Pracovní skupina Architektura národní datové infrastruktury

Pracovní verze k 22. 9. 2021; vypracoval David Antoš (david.antos@cesnet.cz) a kol.

1 Úvod

Jedním z hlavních pilířů implementace EOSC v ČR je vybudování národní repozitářové platformy pro vědecká data. Ta má sloužit jako jeden ze základních prvků infrastruktury pro vědecká data v ČR, spolu s prostředím pro ukládání nestrukturovaných dat, výpočetním prostředím a prostředím pro spolupráci.

Východiska: e-infrastruktura v ČR provozuje řadu datových úložišť v celkovém objemu desítek PB. To zahrnuje zejména úložiště souborová nebo objektová s minimální podporou ukládání metadat. Část úložné kapacity je určena pro data přímo zpracovávaná ve výpočetním prostředí a je s ním silně provázána (např. geografickou blízkostí, přímým přístupem), část pro data spíše archivního a dlouhodobějšího charakteru. Významné množství vědeckých dat je současně uloženo v řadě oborových úložišť na národní i mezinárodní úrovni, případně drženo vědeckými týmy nebo i jejich institucemi s omezenou viditelností a dostupností mimo bezprostřední autory a uživatele.

Jedním z klíčových cílů implementace EOSC v ČR je vybudování národního prostředí pro ukládání a zpracování vědeckých dat, které výrazně sníží současnou fragmentaci, bude integrováno do jednotné infrastruktury a bude poskytovat flexibilní platformy pro pokrytí specifických potřeb vědeckých skupin a uživatelských institucí.

Pracovní skupina Architektura národní datové infrastruktury (PS ANDI) má za cíl návrh a architektonický dohled nad implementací národní datové infrastruktury, specificky nad platformou pro datové repozitáře a bezprostředně navazující služby, jako je technická implementace zacházení s persistentními identifikátory, sběr a indexace metadat, vyhledávání v metadatech, binárně spolehlivé ukládání datových sad a integraci těchto funkcionalit do národní e-infrastruktury jako celku. Samotné funkční a kvalitativní požadavky na národní datovou infrastrukturu budou převážně formulovány v ostatních pracovních skupinách a ve spolupráci s nimi. PS ANDI rozpracovává tyto požadavky do podoby návrhu implementace a architektury systému s ohledem na škálovatelnost a výkonnost řešení pro objemy dat v řádech přesahujících desítky petabajtů, zajištění integrity a dostupnosti dat a metadat, integraci jednotlivých komponent na technické úrovni a s ohledem na ekonomiku řešení.

2 Cíle

Pracovní skupina se bude zabývat zejména následujícími komponentami národní datové infrastruktury:

- **Národní repozitářová platforma** má poskytovat funkcionalitu obecného národního repozitáře (tj. ukládání dat opatřených metadaty a zejména persistentními identifikátory). NRP bude úzce propojena s národní e-infrastrukturou, s jejímž rozvojem musí být koordinována. NRP bude umožňovat vytváření specifických instancí repozitářů pro konkrétní účely, např. pokrývající metadatová a vyhledávací specifika jednotlivých institucí či odborných komunit (až do úrovně oborově vědních clusterů). Přitom tyto komunity odstíní od nutnosti provozovat vlastní infrastrukturu, předpokládá se ale, že odborné komunity budou do vytváření specifických instancí repozitářů vkládat úsilí nutné právě k vytvoření podpory specifických metadat a specifických workflow.

Předpokládáme, že jednou z instancí NRP bude i obecný národní repozitář pro data bez specifické oborové či institucionální příslušnosti.

- **Metadatový adresář vědeckých dat** bude agregovat metadata ze všech repozitářů, které budou tvořit národní datovou infrastrukturu (zejména všech instancí v národní repozitářové platformě a oborových i dalších repozitářů). Bude tak sloužit jako přístupový bod pro uživatelská vyhledávání (zajišťuje tak vyhledatelnost dat).
    - Zároveň bude vhodným způsobem napojený také na mezinárodní systémy jako např. OpenAIRE, mezinárodní oborově-vědní repozitáře aj.
    - Pro přístup k datům budou poskytována také strojově využitelná rozhraní (API).
    - Bude nastavena interoperabilita s IS VaVaI.

Pracovní skupina bude vytvářet/rozvíjet:

- Technickou architekturu škálovatelné multi-tenant repozitářové platformy pro vědecká data a publikace (desítky až stovky PB).
- Standardy pro interoperabilitu repozitářů na technické úrovni.
- Způsoby zajištění požadované kvality služby.
- Technické ošetření práce s citlivými daty.
- Integraci repozitářů do národní datové infrastruktury, aby mohlo docházet k efektivnímu předávání metadat o hotových živých datech (např. data přímo z přístrojů) a datových sad samotných do repozitáře/repozitářů, kdy metadata budou v maximální možné míře automaticky generována (to vše při zachování plné kontroly vlastníka dat nad jejich zpřístupňováním). Zde se předpokládá velmi úzká spolupráce s pracovní skupinou metadatového adresáře.
- Strategie dlouhodobého uchování binárních dat s ohledem na ekonomiku provozu
- Ve střednědobém a dlouhodobém horizontu nástroje pro LTP (Long-Term Preservation)
- Technickou architekturu Metadatového adresáře vědeckých dat, opět v úzké spolupráci s příslušnou pracovní skupinou.
- Sadu doporučení a standardů pro realizaci repozitářů oborově-vědních clusterů
    - jako instancí Národní repozitářové platformy,
    - pro integraci stávajících repozitářů do systému sběru metadat,
    - pro integraci stávajících repozitářů do systému přenosu vlastních dat (technická interoperabilita).
- Architekturu systému přidělování PID.
- Návaznost na ostatní datové služby národní e-infrastruktury.
- Návaznost na AAI architekturu národní e-infrastruktury.
- Způsoby technické implementace principů jako „nad daty má plnou kontrolu jejich původce“.
- Navrhuje priority implementačních prací jednotlivých komponent s ohledem na dostupnou vývojovou kapacitu.

Cílem pracovní skupiny není vytváření nových standardů a postupů, pokud to není nezbytně nutné (důležitým principem je „co je možno převzít, nechť je převzato“). Pracovní skupina tedy bude sledovat evropské a světové trendy v oblasti repozitářů a spolupracovat s relevantními komunitami, jako jsou zejména evropské e-infrastruktury, Zenodo, OpenAIRE, DataCite, CrossRef a další. Výstupy skupiny musí být v souladu s politikami EOSC (jako např. PID Policy and Architecture, kompatibilita s AAI principy).

V době, kdy části infrastruktury budou v provozu, by se pracovní skupina měla transformovat do technického poradního orgánu pro provoz a rozvoj infrastruktury.

3 Výstupy a jejich aplikace

Návrhy architektury budou důležitým poradním vstupem pro implementaci systémů v rámci e-infrastruktury. Dohodnuté standardy budou pro infrastrukturu závazné po schválení steering committee.

4 Členství, předpokládaní členové, způsob fungování

Pracovní skupina je otevřena všem zájemcům, počet členů není omezen. Předpokládá se nicméně, že zájemci by měli mít zkušenost s budováním infrastrukturních služeb minimálně na úrovni vědecké instituce nebo většího projektu, optimálně v roli systémových architektů. Bylo by vhodné, aby byly zastoupeny instituce, oborově vědní clustery a projekty či instituce, které již provozují významné repozitáře.

Očekává se, že skupina bude spolupracovat s relevantními partnery v ČR i v zahraničí. Primárními národními partnery pro intenzivní spolupráci budou zejména další pracovní skupiny projektu.