

Analýza aktuálního stavu kapacit pro implementaci EOSC v ČR

Technologické centrum AV ČR

Projekt sdílených činností „Strategická inteligence pro
výzkum a inovace“ (MS2104)

Cíl a obsah studie

Cíl: Zmapování výzkumného terénu v ČR v oblasti připravenosti na implementaci iniciativy EOSC pro potřeby nastavení optimálního mixu podpory z výzev OP JAK.

Kap. 1 – Stručné shrnutí

Kap. 2 – Úvod

Kap. 3 – Evropský rámec EOSC a zapojení ČR

Kap. 4 – Rozvoj EOSC v EU

Kap. 5 – Předpoklady pro implementaci EOSC v ČR

Kap. 6 – Shrnutí a závěry

Kap. 7 – Přílohy

3) Evropský rámec EOSC a zapojení ČR

Evropské partnerství EOSC

Strategická výzkumná a inovační agenda EOSC (SRIA)

EOSC Tripartite Governance:

- EK
- zúčastněné země - EOSC Steering Board
- výzkumná komunita - EOSC Association a EOSC Association Task Forces

Čeští zástupci v EOSC Asociatin:

- CESNET, MU (CERIT-SC), VŠB-TUO (IT4Innovations), NTK
- Ústav fyzikální chemie J. Heyrovského AV ČR (jako pozorovatel)

Mezinárodní projekty (na bázi sdružování velkých výzkumných infrastruktur):

- např. EOSC-Life, PaNOSC, EOSC Synergy, EOSC Future, EOSC-Hub, EUROCC, EGI ACE, nově HPLT aj.

4) Rozvoj EOOSC v EU, situace a praxe ve vybraných zemích

Evropské partnerství EOOSC

- společné investice v období 2021-2027 1 mld. EUR

Země v pokročilé fázi:

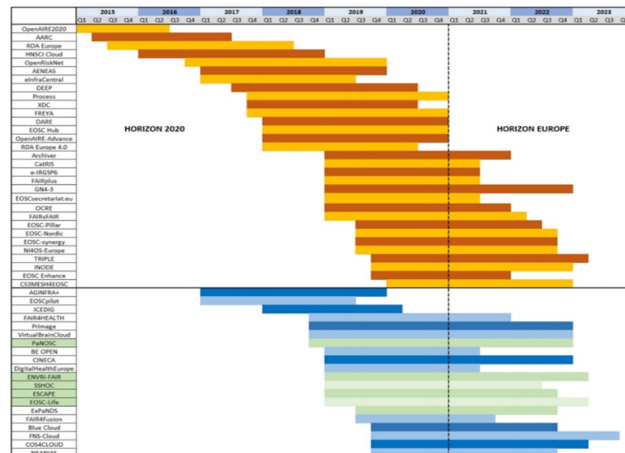
- Německo - 2021-2030 900 mil. EUR na EOOSC prostř. NFDI (Nationale Forschungsdateninfrastruktur)
- Nizozemsko - 2019 40 mil. EUR a každý další rok 20 mil. EUR na e-infrastruktury
- Francie

Země v počáteční fázi:

- Rakousko, Portugalsko, Švédsko

ČR je někde uprostřed

- Výhody: zkušenosti některých infrastruktur, rozvinutá e-infrastruktura, systematický přístup k budování EOOSC, ujasněné financování
- Nevýhody: malá informovanost většiny potenciálních uživatelů, nedostatek lidských zdrojů pro systematickou práci s daty



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

Zdroj: e-INFRA.cz

5.1) Průzkum vědeckých repozitářů v ČR (Bartošek a kol. 2020, ÚVT MÚ)

Naprostou většinu repozitářů bylo možné charakterizovat jako **repozitáře s převážně textovým obsahem** (37), uchování a zpřístupnění výzkumných dat se primárně týkalo pouze dvou repozitářů (Český sociálně vědní archiv a LINDAT/CLARIAH).

Jako nejčastější výzvy a problémy při provozu repozitářů byly identifikovány:

- nedostatek technického / odborného personálu či know-how (18),
- neujasněná či chybějící Open Access / FAIR data / Open Science politika (16),
- nedostatečné financování (11),
- nedostatečný zájem ze strany vedení instituce (5),
- nedostatečné technické zázemí instituce (4).

5.2) Přehled vybraných rozvinutých repozitářů a infrastruktur v ČR

- ELIXIR – biologické data
- LINDAT/CLARIAH – data z oblasti počítačového zpracování jazyka
- Český sociálněvědní datový archiv – sociálněvědní data
- GENASIS – data o kontaminaci životního prostředí chemickými látkami
- BBMRI – data o biobankách a biomolekulárních zdrojích
- Ústav molekulární a translační medicíny LF UP v Olomouci (IMTM) – medicínská data
- CERN Open Data Portal – data z oblasti částicové fyziky, otevřený repositář Zenodo

5.3) Rámec a aktuální stav přípravy NDI EOSC

Systémové projekty:

IPs EOSC-CZ - vytvoření společného zázemí pro implementaci datově orientovaných součástí strategie Open Science v ČR (eINFRA CZ)

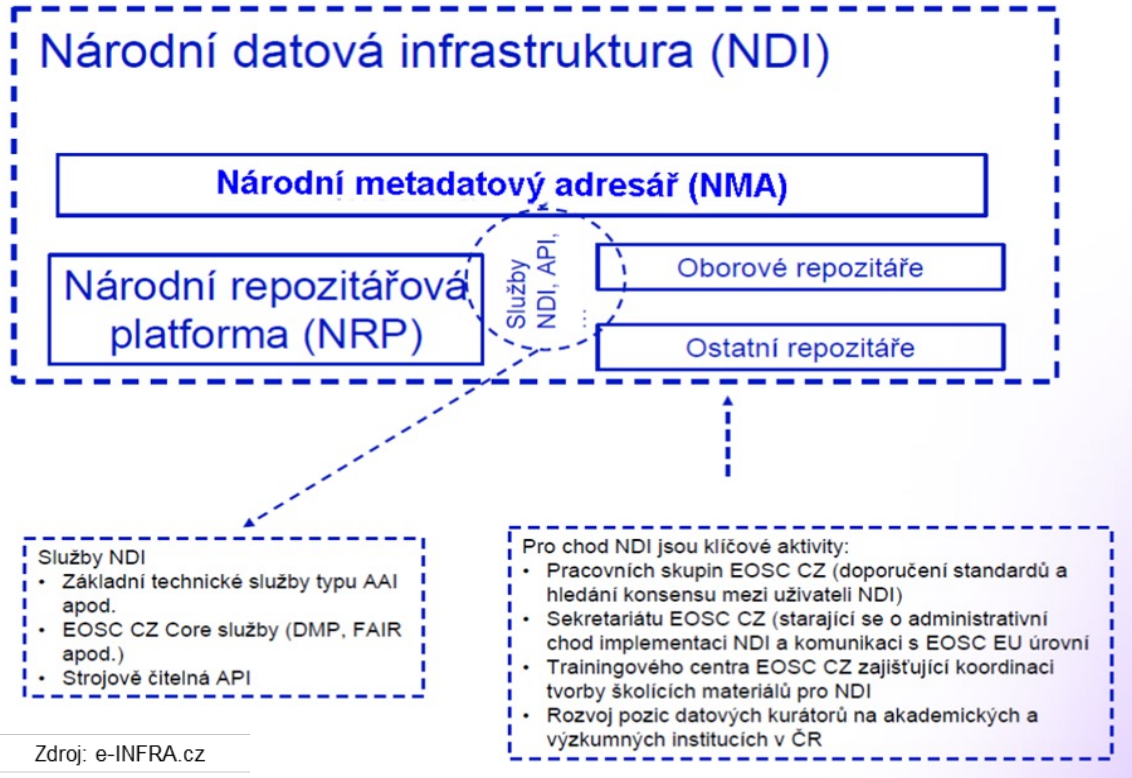
- Sekretariát EOSC-CZ, expertní, informační a školicí zázemí
- Implementace a zajištění provozu klíčových společných komponent NDI - Národního metadatového adresáře (NMA), autentizačních a autorizačních služeb, zázemí pro perzistentní identifikátory (PID), nastavení kyberbezpečnostních politik a řešení incidentů.

IPs CARDS (Czech Academic and Research Discovery Services) - vytvoření unifikovaného prostředí pro sdílení a efektivní správu informačních zdrojů (NTK)

- Zřízení centrální platformy nové generace (PNG) pro správu všech typů informačních zdrojů společné pro knihovny všech VŠ a dalších VO i knihovny s celostátní působností
- Komplexní metodická podpora práce s metadaty výzkumných dat a s perzistentními identifikátory

Oba projekty (IPs) tvoří zázemí pro realizaci **dalších plánovaných výzev Open Science** v rámci OP JAK. Ty by měly podpořit implementaci federované národní datové infrastruktury EOSC na úrovni výzkumných institucí, resp. konsorcií těchto výzkumných organizací.

5.3) Rámec a aktuální stav přípravy NDI EOSC



Zdroj: e-INFRA.cz

Průřezové pracovní skupiny:

- Metadata
- Architektura NDI
- Základní služby
- Vzdělávání a lidské zdroje

Tematické pracovní skupiny:

- Bio/Health/Food
- Environmental Sciences
- Physics
- Material Sciences and Technology
- Data Management for Artificial Intelligence and Machine Learning
- Humanities and the Arts
- Social Sciences
- Sensitive Data

5.4) Předpoklady pro efektivní práci s metadaty výzkumných dat

- **Vytvoření sítě odborníků**, která může poskytnout širší podporu a rozsah znalostí (zejména oborově specifických standardů).
- Předpokladem pro naležitelnost a interoperabilitu (meta)dat (první dva atributy FAIR principů) je kvalita a synchronizace metadatového popisu. Tu by měl na základní úrovni zajistit společně definovaný **metadatový model pro popis výzkumných dat**.
- Předpokladem pro efektivní interoperabilitu repozitářů se zachováním sémantiky je jejich včasná evidence včetně technických parametrů, informací o použitých metadatových schématech, ontologiích apod. Pro zajištění této interoperability bude vytvořen **Národní katalog repozitářů (NKR)**.

5.5) Průzkum očekávání mezi členy pracovních skupin EOSC

Očekávané komponenty v r. 2025 (zástupci 15 entit):

- **Základní metadatový adresář**, který agreguje metadata z repozitářů v Národní repozitářové platformě a umožňuje výměnu metadat s mezinárodními infrastrukturami.
- Pro ukládání záznamů do repozitářů jsou nezbytné **persistentní identifikátory**. Infrastruktura bude proto provozovat technický systém pro jejich přidělování.
- **Národní repozitářová platforma** bude sloužit pro vlastní ukládání dat do repozitářů. To zahrnuje provoz hardwarové infrastruktury, provoz **catch-all instance obecného datového repozitáře**, dále pak **systému pro vytváření tenantů repozitářů** pro potřeby uživatelských skupin a institucí.
- Optimální by byla dostupnost **prostředí pro citlivá data** (osobní, medicínská).
- Celý systém bude **správou identit uživatelů** navázán na národní e-infrastrukturu.
- K použití Národní repozitářové platformy a zejména vytváření instancí/tenantů v ní prostřednictvím API bude k dispozici **dokumentace a také pokročilá uživatelská podpora**.
- Budou k dispozici **nástroje pro přenosy dat do repozitářů**.
- Celý systém bude plně **integrován do dohledu v oblasti kyberbezpečnosti**.

5.6) Typologie pracovních pozic pro implementaci EOSC

- **Vědec** – Každý vědec by měl mít základní dovednosti pro práci s vědeckými daty.
- **Data scientist** – Poučenější vědec se širší znalostí ve specifických oblastech, jako jsou statistika, grafika, fotografie, text mining, prostorové zobrazení apod.
- **Data steward** – Správce dat, metadat a jejich zpracování.
- **Data curator** – Specialista, který pečuje o IT archiv, definuje datové standardy a spravuje repozitář (validace formátů, správa a workflow).
- **Datový právník** – Právník pro oblast IT a dat.

5.7) Rozhovory s experty

Uživatelskou skupinu respondentů tvořili:

- **prof. RNDr. Jan Hajič, Dr.** (UFAL MFF UK), zástupce infrastruktury LINDAT/CLARIAH-CZ a vedoucí oborově-vědní pracovní skupiny Humanities and Arts
- **RNDr. Jiří Chudoba, Ph.D.**, FZÚ AV ČR, zástupce oborově-vědní pracovní skupiny Physics
- **prof. RNDr. Jana Klánová, Ph.D.**, (RECETOX MUNI) zástupkyně infrastruktur RECETOX a EIRENE a vedoucí oborově-vědní pracovní skupiny ENVRI
- **doc. RNDr. Jiří Vondrášek, CSc.** (ÚOCHAB AV ČR), zástupce infrastruktury ELIXIR CZ a vedoucí oborově-vědní pracovní skupiny Bio/Health/Food

Expertní skupinu tvořili:

- **RNDr. David Antoš, Ph.D.** (CESNET), vedoucí pracovní skupiny Architektura NDI
- **PhDr. Radka Římanová, Ph.D.** (Ústřední knihovna UK), vedoucí prac. skup. Vzdělávání a lidské zdroje

Skupinu stakeholderů tvořili:

- **doc. MUDr. Marian Hajdúch, Ph.D.** (Ústav molekulární a translační medicíny Lékařské fakulty UP v Olomouci), zástupce infrastruktury EATRIS-ERIC, člen Rady pro výzkum, vývoj a inovace
- **RNDr. Jan Hrušák, CSc.** (ÚFCH J. Heyrovského AV ČR), zástupce ESFRI
- **prof. RNDr. Luděk Matyska, CSc.** (ÚVT MUNI), zástupce infrastruktury e-EINFRA CZ
- **Ing. Martin Svoboda**, ředitel NTK, zástupce projektu CARDS (IPs CARDS)

6) Shrnutí a závěry

- Potřeba **využití zkušeností** a kapacit stávajících infrastruktur
- Potřeba **důsledné propagace** myšlenky a výhod EOSC.
- Návrh struktury EOSC jako výsledek **širší odborné diskuse**.
- Výše předpokládané **podpory z OP JAK** se jeví jako odpovídající, především ve srovnání s Německem a Nizozemskem.
- Společné **budování federované infrastruktury** pro ukládání anotovaných dat a podpora užívání standardizovaných metadatových schémat.
- Od počátku zavádět postupy **užívání standardizovaných metadatových schémat**.

6) Shrnutí a závěry

- Priorita investic do **vzdělávání a rozvoje lidských zdrojů**, osvěta v oblasti používání data management plánů, výchova datových specialistů.
- **Využití již existujících řešení** v rámci oborově-vědních a institucionálních klastrů.
- Vytvoření **NRP jako multi-tenant platformy** pro „každého“.
- Respektování specifik v oblasti **citlivých dat** napříč různými obory.
- Soulad s principy **EOSC Interoperability Framework**.
- Zajištění přístupu ke **standardizovaným perzistentním identifikátorům**.
- Aktivní **zapojení zájemců do práce EOSC Association Task Forces**.
- **Širší zapojení českých institucí a týmů do mezinárodních projektů**.
- Otázka dlouhodobé udržitelnosti bude aktuální po r. 2025.

7) Přílohy

- 7.1 Klíčové principy pro implementaci a provoz EOSC
 - 7.1.1 Reflexe principů FAIR
 - 7.1.2 EOSC Interoperability Framework
 - 7.1.3 Persistentní identifikátory
 - 7.1.4 Data management plány
- 7.2 Rozhovory s národními experty
- 7.3 Slovníček pojmů implementace EOSC v ČR
- 7.4 Seznam zkratk

Analýza aktuálního stavu kapacit pro implementaci EOSC v ČR

Děkuji za pozornost.

Martin Faťun
fatun@tc.cz