvalitativní údaje s vysokou vypovídající a prediktivní hodnotou. Jde o kvalitativní údaje jako např. ambice vedení podniku, trhy působnosti podniku, technologická pozice, pozice v globální produkční síti, počet implementovaných inovací apod. Při rozšíření počtu sledovaných podniků, bude možné vytvářet dopadové hodnocení (všech národních) programů vysoké kvality a vysoké vypovídající hodnoty. V současné době TA ČR ve spolupráci s kraji ČR byly zajištěny informace pro přibližně 700 podniků.

3. Propojení datové základny pro hodnocení dopadu TA ČR

Z důvodu povahy dat (individuální podniky, rozšiřování proměnných a přiřazování na základě IČO) a individuálního přístupu k jednotlivým hodnocením dopadu je vhodnější (levnější a méně časově náročné) si konkrétní datovou základnu (dataset) sestavit zvlášť pro každé hodnocení zvlášť. Programování řešení v datové základně vyžaduje dalšího pracovníka mimo výzkumníka sestavující kontrafaktuál. Nicméně zpracování datové základny v projektu PROEVAL umožňuje exportovat všechny podpořené podniky (z databáze ISVAV) a základní informace z jiných datových zdrojů, které jsou k nim přiřazovány na základě IČO. Díky tomu jsou základní informace o podpoře a charakteristikách podniku lehce dostupné.

Tabulka 3 Data z datové základny pro subjekty z ISVAV

Data	Zdroj dat
IČO	ISVAV/RES
Název	ISVAV/RES
Adresa	RES
Kraj	RES
Právní forma dle RES	RES
Právní forma dle RES, 2. úroveň	RES

Právní forma dle IS VaVaI	ISVAV
Datum vzniku subjektu	RES
Statutární orgán	Obchodní rejstřík
Výzkumná organizace	Zatím nezpracováno
ISEKTOR	RES
Hlavní NACE	RES
Hlavní NACE, 2. úroveň	RES
Počet zaměstnanců	RES
Podpora 2007-2026	ISVAV

Zdroj: STARFOS export vyhledávání subjektů

Finanční údaje nebyly zahrnuty z důvodu licenčních ujednání s BISNODE a proto je jejich zajištění nutné z jiných zdrojů.