Problematika tvorby SIP balíčků

Bc. Jiří Bernas

Současný stav

- SIP je definován přílohami č. 2 a č. 3 NSeSSS
- Výklad je upřesněn Návrhem technických pravidel pro tvorbu SIP zpracovaného Národním archivem
- Historicky vznikly tři druhy SIP:
  - XML libovolného jména s komponentami uloženými pomocí *<mets:FContent>*
  - ZIP obsahující XML libovolného jména s komponentami zachycenými pomocí *<mets:FLocat>*
  - ZIP obsahující *mets.xml* s komponentami zachycenými pomocí *<mets:FLocat>* a uloženými v adresáři komponenty

Problémy se SIP

- SIP každého výrobce vypadal jinak
- Mnoho SIP neprošlo prostou validací proti schématu
- Nebyla brána v potaz příloha č. 3 NSeSSS, zejména část *Popis prvků schématu XML za účelem vytvoření datového balíčku SIP* (str. 34 a následující)
- Údaje v SIP jsou v nesouladu
  (patrné je to zejména u skartačních lhůt)
- Mnoho prvků zůstává nevyplněných
- Někteří dodavatelé eSSl a původci vytvářeli dva typy SIP (pro skartační řízení a pro předání do archivu)
- Obtížně se zpracovávají SIP s velkými komponentami

Problémy se SIP - příklady

- Nevyplnění atributu *schemaLocation* v *<mets:mets>*

  uvedeno: *<mets:mets OBJID=„ID_123456" LABEL="Datový balíček pro předávání dokumentů a jejich metadat do archivu - Submission Information Package (SIP)“>*

  správně: *<mets:mets xmlns:mets="http://www.loc.gov/METS/" xmlns:nsesss="http://www.mvcr.cz/nsesss/v2" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xlink="http://www.w3.org/1999/xlink" OBJID="BBM_123" LABEL="Datový balíček pro předávání dokumentů a jejich metadat do archivu - Submission Information Package (SIP)" xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/mets.xsd http://www.mvcr.cz/nsesss/v2 http://www.mvcr.cz/nsesss/v2/nsesss.xsd">*

Problémy se SIP - příklady

- Rozpor mezi <*mets:structMap*> a <*nsesss:MaterskeEntity*>

- Podle <*mets:structMap*> je spis zatříděn
  - Spisový plán *Spisový plán 2009-2012*
    - *Věcná skupina Porady*
      - *Věcná skupina Porady vedení*
- Podle <*nsesss:MaterskeEntity*> je spis zatříděn
  - Spisový plán *Spisový plán 2009-2012*
    - *Věcná skupina Porady vedení*
      - *Věcná skupina Porady*

Problémy se SIP - příklady

- Duplicitní identifikátory eSSl

- Použit stejný identifikátor eSSl <*nsesss:Identifikator*> pro dokument a jeho skartační režim
- Identifikátor eSSl má být jedinečný v celé eSSl

Problémy se SIP - příklady

- Chybné stanovení roku skartační operace

- Příklad
  - Spis byl uzavřen v roce 2010
  - Skartační znak a lhůta je S/5
  - Skartační událost je uzavření spisu
  - Rok skartační události: 2015 (správně 2010)
  - Rok skartační operace 2020 (správně 2016)

Problémy se SIP - příklady

- Nelogické údaje

- Příklad:
  - Dokument vytvořen před svým vyřízením (založen v roce 2012 a vyřízen v roce 2011
  - Věcná skupina založena v roce 1753
  - Věcná skupina založena před založením spisového plánu
  - Dokument (spis, díl) založen mimo určené časové období (období, po které eSSl přiděluje pořadová čísla, typicky rok)

Řešení

- Sjednocení výkladu NSeSSS - Návrh technických pravidel pro tvorbu SIP, validátor SIP
- Návrh zapracován do novely NSeSSS
- Základní principy:
  - Popis prvků schématu XML za účelem vytvoření datového balíčku SIP v příloze č. 3 NSESSS je závazný
  - Pro uložení komponenty je preferován *<mets:FLocat>*
  - Struktura dokumentu musí být zachycena v *<mets:structMap>* a to do úrovně komponenty
  - Jednotná podoba SIP se souborem *mets.xml* a adresářem *komponenty*, vše „zabaleno v zipu“
  - Rozlišení SIP pro skartační řízení a pro předání do archivu (SIP pro skartační řízení se vytváří ze SIP pro předání do archivu vypuštěním *<mets:fileSec>*

Změny

- Pravidla pro podobu balíčku SIP byla doplněna o textovou část, která se stala součástí XX. kapitoly NSESSS.
- Příloha č. 3 vychází z nejnovější verze METS
- Příloha č. 3 popisuje podmínky použití prvků schématu METS, aby bylo jasné, jak se má vykládat slovní vyjádření, kterým se upravuje použití METS pro účely balíčku SIP
- Nahrazuje se prvek FContent prvkem FLocat
- Upravuje se vazba prvku file v části fileSec na prvek komponenta v sekci dmdSec (pomocí atributu DMDID na atribut ID a odstraněním atributu OWNERID).
- Upravuje adresářová a souborová struktura balíčku SIP. Povolena je komprimace celé struktury do jednoho souboru.
- Zavádí se rozlišení balíčku SIP určeného pro skartační řízení a balíčku SIP pro přejímku vybraných archiválií

Změny

- Definuje se kódování dokumentu XML výhradně v UTF-8 bez BOM (Byte order mark).
- Upravuje se použití prvku subjekt (agent) pouze pro původce
- Opravuje se chyba ve jménu atributu ARCHIVES
- Odstraňuje se duplicitní definice na jmenný prostor schématu NSESSS.
- Množina kryptografických algoritmů pro provedení kontrolního součtu se omezuje na vybrané algoritmy rodiny SHA-2.
- Jednoznačně se upravuje vazbu prvku div v části structMap na entity/objekty v sekci dmdSec (pomocí atributu DMDID na atribut ID).
- Prvek div ve structMap se doplňuje ho o typ komponenta

Změny

- V příloze č. 2 je nově definován prvek KonecnaVerze pro indikaci komponenty, která odpovídá konečné verzi dokumentu.
- V příloze č. 2 je definován prvek VystupniFormat pro indikaci uložení komponenty ve výstupním datovém formátu.
- Příloha č. 2 upravuje prvek DataceVyrazeni, aby nebyl povinný ve všech případech, ale jen v těch, kdy jde o základní entitu (tedy nikoli rodičovskou entitu).
- Umožňuje se elektronické podepisování balíčků SIP pouze podle standardu XAdES.
- Upravuje se délka prvků Název a Komentář na neomezenou.

Validátor SIP

- http://digi.nacr.cz:8080/ValidatorSIP
- http://digi.nacr.cz:8080/ValidatorSIP_test (testovací)
- Validuje SIP pro předávání dokumentů a jejich metadat do archivu
- Kontroluje zejména technickou kvalitu SIP
- Postupně budou přidávány i testy obsahu (v současnosti je v testu kontrola, zda nejsou použity duplicitní identifikátory a zda je vyplněn název dokumentu
- K dispozici je diagram, jak kontrola probíhá

Závěry

- Většina chyb vyplývá z nedůslednosti při sestavování a z nejasného výkladu
- Technická úroveň SIP se zlepšuje
- Postupně budou ve validátoru přibývat testy obsahu
- I ve validátoru se může objevit chyba, tu je třeba rozporovat (helpnda@nacr.cz)

Děkuji za pozornost

http://digi.nacr.cz

nda@nacr.cz