# Bod 3 programu spoločného zasadnutia PS 2 a PS 49. február 2022

(PS 4 – T54)

**Návrh na používanie špecifikácie Canonical XML 2.0 pri výpočte digitálneho odtlačku doručovanej správy v doručenkách**

Na 15. zasadnutie pracovnej skupiny PS 4 bol predložený návrh Národnej agentúry pre sieťové a elektronické služby na používanie špecifikácie Canonical XML 2.0 pri výpočte digitálneho odtlačku doručovanej správy v doručenkách.

Na základe diskusie boli sformulované nasledujúce možnosti riešenia:

* + - * celý MessageContainer kódovaný ako Base64 v rámci Sk-Talk,
			* zachovať uvádzanie odtlačku v doručenke (bez zmeny a ďalej preveriť možnosť overovania),
			* vypustiť digitálny odtlačok z doručenky,
			* zaviesť Canonical XML 2.0 pre účely doručeniek (v takom prípade je potrebná open source knižnica).

Členovia pracovnej skupiny sa uzniesli, že uvedené možnosti budú predmetom ďalšej diskusie.

V nadväznosti na závery z 15. zasadnutia PS 4 zhrnul pán Ján Ulický nasledovné vysvetlenie problému:

Vysvetlenie problémov použitia normy Canonical XML 1.x budeme pre lepšie pochopenie ilustrovať na príkladoch, spolu s vysvetlením, prečo je v rámci UPVS prostredia takmer nemožné použitie Canonical XML v 1.x v kontexte scenára výpočtu digitálneho odtlačku v elektronickej doručenke. Všetky uvedené príklady XML sú hypotetické, snažia sa ilustrovať vlastnosti Canonical XML na minimálnom XML pre lepšie pochopenie.

Základná úloha normy Canonical XML pri počítaní digitálneho odtlačku spočíva v tom, že transformuje vstupné XML na základe deterministických pravidiel do formy, ktorá zachováva sémantickú zhodu s pôvodným XML a má zároveň zabezpečiť, aby dve sémanticky zhodné XML, ale inak „formátované“, boli po aplikovaní transformácie tzv. byte-equal, čo je základný predpoklad, aby digitálny odtlačok pre obe XML bol zhodný.

To, kde sa normy Canonical XML líšia, je práve výklad pojmu „sémantická zhoda“.

Ďalej v texte pri každej vlastnosti uvedieme 2 zápisy XML, o ktorých tvrdíme, že je pre prostredie UPVS vhodné, aby ich norma Canonical XML za účelom výpočtu digitálneho odtlačku (najmä v kontexte odtlačku uvádzaného v elektronickej doručenke) považovala za sémanticky zhodné.

Na ilustráciu na začiatok uveďme príklad, ktorý aj norma Canonical XML v1.x považuje za sémantickú zhodu.

**Vlastnosť 0 – poradie atribútov**

Majme takéto dva zápisy XML:

Forma 1:

<Formular>

  <Spolocnost typ="sro" krajina="sk">NazovFirmy</Spolocnost>

</Formular>

Forma 2:

<Formular>

  <Spolocnost krajina="sk" typ="sro">NazovFirmy</Spolocnost>

</Formular>

**Popis:** Obe formy sa líšia len poradím zápisu atribútov typ a krajina, názvy elementov, názvy atribútov aj hodnoty sú zhodné.

**Prečo je dôležitá sémantická zhoda:**

Jeden zo spôsobov práce s XML je deserializácia (následne serializácia) XML do dátového objektu a následná práca s ním. Pri tejto operácii sa typicky stratí informácia a poradí atribútov.

**Výsledok:** Obe normy Canonical XML 1.x aj Canonical XML 2.0 vypočítajú pre oba zápisy zhodný digitálny odtlačok.

Ďalej v texte budeme uvádzať už len vlastnosti, kde norma Canonical XML v1.x zlyháva a pre uvedené dvojice by digitálny odtlačok po transformácii Canonical XML v1.x bol rozdielny, ale napriek tomu by pre prostredie ÚPVS bolo veľmi vhodné, aby tomu tak nebolo. Pre všetky uvedené dvojice by transformácia podľa normy Canonical XML v2.0 smerovala k rovnakému digitálnemu odtlačku pre obe uvedené XML príklady.

**Vlastnosť 1 – biele znaky v texte medzi elementami**

Majme takéto dva zápisy XML:

Forma 1

<Formular>

  <Spolocnost typ="sro" krajina="sk">NazovFirmy</Spolocnost>

</Formular>

Forma 2

<Formular><Spolocnost typ="sro" krajina="sk">NazovFirmy</Spolocnost></Formular>

**Popis:** Obe formy sa líšia len v bielych znakoch medzi elementami. V zásade tu ilustrujeme, dve veľmi využívané formy – prvá forma predstavuje tzv. pretty-print, kde je XML formátované tak, že každá úroveň XML je na novom riadku a odsadená medzerami (tabelátormi) na základe hĺbky v strome. Druhá forma je minimalizované XML, tzv. linearizované.

**Prečo je dôležitá sémantická zhoda:**

Pri práci s XML v ľubovoľnom jazyku/platforme sa vždy využíva nejaký XML parser na to určený a závisí od konkrétnej implementácie, či a ako je schopná vôbec riešiť takéto dve formy zápisu. Aj ten trend sa zmenil – kedysi bolo aj tých pár bajtov, ktoré zaberali biele znaky zbytočná redundancia, ktorá mohla predlžovať prenos/spracovanie, teraz sa už skôr preferuje pretty-print forma, ktorá je pre človeka oveľa čitateľnejšia. Dôležité je, že po obsahovej stránke (hodnoty atribútov, elementov) sú obe formy zhodné preto by bolo veľmi vhodné, aby to rešpektovala aj norma Canonical XML.

Treba si uvedomiť, že problém môže byť nielen pri spracovaní správy v rámci UPVS, kde prechádza rôznymi modulmi, ale aj na strane odosielateľa, ktorý pošle sktalk správu, typicky cez nejaké SDK, ktoré rieši SOAP komunikáciu (ktoré to napr. pošle linearizované) a následne si do registratúry odloží správu v pretty-print forme pre jednoduchšiu vizualizáciu a následne pri použití Canonical XML v 1.x nebude sedieť digitálny odtlačok v doručenke voči odtlačky vypočítanom z jeho odloženého originálu.

**Výsledok:** Použitie Canonical XML 1.x bude viesť k rôznym digitálnym odtlačkom pre oba zápisy. Použitie normy Canonical XML 2.0 vypočíta pre oba zápisy zhodný digitálny odtlačok.

**Vlastnosť 2 – umiestnenie namespace deklarácie**

Majme takéto dva zápisy XML:

Forma 1

<Formular xmlns:xsi="<http://www.w3.org/2001/XMLSchema-instance>">

  <Spolocnost typ="sro" krajina="sk" xsi:nil="true" />

  <Kontakt>

    <Meno>Meno Priezvisko</Meno>

    <Titul xsi:nil="true" />

  </Kontakt>

</Formular>

Forma 2

<Formular>

  <Spolocnost xmlns:xsi="<http://www.w3.org/2001/XMLSchema-instance>" typ="sro" krajina="sk" xsi:nil="true" />

  <Kontakt>

    <Meno>Meno Priezvisko</Meno>

    <Titul xmlns:xsi="<http://www.w3.org/2001/XMLSchema-instance>" xsi:nil="true" />

  </Kontakt>

</Formular>

**Popis:** Obe formy sa líšia len v tom, kde je uvedená deklarácia XML namespace s prefixom XSI – pritom oba zápisy sú správne a sémanticky zhodné.

**Prečo je dôležitá sémantická zhoda:**

Treba si uvedomiť, že pri výpočte digitálneho odtlačku uvádzaného v elektronickej doručenke, sa tento počíta z XML elementu MessageContainer. Ten je ale súčasťou obálky SkTalkMessage (ktorá je navyše pri zasielaní na ÚPVS Receive endpoint ešte súčasťou obálky SOAP správy).

To znamená, že obe strany (t. j. aj odosielateľ elektronickej doručenky pri tvorbe elektronickej doručenky aj odosielateľ pôvodnej správy pri kontrole elektronickej doručenky za účelom potvrdenia, že doručovaná správa bola sémanticky zhodná s tou, ktorú reálne odoslal) čelia úlohe ako vybrať z Sk-Talk správy element MessageContainer za účelom výpočtu digitálneho odtlačku. V príkladoch figuruje namespace s prefixom xsi, ktorý je jedným zo štandardných namespace používaných pri XML, a tak je typické, že odosielaná Sk-Talk správa vyzerá takto (uvádzame zjednodušený zápis, ktorý nezodpovedá schéme, chceme len ilustrovať využite xsi prefixu):

<SKTalkMessage xmlns:xsi=<http://www.w3.org/2001/XMLSchema-instance> xmlns=<http://gov.sk/SKTalkMessage>>

  <Body>

    <MessageContainer xmlns=<http://schemas.gov.sk/core/MessageContainer/1.0>>

      <Object>

        <Form xmlns=<http://ns>>

          <Value xsi:nil="true" />

        </Form>

      </Object>

    </MessageContainer>

  </Body>

</SKTalkMessage>

Dôležitý aspekt je, že namespace s prefixom xsi je uvádzaný hneď v root element Sk-Talk správy (a býva často uvádzaný vždy, teda aj bez ohľadu na to, či je xsi prefix v XML vôbec použitý).

Teda pri výpočte digitálneho odtlačku je treba rozhodnúť, kde vo vybranom elemente MessageContainer bude tento namespace deklarovaný.

Uvedené zápisy zároveň demonštrujú dve základné stratégie, ktoré XML frameworky využívajú:

* V prvom zápise sa v root elemente XML elementu, ktorý sa „vyberá“, deklarujú všetky namespacy, ktoré sú deklarované v rodičovských elementoch (v prípade MessageContainer-a sú to elementy Body a SKtalkMessage). A teda namespace s prefixom xsi by sa deklaroval v elemente MessageContainer.
* V druhom zápise sa namespace deklaruje až v tom elemente, ktorý daný prefix „používa“. V takom prípade by bol namespace pre xsi prefix deklarovaný až v XML elemente Value. Táto druhá stratégia znamená, že xsi prefix môže byť deklarovaný opakovane, tak ako to ilustruje aj zápis 2 v príklade v úvode.

V každom prípade, umiestnenie deklarácie nijako nemení obsahovú stránku XML, a teda oba zápisy by mali byť sémanticky zhodné.

**Výsledok:** Použitie Canonical XML 1.x bude viesť k rôznym digitálnym odtlačkom pre oba zápisy. Použitie normy Canonical XML 2.0 vypočíta pre oba zápisy zhodný digitálny odtlačok.

**Vlastnosť 3 – prefixy**

Majme takéto dva zápisy XML:

Forma 1

<Formular xmlns="<http://ns>">

  <Spolocnost typ="sro" krajina="sk">NazovFirmy</Spolocnost>

</Formular>

Forma 2

<p:Formular xmlns:p="<http://ns>">

  <p:Spolocnost typ="sro" krajina="sk">NazovFirmy</p:Spolocnost>

</p:Formular>

**Popis:**

V tomto prípade, majú XML elementy definovaný aj namespace. Každý XML element je jedinečne identifikovaný kombináciou Name a Namespace. V oboch zápisoch teda zhodne figurujú elementy „Formular“ s namespace „http“//ns“ a „Spolocnost“ s namespace http://ns.

Rozdiel v zápisoch spočíva v tom, že v prvom sa používa tzv. default prefix (neuvedený) a v druhom zápise sa používa explicitne uvedený prefix. Nič to ale nemení na fakte, že pre oba zápisy sú celé názvy elementov (Name+Namespace) zhodné a teda obe XML by mohli byť chápané ako sémanticky zhodné pre účel počítania digitálneho odtlačku.

**Prečo je dôležitá sémantická zhoda:**

Opäť to súvisí s tým, že MessageContainer, z ktorého sa odtlačok počíta je vnútorným elementom Sk-Talk správy (a teda ešte aj SOAP obálky pri komunikácii s ÚPVS). Navyše Sk-Talk správa prechádza spracovaním a napr. vyššie spomenutá metóda pre prácu s XML deserializácia, následná serializácia aj informáciu o prefixoch v pôvodnom XML stratí. Bolo by preto vhodné, aby aj takéto zápisy boli pre účely výpočtu digitálneho odtlačku do elektronickej doručenky považované za sémanticky zhodné.

**Výsledok:** Použitie Canonical XML 1.x bude viesť k rôznym digitálnym odtlačkom pre oba zápisy. Použitie normy Canonical XML 2.0 vypočíta pre oba zápisy zhodný digitálny odtlačok.

**Vlastnosť 4 – efektivita**

Tentokrát sa zaobídeme bez XML, pretože reč bude o efektivite implementácie normy Canonical XML 1.x versus Canonical XML 2.0, najmä čo sa týka RAM pamäte potrebnej na transformáciu do formy, z ktorej sa počíta digitálny odtlačok. Samozrejme tento fakt je menej dôležitý v kontexte scenára, že si prijímateľ elektronickej doručenky, ktorý dostáva rádovo jednotky správ, potrebuje overiť, že správa, ktorú odoslal, bola reálne aj doručovaná. Argument je dôležitý najmä pre samotné ÚPVS, ktoré vystavuje doručenky v státisícoch denne a zároveň aj poskytuje funkcionality výpočtu digitálneho odtlačku pre integrované subjekty.

Norma Canonical XML 1.x je vo svojej podstate navrhnutá tak, že pri transformácii je nutné alokovať minimálne 2x veľkosť XML.

Oproti tomu norma Canonical XML 2.0 bola už vo svojej postate koncipovaná tak, že transformácia prebieha tzv. streamovo (dokonca je to požiadavka pre implementáciu). XML sa tak chápe ako postupný zoznam XML nodov, ktoré sa postupne transformujú do výslednej podoby a tým pádom je požadovaná pamäť pre výpočet digitálneho odtlačku nezávislá od veľkosti XML.

Aktuálne je na ÚPVS limit na správu 50 MB s víziou zväčšiť tento limit (v prvom kroku sa hovorí o 200 MB).

V prípade normy Canonical XML 1.x by výpočet digitálneho odtlačku pre hypotetický MessageContainer s celkovou veľkosťou 200MB vyžadoval viac ako 400MB.

Oproti tomu výpočet digitálneho odtlačku podľa normy Canonical XML 2.0 vyžaduje alokáciu v stovkách kB.

Mikroslužba pre výpočet digitálneho odtlačku v prípade použitia normy Canonical XML 2.0 môže spokojne bežať na najnižšej konfigurácii s 1GB pamäte a dokáže obslúžiť stovky paralelných požiadaviek na výpočet odtlačku aj pri najhoršom scenári, čo sa veľkosti týka. Bottleneckom sa stane skôr sieťová priepustnosť.

**Resume**

Celá argumentácia ohľadne výhod normy Canonical XML 2.0 bola vedená v kontexte scenára výpočtu digitálneho odtlačku správy uvádzanej v elektronickej doručenke, kde pri tomto scenári existuje viacero objektívnych dôvodov, prečo je staršia norma Canonical XML 1.x prakticky nepoužiteľná:

* XML element, z ktorého sa digitálny odtlačok počíta, je súčasťou obálky, a teda je nutná „transformácia“, ktorá element z obálky vytiahne,
* správa je odosielaná cez SOAP, ktorý je rovnako XML based,
* Sk-Talk správa prechádza spracovaním cez rôzne moduly.

Všetko je to navyše závislé na implementácii jednotlivých knižníc, ktoré môžu implementovať vyššie uvedené transformácie rôzne.

Preto je relevantné uvažovať o začatí využívania normy Canonical XML 2.0 pre tento scenár. Digitálny odtlačok je veľmi dôležitou súčasťou elektronickej doručenky a zvyšuje dôveryhodnosť celého elektronického doručovania, preto by som možnosť neuvádzania digitálneho odtlačku neodporúčal.

Pri iných scenároch – ako napr. použitie v rámci podpisového kontajnera, mnohé dôvody uvedené vyššie odpadajú a používanie normy Canonical XML 2.0 vôbec nie je potrebné.