Správa dat

Péče o data, aby se neztrácela a dala použít

# O co jde a k čemu to je

Práce s kvantitativními daty je nedílnou součástí analytických činností ve veřejné správě. **Data management** pokrývá soubor činností, které zaručují, že data jsou přehledně uložena, popsána a zpřístupněna dalším uživatelům ve formě datových produktů.

**Případová studie**

Tento návod jsme vyvinuli na základě práce s oddělením výzkumu a evaluací na Agentuře pro sociální začleňování.

[Číst případovou studii ⟶](../pripadove-studie/asz)

|  |
| --- |
| Správa dat!? Pojmy a záběr tohoto návodu |
| V tématu správy dat typicky pod pojmem *data management* (DM) potkáte dvě různé věci, podle toho, z jakého prostředí nebo profese vychází.  **(Research) data management**: správa dat v průběhu výzkumného projektu. K tomu pro výzkumníky existuje sada metodik a [technických nástrojů](https://ds-wizard.org/), mimo jiné i jako pomůcky ke splnění povinností, které na výzkumníky čím dál častěji kladou grantové agentury s cílem podpořit dostupnost a znovuvyužití dat z výzkumu.  Některé z těchto metodik a nástrojů se vám mohou hodit např. ve větších projektech, kde sbíráte nebo generujete nová data. Obecně užitečnou pomůckou je [**sada principů FAIR**](https://www.go-fair.org/fair-principles/): data by měla být nalezitelná (**F**indable), dostupná (**A**ccessible), propojitelná (**I**nteroperable) a znovu použitelná (**R**eusable) - viz @wilkinsonFAIRGuidingPrinciples2016.  **Data Management** jako role a funkci ve velkých organizacích související se správou dat ve velkých IT systémech. Zde se dominantně využívá tzv. Data Management Book of Knowledge [@damainternationalDAMADMBOKDataManagement2017], kompendium znalostí a principů pro různé oblasti správy dat. I zde se dá inspirovat, ale DM v tomto smyslu je celá profese se svými certifikacemi a aplikací především ve světě velkých systémů.  V obou světech se pak můžete setkat i s příbuzným pojmem *data stewardship* – pojem klade důraz na dlouhodobou péči o data i jménem jeho budoucích uživatelů.  **V tomto návodu** pracujeme s pojmem data management volněji: péče o data v rámci analytických projektů, která se vám typicky dostanou do rukou: často data přebraná odjinud, pokud sbíraná, tak spíš v menším rozsahu. Zde je typicky třeba dosáhnout základní katalogizace, dokumentace a sdílení dat v týmu nebo organizaci.  Pojem správa dat pak nenavazujeme na právní pojmy správce dat ve smyslu legislativy ochrany soukromí nebo informačních systémů. |

# Sebediagnostika: je to pro mě?

Tento návod se vám bude hodit, pokud pro vás platí něco z následujícího:

* Náš útvar pravidelně zpracovává datové výstupy pro interní i externí využití.
* Pracujeme s datovými zdroji od různých poskytovatelů, které následně kompilujeme do ucelených data setů
* O data se stará jeden člověk nebo několik málo lidí v týmu, ale chtěli bychom, aby zbytek týmu mohl být při využivání dat samostatnější

# Jak na to

## Co je cílem

* zpřístupnit a zpřehlednit data širší skupině lidí (typicky uvnitř týmu/organizace)
* mít jasno v tom, odkud se která data vzala a jak je upravujeme pro další práci
* zdokumentovat jednotlivé datové sady: co v nich je a co znamenají jednotlivé položky?

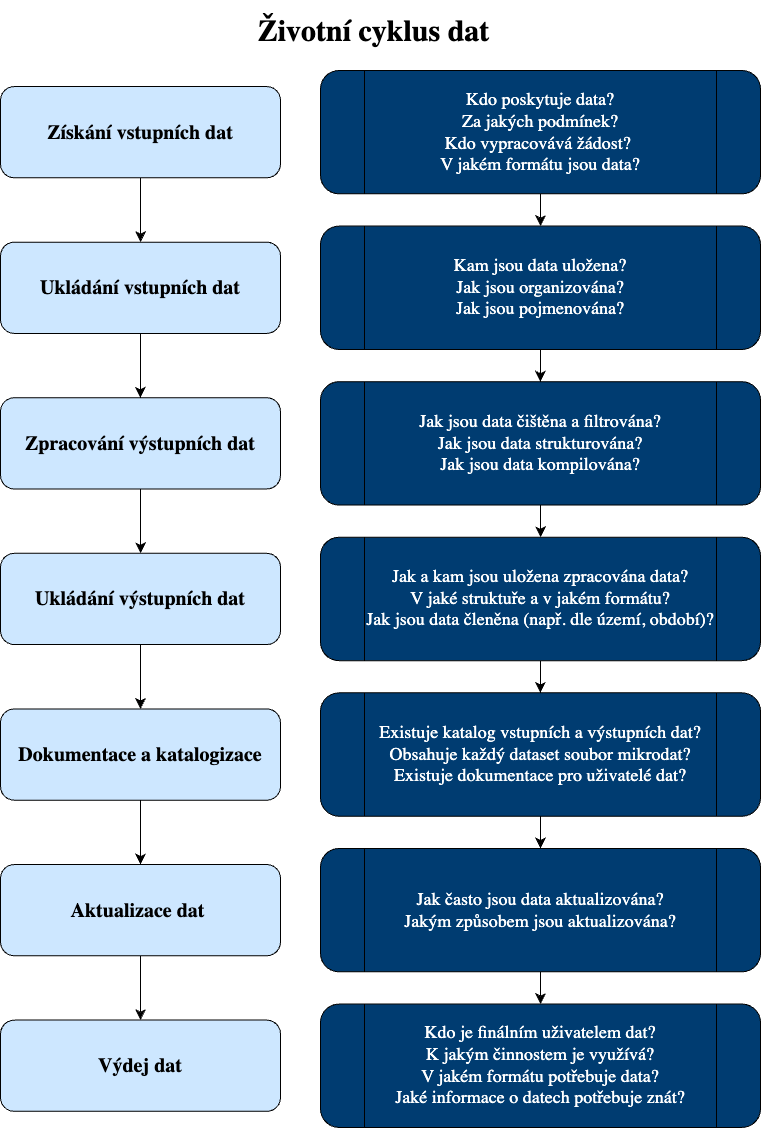
## Zmapujte životní cyklus vašich dat

Nejprve rozlište a popište kroky vedoucí od získání (či generování) dat k finálnímu datovému produktu určený ke sdílení. Těmto krokům se říká [životní cyklus dat](https://openscience.cuni.cz/OSCI-61.html).

Vytvořte tabulku, kde řádky představují každý krok v životním cyklu dat. Pro každý krok popište co se děje a kdo to vykonává.

* Získání vstupních dat (či jejich sběr)
* Ukládání vstupních dat
* Zpracování výstupních dat (čištění, strukturování, výběr, kompilace)
* Ukládání výstupních dat
* Dokumentace a katalogizace
* Aktualizace výstupních dat
* Výdej dat

Můžete také využít schéma životního cyklu dat. Níže najdete generické schéma, které můžete následně upravit podle vašich potřeb.



## Zeptejte se finálních uživatelů, jak data využívají a pro jaké účely

Předtím než navrhnete způsob, jak data poskytnout finálním uživatelům, zeptejte se jich, jakým způsobem hodlají datové produkty využít a pro jaký účel. Zároveň identifikujte případné obavy a překážky, které mohou vzniknout.

*Uživatelem* – skutečným nebo kýženým – často budou další členové týmu nebo další lidé ve vaší organizaci. Můžou jím ale být také externí subjekty, kterým data předáváte – jiné organizace, výzkumníci aj.

Prakticky:

1. udělejte si seznam všech, kteří by měli nebo chtěli přistupovat k datům, o která se staráte.
2. s každým z nich - nebo aspoň těmi nejzásadnějšími - si pár minut pohovořte. Zeptejte se na nějaký nedávný příklad, kdy s daty pracovali, a zkuste identifikovat, co jim brání, jakým nejistotám čelí a jaký způsob přistupování k datům by jim vyhovoval.
3. v dalších krocích pak můžete s těmito stejnými lidmi opět v pár minutách otestovat způsob přistupování k datům, případně i širší systém správy dat.

## Strukturujte vaše úložiště dat

Při správě datových úložišť rozlište surová data, back-end a front-end:

* úložišťe surových dat: datové sady, které obdržíte od poskytovatele uložte v jejich surové formě spolu s dokumentací, kterou vám poskytovatel dodal.
* back-end: datové sady, které jsou strukturované podle vašich potřeb (např. sloučením několika datových sad)
* front-end: datové sady, které jsou určené k využití dalšími uživateli (interní i externí)

Prakticky:

Pokud nevidíte technickou potřebu mít vše zachycené v databázích, začněte prostě tím, že datové soubory dobře uspořádáte na sdíleném disku.

Postarejte se také o konzitentní názvy souborů - viz [How to name files](https://speakerdeck.com/jennybc/how-to-name-files-the-sequel): aby je přečetl stroj i člověk a aby se správně řadily

Případně k tomu vytvoříte rozcestník v dokumentu, na intranetu, v Sharepointu aj. Začněte těmi sadami a soubory, které skutečně používáte.

## Zdokumentujte svá data

Každá datová sada má být dokumentovaná tak, aby bylo možné dohledat základní popis datových sad a jejich proměnných.

Dokumentujte

1. **datové sady** - ať už jako jednotlivé soubory, listy v excelu, nebo např. datové sady, které dostáváte či produkujete pravidelně
2. **proměnné** v každé datové sadě

Lze rozlišit dva typy dokumentace dat:

* metadata: soubor popisů pro jednotlivé sady a datové proměnné
* dokumentace pro uživatele: soubor informací, které pomohou uživatelům v orientaci v datových sadách a jejich využití.

|  |
| --- |
| **Pojem: metadata**  Data o datech – strukturovaná data, která popisují jiná data. Kromě popisu slouží i k zajištění interoperability. Dobře zanesená metadata umožňují strojům si data předávat a propjovat automatizovaně.  *Metainformace*: širší množina, může zahrnovat číselníky, klasifikace aj. – viz [databázi metainformací ČSÚ](https://www.czso.cz/csu/czso/databaze_metainformaci). |

### Datové sady a soubory

Prakticky:

1. V **jedné tabulce/seznamu** zachyťte datové soubory, které jsou “na vstupu”: váš sběr, vyžádaná data od jiné organizace, export z nějakého systému, atd.
2. Hned vedle dat pokud možno uložte jakoukoli **dokumentaci**, kterou k souborům máte - včetně např. dokumentu, kterým jste o data požádali a průvodní informace od poskytovatale. Pokud jsou na vstupu data od ČSÚ, uložte si jejich dokumentaci. Zachyťte i jakákoli omezení využití, která se k souboru vážou (důvěrnost, zveřejnitelnost).
3. V tabulce zachyťte **metadata o souboru**: původ dat, datum přijetí, přesnou cestu k souboru, kontakt na tvůrce/správce, časové pokrytí dat, odkaz na komunikaci s původcem dat nebo žádost o data
4. Ve **druhém seznamu** zachyťte datové soubory, které na základě těch vstupních vytváříte vy. Často půjde o sumáře, propojení tabulek, výběry… Zde zachyťte
5. Pokud vytváříte ještě **další produkty “na výdeji”** (pravidelné exporty pro určité uživatele, náhledy, dashboardy, atd.) zachyťte je také. Obzvláště tuto kategorii je dobré mít např. v rozcestníku na intranetu

|  |
| --- |
| Doporučené položky metadat o datových souborech (rozbal kliknutím) |
| * Hvězdičkou označeny základní položky, které byste měli znát a zaznamenat u každé datové sady.  | Položka | Vysvětlení | | --- | --- | | Název\* | Název datové sady | | Popis\* | Popis datové sady | | Kód datové sady | Kód (např. v katalogu poskytovatele)) | | Zdroj - původce dat\* | Od koho jsme data dostali? | | Zdroj - typ zdroje | Např. jedná se o administrativní, nebo statistická data? | | Zdroj - kontakt\* | Kontakt na straně poskytovatele | | Interní kontakt / vlastník | Kdo u nás o těchto datech ví nejvíc / stará se o ně? | | Odkaz na soubor\* | Na disku, Sharepointu atd. | | Odkaz na dokumentaci od zdroje\* | Měla by být uložená blízko dat, popř. odkaz na veřejnou dokumentaci | | Omezení využití\* | Co víme, že s daty nesmíme dělat? | | Odkaz na vlastní dokumentaci | Pokud máme | | Frekvence změny / sběru | Jak často data dostáváme / jak často se aktualizují? Popř. jde o ad hoc sběr? | | Časová granularita | Jaké časové úseky jsou v datech zachyceny? | | Míra agregace | Jak jsou data agregována? (kategorie, území) | | Záběr pokrytí | území, populace aj. | | Vztah k jiným sadám | Které sady poskytují podobná data? | | Kde jsme data využili? | Pro back-end zpracovaná data | | Detailní věcná dokumentace - odkaz | Pokud máme |   Pokud vám některé položky nedávají smysl, vypusťte je – nevyplňujte nic pro forma.  Inspirováno metadatovým standardem [Dublin Core](https://www.dublincore.org/specifications/dublin-core/dcmi-terms/). Tam najdete i další vlastnosti datových sad, které může mít smysl dokumentovat. Popř. můžete využít i sadu ze standardu [Schema.org](https://schema.org/Dataset). |

### Proměnné v datech

Následně u zásadních datovýh souborů zdokumentujte klíčové proměnné. Je na vás, jestli tuto dokumentaci budete držet u zdrojových souborů, u zpracovaných (back-end) dat nebo v odděleném souboru.

Optimální je, aby tento metadatový list (codebook) byl co nejblíže datům. Pokud data udržujete v Excelu, dejte jej na první dokumentační list sešitu.

|  |
| --- |
| Doporučené položky metadat o proměnných / sloupcích dat (rozbal kliknutím) |
| * Hvězdičkou označeny základní položky, které byste měli znát a zaznamenat u každé proměnné.  | Položka | Vysvětlení | | --- | --- | | Název\* | Buď slovní, nebo technický (Počet obyvatel nebo pocet\_obyvatel) - u všech proměnných konzistnentně | | Kód proměnné\* | Vlastní nebo od zdroje - technický název, pokud jako hlavní název používáte slovní | | Popis\* | Základní popis proměnné | | Zdroj\* | Odkud položka pochází? Sběr, zadání v IT systémy, výpočet… | | Datový typ\* | číslo, text, datum, čas, binární hodnota, procento | | Kód ukazatele | Pokud se jedná o ukazatel a kód existuje (např. od ČSÚ) | | Jak jsou označena chybějící data\* | lze vyřešit na úrovni datasetu, ale může se lišit | | Jednotka\* |  | | Použitý číselník | Pro proměnné kódované číselníkem: klasifikace, území apod. zaznamenejte kód a správce číselníku | | Běžné rozpětí |  | | Odkaz na dokumentaci | Pokud je k proměnné zvláštní dokumentace | | Poznámka | Jakékoli další informace | |

**Alternativy**

*Katalog datových položek* (tzv. data dictionary)

Alternativní přístup je vytvořit velký katalog všech datových položek, které v organizace používáte, a k nim zanést jejich původ a metadata.

To se hodí zvlášť v případě, že používáte nějakou soustavu indikátorů, popř. vytváříte vlastní indikátory a indexy.

Dokumentace jednotlivých proměnných pak vypadá podobně, ale všechny proměnné jsou zdokumentované na jednom místě a dokumentace proměnných se odkazuje na dokumentaci zdrojových datových sad a souborů.

*Správa metadat*

U větších systémů, kde se data přebírají, propojují a zpracovávají automaticky, lze zavést systém správy metadat, kdy metadata vstupují přímo do zpracování dat a výsledný prezentační systém (např. dashboard nebo datová aplikace) může datovou dokumentaci zobrazovat vedle dat. To ale vyžaduje pokročilejší technologie a správu dat napříč organizací.

Také je možné stavět na principech propojených dat ([linked data](https://www.w3.org/wiki/LinkedData)), to se ale opět týká spíš správců velkých datových fondů.

*Integrace dat*

Pro některé situace a využití může být vhodné data z externích zdrojů pravidelně a co nejvíce automatizovaně integrovat do jednotné databáze. Příkladem je PAQ Research, kde většinu externích dat na základě šablon a automatizovaných skriptů načítají do jedné databáze.

## Nastavte proces pro výdej dat k dalšímu využití

Na základě informací od finálních uživatelů navrhněte způsob, jakým budete předávat vaše datové produkty.

Může jít třeba o excelovou tabulku, ze které budou uživatelé přímo čerpat, nebo o sofistikovanější rozhraní (např. databáze nebo dashboardy).

Pro přípravu dat k využití v organizaci se vám může hodit návod [Analýza dat moderně ⟶](analyza-dat.qmd)

A pokud vaše řešení bude zahrnovat třeba dashboard nebo webové rozhraní, nahlédněte do návodu [Uživatelské testování ⟶](user-testing.qmd)

## Udržujte pořádek i uvnitř datových souborů

Při správě dat – především v Excelu – je dobré se držet několika zásad:

* Na jednom listu pouze jedna sada dat
* Na listu s hrubými daty nejsou žádné vzorce ani výpočty
* V jedné buňce pouze jeden údaj
* Tedy: data v tzv. tidy formátu (@wickhamTidyData2014; 6. kapitola @wickhamDataScienceImport2023)
* Všechna data zanesena v buňce, nikoli formátováním
* Konzistentní názvy listů
* Konzistentní názvy proměnných
* Konzistentní názvy proměnných
* Smysluplné kódování hodnot (např. proměnná muz s hodnotami 0 a 1 je lepší než pohlavi s hodnotami 0 a 1)
* Konzistentní kódování chybějících hodnot

Data ukládejte s vypnutými filtry, s kurzorem na listu dokumentace a v datových listech vlevo nahoře.

Více viz článek pro praktiky @bromanDataOrganizationSpreadsheets2018 a metodiky analytické profese britské státní služby [@analysisfunctioncentralteamonsCreatingSharingSpreadsheets2021; @analysisfunctioncentralteamonsReleasingStatisticsSpreadsheets2021].

# Na co si dát pozor aneb tipy & triky

## Neutopte se v diskusích a snaze o dokonalost

Zásadní je někde začít. Dlouhé diskuse o tom, která data jsou hrubá a která už zpracovaná, nebo jestli evidovat každou novou verzi té stejné datové sady, vám nejspíš nepomůžou. Dokonalý systém neexistuje.

Začněte něčím jednoduchým – třeba jen zachyťte, jaká data často využíváte.

## Rutina, nejen velký třesk

Kromě jednorázového úklidu a zdokumentování péče o data vyžaduje i **změnu běžné rutiny**, např. zvyknout si při přebírání dat odjinud tato data zanést do dokumentace. Ano, chvilku to zabere, ale konzistentní postupy se vyplatí, protože uspoříte čas při potýkání se s nepořádkem později a posílíte schopnost svého týmu data správně využít.

## Obsahová, nejen technická dokumentace

Do péče o data a jejich dokumentaci zatáhněte ty, kdo s daty pracují nebo dobře znají prostředí, odkud se data generují. V dokumentaci můžete zachytit znalosti o specifikách daných dat, insider znalosti o tom, jak se sbírají, a zkušenost s prací s nimi: na jaké nečekané vlastnosti dat jsme přišli při práci s nimi.

Tím také posílíte **péči o znalosti** svého týmu: při dokumentaci dat můžete zachytit informace, které by jinak zůstaly ukryty v hlavách vašich kolegyň a kolegů a na které by ostatní možná museli později přicházet znova.

## Ještě větší transparentnost

U projektů, kde je důležité mít veřejně dostupný záznam o provedené analýze, se může hodit projekt archivovat v repozitáři např. na Zenodo nebo OSF – tím zdokumentujete nejen data, ale i celý projekt; je to forma knowledge managementu viz [(-> návod)](knowledge-management.qmd).

|  |
| --- |
| Příklad |
| V Agentuře pro sociální začleňování se zaměřili na porozumění tomu, jak vznikají a co znamenají administrativní data, která používají.  Začali u dat z dávkových systémů, která Agentura přebírá od Úřadů práce pro analytické účely. Posbírali, co všechno o těchto datech vědí interně výzkumníci ve dvou různých týmech, včetně zjišťěných much a chytáků, a pak se vydali na Úřad práce zeptat se, jak se vlastně data zadávají, aby věděli, jak je interpretovat.  Takto zachycené znalosti můžete zpřístupnit třeba na interní wiki nebo v běžném dokumentu ve složce vedle dat. |

# Další zdroje

České pojednání o tématu správy dat najdete v knize @krejciCestyDatumZdroje2012, základní principy najdete i na webu [Českého sociálně vedního datového archivu](https://archiv.soc.cas.cz/cz/prace-s-daty/management-dat-v-prubehu-vyzkumneho-projektu); tamtéž i [návod na tvorbu tzv. plánu správy dat](https://archiv.soc.cas.cz/cz/prace-s-daty/data-management-plan-pro-socialni-vedy) pro společenskovědní projekty. Užitečné mohou být i [prezentace z nedávného semináře ČSDA k tématu](https://archiv.soc.cas.cz/cz/vyuka-a-kurzy/poradane-akce/event/21-open-science-a-sdileni-socialnevednich-dat-v-praxi-data-management-plan-ochrana-respondentu-a-legalni-aspekty-sdileni-dat).

Příklad dlouhodobé správy dat pro další využití je [Český sociálně vední datový archiv (ČSDA)](https://archiv.soc.cas.cz/cz/), kde v [prohlížeči datových sad](http://nesstar.soc.cas.cz/webview/) najdete i způsob zaznamenání metadat, ze kterého se lze inpirovat.

K dokumentaci se mohou hodit [standardy tzv. Dublin core metadat](https://www.dublincore.org/) - sice působí technicky, ale když je dokážete využití, nebudete vynalézat to, co už někdo vymyslel.

Alternativní přístup k dokumentaci dat a související nástroje nabízí [Data Documentation Initiative](https://ddialliance.org/).

Pragmatický standard dokumentace tabulkových datových sad [používá britská státní správa](https://www.gov.uk/government/publications/recommended-open-standards-for-government/using-metadata-to-describe-tabular-data-for-publishing).

K dokumentaci datasetů viz @gebruDatasheetsDatasets2021, též [kap. 10](https://tellingstorieswithdata.com/10-store_and_share.html) v @alexanderTellingStoriesData2023.

Kompletní průvodce pro správu dat v kontextu výzkumu je [Data Management Expert Guide](https://dmeg.cessda.eu/) od CESSDA.

## Odkazované zdroje a literatura