Data a vědecká komunikace

CORE042: Data – odpověď na základní otázku života, vesmíru a vůbec...
7. přednáška

Michal Lorenz <lorenz@mail.muni.cz>
Katedra informačních studií a knihovnictví – Filozofická fakulta MU

Filozofie dat

- **z lat.** dare – dávat, „danosti“ abstrahované z fenoménem
- první použití pojmu 1646 v **teologii**, později v matematice
  1) množina principů přijatých jako základ argumentu
  2) fakta, konkrétně ta vědecká
- v 18. st debatován singulár x plurál
- **běžný význam**:
- fakta ve formě vědecké evidence získané z experimentů, pozorování a dalších způsobů bádání
- **materiální forma**: čísla, texty, symboly, obrázky, zvuk, elektromagentické vlny, prázdnota či ticho

MUNI

Role dat v poznání

- syrová data (raw data) - **epistemický omyl**, neexistující entita
- data nejsou pravdivá, ani nejsou realitou, jsou rétorickou kategorií
- *data nerostou na stromech, ani je netěžíme ze země, konstituují fakta, kterým předchází, domnělá evidence (Buckland)*

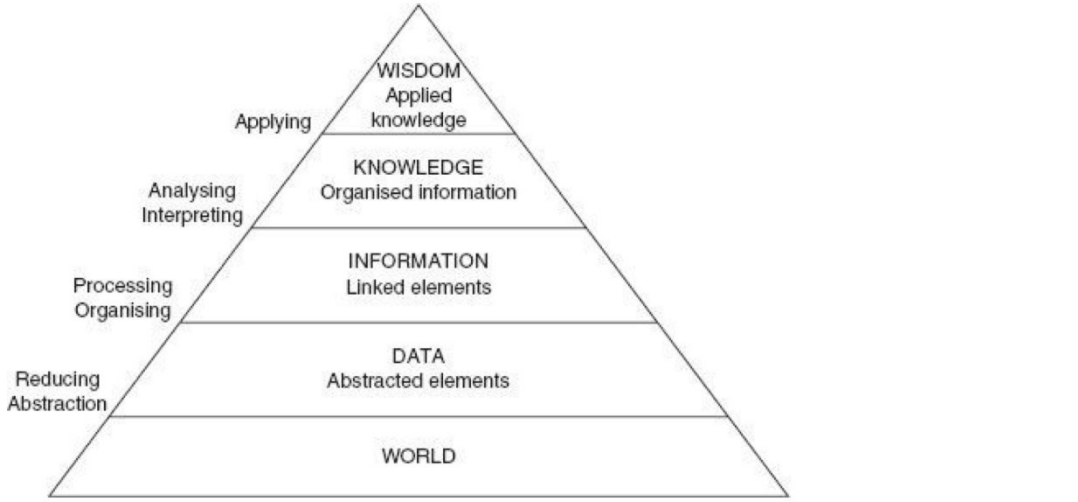
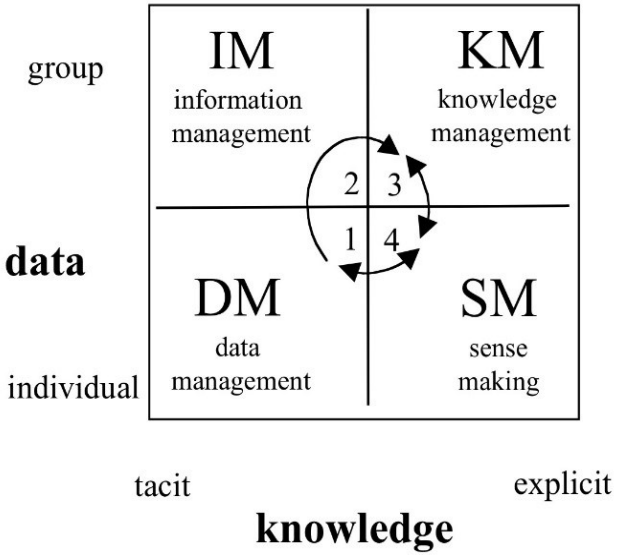


Figure 1.1 Knowledge pyramid (adapted from Adler 1986 and McCandless 2010)



převzato z: BAKER, K. S. & BOWKER, G. C. (2007). Information ecology: Open system environment for data, memories, and knowing. *Journal of Intelligent Information Systems*. **29**(1), 127-144. DOI: 10.1007/s10844-006-0035-7.

MUNI

Typy dat

- kvantitativní, kvalitativní
- **kvantitativní** – konzistentní numerický zápis, rozsáhlá, spojená s fyzickými vlastnostmi fenoménu (délka, výška, vzdálenost, váha…)
- nebo reprezentativní a spojená s nefyzickými vlastnostmi fenoménu (sociální třída, dosažené vzdělání, pořadí kvality života…)
- čtyři úrovně měření

Levels of measurement	Definition	Example
Nominal data	Categorical in nature, with observations recorded into discrete units.	Unmarried, married, divorced, widowed
Ordinal data	Observations that are placed in a rank order, where certain observations are greater than others.	Low, medium, high
Interval data	Measurements along a scale which possesses a fixed but arbitrary interval and an arbitrary origin. Addition or multiplication by a constant will not alter the interval nature of the observations. Data can either be continuous (e.g., time or length) or discrete (e.g., counts of a phenomenon) in nature.	Temperature along the Celsius scale
Ratio data	Similar to interval data except the scale possesses a true zero origin, and multiplication by a constant will not alter the ratio nature of the observations.	Exam marks on a scale of 0–100

- **kvalitativní** – nenumerická, jako např. umění, video, zvuk..
- mohou být kvantifikována, ale dochází k významné redukci ztrátě bohatosti dat
- významný postup ve výpočetní analýze pomocí strojového učení a vytěžování dat

MUNI

Typy dat

– strukturovaná, polostrukturovaná, nestrukturovaná

strukturovaná – snadno organizovatelná, zpracovatelná a vyhledatelná, datový model, př. relační databáze

polostrukturovaná – nepravidelná a flexibilní struktura, nevhodná pro relační databáze, ale tagovaná a s přiměřeně konzistentním souborem polí, př. webová stránka v tagovaném XML

nestrukturovaná – bez identifikovatelné společné struktury, kvalitativní, převoditelná na strukturovaná data, př. textové vyprávění, fotky

MUNI

Typy dat

- prvotní, doplňková, pomíjivá a derivovaná data – dle způsobu generování
- **prvotní (captured)** – přímé, záměrné měření (př. pozorování, průzkum, experiment, laboratoř, tvorba záznamů, kamery, skenery, sensory..)
- **doplňková (exhaust)** – vytvářené zařízením jako vedlejší produkt (př. elektronická pokladna –monitoruje zásoby, výkon pokladního, nákup zákazníka..)
- **pomíjivá (transient)** – neprozkoumaná a nezpracovaná data, protože jsou moc obsáhlá, nestrukturovaná, nákladná na uložení, chybí techniky na získání hodnoty či mají malé strategické využití, faktoidy (př. zdravotníci mažou až 90% generovaných dat real-time videí)
- **derivovaná (derived)** – produkovaná dodatečným zpracováním zachycených (př. primární sčítání aut, derivovaná počet sčítání, sčítání za hodinu apod.)

MUNI

Typy dat

- prvotní data často vstup do modelu, derivovaná data výstup z modelu
- derivovaná data tvořena za nějakým účelem
- primární data mohou být zpracována na několika úrovních – př. NASA Earth Observing System má 6 úrovní

Table 1.2 The six levels of data of NASA's Earth Observing System

Data level	Description
Level 0	Reconstructed, unprocessed instrument and payload data at full resolution, with any and all communications artefacts (e.g., synchronisation frames, communications headers, duplicate data) removed.
Level 1A	Reconstructed, unprocessed instrument data at full resolution, time-referenced, and annotated with ancillary information, including radiometric and geometric calibration coefficients and georeferencing parameters computed and appended but not applied to Level 0 data.
Level 1B	Level 1A data that have been processed to sensor units.
Level 2	Derived geophysical variables at the same resolution and location as Level 1 source data.
Level 3	Variables mapped on uniform space-time grid scales, usually with some completeness and consistency.
Level 4	Model output or results from analyses of lower-level data (e.g., variables derived from multiple measurements).

Source: Adapted from https://earthdata.nasa.gov/data/standards-and-references/processing-levels

Source: Adapted from https://earthdata.nasa.gov/data/standards-and-references/processing-levels

MUNI

Typy dat

- primární, sekundární a terciární data
- **primární** – tvořena výzkumníkem a jeho nástroji
- **sekundární** – dostupná ostatním k opětovnému použití, generovaná někým jiným, primární data někoho jiného
- **terciární** – forma derivovaných dat – součty, kategorie, statistické výsledky, často zveřejněna statistickými instituty kvůli důvěryhodnosti

  př. sčítání lidu v Irsku: sekundární data mohou být uvolněna až po 100 letech, ale uvolněno shrnutí – součty a kategoriální data
- vědci shromažďují primární data – přizpůsobena zaměření a potřebám vědců
- často kombinují primární, sekundární a terciární data – produkce hodnotnějších odvozených dat, význam sekundárních a terciárních dat – využití pro replikování výzkumů, budování větších a sofistikovanějších datových sad

MUNI

Typy dat

- indexikální, atributivní, meta- a pardata
- **indexikální** – umožňují identifikaci a propojení, jedinečné identifikátory (př. číslo pasu, kreditní karty, výrobní číslo, identifikátor digitálního objektu, IP adresa, objednací číslo apod.)
- umožňují propojit velké množství neindexikálních dat a jejich sledování, kombinování, vyloučení apod.
- **atributivní** – reprezentují aspekty fenoménu (př. osoba – indexikální data otisk prstu, DNA sekvence, atributivní – věk, pohlaví, výška, váha, krevní skupina, barva vlasů apod.)
- **metadata** – strukturované informace popisující, vysvětlující, lokalizující či jinak zjednodušující vyhledání, využití nebo správu informačních zdrojů (NISO)

  typy metadat: NISO: deskriptivní, strukturální, administrativní

- **paradata** – data o generování a shromažďování dat (př. čas a konání rozhovorů apod..)

MUNI

Paradigmata vědy

- Před tisícem let:

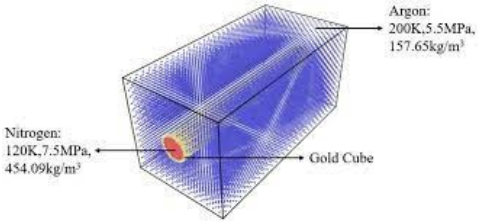
  věda **empirická,** *popisovala přírodní fenomény*
- Posledních několik stovek let:

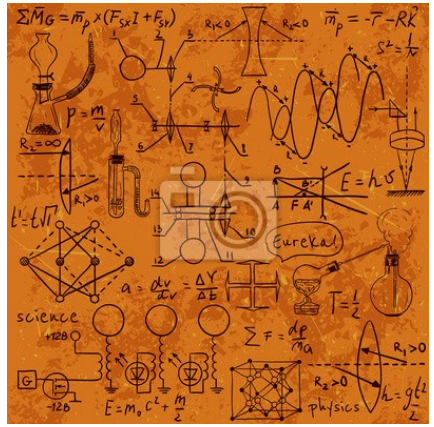
  teoretické odvětví, *používání modelů, zevšeobecňování*
- Posledních několik desetiletí:

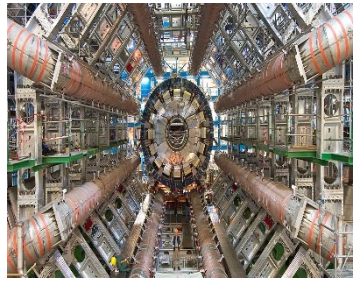
  výpočetní odvětví

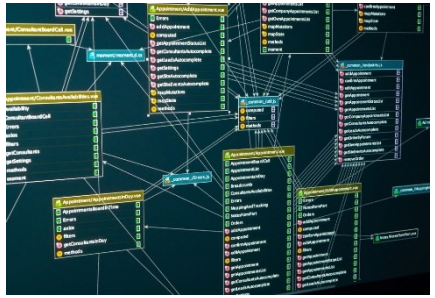
  simulace komplexních fenoménů
- Dnes: **průzkum dat** (eScience)

  sjednocení teorií, experimentů a simulací
- data zachytávaná zařízeními a nástroji či simulovaná
- zpracovaná softwarem, uložená v počítači,
- analyzovaná vědci v databázích a statisticky









MUNI

Vědecká komunikace

- **úzké pojetí**
- formální i neformální činnost související s využíváním a šířením informací prostřednictvím veřejných a soukromých kanálů
- **široké pojetí**
- růst vědecké informace
- vztahy mezi výzkumnými oblastmi a obory
- srovnání komunikačních činností, informačních potřeb a použití mezi obory a uživatelskými skupinami
- vztahy mezi formálními a neformálními aspekty komunikace

MUNI

Vědecká komunikace - trendy

- **e-science** (cyberscience)
- výpočetně náročná věda v distribuovaném síťovém prostředí. Používá velké sady dat a grid computing
- vědecko-výzkumná činnost ve virtuálním prostoru, generovaná pomocí počítačů připojených k síti a ICT
- **gridové technologie** – distribuované klasterové počítání, paralelní zpracovávání výpočetních úloh
- **věda 2.0** – on-line publikování, upravování a spolupráce vědců pomocí technologií webu 2.0 (blogy, tagy a sociální sítě)
- **repozitář** – úložiště, zařízení pro uskladnění akademických prací a dat

MUNI

Měření vědecké komunikace

- hlavním prostředek: **publikace** výsledků vědecké činnosti
- publikuj nebo zhyň (publish or perish)
- dnes požadováno dodání evidence, jak byly zjištění odvozeny: **doložení daty**
- **scientometrie -** Sola de Price, aplikace statistických metod na vědu a její růst, původ bibliometrie, impakt faktor, H-index
- **bibliometrické zákony**: Bradfordův zákon rozptylu, Lotkův zákon vědecké produktivity, Zipfův zákon výskytu slov, Priceův zákon vědecké produktivity, Garfieldův zákon koncentrace, Senguptův zákon bibliometrie
- hyperautorství a smrt autora
- epistemické kultury – jak vědci tvoří znalosti? Studium domén, jejich praktik a zvyků

MUNI

Malá x velká věda

- **Velká věda** –internacionální, kolaborativní, výzkumné infrastruktury, neviditelná univerzita (komunitní vztahy, výměna informací
- **Malá věda** – heterogenní metody, data, lokální kontrola a analýza

- **Velká data** – data velkého rozsahu (relativní), logistické problémy s manipulací a managementem, standardizace a homogenita
- Data jsou velká: množstvím, varietou, rychlostí, kombinací, všudypřítomností
- **Malá data** – data vhodná pro lidi a jejich porozumění, heterogenita dat, neviditelná univerzita

MUNI

LINDAT/CLARIAH-CZ

- Digitální humanitní vědy
- Digitální výzkumná infrastruktura pro humanitní, jazyková a umělecká data

  https://lindat.cz/
- Katalog platforem a digitálních sbírek na FF MU: https://digital-humanities.phil.muni.cz/vyzkum-a-projekty/katalog-platforem

MUNI

Literatura

- Borgman, Ch. L., (2015). *Big data, little data, no data: scholarship in the networked world*. Cambridge: The MIT Press.
- Bowker, G. C., 2000. Biodiversity Datadiversity. *Social Studies of Science*, 30(5), 643-683. DOI 10.1177/030631200003000
- Cronin, B. Scholarly communication and epistemic cultures. *New Review of Academic Librarianship*, 9:1, 1-24. DOI 10.1080/13614530410001692004
- Gitelman, L. (Ed.), 2013. "Raw data" is an oxymoron. Cambridge, Massachusetts: MIT Press.
- Hey, T., Tansley, S., Tolle, K. (Eds.), 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Redmond: Microsoft Research.

MUNI

Literatura

- Kitchin, R., 2014.*The data revolution: big data, open data, data infrastructures & their consequences*. Los Angeles: Sage.
- Kitchin, R., Lauriault, T. P. a McArdle, G. (Eds.), 2018. *Data and the city. Regions and cities*. London: Routledge, Taylor & Francis Group.
- Price, D. J. de Solla, 1965. Little science, big science. New York: Columbia University Press.

Price, D. J. de Solla, 1978. *Science since Babylon*. New Haven: Yale University Press.

MUNI