European Open Science Cloud (EOSC)

Luděk Matyska, e-INFRA CZ (CERIT-SC a CESNET)

EOSC – Co je to?

- „Pohádka“ o světě, ve kterém jsou výzkumná data
  - Pečlivě produkovaná (kontrola kvality při vzniku)
  - Pečlivě anotovaná (víme, co představují)
  - Pořádně uložená (víme, že se neztratí a budou s námi dlouho)
  - Dobře indexovaná (víme, že je najdeme)
  - Kontrolovaně přístupná (víme, že se k nim dostaneme, když smíme)
    - Pokud možno, není třeba žádná kontrola přístupu
  - Drží stopu původu (víme, kdo je vytvořil a jak)
- a je možné je kombinovat, analyzovat a dále s nimi pracovat
- **EOSC = Rychlejší a efektivnější cesta k novým vědeckým výsledkům**

EOSC – Proč to už nemáme?

- Péče o data je nákladná
  - Většina popsaných kroků není automatizována
  - Z pohledu „autora“ dat se většinou jedná o „zbytečné“ kroky
    - Občas ani kvalita není prvoplánová (srovnejte s „pilotním kódem programů“)
- A nikdo ji (zatím) nechce platit
  - Vědci jsou oceňování za články, ne přímo za to, že generují data
  - Dlouhodobá péče o data přesahuje konec projektů
- Není shoda na metadatech
  - Standardy jsou „pohyblivý terč“
- **Současná data jsou fragmentovaná a určitě ne FAIR**

EOSC – Co ještě chybí?

- Motivace vědců
  - Proč se mám starat o data, která pak použije někdo jiný?
  - Jak se ochráním proti tomu, že mi někdo „ukradne“ výsledky, protože data zpracuje rychleji?
  - Jak se mi vrátí investice do péče o data v mém kariérním růstu
    - Když pro postup potřebuji publikace a citace?
- Nedůvěra
  - Kde mám jistotu, že někým vystavená data budou kvalitní?
    - Junk publikace z Open Access prostoru jsou dobrým varováním
- **Není to celé jen trik IT komunity vytáhnout další peníze, které pak chybí v „pořádné“ vědě?**

EOSC – Co máme?

- Nadšení
  - Ta „pohádka“ na prvním slidu je docela podmanivá
  - Evropská komise není jediný, kdo „věří“ v Open Science
- Motivaci – Aktuální zkušenosti s COVID-19
  - Ukázalo se, že pro skutečně seriózní, cílený výzkum je sdílení dat naprosto kritické
  - A současné systémy jsou přinejlepším extrémně těžkopádné a spíše reálně nepoužitelné
    - Najednou se řeší i takové „drobnosti“, jak dostat na jedno místo výsledky z více výzkumných laboratoří, aby je bylo možné společně zpracovat
    - Různé formáty, různá úroveň „kvality“, různá struktura dat, různá pravidla zpřístupnění, …

EOSC a e-infrastruktury

- Včera zaznělo

EOSC is a data-infrastructure and could be seen as a twin sister/brother of the European e-infrastructure organizations

- Co si pod tím představit?
  - Podmiňovací způsob je podezřelý: může tomu tedy být i jinak?
  - V čem se to bratrství/sesterství má projevit?
- e-Infrastruktury jako **podvozek**?
  - EOSC mluví o „Marketplace“ s implicitním očekáváním zapojení komerčních poskytovatelů – budou schopny e-infrastruktury mít kompetitivní nabídku?
  - A co je vůbec „**kompetitivní nabídka**“?

EOSC a e-infrastruktury

- Péče o základní služby
  - Autentizační a autorizační infrastruktura
  - Permanentní identifikátory
  - …
- Kapacita pro data, ale i jejich zpracování
  - Dlouhodobé know-how práce v distribuovaném/federovaném prostředí
  - Nejasná role superpočítačových center
    - Stanou se primárními poskytovateli kapacity v kombinaci s komerčními nabídkami?
  - Centralizace versus distribuovaná federace
    - Na národní i EU (světové) úrovni
  - Vhodný formát dlouhodobé spolupráce se stále hledá

EOSC a „data“

- Co jsou „data“ pro EOSC
  - Jakékoliv digitální objekty
- Publikace – cíleny již v Open Access aktivitách (Plan S)
- Výzkumná data – výsledky měření, výpočtů, analýz, …
- Software a workflow – čím jsou data zpracovávána
- Něco dalšího?

- Zřetelná extrémní heterogenita digitálních objektů
  - O implementaci EOSCu se nemůže starat jen jeden subjekt, musí to být součást celkového vědeckého prostředí (=změna k Open Science)

EOSC a datová infrastruktura

- Co je to datová infrastruktura?
  - Fyzická úložiště
    - Až po Long Term Preservation úroveň
  - Metadatové adresáře
    - „Jednotný" popis dat umožní vyhledávat, ale má i celou řadu dalších funkcí
  - Permanentní identifikátory
    - Úroveň detailu (minimální adresovatelný objekt, struktura adres (rodné číslo), …)
  - Řízení přístupu
- Co z toho už dnes zvládnou e-infrastruktury?
  - Ale stejné znalosti/know-how mají i další subjekty, např. velké výzkumné (tematické) infrastruktury

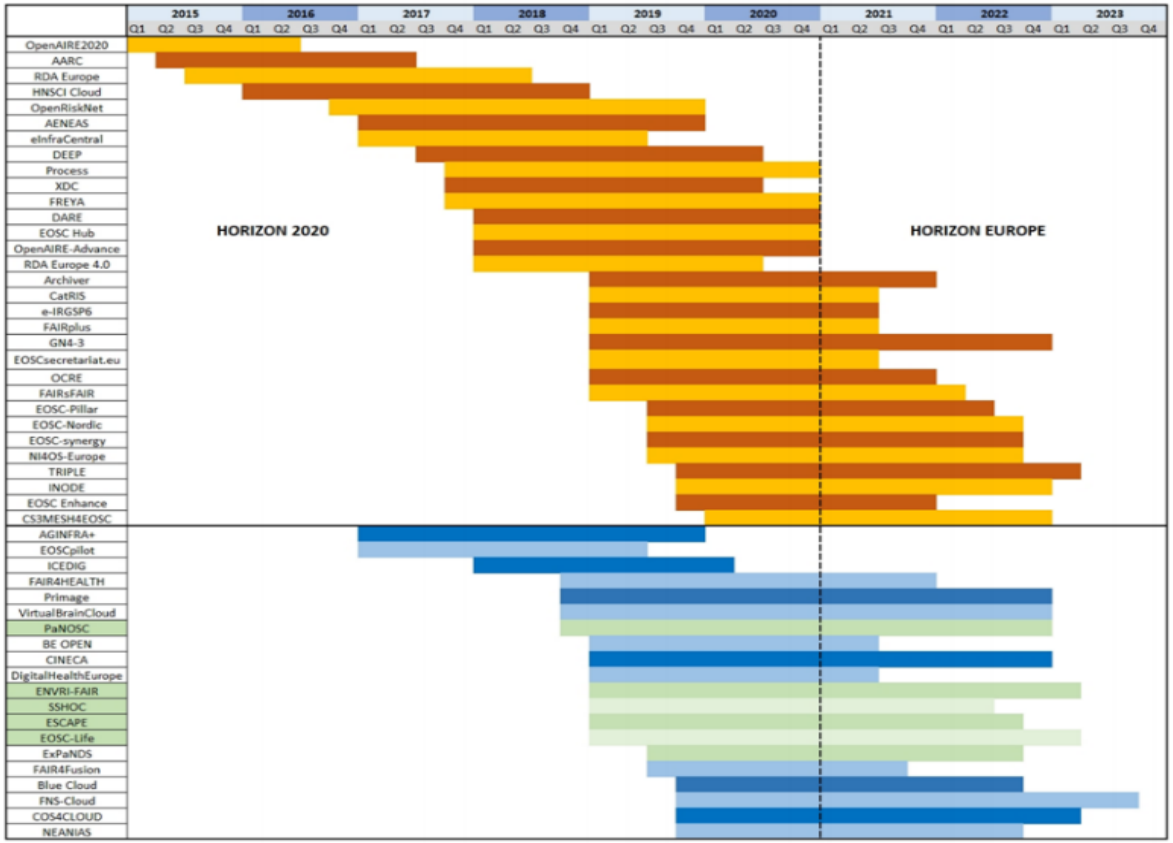
Architektura datové infrastruktury?

- Centralizované řešení
  - Nereálné – „Jeden nevládne všem“
  - Na druhé straně „přirozená“ cesta (viz Google)
- Plně distribuované nezávislé ostrůvky nejsou EOSC
  - Současná situace s vysokou mírou fragmentace
  - I kdyby se vyřešil samotný problém (dlouhodobého) uložení dat
- Distribuované federované uspořádání
  - Nezávislý vznik i rozvoj „ostrůvků“
  - Technologická interoperabilita
  - Sdílená metadatová vrstva
  - Sdílené základní služby (PID, AAI, …)

EOSC a e-INFRA CZ – nyní

- Formálně
  - Všechny tři součásti jsou členy EOSC AISBL
  - **CESNET je *mandated organization***
- E-INFRA CZ největší český partner v EOSC-related projektech, financovaných EK
  - Původně CESNET hlavním partnerem
  - Příprava budování zázemí EOSC
    - Včetně vývoje software
    - DEEP DataCloud, EOSC-Hub, EOSC SYNERGY, PaNOSC, nyní EGI ACE a EOSC FUTURE
  - Postupně zapojen i CERIT-SC/MU
    - Napojení na vědy o živé přírodě a lékařství
    - EOSC-Life, ConcePTION, CINECA
  - A zájem i ze strany IT4Innovations
    - Překlenovací technologie

EOSC a projekty EK



Key: Orange shades indicate coordination, best practice or technology projects. Blue shades indicate thematic science-based projects. Green shades indicate ESFRI cluster projects.

EOSC a e-INFRA CZ – nyní

- Technicky
  - MetaCentrum – distribuovaná výpočetní kapacita
    - Know-how v celé řadě oblastí
    - Cloudy, kontejnery, workflow, …
  - Datová úložiště
  - Vysokorychlostní počítačová síť pro efektivní přesun data
  - IT4Inoovations a další kapacita
- Spolupráce s infrastrukturami
  - Např. ELIXIR CZ; sensitive cloud, ELIXIR a BBMRI AAI,
- Specifické služby – **LS AAI**
- Prostřednictvím CERIT-SC/MU přímé napojení na OA (OpenAIRE AMKE)

EOSC a e-INFRA CZ – budoucnost?

- Explicitní budování národní datové infrastruktury
  - Jako partner, ne monopolista
- Odpovědnost za některé (klíčové, společné, …) elementy
  - AAI již nyní
  - Kdo se postará o PIDs?
- Existující repozitáře jsou zpravidla v péči jiných subjektů
  - Velké výzkumné infrastruktury, vysoké školy a další výzkumné instituce, …
- Budoucí architektura datové infrastruktury musí tyto části propojit
  - Role **metadatového adresáře**

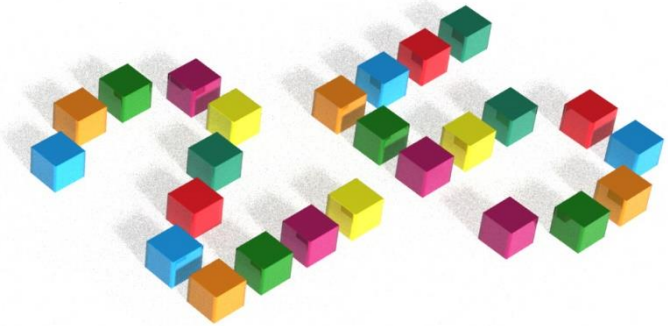
EOSC a lidé

- Péče o data není plně automatizovatelná
  - Datové kurátorství
- Data Scientists, Data Curators, …
  - Nové odbornosti
  - Pomezí správy dat a příslušných vědních oblastí
- Nové pracovní příležitosti
  - Komplexní správa dat musí jít s daty a jejich sémantikou, ne s technologiemi
- Zastřešující role centrálních institucí
  - Analogie role NTK u publikací (OA)

EOSC jako příležitost

- Pochopení hodnoty dat a dalších digitálních objektů
  - Řízená persistence
- Nové kariérní příležitosti
  - I úprava těch stávajících – ceněná péče o data
- Nové výzkumné výsledky
  - Větší a kvalitnější datová základna

- Posílení spolupráce výzkumných subjektů
  - „Přátelštější“ vědecké prostředí
  - Méně (umělé, neproduktivní) soutěže



Děkuji za pozornost!