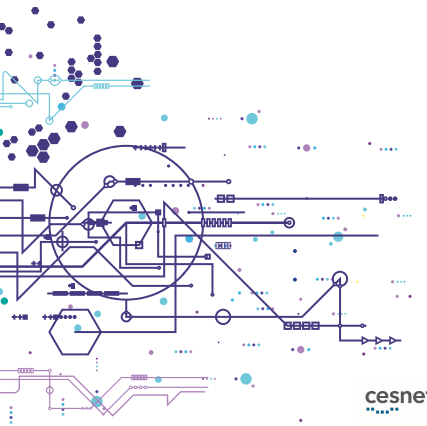
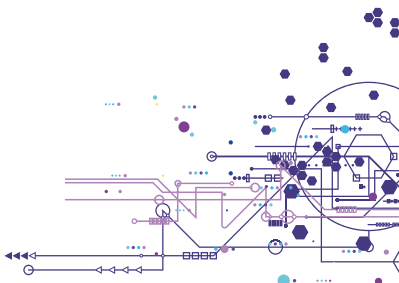
NOVINKY V DATOVÝCH ÚLOŽIŠTÍCH

David Antoš

10. 5. 2022





O čem to bude

- v infrastruktuře máme
  - úložiště přímo připojená k výpočetním zdrojům
  - *úložiště nezávislá na výpočetních zdrojích*
    - mj. dlouhodobé úložiště pro uživatele MetaCentra a IT4I
- stále platí
  - úložiště CESNET jsou umístěna v ČR
  - jsou provozována v rámci komunity
  - data patří výhradně vlastníkům

Případy užití

- nestrukturovaná data
  - předávání souborů
  - sync'n'share
  - zálohy
  - archivace
  - sdílení dat
- dlouhodobé garantované úložiště
- dlouhodobé uchovávání FAIR vědeckých dat

Služby úložišť

- jednorázové zasílání souborů
  - FileSender, `https://filesender.cesnet.cz`
  - úschovna pro předání, až 500 GB
- sync'n'share (Enterprise File Synchronisation and Sharing)
  - ownCloud, `https://owncloud.cesnet.cz`
  - synchronizace na server
  - sdílení konkrétnímu uživateli/linkem
  - klienti pro mobilní platformy
  - nástroje pro federování sync'n'share systémů
    - Science Mesh
    - navázání důvěry mezi uživateli „na pozvání“

Přístup k souborům

- NFSv4, rsync, scp, FTPS, CIFS, Globus
- hierarchické úložiště du4
- diskové pole du5
  - oboje na konci životnosti

Zřízení přístupu

- uživatelé MetaCentra mají připojené du4
  - a přímo přístupný prostor na něm pro archivy
  - nepoužívejte přímo pro výpočty
- NEBO:
- pokud ukládáte jen individuální data
  - v řádu jednotek TB
  - na souborovém systému
    - přesune se na S3
  - stačí se jen zaregistrovat na `https://du.cesnet.cz`
  - vyžaduje se ověření uživatele z akademické instituce
  - členství se po roce prodlužuje
  - VO Storage

Správa uživatelů pro náročnější

- NEBO:
- nestačí VO Storage? založíme vám virtuální organizaci!
- VO je skupina uživatelů se společným zájmem, kteří vystupují jako celek
- VO má správce, který
  - s námi domlouvá parametry úložiště
  - řídí přístup uživatelů k němu
- dohoda o poskytování zdrojů mezi VO a správcem zdroje
  - popisuje nastavení technických parametrů
- pro souborové systémy i objektová úložiště

Objektová úložiště

- Ceph clustery
- poskytujeme
  - blokové zařízení RBD (spíše pro zálohování)
  - ukládání objektů do S3 (univerzální)
- přístup pomocí tokenů
- klienti pro operační systémy
- nad RBD lze vytvářet LUKS kontejnery
- v přípravě: web pro řízení přístupových tokenů

Komunitní budování infrastruktury

- zapojování zdrojů uživatelských skupin do e-infrastruktury
- podobně jako v MetaCentru
- předpokládá se objektové úložiště jako technologický základ
- pod jednotnou správou
- uživatelům bezprostředně přinese
  - zjednodušení správy
  - off-site repliku dat
  - primární pracovní replika dat zůstává lokální
- pokud plánujete data ve vyšších stovkách TB, ozvěte se

Stěhování

- souborové systémy na HSM (du4)
  - plánovaná životnost do konce 2022
- stěhování
  - na objektové úložiště
  - MetaCentrum: zařídíme
  - VO Storage: zařídíme
  - ostatní VO: budeme postupně oslovovat správce VO
- zálohy se typicky jen pošlou jinam
- archivy je třeba přestěhovat
  - sada návodů, support
- následovat bude vyřazení pole v Jihlavě (du5)

Další objektová úložiště

- 2022: MENDELU
- 2023: ELI
- další financování z OP JAK
  - e-infrastruktura
  - implementace EOSC

Dlouhodobá garantovaná úložiště

- repozitáře pro garantované dlouhodobé uložení dat
  - binárně spolehlivé úložiště
  - které dostane archivní informační balíček (OAIS)
  - bude ho periodicky kontrolovat
  - to vše je třeba zapsat do auditních zpráv
  - a umět reagovat na nalezení chyby dat
    - opravit z jiné repliky
- možné pojištění na ztrátu dat (nadstandard)
- jako spolehlivé úložiště pod plný OAIS repozitář/spisovnu/…

Implementace EOSC

- významný vývoj standardů vědeckého publikování ve vztahu k datům
  - standardní ukládání souborů a objektů postačuje pro „provozní data“
  - dokumentované dostupné datové sady začínají být nutností
- požadavky na vědce vs. možnosti infrastruktury
  - co je dostatečně univerzální služba?

Hnutí FAIR dat

- zformulováno 2014
- (téměř současně se formoval ELIXIR)
- TL;DR průlet FAIR:
  - Findable: data mají PID a prohledávatelná metadata
  - Accessible: dostupné standardním protokolem
  - Interoperable: formální jazyk pro reprezentaci znalostí, slovníky, reference
  - Reusable: jasná licence, jasný původ dat, komunitní standardy
- podrobně `https://www.force11.org/group/fairgroup/fairprinciples`
  - s pasážemi pro fanoušky ontologií a filosofování o reprezentaci znalostí
- za „daty“ ve FAIR lze vidět i algoritmy, nástroje a workflow

Co FAIR je a není

- FAIR jsou stručné doménově i technologicky nezávislé principy
- FAIR není standard
- FAIR $\not\Rightarrow$ veřejně přístupný
- FAIR $\not\Rightarrow$ dostupný zdarma
  - jen má být jasné, jak se věci mají
- FAIR je svatý grál
  - „nad čím bychom se měli zamyslet“
  - ne „vše z toho je třeba splnit dokonale“
- FAIR je proces
  - level of FAIRness, FAIRification

Technická implementace

- co potřebujeme pro implementaci?
- minimalisticky vystačíme s persistentními identifikátory
  - to je zvládnutá technologie
- a katalogy metadat
  - s ontologiemi a jiným spiritismem
- aplikace řídící workflow se hodí
- uchopitelný technický základ pro ukládání dat: repozitář

Repozitář

- úložiště dat [případně publikací]
  - ve více či méně kontrolovaných formátech
- opatřených metadaty včetně persistentních identifikátorů
- umožňující podle metadat vyhledávat
- obvykle přístupný přes webové rozhraní a API
- obvykle propaguje metadata do agregátorů a vyhledávačů
- obvykle s řízením životního cyklu záznamu a přístupových práv

Architektura EOSC v ČR

- (technocentrický pohled)
- vytvoření a rozvoj národní datové infrastruktury
  - prostředí pro ukládání, zpracování a zpřístupnění FAIR dat
  - repozitářová platforma
- základní metadatový adresář
  - pro vyhledatelnost
- navázané aktivity
  - národní sekretariát EOSC
  - podpora práce s daty: kurátoři/datoví vědci/jejich výchova
  - aktivity oborově-vědních clusterů

Národní repozitářová platforma

- virtuální úložiště s podporou tvorby tenantů/instancí
- možnost brandingu a samostatné správy tenantů
- možnost podpory specifických metadat
  - jak v metadatovém modelu, tak v prezentační vrstvě
- klíčová role řízení přístupu
  - plná kontrola původcem/vlastníkem dat
- předpokládáme, že za 7–8 let budou data v repozitářích převažovat

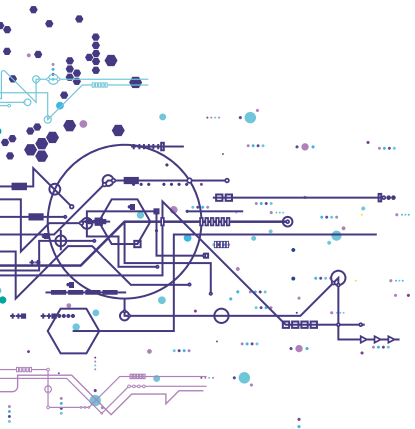
Co zatím máme

- prototypový obecný repozitář
  - `https://data.narodni-repozitar.cz/`
  - aktuálně
    - NTK jako kurátoři všech záznamů
  - cílový stav
    - sebeobslužné odborné kurátorské komunity
    - základ repozitářové platformy
    - jako long-tail tenant
- pracovní skupiny pro formulaci
  - architektury systému
  - zapojení klíčových institucí

Shrnutí

- sada služeb pro ukládání obecných neanotovaných dat
  - souborové systémy
  - objektová úložiště
  - sync'n'share
- služby pro anotovaná/FAIR data
  - prototypový repozitář
  - plány národní repozitářové platformy

e-INFRA CZ



Napište nám
info@einfra.cz

e-infra.cz