Podmínky vytváření nových a úprava stávajících oborových repozitářů v projektu Open Science II

L. Matyska a kolektiv EOSC-CZ, CARDS a NRP projektů

11. října 2024

Terminologie

Národní repozitářová platforma (NRP) je distribuovaný systém pro vytváření instancí repozitářů, budovaný zejména v rámci projektů NRP a IPs EOSC-CZ s podporou IPs CARDS.

Instancí repozitáře (nebo krátce repozitářem, pokud nehrozí nedorozumění) rozumíme konkrétní repozitář vědecké skupiny nebo instituce. Příkladem instance repozitáře je catch-all repozitář pro data https://data.narodni-repozitar.cz/, nebo „repozitář genomických dat obratlovců".

Softwarovým repozitářovým systémem rozumíme softwarový balík, na kterém jsou instance repozitářů postaveny. V rámci NRP jsou k dispozici tři podporované *základní* repozitářové systémy: CESNET Invenio, CLARIN DSpace a ASEP/ARL. (Nezaměňujte termíny instance repozitáře a repozitářový systém.) Repozitáře mohou být postaveny i na jiných, *alternativních repozitářových systémech*. Příkladem alternativního repozitářového systému je Islandora (použita v pilotním repozitáři HUME Lab), nebo třeba vlastní systém, na němž je postavena instance repozitáře pro specifické potřeby konkrétní vědecké skupiny (příkladem je pilotní repozitář pro biodiverzitní herbáře).

Vedle repozitářových systémů NRP poskytuje rovněž prostředí pro jejich běh, tj. vrstvy od hardware po prostředí pro ukládání dat (protokolem S3) a prostředí pro provoz aplikací v kontejnerech (Kubernetes). Pro účely tohoto dokumentu není třeba tuto vrstvu rozlišovat do dalšího detailu, prakticky znamenají dostupnost S3 a Kubernetes jako služby pro provozovatele jednotlivých repozitářových systémů i samotných repozitářů v NDI. Dále budeme o této vrstvě hovořit jako o *prostředí pro ukládání dat a běh aplikací,* případně na techničtější úrovni jako o *S3 + Kubernetes*.

Specifické postavení má *Národní metadatový adresář (NMA)*, což je instance repozitáře výhradně pro automatizovanou agregaci metadat ze všech instancí repozitářů. NMA je budovaný a provozovaný v rámci projektu IPs EOSC-CZ.

Specifické postavení má rovněž *Národní katalog repozitářů* (NKR), který bude evidovat repozitáře a jejich vlastnosti a vznikne v projektu IPs CARDS.

Obecné principy NRP

V rámci NRP chápeme repozitář (instanci repozitáře) v souladu s následující definicí:

„Repozitář je technické, personální a procesní zajištění dlouhodobého úložiště pro ukládání a publikaci citovatelných digitálních objektů."

Zásadně platí, že v rámci NRP musí všechny repozitáře splňovat výše uvedenou definici, musí tedy zejména obsahovat citovatelné záznamy, musí mít webové rozhraní a API pro strojový přístup.

Citovatelným záznamem se rozumí spolehlivé uložení jasně identifikovaného digitálního objektu (např. datové sady). To má dvě základní složky: identifikaci objektu a garanci rozsahu změn. Identifikací objektu je typicky přidělení persistentního identifikátoru (klasicky např. DOI, pokud má

repozitář spíše charakter databázových záznamů, pak vhodný systém jednoznačné identifikace konkrétního záznamu, který má obdobné vlastnosti). Garance rozsahu změn je přesný popis, jaké změny je povoleno provádět u finalizovaného záznamu. Typicky by data záznamu po finalizaci neměla být měněna, u metadat obvykle má smysl připustit doplnění referencí na související položky, jako jsou například vztahy typu Used-By nebo Obsoleted-By. Opravy záznamů by se měly provádět zejména verzováním. Cílem citovatelnosti záznamu je zajistit replikovatelnost vědeckých výsledků. Tím se zejména chceme vyhnout se situaci, kdy různí uživatelé použijí konkrétní datovou sadu v očekávání, že je to stále identická datová sada, přitom tato sada byla změněna, a to třeba i marginální opravou.

To neznamená, že by NRP měla mít jednotnou politiku pro opravy a změny záznamů v repozitářích. Taková politika musí být nastavena dle potřeb a zvyklostí uživatelské komunity. Repozitář ale musí v každém případě nastavit svou politiku tak, aby minimalizovala povolené změny na minimum, preferovala mechanismy jako verzování nebo uložení samostatně identifikovaných opravných balíčků k větším datovým sadám a podobně. Politika repozitáře musí rozsah povolených změn explicitně vymezit a přesně popsat. Systém, který dovoluje libovolné změny uzavřených záznamů a nedává žádné garance ve výše popsaném smyslu, nelze považovat za repozitář.

Na okraj je vhodné uvést, že technická instance repozitáře se v některých repozitářových systémech dá dělit na logické komponenty (obvykle nazývané komunity), které mají zejména samostatně konfigurované řízení přístupu uživatelů a workflow nad sdílenou sadou metadatových modelů. Komunita v rámci repozitáře se tedy chová jako samostatný repozitář (ve smyslu textu projektové výzvy) a měla by tak být ve všech aspektech chápána. Z hlediska organizace repozitářů může dávat smysl agregovat oborově související repozitáře do společných technických instancí, vhodnou granularitu doporučujeme konzultovat s provozovateli příslušného systému. Kdekoli v tomto dokumentu hovoříme o repozitáři, týká se to přirozeně i logického repozitáře v rámci společné technické instance.

Případné další systémy provozované v NRP musí s účelem repozitářové platformy bezprostředně souviset (např. v rámci NRP podporovaný Data Stewardship Wizard, DSW). Pro nástroje tohoto typu je v NRP naplánována kapacita hardwarová i personální.

V NRP naopak není dostupná kapacita obecného úložiště neanotovaných dat pro koncové uživatele, ani výpočetní kapacity pro koncové uživatele, NRP ale zajistí základní propojení na výpočetní prostředky e-INFRA CZ (především MetaCentrum a IT4Innovations). Prostředí pro ukládání dat a běh aplikací v NRP musí sloužit výhradně účelům NRP.

Vyjma samotných softwarových repozitářových systémů a na nich realizovaných instancí pilotních repozitářů jsou nedílnou součástí NRP pomocné služby, jako zejména autentizační a autorizační infrastruktura (AAI), nástroje pro přenosy dat atd.

Druhou základní zásadou provozování systémů NRP je, že všechny budované repozitáře musí v maximální míře používat podporované standardní systémy provozované v rámci platformy. Jedině tak je možné zajistit řádnou podporu ze strany projektů EOSC-CZ, NRP a CARDS. Pokud tyto repozitářové systémy z jakéhokoliv důvodu nevyhovují – příkladem je pilotní repozitář pro herbáře, který je etablovaným nástrojem pro širokou odbornou komunitu v Evropě, používá vlastní repozitářové zázemí a z NRP využívá přímo S3+Kubernetes prostředí – NRP umožňuje postavit i vlastní řešení. V takovém případě pak tvůrce a správce repozitáře musí mít dostatečné odborné znalosti i personální kapacitu na zajištění podobného řešení, z NRP nemůže očekávat další podporu. A samozřejmě je i v takovém případě povinen propojit vytvořený repozitář (nebo nový repozitářový systém) na standardní rozhraní AAI, zajistit odpovídající API a propojení na NMA, služby přenosu dat a případně další nástroje, které budou v NRP postupně nasazovány a kde

u přímo podporovaných repozitářových systémů toto napojení bude zajištěno v rámci projektu NRP.

Ve všech případech by repozitáře měly být zakládány po konzultaci s oborově příslušnou odbornou pracovní skupinou EOSC, která by měla formovat představy o vhodné struktuře, granularitě a oborových metadatových modelech repozitářů v jednotlivých oborech. Protože provoz repozitáře nezbytně vytváří nároky na zřízení i údržbu, repozitáře by neměly být příliš úzce specializované (zejména *nikoli* „repozitář pro katedru“ nebo „repozitář fotografií ústavu archeologického naleziště Dolní Dvorska 1960–1980“, protože v takovém případě bude velmi těžké až nemožné zajistit jejich dlouhodobou udržitelnost).

Základní případy užití NRP

Pro základní přehled úrovně služeb dostupných v rámci projektů NRP a IPs EOSC-CZ, na které mohou navazovat projekty z výzvy OS II, rozlišujeme tři základní případy užití NRP:

I. Vybudování repozitáře s použitím základních repozitářových systémů, tedy jako instanci CESNET Invenio, CLARIN DSpace nebo ASEP/ARL (bod B. výzvy OS II, tedy budování nových repozitářů).
II. Vybudování repozitáře s využitím alternativních repozitářových systémů v odůvodněných případech (bod B. výzvy).
III. Napojování stávajícího repozitáře provozovaného dosud mimo systémy NRP (bod A. výzvy, který se týká výhradně existujících repozitářů).

Obecné poznámky:

- Kdykoli v tomto materiálu hovoříme o zřízení nějaké role, neimplikuje to žádné předpoklady o nutných kapacitách nebo počtu osob. Pro zástupnost se nicméně důrazně doporučuje, aby jednotlivé role byly pokryty několika fyzickými osobami. Na druhou stranu se předpokládá, že jedna fyzická osoba může vůči repozitáři a NRP vystupovat ve více rolích. Části příslušných odpovědností samozřejmě mohou být delegovány na další osoby, typicky správce repozitáře deleguje funkci kurátorů, schvalovatelů a podobně.
- Předpokládá se, že nově vytvářené repozitáře budou mít pilotní fázi pro jejich nastavení a vznik nezbytných podkladů, a poté se dostanou do produkčního provozu.
- Pokud není uvedeno jinak, všechny popisované role musí existovat a být personálně obsazeny po celou dobu produkčního provozu příslušného repozitáře. Správce repozitáře musí být k dispozici během celého životního cyklu repozitáře.
- Stejně tak všechny požadované dokumentace musí existovat po celou dobu produkčního provozu příslušného repozitáře.

Vybudování repozitáře s použitím základních repozitářových systémů NRP

NRP poskytuje zřízení repozitáře pro vědeckou komunitu nebo instituci jako službu. **Projekty NRP, IPs EOSC-CZ a IPs CARDS v takovém případě zajišťují a správce repozitáře tak prakticky „dostává jako službu“ následující:**

1. Konzultace spojené s výběrem vhodného repozitářového systému (ze základních repozitářových systémů, tedy CESNET Invenio, CLARIN DSpace, ASEP/ARL), metodickou a analytickou podporu s určením potřeb a očekávání uživatelské skupiny.
2. Konzultace spojené s výběrem (případně vytvořením) oborového metadatového modelu s ohledem na realizovatelnost v jednotlivých repozitářových systémech.
3. Konzultace spojené s interoperabilitou/mapováním oborového profilu na základní metadatový model pro NMA.
4. Vytvoření instance repozitáře na základě této analýzy (včetně otestování funkcionality ve spolupráci se správcem repozitáře).
5. Provoz všech vrstev NRP od hardware (včetně spolehlivého ukládání dat) až po vytvoření a nastavení příslušné instance.
6. To zahrnuje plnou integraci do všech nezbytných systémů, podrobněji (toto jsou činnosti, u kterých neočekáváme, pokud není explicitně uvedeno jinak, speciální dodatečnou kapacitu na straně správce repozitáře):
    a. Nasazení standardních metadatových profilů[1] a jejich evidenci (vyjma specifických schémat nad rámec možností základních repozitářových systémů).
    b. Technické nastavení sklízení metadat do NMA v souladu se základním metadatovým modelem.
    c. Technické nastavení workflow depozice dat.
    d. Přímo dostupnou implementaci přidělování persistentních identifikátorů, dostupnou metodickou a administrativní podporu (viz identifikatory.cz).
    e. Technické nastavení návazností na e-INFRA CZ AAI systémy a rolí skupin uživatelů v instanci repozitáře.
    f. Prefabrikované podklady pro vytvoření uživatelské dokumentace.
    g. Uživatelskou podporu pro správce repozitáře, L2 a L3 podporu eskalovaných požadavků koncových uživatelů repozitářů (nikoli však L1 koncových uživatelů).
    h. Integraci na prostředí národní e-infrastruktury, speciálně dostupnost nástrojů pro přenosy dat mezi repozitářem, obecnými úložišti a výpočetními zdroji v e-infrastruktuře e-INFRA CZ (MetaCentrum, IT4I, úložiště CESNET, ...).
    i. Konfiguraci logování systémů repozitáře do centrálního logovacího systému.
    j. Zařazení do kyberbezpečnostního dohledu (CESNET-CERTS, FTAS), provádění bezpečnostních a zejména penetračních testů systémů. Handling incidentů a spolupráce s kyberbezpečnostním týmem.
    k. Sběr statistických údajů o provozu a využití systému.
    l. Provozní monitoring.
7. Očekáváme, že repozitáře vytvořené jako instance základních repozitářových systémů NRP budou splňovat technické a organizační požadavky na zázemí důvěryhodných

[1] Každý z repozitářových systémů má konkrétní omezení složitosti a způsobu definice modelu. Správce repozitáře musí zvolit model a dodat jeho popis ve standardizované podobě dle konkrétních požadavků repozitářového systému (např. json v yaml-u pro Invenio). NRP pak zajistí instanci repozitáře s tímto modelem a očekává součinnost se správcem repozitáře /technickým kontaktem) ohledně testování a doladění modelu. Složitější modely, včetně těch, které nejde bez dalších úprav takto připravit, musí mít na straně správce repozitáře dostatečnou technickou a personální kapacitu pro implementaci případných úprav repozitářového systému resp. jeho rozhraní; tyto činnosti a s nimi spojené náklady již NRP nenese, to má kapacitu pouze na využití předpřipravených standardizovaných postupů.

repozitářů z hlediska poskytovatelů dotací v ČR. Při získání kvalifikace důvěryhodného repozitáře pro tyto účely pak bude postačovat splnění podmínek pro vlastní repozitáře (péči o data, …; tato pravidla se teprve připravují), celé technologické zázemí a garanci poskytne NRP.

Od uživatelské komunity se v takovém případě očekává (a tyto činnost je vhodné podpořit z OS II):

1. Zřízení role správce repozitáře. Správce repozitáře je partnerem provozovatele infrastruktury (tedy zejména projektu NRP) pro dohodu o konfiguraci repozitáře ve všech dále popisovaných bodech. Správce repozitáře také nese primární odpovědnost za data v repozitáři ukládaná a za všechna níže popsaná nastavení (která samozřejmě dle potřeby deleguje na další osoby). Správce repozitáře je také informován o provozních událostech repozitáře a repozitářových systémů či celé NRP (aktualizacích, výpadcích apod.). Správce repozitáře rovněž odpovídá za spolupráci s kyberbezpečnostním týmem a za hlášení kyberbezpečnostních incidentů, pokud k nim dojde na úrovni jím spravovaného repozitáře.
2. Zřízení role datového kurátora, který formuje obecná pravidla pro data ukládaná v repozitáři (např. z hlediska délky uchování dle typu záznamu) a rozhoduje o konkrétních datových sadách (např. řeší požadavky na výmaz). Kurátor vystupuje také v roli metadatového specialisty, tj. stará se o harmonizaci metadat v rámci daného repozitáře v souladu se stanovenými oborovými metadatovými profily a interoperabilitu metadat s dalšími systémy (zejména NMA). Podílí se na stanovení oborového metadatového profilu. Tyto role lze od sebe případně oddělit.
3. Stanovení oborových metadatových profilů, které budou v repozitáři k dispozici (ve spolupráci s IPs CARDS a metodiky pro příslušné systémy) v nástroji pro správu metadatových profilů. Stanovení položek metadatových schémat, které budou exportovány do NMA (mapování na základní metadatový model). V případě metadatových profilů překračujících možnosti základních repozitářových systémů rovněž spolupráci na implementaci jejich podpory.
4. Stanovení seznamu licencí dostupných v procesu depozice dat.
5. Stanovení workflow pro depozici dat (např. procesu schvalování záznamů).
6. Stanovení workflow pro přístup k datům (od otevřeného přístupu až po proces zahrnující schválení etickou komisí či komisí pro řízení přístupu).
7. Stanovení rolí uživatelských skupin v repozitáři, např. běžný vkladatel, kurátor, schvalovatel v jednotlivých částech workflow. Navázání těchto rolí na skupiny osob v e-INFRA CZ AAI.
8. Vytvoření uživatelské dokumentace pro repozitář. K tomu budou od NRP k dispozici prefabrikované komponenty popisující základní fungování jednotlivých repozitářových systémů v NRP a doporučené šablony dokumentací. Uživatelská dokumentace nicméně musí pro konkrétní instance repozitáře popisovat používané metadatové modely, workflow vkládání do repozitáře, rozhraní pro vyhledávání, popis rolí skupin osob v repozitáři a podobně.
9. Popis politiky repozitáře, zejména kdy je záznam považován za uzavřený, tato politika má jasně označit, kdy je uložený záznam v repozitáři považován za finalizovaný, a jaké změny jsou přípustné ve finalizovaných záznamech (např. „přidání metadatové položky s odkazem na opravu nebo použití záznamu, ale nic jiného“).
10. Poskytování uživatelské podpory (první úrovně) pro koncové uživatele repozitáře. Infrastruktura NRP poskytne (volitelné) nástroje pro evidenci uživatelských požadavků. Z prostředků NRP jsou pokryty další úrovně uživatelské podpory, což zahrnuje eskalované požadavky, které vyžadují zásah správců infrastruktury, a dále podporu pro samotné správce repozitáře.

11. Předávání údajů do Národního katalogu repozitářů - Registrace repozitáře a jeho parametrů do Národního katalogu repozitářů (NKR), průběžné zasílání aktualizací v případě změn. Evidence zahrnuje i metadatové profily (schémat) a užité řízené slovníky a ontologie. Zasílání informací o repozitáři do NKR by mělo probíhat automatizovaně přes OAI-PMH nebo API.

Vybudování repozitáře provozovaného na zdrojích NRP bez využití základních repozitářových systémů

V nezbytných případech (kdy ani po konzultaci není možné využít žádný ze tří poskytnutých a podporovaných repozitářových systémů) je možno přímo v prostředí NRP provozovat repozitáře postavené nad alternativními implementacemi repozitářových systémů.

V takovém případě uživatelská komunita provozující takový repozitář v zásadě od NRP dostane přístup do prostředí pro ukládání dat a běh aplikací (tedy S3 a Kubernetes), nicméně v takovém případě musí odpovídat za všechny činnosti a nést všechny související náklady spojené s instalací a integrací alternativního repozitářového systému a konkrétní instance repozitáře či repozitářů (pokud jich chce provozovat více) do prostředí NRP a jejího provozu.

Správce repozitáře pak musí také převzít odpovědnost za:

1. Všechny položky popsané jako odpovědnost správce v případě použití základních repozitářových systémů.
2. Zajištění instalace a provozu repozitáře a odpovídajícího softwarového zázemí (zpravidla alternativního repozitářového systému) v prostředí pro běh aplikací a s využitím úložných vrstev NRP.
3. Nasazení oborových metadatových profilů a jejich evidenci (v integraci na systém evidence metadatových profilů), harmonizaci metadat v rámci daného repozitáře a interoperabilitu metadat s dalšími systémy, zejména NMA (může řešit také kurátor repozitáře).
4. Výběr a implementace přidělovaní persistentních identifikátorů ze sady standardně podporovaných, nastavení přidělovaných rozsahů.
5. Technické nastavení sklízení metadat do NMA dle požadavků NMA, mj. v souladu se základním metadatovým modelem.
6. Technické nastavení workflow depozice dat a řízení přístupu k datům.
7. Technické nastavení návazností na e-INFRA CZ AAI systémy a rolí skupin uživatelů v instanci repozitáře.
8. Vytvoření uživatelské dokumentace.
9. Vytvoření dokumentace pro administraci a provoz systému.
10. Uživatelskou podporu pro koncové uživatele na všech úrovních, tedy L1 až L3 s výjimkou požadavků bezprostředně se týkajících provozu a nastavení prostředí pro běh aplikací a ukládání dat (S3 + Kubernetes).
11. Poskytnutí nástrojů pro integraci do prostředí národní e-infrastruktury, zejména pro přenosy dat mezi repozitářem a datovými úložišti i výpočetními zdroji v e-infrastruktuře e-INFRA CZ.
12. Konfiguraci logování systémů repozitáře do centrálního logovacího systému NRP.
13. Zařazení do kyberbezpečnostního dohledu (CESNET-CERTS, FTAS). Spolupráce při realizaci bezpečnostních a zejména penetračních testů repozitáře a s ním bezprostředně souvisejících systémů. Handling incidentů a spolupráce s kyberbezpečnostním týmem. Povinné hlášení kybernetických bezpečnostních incidentů.
14. Nastavení souladu se standardními podmínkami služby, vymezení dalších podmínek v součinnosti s compliance NRP.
15. Sběr statistických údajů o provozu a využití systému.
16. Provozní monitoring.
17. Předávání údajů do Národního katalogu repozitářů - Registrace repozitáře a jeho parametrů do Národního katalogu repozitářů (NKR), průběžné zasílání aktualizací v případě změn. Evidence zahrnuje i metadatové profily (schémat) a užité řízené slovníky a

ontologie. Zasílání informací o repozitáři do NKR by mělo probíhat automatizovaně přes OAI-PMH nebo API.

Správce repozitáře musí rovněž zajistit dostatečnou kapacitu systémových správců a dalšího personálu pro stabilní provoz systému.

Integrace stávajícího samostatně provozovaného repozitáře do prostředí NRP/NDI

Tento případ užití pokrývá situace, kdy se do prostředí NRP/NDI má připojit již existující repozitář provozovaný jako samostatná entita.

Správce repozitáře provozovaného mimo NRP má plnou odpovědnost za jeho provoz od hardware po samotnou službu repozitáře. Aby takový repozitář bylo možno považovat za „napojený na NRP/NDI“, pak musí obecně splňovat stejné podmínky, jako repozitáře v alternativních implementacích přímo provozované v rámci NRP, vyjma toho, že i samotný provoz hardwarových zdrojů, správu systému a kompletní podporu uživatelů zajišťuje rovněž správce repozitáře. To se týká nejen provozu vlastního systému repozitáře a jeho úložiště pro data, ale i dalších komponent, které jsou k jeho provozu nezbytné. Správce takového repozitáře musí zajistit funkcionalitu, odpovídající té, kterou poskytuje NRP; není nijak omezen způsob, jak toho dosáhne (jak to bude implementováno), ale veškerá funkcionalita musí být v přiměřené podobě (přiměřenost primárně určuje správce, ale může být administrátory NRP požádán o doložení).

Základní minimální povinné napojení na prostředí NDI je tvořeno

- Napojením na NMA a poskytováním metadat v souladu se základním metadatovým modelem
- Předáváním údajů do Národního katalogu repozitářů
- Napojením na AAI, které provozuje NRP
- Definovaná API pro přenos dat
- Přidělováním PID (nikoliv nutně DOI)

Rovněž pro tyto repozitáře platí všechna ustanovení, týkající se správců, rolí, nastavení licenčních a dalších politik a podmínek, souvisejících s provozem repozitáře.

> Repozitář musí mít také odpovídající kyberbezpečnostní zázemí a zajištěnu alespoň bazální úroveň monitoringu pro zajištění kvality provozu a sběru údajů o používání systému pro statistické účely. Repozitář musí být rovněž logování – nezbytné pro analýzu kyberbezpečnostních incidentů. Správce repozitáře rovněž odpovídá za dostatečnou míru compliance s právními a dalšími předpisy, ve vazbě na charakter dat.

V rámci technických možností a kapacit jsou naopak služby NRP správci existujícího repozitáře k dispozici, a silně doporučujeme jejich co nejširší využívání. Konkrétní nastavení pak bude třeba řešit specificky pro každý konkrétní repozitář. Dle konkrétní technické situace lze uvažovat i o kombinaci vlastního řešení repozitáře ve smyslu této sekce s použitím některé služby NRP (jako příklad se nabízí např. model, kdy takový repozitář používá S3 v NRP jako jedno z úložišť dat).

Závěrečné poznámky

Cílem tohoto dokumentu je zejména vytvořit uchopitelnou představu o úrovni služby NRP pro jednotlivé případy užití. S ohledem na jeho strukturu jsme nepovažovali za užitečné do něj integrovat časové hledisko, časový plán dostupnosti jednotlivých služeb lze dohledat v časové ose projektu.

Příloha: Strukturovaná konsolidace požadavků do tabulek

Podmínky pro vybudování repozitáře s použitím základních repozitářových systémů NRP

Název podmínky	Zřízení role správce repozitáře
Konzultant	tech-support@eosc.cz
Očekávaná odpovědnost správce repozitáře	Správce repozitáře je partnerem provozovatele infrastruktury (tedy zejména projektu NRP) pro dohodu o konfiguraci repozitáře ve všech dále popisovaných bodech. Správce repozitáře také nese primární odpovědnost za data v repozitáři ukládaná a za všechna níže popsaná nastavení (která samozřejmě dle potřeby deleguje na další osoby). Správce repozitáře poskytuje součinnost provozovateli infrastruktury při nasazování a testování instance repozitáře. Správce repozitáře je také informován o provozních událostech repozitáře a repozitářových systémů či celé NRP (aktualizacích, výpadcích apod.). Správce repozitáře rovněž odpovídá za spolupráci s kyberbezpečnostním týmem a za hlášení kyberbezpečnostních incidentů, pokud k nim dojde na úrovni jím spravovaného repozitáře.
Popis/komentáře	V kontaktu uvádíme mailing list k týmu, který tyto záležitosti řeší.

Název podmínky	Zřízení role datového kurátora
Konzultant	Hana Vyčítalová (metadata@techlib.cz) Anastasia Avdeeva anastasia.avdeeva@ruk.cuni.cz
Očekávaná odpovědnost správce repozitáře	Zřízení role datového kurátora, který formuje obecná pravidla pro data ukládaná v repozitáři (např. z hlediska délky uchování dle typu záznamu) a rozhoduje o konkrétních datových sadách (např. řeší požadavky na výmaz). Kurátor vystupuje také v roli metadatového specialisty, tj. stará se o harmonizaci metadat v rámci daného repozitáře v souladu se stanovenými oborovými metadatovými profily a interoperabilitu metadat s dalšími systémy (zejména NMA). Podílí se na stanovení oborového metadatového profilu.
Popis/komentáře	Datový kurátor může být stejná osoba jako správce repozitáře, ale doporučujeme spíše tyto role oddělit a přenést na různé osoby (velikost úvazků bude záviset na rozsahu a složitosti repozitáře, dat v něm obsažených i na úrovni služeb, které repozitář bude poskytovat). Lze případně oddělit roli kurátora a metadatového specialisty.

Název podmínky	Stanovení oborového metadatového profilu
Konzultant	Jakub Klímek <jakub.klimek@matfyz.cuni.cz> technická implementace exportů do NMA David Antoš <david.antos@cesnet.cz> Interoperabilita se základním metadatovým modelem Hana Vyčítalová metadata@techlib.cz
Očekávaná odpovědnost správce repozitáře	Stanovení oborových metadatových profilů, které budou v repozitáři k dispozici (ve spolupráci s IPs CARDS a metodiky pro příslušné systémy) v nástroji pro správu metadatových profilů. Stanovení položek metadatových schémat, které budou exportovány do NMA (mapování na základní metadatový model). V případě metadatových profilů překračujících možnosti základních repozitářových systémů rovněž spolupráci na implementaci jejich podpory.
Popis/komentáře	

Název podmínky	Konfigurace PID
Konzultant	Hana Heringová <identifikatory@techlib.cz>
Očekávaná odpovědnost správce repozitáře	Výběr přidělovaných persistentních identifikátorů ze sady standardně podporovaných, nastavení přidělovaných rozsahů. Odpovědnost za užité PIDs, integrace PIDs a dodržování best practice v práci s PIDs v souhladu s pokyny Národního centra PID (NTK) a poskytovateli PIDs.
Popis/komentáře	Jedná se zejména o členství v mezinárodních organizacích a konsorciích, které stanovují pravidla, jak s PIDs zacházet – např. přidělování DOI: existující projekty poskytnou metodickou i technickou pomoc, ale formálně odpovědnost nese provozovatel daného repozitáře, ten se musí stát členem konsorcia DataCite, nikoliv poskytovatelé repozitářových systémů – CESNET/UK/KNAV.

Název podmínky	Licence
Konzultant	Pavel Straňák <stranak@ufal.mff.cuni.cz>
Popis/komentáře	Zajištění přidělení licence každému datasetu v procesu jeho publikace.
Očekávaná odpovědnost správce repozitáře	Každý dataset dostupný v repozitáři bude mít v metadatech specifikovanou licenci.

Název podmínky	Workflow pro depozici dat
Konzultant	Pavel Straňák <stranak@ufal.mff.cuni.cz>
Očekávaná odpovědnost správce repozitáře	Stanovení workflow pro depozici dat (např. procesu schvalování záznamů). Stanovení workflow pro přístup k datům (od otevřeného přístupu až po proces zahrnující schválení etickou komisí či komisí pro řízení přístupu).
Popis/komentáře	

Název podmínky	Role uživatelských skupin
Konzultant	Peter Lényi <lenyi@ics.muni.cz>
Očekávaná odpovědnost správce repozitáře	Stanovení a implementace rolí a oprávnění v rámci uživatelských komunit v instanci repozitáře, např. běžný vkladatel, kurátor, schvalovatel. Navázání těchto rolí na skupiny osob v e-INFRA CZ AAI. Nastavení a implementace dalších pravidel řízení přístupu skupin uživatelů k datům, např. sdílení konkrétních dat pouze s vybranými uživateli.
Popis/komentáře	

Název podmínky	Uživatelská dokumentace repozitáře
Konzultant	Anastasia Avdeeva <anastasia.avdeeva@ruk.cuni.cz>, Jan Kolouch <jan.kolouch@cesnet.cz> Pavlína Špringerová <springerova@ics.muni.cz>, <ux@eosc.cz>
Očekávaná odpovědnost správce repozitáře	Vytvoření uživatelské dokumentace pro repozitář. K tomu budou od NRP k dispozici prefabrikované komponenty popisující základní fungování základních repozitářových systémů v NRP a doporučené šablony dokumentací. Uživatelská dokumentace nicméně musí pro konkrétní instance repozitáře popisovat používané metadatové modely, workflow vkládání do repozitáře, rozhraní pro vyhledávání, popis rolí skupin osob v repozitáři a podobně.
Popis/komentáře	

Název podmínky	Politiky repozitáře
Konzultant	Jan Kolouch <jan.kolouch@cesnet.cz>, Anastasia Avdeeva <anastasia.avdeeva@ruk.cuni.cz>
Očekávaná odpovědnost správce repozitáře	Popis politiky repozitáře, zejména kdy je záznam považován za uzavřený, tato politika má jasně označit, kdy je uložený záznam v repozitáři považován za finalizovaný, a jaké změny jsou přípustné ve finalizovaných záznamech (např. „přidání metadatové položky s odkazem na opravu nebo použití záznamu, ale nic jiného“).
Popis/komentáře	

Název podmínky	Ustavení uživatelské podpory první úrovně (L1) pro repozitář
Konzultant	Jan Kolouch <jan.kolouch@cesnet.cz>
Očekávaná odpovědnost správce repozitáře	Poskytování uživatelské podpory (první úrovně) pro koncové uživatele repozitáře. Infrastruktura NRP poskytne (volitelné) nástroje pro evidenci uživatelských požadavků. Z prostředků NRP jsou pokryty další úrovně uživatelské podpory, což zahrnuje eskalované požadavky, které vyžadují zásah správců infrastruktury, a dále podporu pro samotné správce repozitáře.
Popis/komentáře	

Název podmínky	Předávání údajů do Národního katalogu repozitářů
Konzultant	Petra Černohlávková <petra.cernohlavkova@techlib.cz>
Očekávaná odpovědnost správce repozitáře	Registrace repozitáře a jeho parametrů do Národního katalogu repozitářů (NKR), průběžné zasílání aktualizací v případě změn. Evidence zahrnuje i metadatové profily (schémat) a užité řízené slovníky a ontologie.
Popis/komentáře	Zasílání informací o repozitáři do NKR by mělo probíhat automatizovaně přes OAI-PMH nebo API.

Podmínky pro vybudování repozitáře provozovaného na zdrojích NRP bez využití podporovaných repozitářových systémů

Správce repozitáře s vlastní implementací repozitářového softwaru musí splnit všechny podmínky předchozí sekce, a dále ještě následující podmínky spojené s příslušným repozitářovým systémem:

Název podmínky	Instalace a zprovoznění systému
Konzultant	tech-support@eosc.cz (David Antoš david.antos@cesnet.cz)
Očekávaná odpovědnost správce repozitáře	Zajištění instalace a provozu repozitáře a odpovídajícího softwarového zázemí (zpravidla alternativního repozitářového systému) v prostředí pro běh aplikací a s využitím úložných vrstev NRP.
Popis/komentáře	

Název podmínky	Implementace oborových metadatových profilů
Konzultant	Jakub Klímek <jakub.klimek@matfyz.cuni.cz>
Očekávaná odpovědnost správce repozitáře	Nasazení oborových metadatových profilů a jejich evidenci (v integraci na systém evidence metadatových profilů).
Popis/komentáře	Oborové modely by měly být mapovatelné na základní metadatový model.

Název podmínky	Sklízení metadat do NMA
Konzultant	David Antoš <david.antos@cesnet.cz> Interoperabilita se základním metadatovým modelem Hana Vyčítalová metadata@techlib.cz
Očekávaná odpovědnost správce repozitáře	Technické nastavení sklízení metadat do NMA dle požadavků NMA protokolem OAI-PMH (výběrem z podporovaných serializací metadatových formátů) v souladu se základním metadatovým modelem.
Popis/komentáře	

Název podmínky	Implementace workflow pro depozici dat v repozitáři
Konzultant	Pavel Straňák <stranak@ufal.mff.cuni.cz>
Očekávaná odpovědnost správce repozitáře	Technické nastavení workflow depozice dat a řízení přístupu k datům.
Popis/komentáře	

Název podmínky	Implementace a konfigurace PID
Konzultant	Hana Heringová <identifikatory@techlib.cz>
Očekávaná odpovědnost správce repozitáře	Výběr a implementace přidělovaní persistentních identifikátorů ze sady standardně podporovaných, nastavení přidělovaných rozsahů. Odpovědnost za užité PIDs, integrace PIDs a dodržování best practice v práci s PIDs v souladu s pokyny Národního centra PID (NTK) a poskytovateli PIDs.
Popis/komentáře	V případě implementace alternativních repozitářů správce odpovídá za technickou implementaci přidělování PID v provozovaném systému. Provozovatel daného repozitáře nese odpovědnost také za správné užití PID, integraci PIDs a dodržování best practice v práci s PIDs v souladu s pokyny Národního centra PID (NTK) a poskytovateli PIDs. Pokud chce přidělovat DOI v rámci alokace pro ČR, musí se organizace zodpovědná za provoz repozitáře stát členem konsorcia DataCite.

Název podmínky	Licence
Konzultant	Pavel Straňák <stranak@ufal.mff.cuni.cz>
Popis/komentáře	Zajištění přidělení licence každému datasetu v procesu jeho publikace.
Očekávaná odpovědnost správce repozitáře	Každý dataset dostupný v repozitáři bude mít v metadatech specifikovanou licenci. (asi to bude podmínkou v NMK, ale tam ještě nejsme)

Název podmínky	Integrace e-INFRA CZ AAI
Konzultant	Peter Lényi <lenyi@ics.muni.cz>
Očekávaná odpovědnost správce repozitáře	Napojení instance repozitáře na autentizaci uživatele přes e-INFRA CZ AAI, preferovaně protokolem OIDC. Stanovení a implementace rolí a oprávnění v rámci uživatelských komunit v instanci repozitáře, např. běžný vkladatel, kurátor, schvalovatel. Navázání těchto rolí na skupiny osob v e-INFRA CZ AAI. Nastavení a implementace dalších pravidel řízení přístupu skupin uživatelů k datům, např. sdílení konkrétních dat pouze s vybranými uživateli. Stanovení a implementace životního cyklu uživatele v repozitáři, včetně odstraňování neaktivních uživatelů.
Popis/komentáře	Napojení je přes OIDC, případně je možné použít SAML v odůvodněných případech.

Název podmínky	Uživatelská dokumentace
Konzultant	Jan Kolouch <jan.kolouch@cesnet.cz>, Anastasia Avdeeva <anastasia.avdeeva@ruk.cuni.cz>, Pavlína Špringerová <springerova@ics.muni.cz>, <ux@eosc.cz>
Očekávaná odpovědnost správce repozitáře	Vytvoření uživatelské dokumentace pro koncové uživatele repozitáře zahrnující rovněž základní popisy příslušné implementace.
Popis/komentáře	

Název podmínky	Interní dokumentace
Konzultant	<tech-support@eosc.cz>, Anastasia Avdeeva <anastasia.avdeeva@ruk.cuni.cz>
Očekávaná odpovědnost správce repozitáře	Vytvoření dokumentace pro administraci a provoz systému.
Popis/komentáře	Předpokládáme, že dokumentaci může potřebovat KA2 NRP při spolupráci na integraci do prostředí S3 a Kubernetes a rovněž při řešení případných provozních problémů.

Název podmínky	Ustavení a provoz uživatelské podpory
Konzultant	Jan Kolouch <jan.kolouch@cesnet.cz>
Očekávaná odpovědnost správce repozitáře	Ustavení systému podpory pro koncové uživatele na všech úrovních, tedy L1 až L3 s výjimkou požadavků bezprostředně se týkajících provozu a nastavení prostředí pro běh aplikací a ukládání dat (S3 + Kubernetes). Uživatelská podpora musí být dostupná po celou dobu existence repozitáře.
Popis/komentáře	

Název podmínky	Nástroje pro přenosy dat z a do prostředí národní e-infrastruktury
Konzultant	Jiří Sitera <sitera@civ.zcu.cz>
Očekávaná odpovědnost správce repozitáře	Poskytnutí a údržba nástrojů pro integraci do prostředí národní e-infrastruktury, zejména pro přenosy dat mezi repozitářem a datovými úložišti i výpočetními zdroji v e-infrastruktuře e-INFRA CZ.
Popis/komentáře	V závislosti na API poskytovaném implementací repozitáře a s ohledem na potřeby uživatelské komunity to mohou být např. integrace - Pro CLI nástroje ve výpočetních prostředích (MetaCentrum, IT4I) - Portálových systémů, např. Galaxy - Staging do výpočetního prostředí spuštěný z repozitáře ("na tlačítko") - Integrace s Jupyter notebooky

Název podmínky	Provozní logging
Konzultant	Radko Krkoš <Krkos@cesnet.cz>, Andrea Kropáčová <andrea.kropacova@cesnet.cz>
Očekávaná odpovědnost správce repozitáře	Konfigurace logování systémů repozitáře do centrálního logovacího systému NRP, s ohledem na nastavení logovacích parametrů provozovaných systémových a aplikačních služeb pro podporu strojového zpracování pro analytické zpracování provozních záznamů Aplikace konfiguračních změn dle doporučení CESNET SOC. Transportní protokol (API) je Syslog (RFC 5424). Monitoring úplnosti a konzistence logování.
Popis/komentáře	Aplikace změn konfigurace v průběhu života repozitáře (nepředpokládá se často) a monitoring jsou průběžné aktivity.

Název podmínky	Kyberbezpečnost
Konzultant	Andrea Kropáčová <andrea.kropacova@cesnet.cz>, Radko Krkoš <Krkos@cesnet.cz>>
Očekávaná odpovědnost správce repozitáře	Zařazení do kyberbezpečnostního dohledu CESNET-CERTS – oznámení IP adres, na kterých je repozitář provozován, kontaktních údajů na odpovědné osoby, hlášení plánovaných atypických aktivit (velké nárazové přenosy dat, mnoho spojení, vlastní penetrační a jiné testování apod.), Napojení přístupového prvku repozitáře do infrastruktury monitoringu síťových toků FTAS (konfigurace exportu netflow dat) Reakce na incidenty a poskytování součinnosti kyberbezpečnostnímu týmu CESNET-CERTS při řešení incidentů, Povinné hlášení kybernetických bezpečnostních incidentů CESNET-CERTS. Systematický monitoring technických zranitelností a reakce na ně (vulnerability management).
Popis/komentáře	Spolupráce na IH, hlášení CERTS a monitoring zranitelností jsou průběžné činnosti vykonávané v průběhu celé doby života repozitáře. Je nutné, aby repozitář provozoval správce znalý základních principů správy OS a logování.

Název podmínky	Compliance
Konzultant	legal@eosc.cz, Marcela Pospíšilová <pospisilova@cesnet.cz>
Očekávaná odpovědnost správce repozitáře	Nastavení souladu se standardními podmínkami služby, vymezení dalších podmínek v součinnosti s compliance NRP.
Popis/komentáře	Jde primárně o procesně legální činnosti

Název podmínky	Statistiky o provozu
Konzultant	Jan Kolouch <jan.kolouch@cesnet.cz> Pavlína Špringerová <springerova@ics.muni.cz> <ux@eosc.cz>
Očekávaná odpovědnost správce repozitáře	Nastavení systému pro sběr statistických údajů/dat o provozu a využití systému. Sběr a poskytování těchto údajů po celou dobu existence repozitáře.
Popis/komentáře	Nepředepisujeme konkrétní API pro tento systém, spíše předpokládáme domluvu se správcem alternativního repozitáře, co obsahuje jím zvolená varianta repozitáře.

Název podmínky	Provozní monitoring
Konzultant	Jan Kolouch <jan.kolouch@cesnet.cz> Pavlína Špringerová springerova@ics.muni.cz><springerova@ics.muni.cz>, <ux@eosc.cz>
Očekávaná odpovědnost správce repozitáře	Nastavení provozního monitoringu služeb repozitáře.
Popis/komentáře	Jde o požadavek na repozitář, ale složitost souvisí s tím, zda alternativní repozitář, vybraný správcem, obsahuje vlatsní monitorovací systém či nikoliv.

Název podmínky	Předávání údajů do Národního katalogu repozitářů
Konzultant	Petra Černohlávková <petra.cernohlavkova@techlib.cz>
Očekávaná odpovědnost správce repozitáře	Registrace repozitáře a jeho parametrů do Národního katalogu repozitářů (NKR), průběžné zasílání aktualizací v případě změn. Evidence zahrnuje i metadatové profily (schémat) a užité řízené slovníky a ontologie.
Popis/komentáře	Zasílání informací o repozitáři do NKR by mělo probíhat automatizovaně přes OAI-PMH nebo API.

Výše uvedený seznam, s přihlédnutím ke komentářům v jeho textové verzi, může sloužit rovněž pro případy napojování samostatného repozitáře do prostředí NDI.

Zbytek dokumentu tvoří úvodní zadání od týmu Univerzity Karlovy připravujícího OSII pro referenci:

Podmínky vytváření nových a úprava stávajících oborových repozitářů v projektu Open Science II

Splnění povinnosti napojení na NRP a NMA a jejich integrace do společně budované NDI

Univerzita Karlova žádá řešitele IPs CARDS a projektu NRP o doplnění základních parametrů pro přípravu a definování výstupů připravovaných v projektu Open Science II (OS II).

Dokument bude poskytnut všem potencionálním příjemcům (bez ohledu na předpokládaný statut „partnera“ nebo „příjemce minizáměru“) ve struktuře Tematických (oborových) PS EOSC.

Termín dodání: 11. října 2024, konzultace k vyplňování irena.velebilova@ruk.cuni.cz

Úvod

Výzva OS II ukládá povinnost žadateli (s. Výzvy Open Science II, https://opjak.cz/wp-content/uploads/2024/06/Vyzva_Open_Science_II_web.pdf s. 3 a dále):

Aktivita 2. Rozvoj existujících repozitářů, budování nových oborových/mezioborových repozitářů a konsolidace dat

Cílem aktivity je rozvoj, konsolidace, případně vytváření tematických (oborově-vědních či mezioborových) repozitářů výzkumných dat, jejich povinné napojení na Národní repozitářovou platformu (dále též "NRP“) a Národní metadatový adresář (dále též “NMA“) a jejich integrace do společně budované NDI za účelem konsolidace a zpřístupnění výzkumných dat v ČR. Konkrétně je předmětem aktivity:

A. rozvoj repozitářů existujících mimo NRP v době vyhlášení výzvy: rozšíření komunit, které tyto repozitáře používají; zviditelnění repozitářů i mimo primární cílovou skupinu (v rámci oborového/tematického clusteru i mimo něj).

B. vznik nových oborově a mezioborově zaměřených repozitářů: konsolidace uživatelských komunit, které zatím operovaly bez řádně spravovaných repozitářů; zviditelnění repozitářů i mimo primární cílovou skupinu (v rámci oborového/tematického clusteru i mimo něj). Vznik nových oborově/mezioborově zaměřených repozitářů se předpokládá primárně formou repozitářů v NRP, vytváření nového vlastního samostatně spravovaného repozitáře musí být explicitně zdůvodněno.

Cíl sběru údajů

- Univerzita Karlova vychází ze situace, že standardizace parametrů NRP, NDI a NMA je jedním z cílů/náplní projektu CARDS a NRP a nejsou nyní dostupné.
- Pro přípravu projektu OS II je nutné rámcové, základní podmínky definovat již nyní, protože ovlivňují počet a náplň připravovaných výstupů v OS II.
- Pokud řešitelé CARDS; NRP předpokládají, že do procesu standardizace se zapojí i řešitelé projektu OS II (např. zajištěním účastníka na pracovní skupině připravující standard, oponenturou návrhů, ověřením prototypu, …), pak je třeba uvést takový předpoklad do popisu podmínky, jako jednu z činností, se kterou má výstup „oborový repozitář“ (dále jen oborový repozitář) počítat.

- Cílem sběru údajů je vytvoření „checklistu“ podmínek pro oborový repozitář OS II, aby byl připraven splnit podmínku napojení na NRP a NMA a byl zahrnut do NDI.
- Každou podmínku zpracujte, prosím, samostatně v níže uvedené struktuře (ideálně přiložené tabulce).
- Příklady „podmínky“ pro oborové repozitáře OS II: např. autentizace; zajištění pozice kurátora repozitáře, systémového správce, …; zajištění provozu helpdesku L1, L2; L3, ověření návrhu standardu pro službu nebo mikroslužbu, pořízení infrastruktury – HW, SW, zpracování metodiky a návodů pro helpdesk; užití persistentních identifikátorů: DOI, Handle, ORCID, kyberbezpečnostní standardy, formát metadatového schéma...

Struktura šablony dokumentu (vysvětlivky):

- Název podmínky: jednoznačný stručný název podmínky.
- Konzultant k podmínce, jméno, email, předpokládáme, že půjde o řešitele klíčových aktivit NRP; CARDS.
- Popis podmínky:
    - Věcný popis: Strukturovaný, věcně konkrétní text, který srozumitelně a konkrétně popíše věcnou náplň podmínky. Včetně případného návrhu, jak podmínku splnit, (bez ohledu na to, zda tyto podmínky zajistí řešitel oborového repozitáře z prostředků projektu OS II nebo prostředků jiných).
    - Existující standard: Pokud existuje k podmínce odkaz na publikovaný standard (např. AAI, pak uvést odkaz).
    - Ze strany NRP; CARDS bude připraveno pro splnění podmínky: co konkrétního může očekávat oborový repozitář, že bude připraveno ke splnění podmínky ve výstupu NRP; CARDS, tj. na co nemusí plánovat prostředky.
- Typ repozitáře: specifikovat, zda se podmínka týká*:
    - A = rozvoj repozitářů existujících mimo NRP;
    - B = vznik nového repozitáře v rámci instancí NRP**;
    - C = vznik nového repozitáře mimo instance NRP**.

      *Lze uvést všechny typy repozitářů.

      **Pozn. za instanci NRP považujeme: Invenio; Clarin DSpace, ASEP a Repozitář pro biodiverzní data).

- Projekt: z kterého projektu CARDS; NRP; jiný (doplňte).
    - Výstup projektu: název souvisejícího výstupu projektu: výstup CARDS; NRP související s definovanou podmínkou (viz tabulka výstupů a provazeb projektů OS https://www.eosc.cz/media/3729597/vystupy-a-provazby-projektu-open-science.xlsx). Lze uvést i číslo řádku z této tabulky výstupů.
    - Termín pro OS II: datum, kdy bude výstup projektu CARDS; NRP připraven pro řešitele OS II (je-li známo).
- Poznámka: prostor, který nesmí nikdy chybět

Název podmínky		
Konzultant	Jméno:	Email:
Věcný popis podmínky		
Existující standard		

Co bude ze strany NRP/ CARDS připraveno pro splnění podmínky		Přibližný termín ukončení přípravy pro splnění podmínky	
Typ repozitáře	A = rozvoj repozitářů existujících mimo NRP	B = vznik nového repozitáře v rámci instancí NRP	C= vznik nového repozitáře mimo instance NRP
Projekt	CARDS	NRP	Jiný:
- Výstup projektu			
- Termín pro OS II			
Poznámka			