MUNI
ICS

Digitální svět: technologie, potenciál, rizika

Otevřená věda a Dlouhodobé uchování digitálních informací

Miroslav Bartošek, ÚVT MU
bartosek@ics.muni.cz

Přednášející

Miroslav Bartošek

- ÚVT MU, Knihovnicko-informační centrum MU
- Automatizace knihoven
- Digitální knihovny
- Open Science

MUNI ICS

Obsah přednášky

1. **Otevřená věda**
   - Historie vědecké komunikace
   - Open Science
   - Open Access
   - FAIR data
   - Human Genome project
2. **Dlouhodobé uchování digitálních informací**
   - Klasické přístupy k uchování informací
   - Nosiče informací a jejich životnost
   - Digitální informace – optimismus a vystřízlivění
   - Digital Preservation
   - Model OAIS x Osobní ochranné strategie
   - Doba digitálního temna?

MUNI
ICS

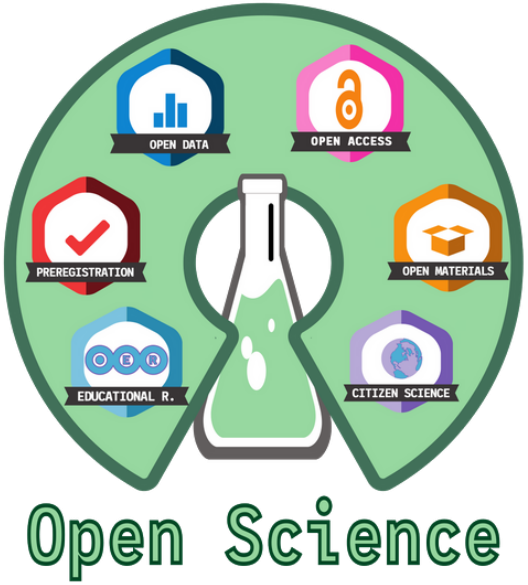
1.
Otevřená věda

Otevřená věda / Open Science

- Využít digitální technologie pro **lepší spolupráci a komunikaci** ve vědě!

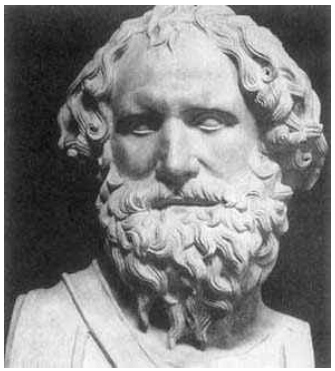
Game changers:

- Informace v digitální podobě
- Globální digitální komunikace



MUNI ICS

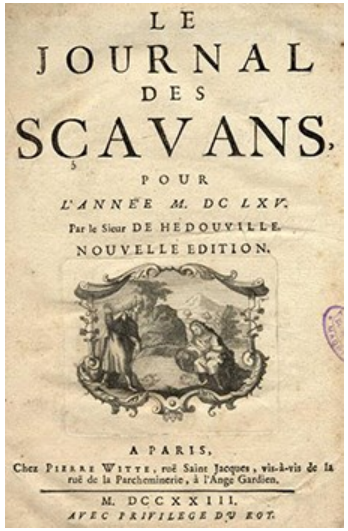
1.1 Historie vědecké komunikace

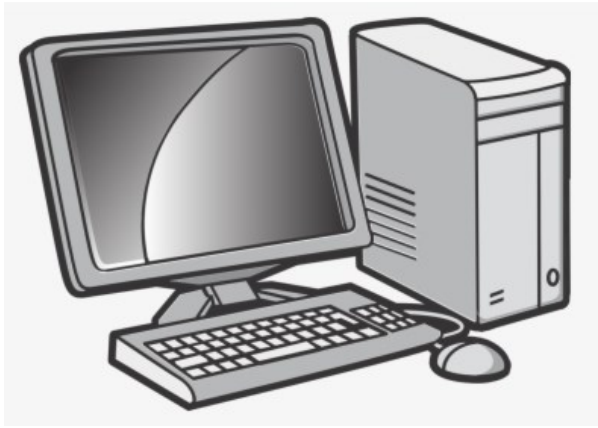


- **Starověk**: cesty, rukopisy, dopisy, Alexandrijská knihovna
- **1440**: knihtisk (Johannes Guttenberg), knihy
- **1665**: učené společnosti a <u>první vědecké časopisy</u>
  - Journal des sçavans – 5. ledna 1665 (Ludvík XIX – král Slunce)
  - Philosophical Transactions of the Royal Society – 6. března 1665 (Charles II)

- **20st:** vědecké konference (rychlé cestování)
- **1983**: Internet – globální digitální komunikace
- **1989**: World-wide-web (Tim Berners-Lee)
- **2002**: **Budapest Open Access Initiative**

MUNI ICS





osobní komunikace

tištěné časopisy

glob. digitální komunikace

MUNI ICS

1.2 Problémy současné vědecké komunikace

Klíčové stále **vědecké časopisy**

(„objektivní?“ měřítko pro hodnocení kvality)

- Finanční neudržitelnost
- Monopolistické praktiky vydavatelů
- Záplava balastu („publish-or-perish“)
- Pomalá komunikace
- Nízká dostupnost
- Přístup jen pro bohaté

Co s tím?

- Vědec napíše a **zdarma předá** vydavateli →
- Vydavatel zrediguje*, upraví pro tisk vytiskne a **prodá za předplatné** →
- Vědec zaplatí a čte

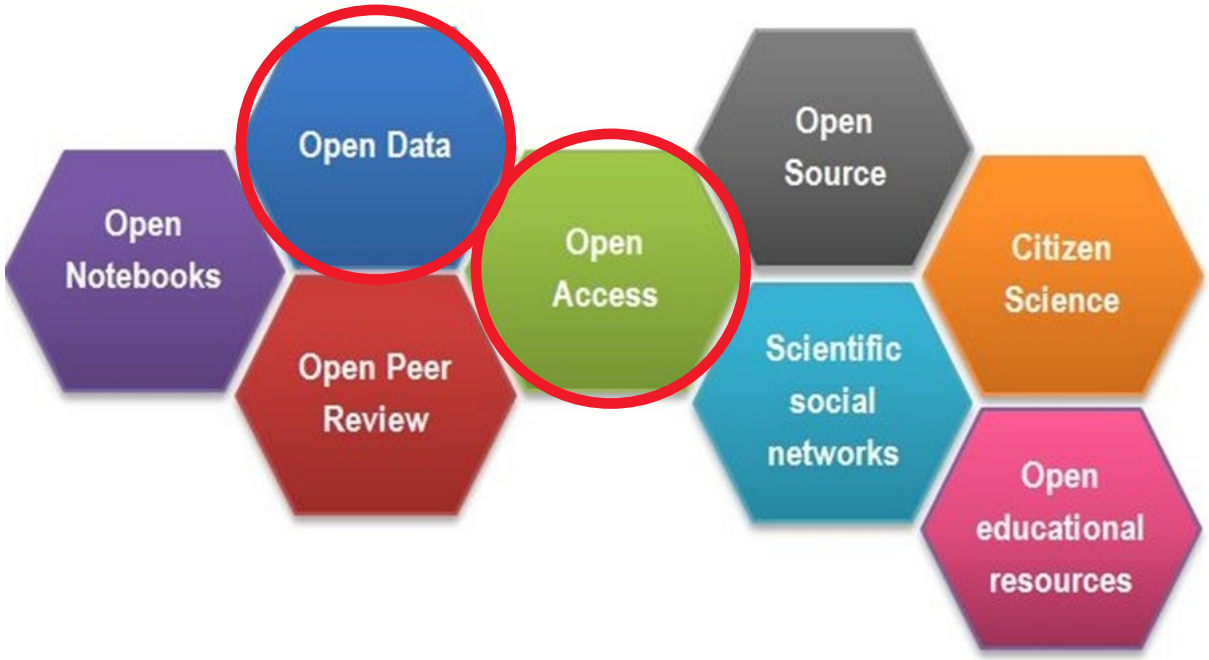
- *Recenzní řízení (peer-review) (provádí vědci, obvykle zdarma)

Společnost platí 2x: 1.výzkum, 2.publikace

Otevřená věda = publikace a data

1.3 Open Science

- Využít celosvětově dostupnou digitální komunikaci a ochotu vědců sdílet (zdarma) výsledky jejich bádání pro lepší komunikaci vědy
- Open Science is the idea that scientific knowledge of all kinds should be openly shared as early as is practical



MUNI ICS

Vědecké publikace – Open Access

1.4 Open Access

- Počátek 2002 – BOAI
- „Literatura, která by měla být volně dostupná online, je ta, kterou vědci poskytují světu, aniž by za ni očekávali platbu. Primárně tato kategorie zahrnuje recenzované časopisecké články; patří sem ale i nerecenzované preprinty, které vědci mohou chtít nabídnout online pro připomínkování nebo jako upozornění kolegům na důležité výzkumné poznatky.
  Pojmem „otevřený přístup“ k této literatuře myslíme její volnou dostupnost na veřejném internetu umožňující libovolnému uživateli číst, stahovat, kopírovat, distribuovat, tisknout, prohledávat nebo vytvářet odkazy na plné texty těchto článků, sklízet je pro potřeby indexace, předávat je jako data pro software, nebo používat je k jakýmkoliv jiným legálním účelům bez finančních, právních nebo technických omezení s výjimkou těch, která jsou neoddělitelnou součástí získání přístupu k internetu samotnému. Jediným omezením na reprodukci a distribuci a jediným uplatněním autorsko-právní ochrany v této oblasti by mělo být poskytnout autorům kontrolu nad integritou jejich prací a právo na řádné uznání a uvedení autorství.“
- Idea jasná, ale cesta k realizaci dlouhá a komplikovaná
- Přes 40.000 vědeckých časopisů (uzavřené / v transformaci / OA)
- V současnosti jen kolem 30 % článků vychází v OA

1.4 Open Access

- **Volná dostupnost kvalitních vědeckých publikací v e-podobě**
- **Kdo zaplatí náklady na publikování**?
  (redakce, peer-review, vydání, distribuce, archivace)
  - **Gold OA**
    - Platí autor (APC – Article Processing Charge)
    - Čtenáři bezplatný přístup a neomezené využití ihned po vydání
  - **Green OA**
    - Finální publikace platí čtenář (předplatné)
    - Pracovní verze publikace zveřejní autor k bezplatnému přístupu v repozitáři (př. Arxiv.org, repozitář MUNI, Zenodo)
  - **Platinum OA**
    - Platí vydavatel (výzkumné organizace, učená společnost)
    - Autoři i čtenáři zdarma

MUNI ICS

Výzkumná data – FAIR

1.5 Výzkumná data

Proč je důležité otevírat nejen publikace ale i výzkumná data?

- Ověření správnosti výsledků
  - kontrola (nesprávné postupy, pominutí nevhodných dat, falšování)
- Reproducibilita vědy
  - možnost opakovat experiment a porovnat míru shody výsledků
- Znovuvyužití dat
  - úspora (neopakovat stejné drahé experimenty)
  - jedinečnost (data, které již nelze nikdy získat)
  - využití nepoužitých dat (snímek širšího okolí sledované hvězdy)
  - využití existujících dat v novém kontextu a pro nové účely
- Urychlení inovačního cyklu, přístup veřejnosti, …

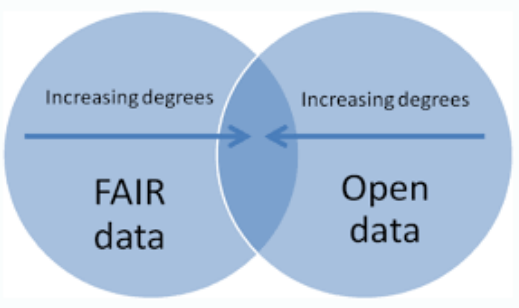
1.5 Specifika/složitost výzkumných dat

- Nelze požadovat okamžitý přístup (právo prvního využití)
- Nelze vždy otevřít (citlivé osobní nebo komerční údaje)
- Velmi velký rozsah (i TB, tisíce souborů)
- Velká variabilita formátů a forem (často netextové)
- Rozdílné oborové standardy (pokud vůbec existují)
- Různé třídy dat: Raw data – Zpracovaná data – Analyzovaná data
- Velká pracnost se zpřístupněním dat někomu jinému (uspořádání, popis)
- Málo prozkoumaná oblast
  - důvěryhodnost, úplnost, kvalita, vlastnictví, dlouhodobé uchování, kurátorství, …
  - ocenění akademickou komunitou?

Soudobý trend: pojďme data otevírat! (i přes tu velkou složitost)
Projekt EOSC-CZ (OP JAK 2023-2028, ÚVT MU)

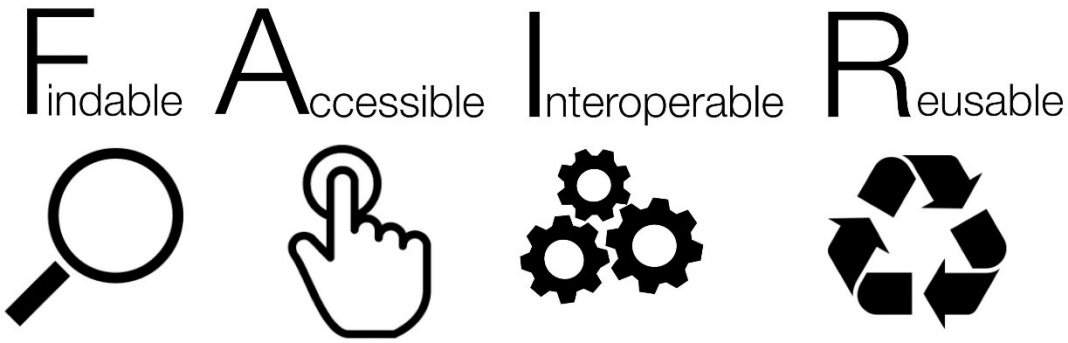
MUNI ICS

1.5 Jak data otevírat – FAIR



„As Open as Possible, As Closed as Neccessary“

- **Findable** – **nalezitelná** (globální identifikátory, registrovaná metadata)
- **Accessible** – **dostupná** (MD i data srozumitelná lidem i strojům, důvěryhodný repozitář
- **Interoperable** – **interoperabilní** (zavedené otevřené standardy, strojově zpracovatelná)
- **Reusable** – **znovupoužitelná** (jasná licence, přesná data o původu, reproducibilita)



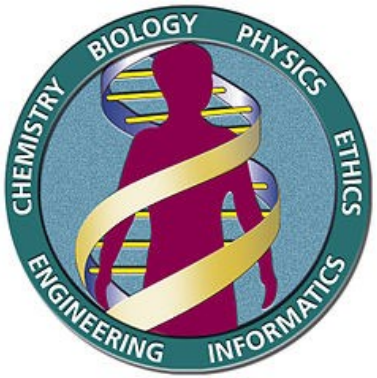
MUNI ICS

Za tajemstvím lidského genomu

https://www.stoplusjednicka.cz/cesta-za-tajemstvim-lidskeho-genomu

MUNI ICS

1.6 Human Genome Project (HGP)



- **Projekt mapování lidského genomu**
- Přečíst kompletní genetickou informaci (DNA) člověka!
  (sekvence **3,1 mld nukleotidů**, A,G,C,T)
- 1990-2003, ambiciózní, srovnáván s projekty Manhattan, Apollo
- 3 mld USD, mezinárodní (20 laboratoří z **USA**, UK, JP, FR, DE, CH)
- Obrovský objem prací, nezvládnutelný bez nových postupů a technologií
- Získaná data a technologie otevřít komukoliv
- Očekáván obrovský přínos pro medicínu, genetiku, molekulární biologii, další
- Etické, společenské a právní otázky
- Základní metoda: **sekvenace DNA**
  (vynalezena teprve pár let před projektem)

MUNI ICS

1.6 Sekvenace DNA

Adenin
Cytosin
Guanin
Thymin

- Zjišťování pořadí nukleových bází („písmen“ A,C,G,T) v krátkých sekvencích DNA pomocí biochemických metod a počítačového zpracování
- Sangerova metoda sekvenování (1977)
  - Část DNA rozsekáme na malé úseky (tisíce písmen), ty přečteme a seskládáme
  - Zjednodušený příklad (dle Storchová Z: Homo sapiens sapiens: přečteno! Vesmír 97, 2000/8, 427-429)
    - Chcete přečíst větu „Tak dlouho se chodí se džbánem pro vodu, až se ucho utrhne.“
    - Neznáte ale jazyk, takže se nemůžete domýšlet, a umíte přečíst vždy jen pár znaků
    - Celý text rozdělíte na malé úseky, náhodně. Získáte např. toto:
    - Na počítači vyhledáte překrývající se úseky (např. „**trhn**“ a „**hne.**“)
    - Seřadíte z toho kratší části a nakonec celou větu

e chod	louho	ž se u
vod	o se	hodí
e džb	trhn	í se d
trhn	hne.	ak dl
u, a	až	s dlou
vod	ro vo	u, a

  - **U lidského genomu má ta věta celkem 3.1 miliardy znaků!**

(pokud bychom ji přepsali do běžného textu knihy A4, dostaneme sloupec knih 30 metrů vysoký!)

MUNI ICS

1.6 Good guys vs Bad boys!

- Postup prací HGP byl velmi zdlouhavý, pomalý
- 1998: Craig Venter odešel z projektu a založil komerční firmu Celera Genomics
- Cíl: **předběhnout HGP, získat patenty na geny a prodávat je zájemcům!** (financování od farmaceutických firem, soukromých investorů)
- Nové zjednodušené postupy sekvenace – ne tak přesné, ale rychlejší
- Obrovská rivalita (nepřátelství) a soutěžení mezi oběma týmy
- Remíza: 2000 zveřejnili společně pracovní verzi genomu (finální 2003) (de-facto porážka firmy Celera)

- 2013: Nejvyšší soud USA: DNA je produktem přírody a nelze ji patentovat!

1.6 Výsledky a dopady HGP

- **Velký úspěch**: zmapován s vysokou přesností kompletní genom člověka!
  (historie genetiky rozdělena na „před“ a „po“)
- Vznik celých nových vědních oborů (bioinformatika, computational genomics, …)
- Rozsáhlé veřejně přístupné databáze a genové banky (GenBank, …)
- Obrovské zrychlení a zlevnění sekvenování (dnes celý člověk za pár hodin a pár set USD)
- Pokroky ve zpracování velkých objemů dat (dnešní sekvenátory TB dat/den, viz CEITEC-MU)
- Sekvenovány genomy velkého množství organismů (i vyhynulých – neandrtálec, mamut)
- Rozvoj poznání v mnoha oblastech (původ a vývoj druhů, migrace, …)
- **Ale:**
  - Genomu dosud až tak nerozumíme, je mnohem komplikovanější, než jsme si mysleli
    (Pačes: „Jsme na tom stejně, jako bychom přečetli celou knihu v portugalštině a neuměli portugalsky“)
  - „Odpadní“ část DNA (nekóduje geny, 98 % DNA) hraje mnohem větší roli, než jsme si mysleli
  - Některá očekávání se zatím nenaplnila, nebo jen z části (personalizovaná medicína)

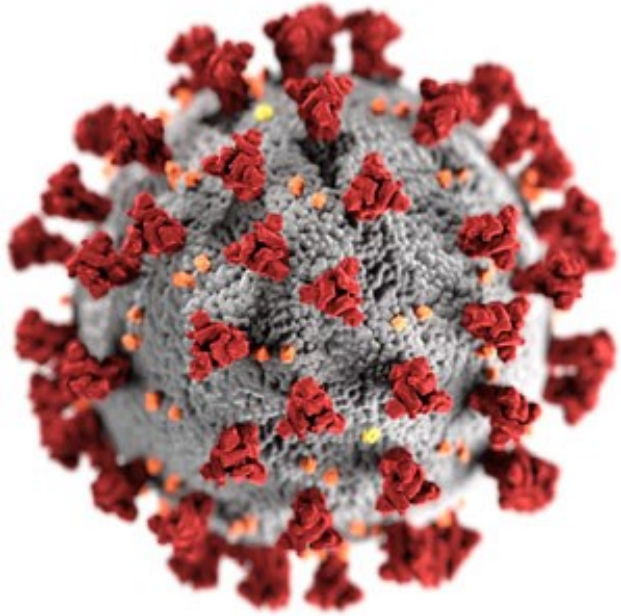
MUNI ICS

1.6 Etické a společenské otázky

- Ochrana (vysoce citlivých) osobních údajů
- Patentování (uzavírání) informací
- Psychologické aspekty
- Genetické inženýrství (dítě na přání)
- Eugenika
- Dostupnost benefitů jen pro někoho (bohaté)

1.7 Od HGP po COVID-19

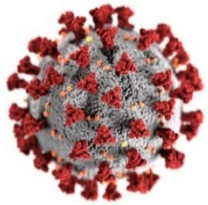
– 11. ledna 2020. Čínské úřady sledují ohnisko nového respiračního onemocnění ve Wu-chanu. Světová veřejnost o něm ví týden a věří, že co v Číně začalo, to v Číně taky skončí.



Boček Jan. *‚A teď se vám doktor podívá do šroubovice.‘ Pandemie urychlila nástup genomiky a medicíny na míru*. iROZHLAS.CZ, 26.4.2021. https://www.irozhlas.cz/veda-technologie/veda/vime-co-bude-po-covidu_2104261215_jab

MUNI ICS

1.7 Od HGP po COVID-19



- 11. ledna 2020. Čínské úřady sledují ohnisko nového respiračního onemocnění ve Wu-chanu.
- Ve stejný okamžik už vědci znají genetickou strukturu viru.
- Neznámá biotechnologická firma Moderna začala s US úřady okamžitě plánovat, co dál.
- O dva dny později byla vakcína hotová (mRNA). Zbytek roku zabraly klinické testy.
- 22. února 2021. Do Wu-chanu přijíždí delegace WHO. Mezi cíli mise je určit původ viru.
- Výsledek pátrání: virus se na člověka přenesl dříve, než se předpokládalo.
- Stejná studie také upozornila, že přenos ze zvířete na člověka vůbec není tak vzácný.
- Ještě před dekádou by vývoj vakcíny i pátrání po původu pandemie vypadaly úplně jinak.
- Zdravotníci připravují systém včasného varování. Cílem je zachytit nové hrozby do týdně.
- Hrozba nové pandemie je trvalá. Co můžeme udělat, je pokusit se tu příští zachytit v několikatýdenním okně před nekontrolovaným rozšířením do světa.
- „Pandemie urychlila nástup genomiky při analýze nakažlivých nemocí o několik let“ (Francis deSouza, prezident firmy Illumina, největšího výrobce sekvenátorů). Covid přinesl éru levného a rychlého sekvenování.

MUNI ICS

MUNI ICS

2.

Dlouhodobé uchování digitálních informací

Uchování digi-info / Digital Preservation

- Uchování a předávání informací = podmínka rozvoje civilizace
- Digitální technologie jsou fajn, ale přináší jeden velký problém:
  - **Dokážeme zajistit uchování/předávání digitální informace napříč věky?**

2.1 Uchování informací

- **Uchovávání informací = důležitý úkol společnosti**
- Klasické „paměťové“ instituce
  - **muzea** (fyzické artefakty)
  - **archivy** (nepublikovaný materiál)
  - **knihovny** (publikovaný materiál)

Alexandrijská knihovna
(295 př.n.l. – 642 ??)

- Základní přístupy k uchování materiálu
  - konzervace
    uchovávání původního artefaktu
    (metoda: obnovování – refreshing)
  - uchování
    uchování informačního obsahu původního artefaktu, i při zániku originálu
    (metoda: migrace)

MUNI ICS

2.1 Uchování informací

Národní archiv
Praha, Chodov

https://www.nacr.cz/o-nas/historie

Moravský zemský archiv
Brno, Bohunice

https://www.asb-portal.cz/wp-content/uploads/images/fotogaleria/

MUNI ICS

2.1 Uchování informací

Národní knihovna ČR
Praha, Klementinum
(Barokní sál)

Moravská zemská knihovna
Brno, Kounicova ul.

Foto : varadikamen.cz

MUNI ICS

2.2 Nosiče informací a jejich životnost

Rozdílné zkušenosti

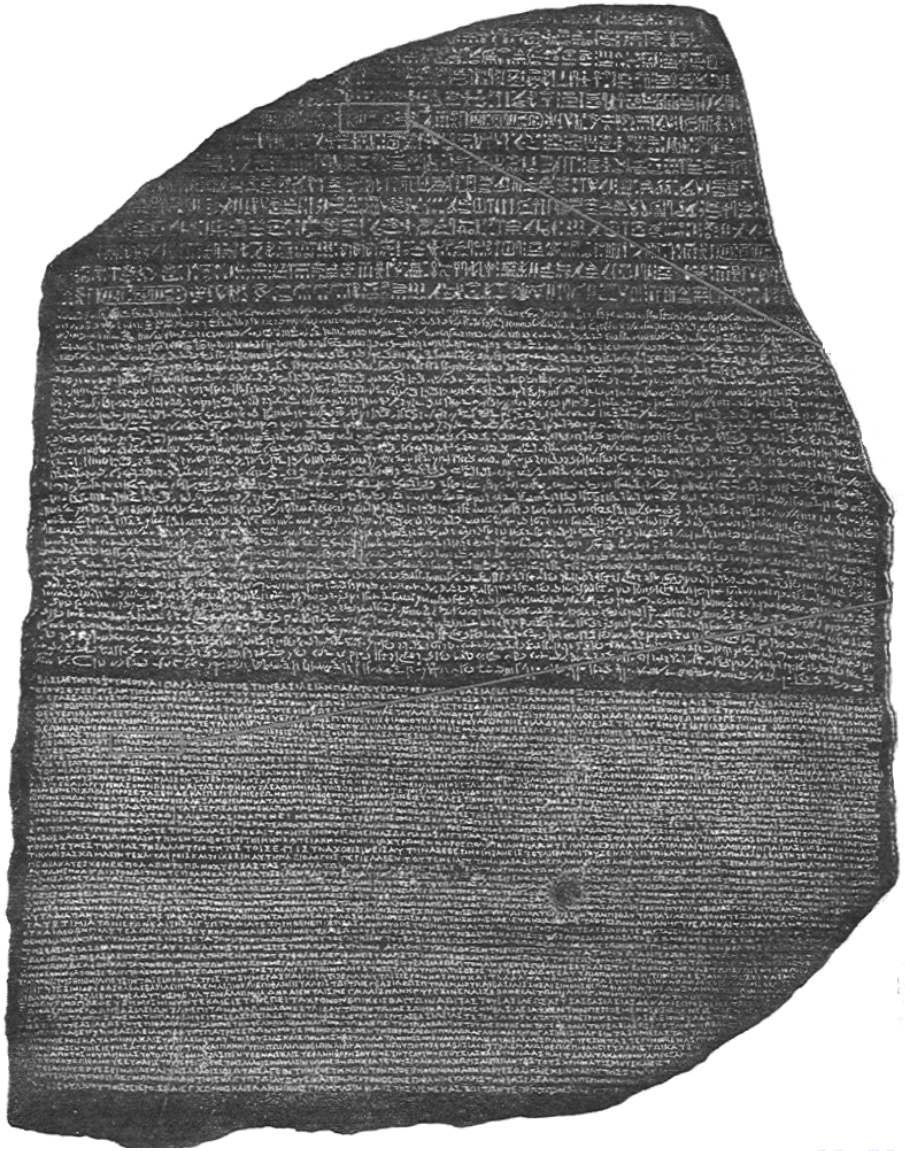
- starověké záznamy (kosti, kámen, hliněné tabulky, papyrus, pergamen, papír) — ~ 4000 let
- fotografické dokumenty (fotografické desky, film, fotopapír) — ~ 200 let (od 1839)
- audiovizuální záznamy (voskové a celuloidové válečky, šelakové desky, LP-desky, magnetické pásky) — ~ 100 let
- elektronické dokumenty (magnetický záznam, optický záznam, SSD) — ~ desítky let

! **Novější nosiče: větší kapacita, ale obvykle kratší životnost!**

MUNI ICS

2.2 Rosettská deska

- Objevena 1799 během Napoleonova tažení do Egypta
- Žulová stéla 114 x 71 cm
- **Text z r. 196 př. n. l.** ve třech různých zápisech:
  - egyptské hieroglyfy
  - egyptské démotické písmo
  - starořečtina
- Champollion: rozluštění hieroglyfů
- Londýn, British Museum



MUNI ICS

2.2 Hodně se dochovalo…

- Jeskynní kresby (Lascaux, Francie, 16 000 let)
- Babylónské hliněné destičky (4 000 let)
- Svitky od Mrtvého moře (cca 2 000 let)
- Starověké rukopisy a staré tisky (inkunábule)
- Antické písemnosti, …
  - Ne vždy se dochoval originál (médium vs informace) (přepisy řeckých děl v klášterech – běh proti času)

V ČR:

- Vyšehradský kodex (cca 1085, korunovace Vratislava II.) (nejstarší psaná památka, součást „korunovačních klenotů“)

MUNI ICS

2.2 …ale hodně také navěky ztratilo

- Originální rukopisy řeckých učenců, různý starobylý materiál…

Ale i novodobé dokumenty:

- Značná část novin na kyselém papíru (konec 19. a poč. 20.st.)
- 50 % filmů ze 40. let
- Marvin Minsky (AI, 60. léta) versus Galileo Galilei (16 st.)
- Originální videozáznam z přistání Apollo 11 na Měsíci (19.7.1969)
- 20 % NASA Viking (první průzkum Marsu, 1976)
- První email (1971), obsah první webové stránky (1990)
- …a mnohé, mnohé další…

! **Křehkost nosiče, chyby/opomenutí, vandalství, přírodní katastrofy, závislost na technologiích, chybějící systém/infrastruktura**

Digitální informace

2.3 Digi-info: Optimismus – a vystřízlivění

– **Digital information is forever. It doesn't deteriorate and requires little in the way of material media.**

Andy Grove, Intel Corp.

– **Digital information lasts forever – or five years, whichever comes first.**

Jeff Rothenberg, RAND, 1995

2.3 V čem je problém s digitální informací?

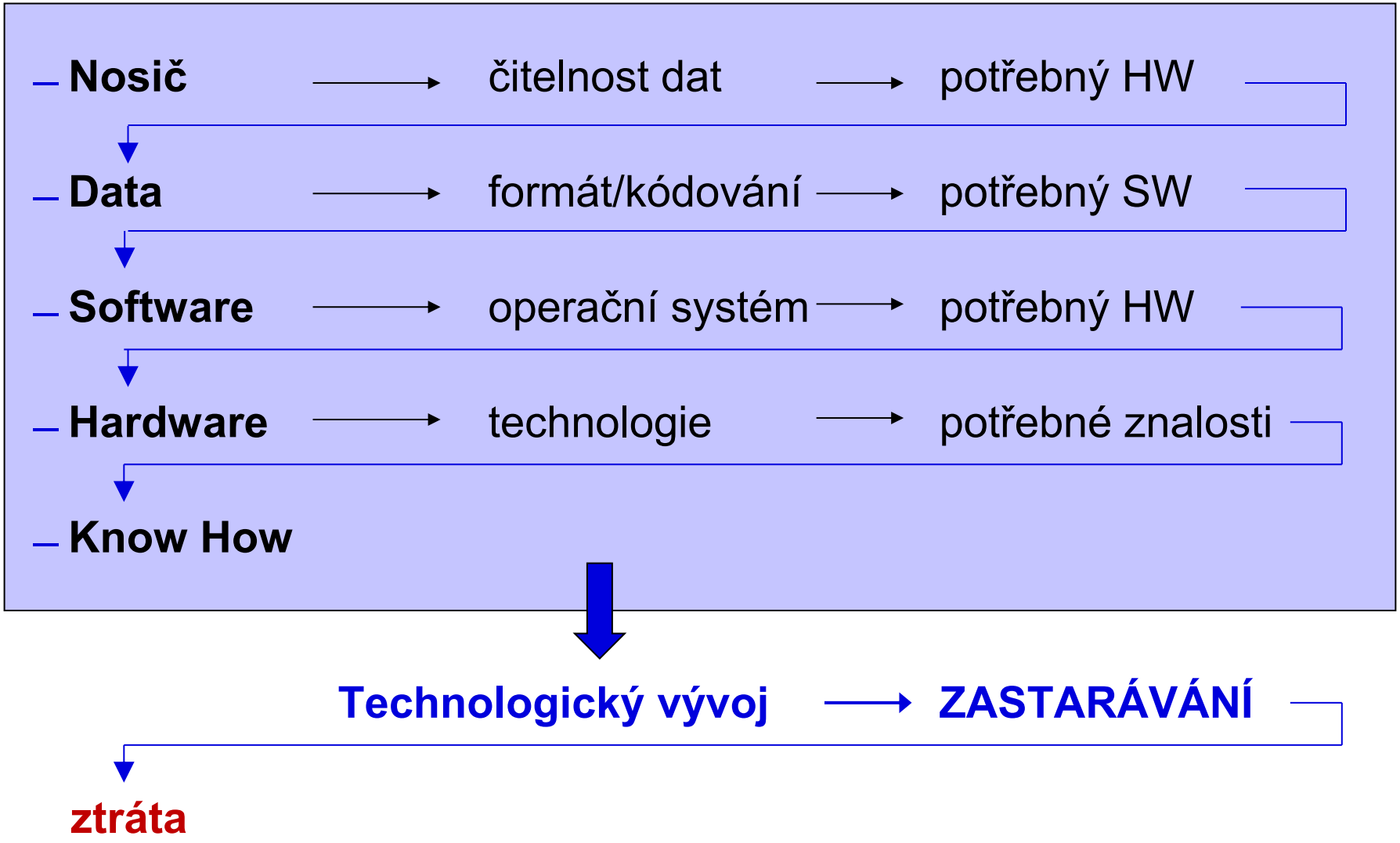
- Přístup a zobrazení digitální informace jsou závislé na technologiích
  (nestačí k tomu lidské smysly)

- **Hrozby pro digitální informaci**

  1. **Křehkost záznamového média**
     (krátká životnost, nízká odolnost vůči změnám)
  2. **Technologické zastarávání**
     (platformní závislost – nosič, formát, software, hardware)

- Další
  - Velký (trvale rostoucí) objem digitálních informací
  - Finanční nákladnost údržby
  - Nezbytné expertní znalosti

MUNI ICS


Nosič
čitelnost dat
potřebný HW
Data
formát/kódování
potřebný SW
Software
operační systém
potřebný HW
Hardware
technologie
potřebné znalosti
Know How
Technologický vývoj
ZASTARÁVÁNÍ
ztráta

2.4 Co je „Digital Preservation“

DP Handbook

– **Digital Preservation**
The goal of digital preservation is the accurate rendering of authenticated content over time.

Digitální uchování kombinuje postupy, strategie a akce zajišťující přesnou reprodukci ověřeného obsahu v průběhu času, a to s ohledem na případná selhání záznamových médií a na probíhající technologické změny.

– Dvě úrovně
- **Bit-level preservation** (dostupnost digitálních dat)
- **Logical preservation** (technologické změny + porozumění obsahu)

Terminologie: Digital Preservation (DP), **LongTerm Preservation** (LTP)

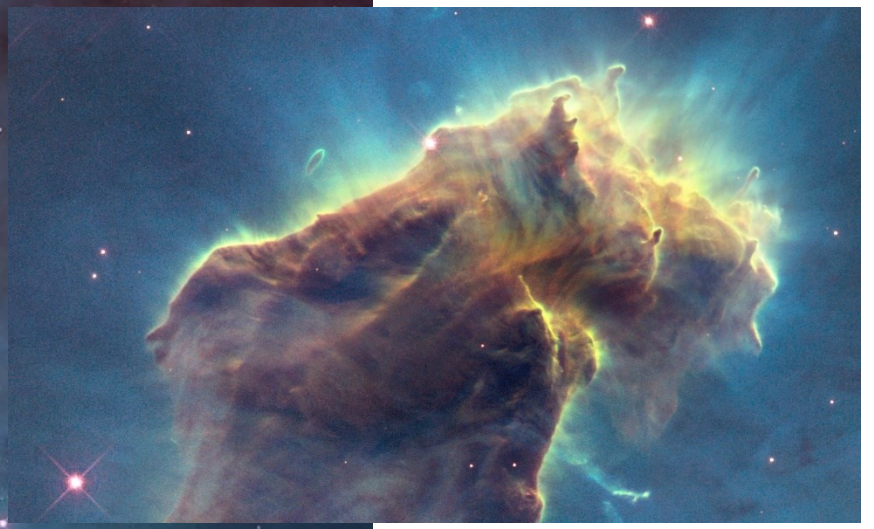
MUNI ICS

2.4 Ochranné strategie

a) **Nosič** (uchování digitálních dat – bitová ochrana)
   - **oživování a replikace**
   - **nový formát**

b) **Informace** (digitální obsah a jeho význam – logická ochrana)
   - **uchování technologického prostředí**
     - technologické muzeum
     - emulace
   - **překonání technologické zastaralosti**
     - migrace
     - encapsulation

! V praxi: kombinace přístupů (+ digitální archeologie, když vše selže)



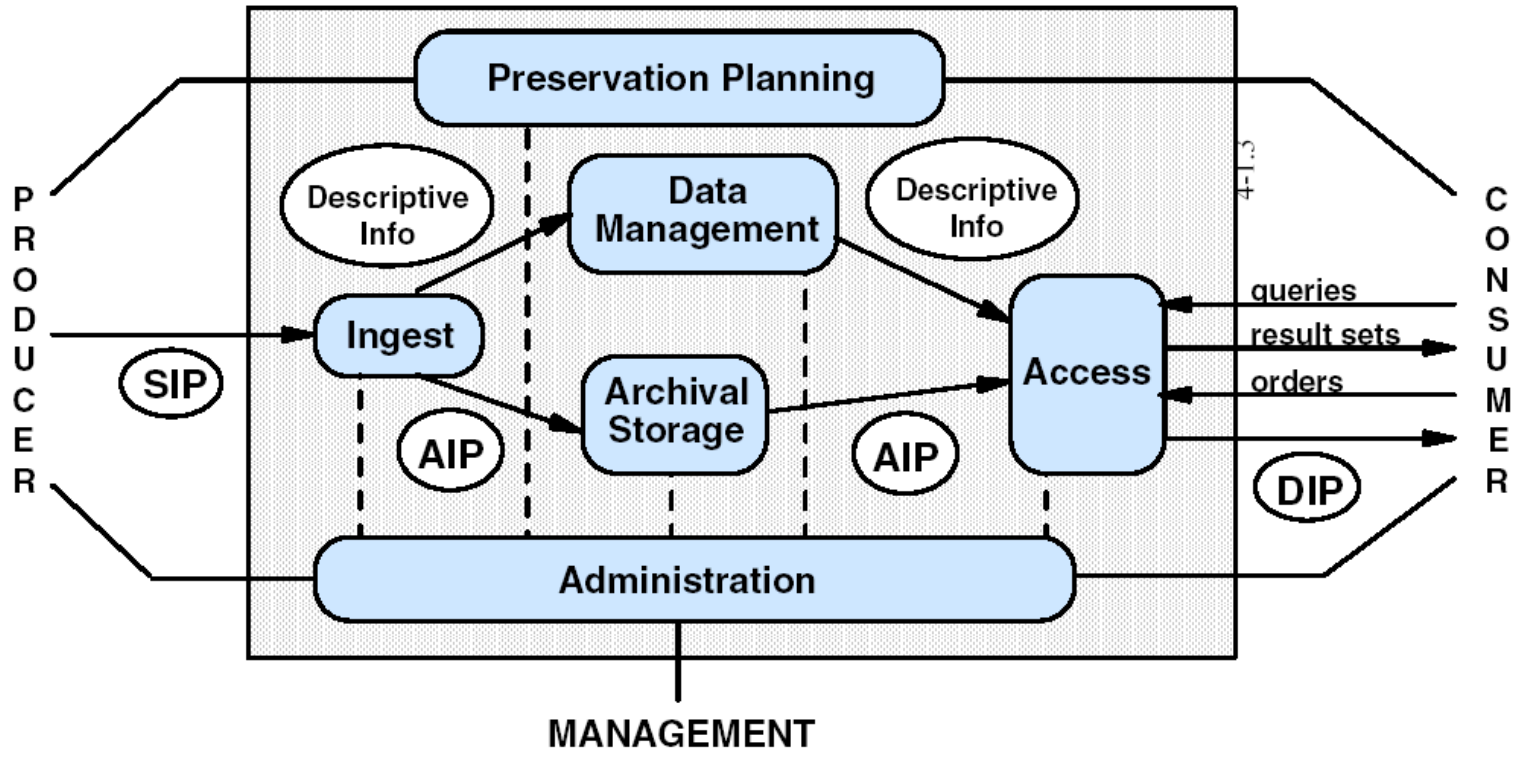
Sloupy stvoření
Hubbleův teleskop, 1995

MUNI
ICS

2.5 OAIS model

- Digitální uchování je složité – jak to řešit „systémově“?
- Koncepce standardizovaného „Digitálního archivu“
- Podnět od kosmických agentur (NASA, ESA, …)
  (obrovské objemy dat, spousta negativních zkušeností)
- 2002: **OAIS** – Open Archival Information System
  - Referenční model pro dlouhodobý Digitální archiv
  - ISO standard (od 2014 i ČSN)
- Pro naše potřeby příliš odborné/obsáhlé
- …ale pár postřehů i pro osobní inspiraci…

2.5 OAIS model



- **SIP** – Submittion Information Package
- **AIP** – Archival Information Package
- **DIP** – Dissemination Information Package

MUNI ICS

2.6 Osobní digitální archivace

- Má smysl starat se „rozumně“ o svá osobní digitální data
- Library of Congress: https://digitalpreservation.gov/personalarchiving/
- **Pár tipů:**
  - Zálohovat: alespoň 2 zálohy na separátních médiích
  - Zálohy uchovávat na vzdálených lokalitách
  - Média označit a držet v bezpečných místech (jako důležité dokumenty)
  - Namátkově ověřit čitelnost médií
  - Alespoň každých 5 let vytvořit nová záložní média
  - Systematicky roztřídit své osobní sbírky (fotografie, audio, video, e-mail, osobní)
  - Vybrat nejdůležitější materiály, rozumně popsat (jména souborů, metadata)
  - Používat rozšířené (otevřené) formáty, důležitá data migrovat na nové

! **Další? Osobní zkušenosti?**

2.7 Chmurné perspektivy?

- Velké množství digitálních informací již dnes nenávratně ztraceno
- Trvale roste množství informací existující pouze v digitální podobě
- Stále se rozšiřuje množství formátů dokumentů a médií
- Informační technologie zastarávají velmi rychle
- Křehkost nosičů digitálních záznamů
- Při vytváření digitálních zdrojů není počítáno s náklady na archivaci
- Snižování rozpočtů pro knihovny a archivy
- Nesmyslné ochranářské trendy omezující dostupnost info (právo být zapomenut)

„There is, at present, no way to guarantee the preservation of digital information. The first line of defense against loss of valuable digital information rests with the creators, providers, and owner of that information. It's every man for himself."

U.S. Commission on Preservation and Access.
Final report of a Task Force on the Archiving of Digital Information. 1996

MUNI ICS

2.7 Chmurné perspektivy?

Analog objects can survive with minimal care for centuries, but no electronic format can hope to persist more than a short while without careful (and perhaps expensive) intervention. There will be no digital equivalent of the Lascaux cave paintings, Mayan stone scripts, Dead Sea scrolls, or other kinds of rediscovered ancient knowledge.

Building Preservation Partnership. The LoC NDIIP Program

MUNI ICS

MUNI
ICS

Diskuse