**Potřebujeme standardizované formáty pro kancelářské dokumenty?**

Jiří Kosek

LISP, FIS VŠE Praha  
nám. W. Churchilla 4, 130 67, Praha 3  
<jirka@kosek.cz>

**Abstrakt**

Po dlouhá léta většina kancelářských aplikací ukládala svá data do proprietárních binárních formátů. Nyní se dvěma nejpoužívanějšími formáty stávají ODF a OOXML, které jsou oba otevřené a založené na XML. První z nich je ISO normou a druhý pravděpodobně také brzy bude. Má smysl, aby vedle sebe existovaly dva tak podobné formáty? Jak tato situace vznikla. Může standardizace formátů zaručit bezproblémovou výměnu dokumentů? Nad historií, současností a budoucností otevřených formátů dokumentů se ve své přednášce zamyslí Jiří Kosek, který se mimo jiné podílí na vývoji formátu DocBook a zpracovával připomínky ČR k návrhu normy OOXML.

**Klíčová slova:** XML, ODF, OOXML, formáty dokumentů, standardizace

V době, kdy jsou všechny počítače propojené pomocí Internetu, nám připadá zcela samozřejmé, že si s ostatními vyměňujeme elektronické dokumenty a můžeme s nimi dále pracovat. Ale i přesto, že informační technologie se vyvíjejí rychle, stále se občas stane, že některé dokumenty si neotevřeme nebo po jejich otevření vidíme něco jiného, než jejich původní autor.

Tyto problémy jsou způsobeny dnešním prostředím, které je heterogenní – uživatelé používají různé operační systémy, na nich běží různé aplikace, … V počítačovém středověku byla situace o poznání jednodušší – nikdo ani neočekával, že by byl schopen pracovat s dokumenty pořízenými na počítačích od jiného výrobce, někdy to dokonce neplatilo pro různé verze počítačů od jednoho výrobce. V unifikovaném prostředí šlo vše hladce, akorát byla fakticky znemožněna výměna dokumentů, což byl z pohledu uživatel zásadní nedostatek.

**1. Historické ohlédnutí za vývojem formátu souborů**

Již od konce 60. let se proto hledaly způsoby, jak zajistit výměnu a sdílení dokumentů v heterogenním prostředí. Vznikaly formáty jako TeX a troff, které se oprostili od platformové závislosti, ale přitom zachovávaly formátování i po přenosu na jiné zařízení. Ještě dále šly obecné značkovací jazyky, ze kterých se později vyvinuly jazyky SGML a XML. Ty se snažily dokument zcela oprostit od popisu formátování a zachycovat pouze význam a strukturu informace. Později vznikly formáty PostScript a PDF, které dokázaly věrně zachytit vzhled dokumentu, nebyly však vhodné pro další editaci.

S masivním nástupem osobních počítačů v 80. letech však začalo být jasné, že masy uživatelů potřebují něco jiného – pohodlný textový editor. Textový editor z principu musí dovolit uživateli provádět všechna obvyklá zvěrstva jako nahodilé změny písma, zarovnání a barvy. Paradigma uživatelského rozhraní textového editoru tak bylo téměř neslučitelné s více strukturovanými formáty jako SGML/XML nebo TeX. Formáty PostScript a PDF se zase příliš nehodí pro editovatelné dokumenty. Každý textový editor tak přicházel s vlastním formátem pro ukládání dat. Při výměně dokumentů bylo potřeba spoléhat na to, že všichni používají stejný produkt, nebo mají alespoň příslušné importní/exportní filtry.

V České republice jsme si na začátku 90. let užívali luxusu „monopolního“ formátu T602. Osobně považuji textový editor T602 v mnoha ohledech za dosud nepřekonaný – zejména z toho důvodu, že příliš nepřevyšoval schopnosti většiny běžných uživatelů.

Vývoj se však nezastavil, přišla grafická uživatelská rozhraní a s nimi další plejáda editorů – MS Word, AmiPro, WordPro, WordPerfect, StarOffice a další. Jediným formátem, kterému v té době rozuměli snad všechny editory, byl formát RTF (Rich Text Format). Tento formát původně vymyslela firma DEC, ale od roku 1990 ho převzala společnost Microsoft a postupně do něj doplňovala nové funkce tak, jak se vyvíjel MS Word. Poslední verze formátu RTF 1.9 byla vydána v lednu 2007.

Formát RTF fungoval docela dobře, trpěl dvěma zásadnějšími problémy. Základem RTF je prostý text. Delší dokumenty s obrázky tak byly po uložení do RTF obludně velké, protože binární data se musela překódovat do textově bezpečné podoby a tím se nafoukla.

Druhým problémem byla nevzdělanost uživatelů. V druhé půli 90. let získal Microsoft na trhu textových editorů téměř monopolní postavení. Hodně mu v tom pomohli samotní uživatelé, kteří dokumenty ukládali do proprietárního binárního formátu DOC, místo formátu RTF. Protože v té době bylo možné formát DOC analyzovat v podstatě jen metodou reverzního inženýrství jeho podpora v ostatních aplikacích nebyla úplně dokonalá. Kdo chtěl mít možnost bez problémů pracovat s dokumenty DOC, tak si raději pořídil MS Office. Masivnější používaní RTF by přitom pro trh textových editorů bylo jistě zdravější.

Díky tomuto mechanismu se postavení MS Office na trhu dále upevňovalo. Podobná situace zavládla i na poli ostatních kancelářských aplikací jako tabulkových kalkulátorů a prezentačních nástrojů. Trojice formátů DOC/XLS/PPT se tak stala „de facto“ standardy pro výměnu kancelářských aplikací.

V uvozovkách by však mělo být spíše slovo „standardy“. Z uživatelského hlediska (pokud byl uživatel vlastníkem MS Office) byly formáty DOC/XSL/PPT v zásadě bezproblémové.[[1]](#footnote-2)1 Tyto formáty však byly problém pro vývojáře třetích stran. Dokumentace k těmto binárním formátům nebyla dlouhou dobu dostupná, a po jejím uvolnění licence nedovolovala využití dokumentace při tvorbě kancelářských aplikací (tedy produktů konkurujících MS Office).

Složité binární formáty nepřály ani aplikacím, které měly například automaticky generovat různé reporty. Spolehlivou cestou bylo jen skriptovat samotný MS Office, což bylo náročné jak finančně (licence), tak na strojový čas. S tímto stavem nebyli vývojáři aplikací spokojení a proto Microsoft začal postupně od roku 2000 přidávat do jednotlivých produktů MS Office alternativní možnost ukládání/načítání dat v XML. Tyto formáty však nikdy nebyly nastaveny jako výchozí, a proto nedošlo k jejich rozšíření mezi běžnými uživateli. Nicméně velké uplatnění našly v místech, kdy bylo potřeba automatizovat práci s kancelářskými dokumenty.

**2. Éra otevřených a standardizovaných formátů přichází**

Přelom milénia byl však ve znamení renesance otevřených formátů. Rozvoj webu a potřeba komunikace dali vzniknout jazyku XML a vytváření výměnných formátů založených na XML se tak postupně stalo nejen v mnoha případech výhodné, ale i módní. Ke starším formátům pocházejícím ještě z dob SGML jako DocBook (počítačová dokumentace), TEI (literatura) a leteckým a vojenským standardům pro dokumentaci se tak začaly přidávat desítky nových formátů. Pro komunikaci mezi aplikacemi se začínají masově používat webové služby, které jsou postaveny na výměně zpráv XML. XML se tak pevně zabydlelo v oblasti výměny strukturovaných dat (kde nahradilo mnohem nákladnější EDI) a strukturovaných dokumentů (zastoupené například formáty DocBook, DITA a TEI).

Mezi hlavními výhody XML bývá uváděna nezávislost na dodavateli a dlouhá životnost dokumentů. Nicméně obojí jsou vlastnosti, které nesouvisí přímo s XML, ale samotná podstata XML vede k tomu, že od něj odvozené formáty tyto výhody nabízejí. Popularita XML znovu přivedla na výsluní standardizaci datových formátů. Standardizace přitom není nic jiného než proces, během kterého se zainteresované strany dohodnou, jak něco dělat jednotným způsobem. Výsledek nemusí být vždy technicky nejlepší řešení, ale to vyváží výhody dosažené interoperability. Navíc většina dnešních standardů je vytvářena jako tzv. otevřené. Znamená to, že jsou veřejně dostupné a že pro jejich implementaci není potřeba platit žádné licenční poplatky.

Pro obecné kancelářské dokumenty však žádný takový formát dlouhou dobu neexistoval. Až v roce 2000 kancelářský balík StarOffice začíná pro ukládání svých dat používat formát XML. Formát je postupně vylepšován a přejímá jej i open-source odnož OpenOffice.org. Na konci roku 2002 je pak v rámci standardizační organizace OASIS použit tento formát jako základ budoucího formátu ODF (Open Document Format). Na vývoji se podílí hlavně firma Sun Microsystems. Microsoft se práce na standardu neúčastní. A není to ani překvapivé – těžko si představit, že by firma s majoritním podílem na trhu akceptovala funkčně zaostalý[[2]](#footnote-3)2 formát aplikace s minoritním podílem na trhu. Takhle svět nefunguje.

V květnu 2005 se ODF stává standardem OASIS a na žádost EU je také postoupeno ke schválení jako ISO norma. Nutno podotknout, že to vše se děje poměrně bez zájmu odborné veřejnosti, protože formát je podporován aplikacemi, které mají pouze zanedbatelný podíl na trhu. Navíc přijetí ODF jako ISO normy je poměrně rozporuplný počin – formát má mnoho nedodělků, které mají být odstraněny až v dalších verzích. Nicméně rychlost procesů uvnitř OASIS a ISO je rozdílná, takže například v současné době máme formáty ODF vlastně dva – ISO ODF 1.0 a OASIS ODF 1.1. Probíhají přitom už i práce na novější verzi ODF 1.2. Jak, kdy a zda vůbec budou novější verze přijaty i jako ISO norma není v tuto chvíli vůbec jasné. Pro uživatele tak vznikla poněkud nepřehledná situace. Osobně si myslím, že se mělo počkat nejméně na ODF 1.2 a až to případně přijmout jako ISO normu.

Přijetí ODF jako normy ISO mělo pozitivní efekty. Rozhýbalo ledy i softwarového giganta Microsoft. Těžko říci, zda větší vliv na to měly doporučení EU, mediální tlak propagátorů open-source, příklon některých vlád k řešením postaveným na mezinárodních normách nebo se Microsoft sám změnil – nejspíše asi kombinace všech těchto faktorů.

Microsoft tak vylepšil XML formáty předchozích verzí MS Office a formát s názvem Office Open XML (OOXML) zasílá na konci roku 2005 do standardizační organizace ECMA, která jej po úpravách o rok později přijímá jako svůj standard. Poté je OOXML zasláno do ISO pro přijetí jako ISO norma. Tento proces ještě neskončil, ale je pravděpodobné, že v únoru 2008 se i OOXML stane mezinárodní normou.

Formáty OOXML a ODF jsou si v mnohém podobné, a může se tak zdát zbytečné, že existují oba dva. Nesmíme však zapomenout, že jsme teprve na začátku cesty vedoucí k případnému vytvoření jednotného formátu. Do začátku je potřeba mít konkurenci, která požene vývoj kupředu. Během posledního roku vzniklo tolik nástrojů pro konverzi mezi ODF a OOXML a pro práci s těmito formáty jako nikdy předtím. Během schvalování OOXML na půdě ISO bude specifikace formátu vylepšena. Vytvářejí se studie, které oba formáty porovnávají a hledají vlastnosti, které nemají v druhém formátu ekvivalent.

Bohužel jsme dnes svědky toho, že standardizace se stala magickým zaklínadlem a v případě kancelářských formátů se dokonce bitva konkurentů dostala do podoby, kdy se soupeří o to, kdo dřív dostane razítko ISO. Formáty a produkty, které je podporují, by přitom měli soutěžit na volném trhu svými schopnostmi a užitečností pro uživatele. U obou formátů by přitom stačilo, kdyby si zachovaly statut standardů OASIS, resp. ECMA a o případné mezinárodní standardizaci by se mělo uvažovat až s několikaletým odstupem na základě praktických zkušeností s oběma formáty.

**3. Samotná standardizace nestačí**

To, že je dnes nějaký formát schválen jako norma je automaticky považováno za dostatečné pro dosažení interoperability. To je však naprosto mylná představa.

Kancelářské formáty dovolují vkládat do dokumentů objekty, jejichž formát není nijak omezen. Pokud tedy chceme definovat nějaké prostředí pro spolehlivou výměnu dokumentů, nestačí říci „bude se používat ODF nebo OOXML“. Musíme ještě přesně definovat jaké objekty mohou být do dokumentů vložené – například povolit jen obrázky v široce akceptovaných formátech.

Dalším problémem jsou fonty – každý uživatel může mít na svém počítači nainstalovány jiné fonty, dokument se tak nemusí vždy zobrazit stejně. Chceme-li dosáhnout věrnějšího zobrazení dokumentu po jeho přenesení na jiný systém je potřeba definovat i množinu fontů, které je dovoleno používat.

Standardizované kancelářské formáty umožňují sdílení dokumentů mezi několika různými aplikacemi od různých dodavatelů. To je dobře a je to vlastně stav, kterého by mělo být dosaženo. Uživatelé si však musí na toto heterogenní prostředí zvyknout. Ani ODF ani OOXML nedefinuje přesné algoritmy pro řádkový a stránkový zlom, nejsou definovány jednotné sady písem a jejich metrik, každá aplikace používá vlastní algoritmus pro dělení slov – ve výsledku tak bude stejný dokument v různých aplikacích zobrazen s drobnými odlišnostmi. Na to uživatelé odchování na softwarové monokultuře nejsou zvyklí.

Velké firmy nebo veřejná správa by si tak měli vytvářet dodatečné směrnice, které obecné standardy v takovýchto detailech dospecifikují. Bez toho nebude nikdy bezproblémová výměna dokumentů zaručena.

Příkladem toho, jak by se to dělat nemělo je [doporučení,](http://www.micr.cz/scripts/detail.php?id=3888)[[3]](#footnote-4)3 které vydalo Ministerstvo informatiky těsně před svým zrušením.

***Doporučení k aplikaci § 4 odst. 2 zákona č. 106/1999 Sb., o svobodném přístupu k informacím, ve znění pozdějších předpisů***

*V § 4 odst. 2 je stanovena povinnost v případě zveřejňování informace v elektronické podobě zveřejnit tuto informaci i ve formátu, jehož specifikace je volně dostupná a použití uživatelem není omezováno.*

*V těchto případech Ministerstvo informatiky doporučuje využívat následující formáty:*

* *Pro dokumenty jako jsou texty (.odt), tabulky a grafy (.ods) a prezentace (.odp):*

*Open Document Format for Office Applications (OpenDocument) v1.0*

* *Pro dokumenty, které jsou zpřístupněny přes internet/intranet nebo použitím webového prohlížeče:*

*The Extensible HyperText Markup Language (XHTML v1.0)*

*HyperText Markup Language (HTML v4.01)*

* *Pro dokumenty obsahující pouze textové údaje, bez potřeby formátovacích funkcí:*

*Plain text format (.txt - např. u e-mailových zpráv, kdy formátování není důležité)*

* *Pro úřední dokumenty, které není potřeba dále upravovat/měnit:*

*Portable Document Format/A-1 (ISO 19005-1:2005)*

*Portable Document Format (PDF Reference v1.5 a vyšší)*

Jaké jsou problémy tohoto doporučení.

1. Preferování formátu ODF, který nemá dostatečnou podporu v aplikacích nainstalovaných u uživatelů. I když je samotný software podporují ODF zdarma, jeho instalace a proškolení uživatelů by stálo nemalé prostředky. Nestojím o to, aby se kvůli tomu zvyšovaly daně.

Navíc současná verze formátu ODF nedefinuje standard pro zápis vzorců v tabulkovém kalkulátoru a je tudíž pro tabulky prakticky nepoužitelná. Prezentace v případě veřejné správy asi není potřeba publikovat v editovatelné podobě, takže v tomto konkrétním případě by bylo ODF používáno pouze pro ukládání textových dokumentů, pro které lze stejně dobře použít RTF s mnohem širší aplikační podporou bez nutnosti instalování nových produktů na straně uživatelů.

1. Není specifikováno, jaké vložené objekty se mohou vyskytovat uvnitř ODF a HTML dokumentů.

Ministerstvo to jistě myslelo dobře, ale podobné dokumenty odtržené od reality ve výsledku pomalou cestu k používání otevřených formátů spíše zbrzdí. Jak by doporučení mělo vypadat v dnešní době:

* *Pro zveřejňování dokumentů primárně upřednostňovat formát HTML, který je nejpřístupnější.*
* *Pokud povaha dokumentu není vhodná pro publikování v HTML (např. mapa), pak použít formát PDF. Formát PDF se může použít i jako alternativní doplnění HTML verze, pokud se jedná o dokument primárně určený pro tisk.*
* *Jiný formát než HTML nebo PDF je povolen pouze pro dokumenty, u kterých se předpokládá jejich následné upravování (šablony, formuláře apod.). Takový dokument musí být zveřejněn nejméně ve dvou z následujících formátů:*
* *ODF*
* *OOXML*
* *RTF*
* *DOC/XSL/PPT*

*Dokumenty nesmí obsahovat vložené binární objekty. Mohou obsahovat pouze obrázky v některém z formátů GIF, PNG, JPEG.*

Takové doporučení by umožnilo bez nutnosti instalování nového software publikovat dokumenty, které se již dnes přečte kdokoliv. Za pár let by se doporučení mohlo upravit a v seznamu dovolených formátů ponechat jen ODF a OOXML a možná do něj přidat nějakého nástupce, který do té doby vznikne a převezme dobré myšlenky obou formátů.

**4. Závěr**

Konečně jsme se dočkali situace, kdy všechny hlavní kancelářské balíky jako svůj výchozí formát pro ukládání používají formát, který je otevřený, dobře zdokumentovaný a založený na XML. To je velká výhra pro uživatele i vývojáře. Uživatel má mnohem větší šance, že dokument otevře i za deset let než tomu bylo v případě binárních formátů. Vývojáři se bude s kancelářskými dokumenty mnohem lépe pracovat v jeho aplikacích, protože může využívat obrovské množství nástrojů pro práci s XML, nemusí se trápit s binárním a nedokonale zdokumentovaným formátem. Ale ani samotné rozšíření formátů ODF a OOXML samo nevyřeší všechny problémy spojené s výměnou dokumentů.

1. 1Pomíjím teď některé drobné problémy s kompatibilitou mezi verzemi MS Office. [↑](#footnote-ref-2)
2. 2První verze formátu ODF nepodporovala ani veškerou funkčnost OpenOffice.org, který však nabízí méně funkcí než MS Office, které je nutné nějak promítnout do formátu souborů. ODF například nedefinuje syntaxi pro zápis vzorců v tabulkové kalkulátoru a prezentace nemůže obsahovat tabulky. [↑](#footnote-ref-3)
3. 3 <http://www.micr.cz/scripts/detail.php?id=3888> [↑](#footnote-ref-4)