УДК 004.62, 004.63

А.В. Шарипанов, І.Є. Щетинін, С.М. Іванов

АНАЛІЗ ФОРМАЛЬНИХ МОДЕЛЕЙ ТА СТАНДАРТІВ ОПИСУ СТРУКТУРОВАНОГО ЕЛЕКТРОННОГО ДОКУМЕНТА В КОРПОРАТИВНІЙ ІНФОРМАЦІЙНІЙ СИСТЕМІ

Постійне прискорення розвитку інформаційних технологій та викликані цим зміни у типових бізнес-процесах організацій породжують нові форми складності електронного документообігу при створенні сучасних корпоративних інформаційних систем. Основним завданням при відображенні структурованих документів у електронну форму є визначення універсального механізму розмітки документа та виділення його окремих інформаційних фрагментів для правильної їх інтерпретації в програмних застосунках. Надається порівняльний огляд існуючих мов опису структурованих електронних документів та стандартів, на яких вони засновані. Визначаються переваги і недоліки їх застосування відповідно до типових задач організації електронного документообігу, що виникають при проектуванні корпоративної інформаційної системи.

Ключові слова: структурований електронний документ, електронний документообіг, корпоративна інформаційна система.

Вступ

Використання розподілених (корпоративних) обчислювальних мереж і інших засобів, притаманних сучасним інформаційним технологіям, породжує нові форми складності електронного документообігу (ЕДО) – це, в тому числі, і розгортання ЕДО в умовах географічно та територіально розподілених робочих груп у рамках одного життєвого циклу обробки документів, пов'язаного з різними аспектами (видами) діяльності.

Основним завданням при створенні формату зберігання структурованих документів є визначення універсального механізму розмітки документа та виділення окремих інформаційних фрагментів для правильної їх інтерпретації у програмному застосунку, який обробляє документ.

Слід зауважити, що обмін структурованими документами може відбуватися не тільки в межах власної організації, а й з великою кількістю інших організацій. Оскільки електронні документи (ЕД) можуть зберігатися в архівах протягом тривалого часу, постає задача забезпечення інваріантності відображення документа відносно використовуваних технологій роботи з ЕД.

Для успішного вирішення цих задач виникає необхідність стандартизувати формат розмітки документа з метою створення умов для ефективного ЕДО між інформаційними системами всіх його учасників. Цього можна досягти, наприклад, за допомогою гармонізації методів і засобів взаємодії між інформаційними системами різних виробників та уніфікації існуючих форматів ЕД.

Одним з основних завдань стандартизації у цій галузі є створення відкритих форматів ЕД.

Відкритий формат – це формат ЕД, що дозволяє переносити цей документ не тільки між різними організаціями, а навіть між різними програмними платформами без спотворення його форми, структури, змісту та продовження роботи з ним. Відкриті стандарти – це стандарти, які дозволяють взаємодіяти різним системам через інтерфейси, встановлені міжнародними консорціумами, які розробляють ці стандарти.

Найважливішою особливістю зберігання документів, що відповідають відкритим стандартам, є те, що для їх обробки використовуються програми які реалізують ці відкриті стандарти. При цьому фо-

рмат дозволяє об'єднати в одному документі як неструктуровані, навіть нередаговані фрагменти, так і елементи структурованої форми. Таким чином, один формат зберігання може об'єднувати в собі всі вищеописані способи представлення документа в інформаційній технології – від плоского неструктурованого документа до жорстко структурованої електронної форми.

Еволюція опису електронних документів

В еволюції опису ЕД можна виділити наступні етапи.

До 1969 р. використовувалася процедура розмітки документів, за якою файл даних містив текстову інформацію, а також інструкції для форматування та відображення цієї інформації на екрані. Навігація файлом здійснювалася методом введення номерів потрібних рядків. У 1969 р. Чарльз Гольдфарб (Charles F. Goldfarb), Эд Мошер (Ed Mosher) та Рей Лорі (Ray Lorie) розробили концепцію мови розмітки GML [1] для вирішення задач відображення текстової інформації. GML була логічним поданням для обґрунтування подальшої обробки тексту, а не лише ще однією альтернативою процедурній розмітці. За результатами подальших досліджень у 1973 р. Чарльз Гольдфарб опублікував звіт, в якому було вказано, що : «...можлива розробка узагальненої мови розмітки, яка стане корисною для більш ніж одного застосунку чи комп'ютерної системи. Така мова обмежить розмітку в середині документу до визначення структури документа та інших атрибутів, наприклад, за допомогою «тегів». Визначення типу компонента буде мати лише те значення, що його буде оброблено так само, як і інші компоненти того ж типу. Проте самі команди обробки не будуть міститись у тексті, оскільки вони можуть відрізнятися у різних застосунках та системах оброблення» [2].

У 1975 р. IBM випустила до промислової експлуатації комерційний застосунок «Document Composing Facility» у якому використовувалася GML як мова опису документів, а в 1978 р. опубліковано «Керівництво користувача GML», яке стало базовим документом для розробки стандарту SGML.

Стандарт ISO 8879:1986 «Стандартна узагальнена мова розмітки (SGML)» було видано у 1986 р. Поправки та доповненнями до стандарту вносилися у 1988, 1996 та 1999 рр. і в такому вигляді він діє до цього часу. Останній раз його було переглянуто у 2008 р. [3].

1989–2000 рр. – Tim Berners–Lee, CERN розробив концепцію розподіленої інформаційної системи для внутрішнього використання в Лабораторії. Система призначалася для централізованого обліку та зберігання всілякої інформації за експериментами, від наукової до адміністративної і мала працювати в режимі з багатьма користувачами [4]. Вважалося, що структура документів не повинна бути ієрархічною, а документи в новій системі мають містити гіпертекстові посилання (термін «гіпертекст» – текстові документи, що пов'язані між собою без будь-яких обмежень – ввів ще в 50-х рр. Ted Nelson). Tim Berners–Lee вже мав досвід розробки гіпертекстових систем зберігання даних («Enquire», початок 80-х рр.), однак у новій системі він пропонував не обмежуватися лише текстовим вмістом, а говорити про мультимедійні документи (для майбутніх реалізацій), які могли би окрім тексту містити графіку, звук та відео. Мовою опису сторінок у новій системі стала HTML – мова розмітки гіпертекстового документа для передачі в комп'ютерній мережі, яка була застосунком SGML. Після реалізації внутрівідомча інформаційна система набула широкого розповсюдження в академічних колах і за межами CERN – поступово почало розвиватися те, що нам нині відомо як World Wide Web. В грудні 1994 р. CERN офіційно передав на подальший розвиток WWW консорціуму W3C, який було створено кількома місяцями раніше і який очолив Tim Berners–Lee. Кількість користувачів WWW, серверів та сторінок на них стрімко зростала, мова розмітки сторінок також еволюціонувала та набувала нових функціональних властивостей. Виникла необхідність стандартизувати мову, що існувала у вигляді неформальних специфікацій.

У 1993 р. один за одним були запропоновані 2 незалежних проекти, на базі яких у 1994 р. була створена робоча група при IETF, яка розробила та опублікувала у 1995 р. рекомендацію RFC1866 «Hypertext Markup Language – 2.0». Номер версії було обрано спеціально, щоб відрізняти нову специфікацію від усіх попередніх. До нової специфікації увійшли функціональні можливості мови станом на 1994 р. [5]. Через конфлікт інтересів подальша робота щодо стандартизації HTML в робочій групі при IETF зупинилася, тому була організована нова робоча група при W3C, завданням якої стала розробка подальших специфікацій HTML. Оновлені версії 3.2, 4.0, 4.01 були опубліковані відповідно у січні 1997 р., грудні 1997 р. та листопаді 1999 р. У 2000 р. за запитом робочої групи із HTML при W3C, IETF випустила специфікацію RFC2854. Ця специфікація не визначала нових стандартів, а була покликана відмінити попередні специфікації IETF стосовно HTML, дати посилання на свіжі специфікації W3C з цієї області та видалити HTML з лінійки стандартів, що підтримуються IETF. В тому ж році на підставі рекомендації W3C для HTML 4.01 опубліковано міжнародний стандарт ISO/IEC 15445:2000 HyperText Markup Language (HTML).

Починаючи з 1996 р. в робочій групі при W3C проводилась адаптація мови SGML для використання в WWW поруч із HTML, результатом якої у 1998 р. став опис мови XML версії 1.0 у вигляді рекомендації W3C. XML успадкувала від SGML велику кількість рис без змін: розподіл фізичної та логічної структури (елементи та сутності), можливість перевірки за допомогою граматик, розподіл даних та метаданих (елементи та атрибути), розподіл обробки та подання (інструкції для обробки), синтаксис із застосуванням кутових дужок. Будучи застосунком SGML, XML має деякі відмінності, а також містить деякі обмеження, які відсутні в SGML: чутливі до регістру імена елементів; використання в іменах символів Unicode окрім символів ASCII; заборонено використовувати не закриті початкові та кінцеві теги; значення атрибутів вказуються напряму (без “ ”), а не як літерали; та інші. З повним переліком відмінностей можна ознайомитися в [6]. За конструкцією документи XML відповідають специфікації SGML. Стандарт було перевидано 5 разів (останнього разу в 2008 р.) з незначними змінами, що дозволило не збільшувати номер його версії. Він на сьогодні актуальний та рекомендується для загального застосування.

У період з 1998 – 2003 роки W3C майже повністю зупинила роботу над HTML та спрямувала зусилля на користь використання XML у веб-технологіях. Але роботи над API HTML продовжували лише окремі виробники веб-браузерів, тим не менш, ці напрацювання пройшли специфікацію та були опубліковані як DOM Level 1 (1998 р.), DOM Level 2 Core и DOM Level 2 HTML (2000–2003 рр.) [7].

Для врахування змін у деяких стандартах, на яких спирався XML 1.0 у 2004 р. видано опис мови XML версії 1.1. (перевидано в 2006 р.) [8]. Порівняно з виданням XML 1.0, що існував на той час, в іменах та назвах атрибутів додано підтримку символів у форматі Unicode 3.2 та наступних версій, змінено підхід до формування імен (в іменах дозволили будь-які символи, що явно не заборонені), уточнено список символів, які розпізнаються як переведення рядка, в документі дозволено використання керуючих символів, які раніше були заборонені, а до цілої низки символів заборонено звернення напряму, тільки за посиланням (кодом). Введено поняття «повної нормалізації документа XML» (канонізації) – переліку обов’язкових правил, яких мають додержуватися творці документів, а оброблюючи програми мають їх перевіряти. Використання повної нормалізації дозволяє проводити порівняння імен, значень атрибутів і текстового вмісту як бінарне порівняння рядків в Unicode. Версію стандарту було змінено, оскільки нововведення торкнулися визначення документів, що раніше вважалися правильно сформованими. Оброблювачі XML-документів версії 1.0, як і раніше мають відкидати документи з новими символами, новими позначеннями кінця рядків та посилань на знову дозволені керуючі

символи. Вказівка на використання XML-документа стандарту 1.1 міститься в заголовній частині визначення документа.

Протягом десяти років (2000–2010) робочою групою W3C видано декілька версій стандарту XHTML (остання версія 1.1). XHTML 1.0 переформулює існуючий стандарт HTML 4.01 мовою XML. Таким чином XML-документи з інших застосунків можуть мати включення XHTML, а документи XHTML можуть містити частини з використанням інших мов розмітки. XHTML 1.0 мав 3 відгалуження: transitional, strict та formset [9]. Відгалуження transitional з деякими застереженнями відповідало HTML 4.01. В процесі роботи помічено, що різні організації змінюють HTML та XHTML, для отримання нової функціональності за допомогою нових тегів. Частіше за все такий підхід призводив до повної несумісності отриманих діалектів. Щоб уникнути втрати сумісності в майбутньому, запропонована концепція модульності мови (XHTML Modularization): мова XHTML розбивається на абстрактні модулі, з яких можна вибрати необхідні при визначенні похідної мови. XHTML 1.0 була перероблена з урахуванням цієї концепції, а стандарт отримав нову версію 1.1 [10]. Також на базі цієї концепції розроблена рекомендація XHTML Basic, де визначено мінімально необхідний перелік модулів мови для використання на різних мобільних пристроях та з іншими агентами, яким не потрібні всі можливості мови. Одним з вдалих застосувань XHTML Modularization можна вважати похідну мову XHTML+RDFa, що використовується при створенні сторінок для Semantic Web. У відповідності до обраної онтології розмічається вміст сторінки і в подальшому смислове наповнення сторінки можна використовувати для різних операцій, наприклад, для поліпшення результатів пошуку. Проводилася розробка версії стандарту 2.0, опубліковано його робочий варіант, однак W3C вирішив зупинити цей проект на користь роботи над HTML 5.0.

Опублікування у 2003 р. запропонованої на зміну Web Forms технології XForms викликало новий інтерес до розвитку самого HTML замість пошуку заміни для нього. Інтерес з'явився в наслідок усвідомлення, що розгортання XML як веб-технології більше підходить для повністю нових сервісів (наприклад, RSS, пізніше Atom), а не заміни вже працюючих технологій (як HTML). Ідею щодо продовження розвитку HTML було розглянуто на семінарі W3C у 2004 р. Mozilla й Opera подали декілька головних принципів роботи HTML5 разом з проектом пропозиції стосовно тільки форм. Пропозиція була відкинута як така, що суперечить раніше обраному напрямку розвитку веб-технологій. Співробітники та члени W3C проголосували за продовження оновлення елементів з використанням XML. Незабаром Apple, Mozilla та Opera спільно анонсували свій намір продовжити роботу над HTML5 під юрисдикцією WHATWG. WHATWG спиралася на декілька принципів: технології мають бути зворотно сумісними, специфікації та реалізації мають збігатися, навіть якщо це потребує змінити специфікацію для відповідності існуючим де-факто реалізаціям та специфікації мають бути настільки деталізованими, щоб одразу отримувати повністю сумісні реалізації, без декомпіляції та вивчення кожного окремого рішення для досягнення цього. Остання вимога означала, що до складу специфікації HTML5 необхідно було залучити вміст інших документів: HTML4, XHTML1 та DOM2 HTML. Також це потребувало включення більшої кількості подробиць, ніж вважалося нормальним на той час. В 2006 р. W3C виявила зацікавленість до HTML5 та в 2007 р. створила робочу групу для співробітництва з WHATWG у розробці специфікації HTML5. Apple, Mozilla та Opera дозволили W3C опублікувати специфікацію (остання версія від 2014 р.) [7], залишивши версію з менш обмеженою ліцензією на сайті WHATWG, яка постійно оновлюється [11].

Перехід до відкритих форматів представлення ЕД з метою обробки документів із різнотипним вмістом поступово відбувався і в середовищі офісних застосунків. Адже універсальність – це головна перевага такого підходу: при пересиланні документа в певному форматі в будь-яку організацію є впевненість, що його зміст

буде однозначно інтерпретовано, а при зверненні до ЕД через тривалий час можна забезпечити його відображення у тому самому вигляді, який він мав під час створення.

Слід зазначити, що спроби розв'язати проблему створення відкритого формату ЕД були зроблені ще до появи мови XML. У 1993 р. фірма Adobe з метою виключення операції проміжного роздрукування документа з типового циклу документообігу випустила застосунок для створення та обміну ЕД – Adobe Acrobat [12]. Разом з цим застосунком запропоновано новий формат опису ЕД, PDF 1.0, специфікація якого одразу була опублікована, а також безкоштовну програму перегляду PDF-файлів – Acrobat Reader. Цей формат базувався на спрощеному варіанті процедурної розмітки мовою PostScript, зберігав всі використані при формуванні документа шрифти, міг включати елементи растрової і векторної графіки (можливість додавати анотації, метадані, мультимедійні фрагменти з'явилась згодом) [13, 14]. Кожний бажаючий міг ознайомитися з повною версією Специфікації та реалізувати її, але право вносити до неї зміни Adobe тривалий час лишала за собою. Згодом виявилися суттєві недоліки такого (хоча і дуже прогресивного на той час) підходу. По-перше, Специфікація як набір правил не містила конкретних варіантів щодо реалізації створення вмісту PDF сторінок. По-друге, Специфікація в деяких частинах була неоднозначною, недостатньо висвітлювала окремі питання, а деякі з них взагалі не розглядала. Це спричинило появу великої кількості документів PDF, які повністю відповідали Специфікації, але могли по-різному (іноді взагалі не правильно) оброблятися існуючим програмним забезпеченням, у тому числі й від Adobe. Такі ситуації враховувалися новими правила, які додавалися до Специфікації, і згодом навіть реалізація Adobe почала відхилятися від опублікованих правил. Оскільки файли PDF мали відображатися однаково на екрані та після роздрукування на папері незалежно від платформи, було прийнято рішення підтримувати всі варіації PDF-файлів у програмі Acrobat Reader. Цей крок спричинив появу ще одного феномену: програма Acrobat Reader (а не формат PDF) стала стандартом "de facto", з яким почали звірятися виробники програмного забезпечення, де підтримувався формат PDF [15].

У рамках продовження політики переходу до відкритих стандартів представлення ЕД, у 2007 р. Adobe передала актуальну на той час специфікацію PDF 1.7 до ISO, де у 2008 р. видано новий відкритий стандарт ISO 32000-1:2008 "Document management – Portable document format – Part 1: PDF 1.7" [16]. На той час стандарт ще містив окремі частини, захищені патентами Adobe (надано публічну ліцензію на їх використання) і реалізованими лише в Adobe, але у наступній версії стандарту, ISO 32000-2:2017 "Document management – Portable document format – Part 2: PDF 2.0" ці частини були повністю вилучені. Нова редакція також описує методи перевірки відповідності стандарту файлів PDF та програм для їх обробки, [17] що має зняти проблему появи діалектів формату та правильного відображення файлів PDF різними застосунками, описану вище.

Вперше підмножина PDF/A (A означає «архівний») була стандартизована на рівні ISO у 2005 р. [18]. Цей формат був розроблений відповідно до специфікації PDF 1.4 та призначався для довгострокового зберігання документів з можливістю відтворювати і обробляти їх у майбутньому із передбачуваним результатом. Програмне забезпечення PDF, що надійно відображає документи у форматі PDF/A, є вільним і доступним.

PDF/A є похідним від формату PDF, з якого виключені деякі особливості, що не підходять для довгострокового архівного зберігання документів. Його реалізовано аналогічно визначенню підмножини PDF/X (сімейство стандартів ISO 15930) для цілей друку і поліграфії.

Стандарт додатково визначає невелику кількість вимог до програмних продуктів, які читають файли формату PDF/A. «Сумісний редактор» має дотримуватися певних правил, включаючи керування кольором, використання впроваджених шрифтів при візуалізації документа і створення

вмісту анотацій, що доступне користувачам.

Стандарт не визначає стратегію зберігання або цілі системи архівування. Він визначає «профіль» – сукупність параметрів для ЕД, які гарантують, що документ може бути відтворений у тому самому вигляді і через декілька років. Ключовий елемент відтворюваності документів у форматі PDF/A полягає у вимозі бути на 100 % самодостатніми. Вся інформація, яка необхідна для того, щоб кожен раз відображати документ у незмінному вигляді, впроваджена в файл. Сюди входить (не обмежуючись лише цим) весь візуальний контент документа: текст, растрові зображення і векторна графіка, шрифти й інформація про колір. Документи формату PDF/A не можуть використовувати зовнішні посилання на контент (як то аудіо, відео, впровадження коду на JavaScript і команд на запуск виконуваних файлів, шрифтові програми або гіперпосилання). Всі шрифти мають бути впроваджені для необмеженого універсального відображення. Це так само стосується і так званих стандартних шрифтів PostScript, таких як Times або Helvetica. Не дозволено шифрування, що зумовлює необмеженість PDF/A, він має бути відкритий та доступний будь-якій людині і програмному продукту, що відтворюють файл. Ідентифікатори користувачів і паролі неприпустимо вбудовувати. Контроль доступу виконується поза форматом файлу системою керування контентом або системою керування записами [19].

Перші спроби застосування мови XML у офісних застосунках зробила фірма Microsoft у 1998 р., в бета-версії свого пакета Office 2000: для векторної розмітки (VML), метаданих, файлу маніфесту в прихованому каталозі з ресурсами HTML-сторінки (thicket manifest) та інформації про візуальне подання (presentational information) [20]. В бета-версії Office XP (2000 р.) додано новий формат файлів електронних таблиць, заснований на XML – spreadsheetML. Цей формат залишився і в офіційному випуску пакета у 2001 р. В 2003 р. вийшов Office 2003 де було представлено новий формат текстових документів, WordprocessingML. Дослухаючись до рекомендацій Європейського Союзу від 24 травня 2004 р. [21] в листопаді 2005 р. Microsoft передала свої XML формати на подальший розгляд та стандартизацію в ECMA International. В квітні 2006 р. технічним комітетом ECMA TC45 (до складу якого входять Apple, Barclays Capital, BP, The British Library, Essilor, The Gnome Foundation, Intel, The Library of Congress, Microsoft, NextPage, Novell, Statoil та Toshiba) було підготовлено проект нового стандарту для публічного ознайомлення, а в грудні 2006 р. остаточний варіант проекту було прийнято як стандарт ECMA–376 «Office Open XML File Formats» та передано для міжнародної стандартизації в ISO. В серпні 2007 р. спроба провести прийняття стандарту за прискореною процедурою ISO (без його опрацювання в підкомітеті) зірвалася – проект не набрав достатньої кількості голосів на обох стадіях голосування [22] та був відправлений на доопрацювання у JTC1. Змінений проект було подано на розгляд за прискореною процедурою у квітні 2008 р., прийнято і опубліковано як стандарт ISO\IEC 29500:2008 «Office Open XML File Formats», у чотирьох частинах. Частини стандарту перевидавалися окремо в 2011, 2012, 2015 та 2016 рр. Подальші версії стандарту ECMA–376 приводились у відповідність із діючим стандартом ISO, його п’ята версія відповідає діючому стандарту ISO\IEC 29500 [23].

StarDivision, виробник офісного пакета StarOffice, почала розробку нового формату файлів на базі XML у 1999 р. з метою позбутися обмежень бінарного формату, впровадити підтримку Unicode та отримати відкритий формат, який змогли б використовувати й інші розробники програмного забезпечення. В тому ж році StarDivision придбана Sun Microsystems, а восени 2000 р. Sun Microsystems запустила новий проект з відкритим кодом, OpenOffice.org, в рамках якого продовжилася робота над відкритим форматом на базі XML. В 2002 р. вийшов OpenOffice.org 1.0, а новий XML-формат документів став для нього форматом за замовченням. З 2002 р. роботу над новим

форматом веде створений при OASIS (беруть участь IBM, Microsoft, Novell, KDE e.V., Nokia Corporation, Oracle, The Boeing Company та інші) технічний комітет. Формат отримав назву «OpenDocument Format», його версія 1.0 прийнята як стандарт OASIS в 2005 р. [24, 25]. В 2006 р. ODF 1.0 було затверджено ISO як ISO/IEC 26300:2006 "Open Document Format for Office Applications (OpenDocument) v1.0". ODF 1.1 затверджено як стандарт OASIS в 2007 р., прийнято та опубліковано як стандарт ISO/IEC 26300:2006/Amd 1:2012 «Open Document Format for Office Applications (OpenDocument) v1.1» в березні 2012 р. Поточна версія стандарту, ODF 1.2, затверджена як стандарт OASIS наприкінці 2011 р. та опублікована 11.01.2012 р. У 2015 р. ODF 1.2 затверджена як стандарт ISO 26300:2015. Підтримку стандарту здійснює Document Foundation. [26].

У 2002 р. в Китаї створена робоча група для розробки проекту відкритого національного стандарту подання офісних документів на базі XML-формату – UOF. Необхідність розробки виходила з міркувань поліпшення сумісності між офісними програмами місцевих виробників та підвищенням їх конкурентоспроможності на внутрішньому ринку. До складу робочої групи увійшли представники місцевих розробників офісних програмних пакетів, системних інтеграторів, користувачів та дослідницькі інститути [27]. В 2005–2006 рр. в університеті Пекіна розроблявся проект з конвертації з UOF до ODF при підтримці спеціалістів з ODF із IBM, опубліковано порівняння обох форматів [28]. Команда проекту також приймала участь у засіданнях робочої групи, яка розробляла стандарт, а внесені нею пропозиції допомогли покращити сумісність форматів UOF та ODF. У травні 2007 р. попередній варіант стандарту прийнято як національний стандарт КНР. Перший офісний пакет, EIOffice 2009, в якому UOF був основним форматом, випущено в липні 2008 р. [29].

Зручність використання специфікації мови XML 1.0 в орієнтованих на обробку даних застосунках стало відправною точкою для робіт у напрямку створення скорочених специфікацій мови, із мінімально необхідним набором мовних конструкцій та одночасно повністю таких, що відповідають вихідній специфікації. У 1999 р. групою експертів в області XML-технологій запропонована така специфікація: MinML. Однак після декількох циклів обговорень її розробники дійшли згоди, що розвиток нової специфікації як підмножини XML є скоріше недоліком, ніж перевагою, оскільки багато вже існуючих програмних засобів виявилися несумісними з новою специфікацією. Кларк Еванс (Clark Evans) запропонував змінити сам підхід до створення мови опису даних для обміну структурованою інформацією: зберегти зручне для людини текстове подання інформації, відмовитися від синтаксису XML в описі елементів, використовувати конструкції мов програмування (списки, асоціативні масиви, скалярні дані) [30]. В 2001 р. на розгляд співтовариства експертів подано попередній варіант специфікації. Виявилося, що в тому ж напрямку працював ще один фахівець, Ingy döt Net. Тому після нетривалих переговорів вони поєднали зусилля в роботі над новою специфікацією, а нову мову було названо YAML. На даний момент актуальна специфікація YAML 1.2 версія 3.

В 2001 р. Дуглас Крокфорд (Douglas Crockford) розробив та почав використовувати текстовий формат для зберігання та обміну структурованими даними, заснований на конструкціях мови програмування JavaScript із стандарту ECMA-262. ECMA Script Language Specification, видання 3. Нову мову JSON було описано в специфікації RFC4627 в 2006 р., а в 2009 р. його опис також було включено в стандарт ECMA-262 “ECMAScript Language Specification” видання 5. З 2002 р. стандарт ECMA-262 офіційно прийнято як стандарт ISO/IEC 16262 ECMAScript language specification [31]. Опис JSON винесено у окремий стандарт ECMA-404 у 2013 р. [32], який передано до публікації як стандарт ISO 21778 22.10.2017 р. [33].

Наведена хронологія та встановлені зв’язки між технологіями дозволяють отримати графічне подання (рис. 1), яке

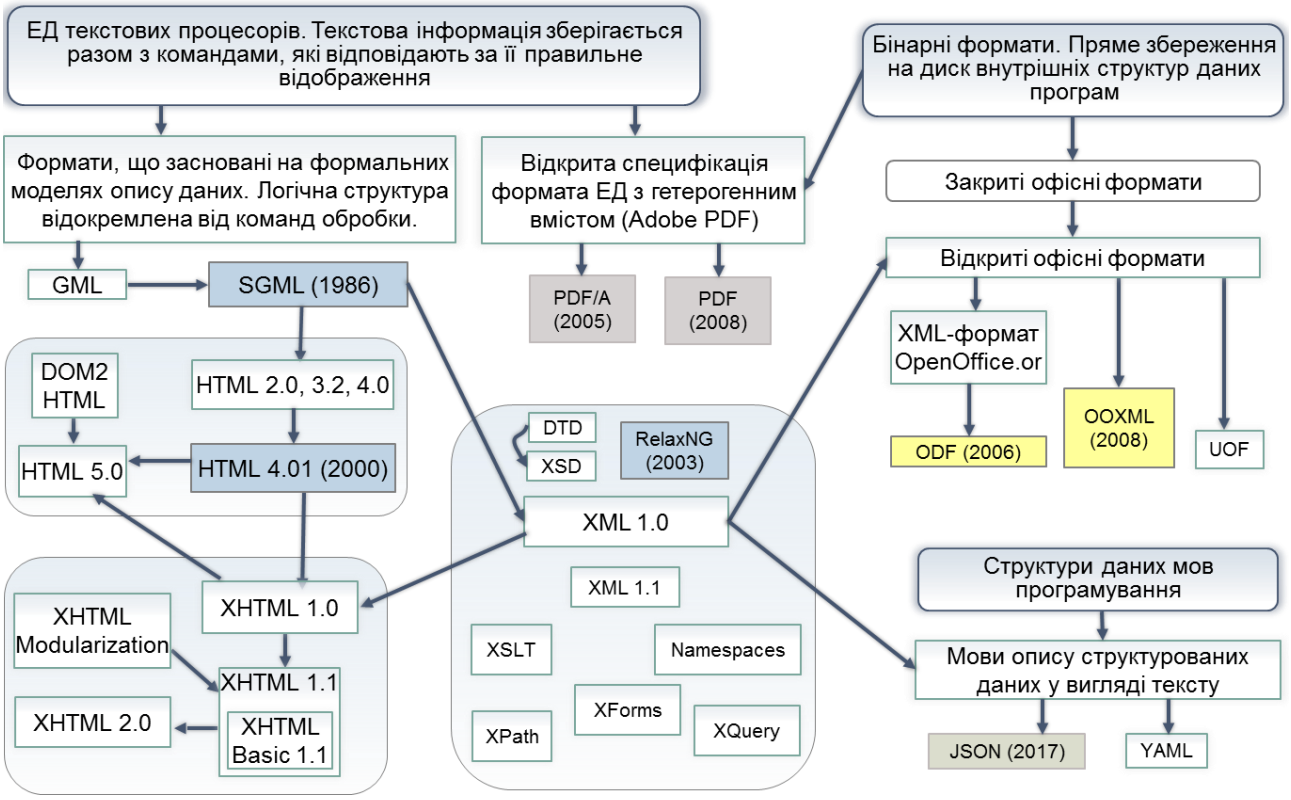

Рис. 1. Еволюція опису ЕД

наочно відображає розвиток структурованого опису ЕД. З рис. 1 видно, що узагальнені мови розмітки уточнювалися під потреби конкретних задач. Також наголосимо, що поняття мови розмітки виникло саме під час розв'язання задачі уніфікованого подання ЕД текстових процесорів на різних платформах. Однак слід зазначити, що не завжди новий стандарт передував реалізації. Прикладом є мова HTML, коли перевірений часом галузевий стандарт "de facto" офіційно проходив стандартизацію в міжнародних установах значно пізніше.

З діаграми видно, що стандартизацію на рівні ISO (рік прийняття першої версії стандарту ISO для технології вказано у дужках) проходять не всі формати та мови опису ЕД. Для більшої частини технологій достатньою є стандартизація на рівні галузевих консорціумів. Такий підхід забезпечує швидке уточнення та оновлення стандартів на підставі змін в основоположних документах, відгуків зацікавлених учасників, незалежних експертів та швидке виправлення знайдених помилок.

Також видно, що в структурованому описі ЕД особливе місце займає технологія XML (мова розмітки у поєднанні з додатковими специфікаціями для роботи з цими документами). Таким чином при розробці нових застосунків замість проектування "з нуля" внутрішнього формату подання даних та реалізації операцій із ним можна використовувати вже існуючу, детально специфіковану, мета-мову і зосередити увагу на розробці схем опису даних свого застосунка. Необхідність в міжнародній стандартизації мови XML відсутня, оскільки вона є підкласом SGML – вже стандартизованої мови, а документи XML відповідають специфікації SGML.

Особливе місце серед стандартизованих форматів опису ЕД займають формати, які не базуються на формальних моделях опису ЕД. Сімейство форматів PDF останніх версій підтримує структурований опис ЕД (анотації, метадані), має всесвітнє визнання, але за більше ніж 20 років існування специфікацій досі не має надійного вирішення проблеми перевірки відповідності PDF-файлів та застосунків, які ці файли створюють, до стандарту. Пропонується встановлювати відповідність стандарту шляхом порівняння результату обробки PDF-файлів з еталонними зображеннями, які є копіями екранів з результатами

роботи еталонної програми відтворення [15]. Тобто не повністю виключається ситуація, коли придатний з точки зору стандарту документ буде не правильно відображений на екрані. Очевидно, такий спосіб програє у надійності методам перевірки відповідності стандарту документів, заснованих на формальних моделях опису. У останньому випадку треба перевіряти документ лише відносно формальної граматики, яка обов'язково входить до специфікації таких стандартів. Ця перевірка може бути легко реалізована у застосунку, який обробляє документи, з метою відкидання файлів, що не пройдуть перевірку. Відносно формату PDF/A також залишається багато питань обробки, які не регулюються стандартом [19]. Тому при використанні цього формату для довгострокового зберігання ЕД доведеться приймати додаткові внутрішні регламенти на рівні корпорації.

Опис закритих офісних форматів мовою XML та створення відкритих стандартів для них визначили новий напрямок в обробці структурно описаних ЕД, в рамках якого стало можливим використовувати існуючі в світі напрацювання по офісним системам для розв'язання задач довгострокового зберігання документів, їх редагування, гарантованого однакового відображення на різних платформах й безболісного обміну інформацією між офісними застосунками конкуруючих компаній, а також вводити в офісні документи розмітку з визначеною користувачем семантикою, тобто створювати структуровані ЕД.

В напрямку розв'язання задач виключно обміну структурованою інформацією також з'явилися свої формати з відмінним від XML синтаксисом та підтримкою основних структур даних мов програмування.

Розглянемо означені вище напрямки більш детально.

Відкриті офісні формати

Як вже було зазначено, спочатку офісні документи зберігалися як пряме відображення структур даних офісних програм на диск. Це дозволяло компактно зберігати дані та швидко завантажувати їх до програми для продовження роботи. Постійне зростання обчислювальної потужності апаратного забезпечення, пропускної здатності комп'ютерних мереж та виникнення нових стандартів опису інформації (а саме мови XML) дозволило розпочати розробку нових відкритих форматів зберігання офісних документів. Метою створення таких форматів є виключення залежності документів від застосунків, надання можливості використовувати новий формат різним розробникам програмного забезпечення у своїх продуктах та на різних платформах (наприклад, в мобільних пристроях), не лише для обробки та довгострокового зберігання документів, а й для вирішення задач автоматизації генерування нових документів з бізнес-даних, автоматичного витягу даних з документів з наступною передачею в інші застосунки і т. ін. З рис. 1 видно, що при створенні нових застосунків, пов'язаних з обробкою та візуалізацією структурованої інформації доцільно використовувати один з актуальних міжнародних стандартів: OOXML чи ODF.

Нажаль UOF, скоріш за все, залишиться лише національним стандартом через низьку зацікавленість у ньому міжнародної спільноти, що в першу чергу зумовлено використанням ієрогліфів у розмітці (рис. 2).

OOXML засновано на моделях даних офісних програм Microsoft, підтримує всі можливості, які існують в їх бінарних форматах зберігання даних. Базовий рівень міжплатформової взаємодії забезпечується повною відповідністю OOXML діючим відкритим стандартам W3C XML 1.0 та XML Namespaces (простору імен XML) – [34–36]. В документах OOXML максимально виключено звернення до властивостей, залежних від поточного оточення. Якщо така властивість використовується, то її значення у системі, в якій створений документ, може бути збережено в документі (наприклад, при використанні константи кольору в документі також можна зберегти і значення, яке вона мала при створенні документа). Доступна можливість використовувати схеми користувача

```
<字:句 uof:locID="t0085">
    <字:句属性 uof:locID="t0086" uof:attrList="式样引用" 字:式样引用="Standard"/>
    <字:文本串 uof:locID="t0109" uof:attrList="标识符">Фрагмент текстового
    документа в UOF.</字:文本串>
</字:句>
```

Рис. 2. Використання ієрогліфів в іменах атрибутів та тегів документа UOF

(визначати простори імен) в форматі XSD та обробляти документи без участі застосунка де вони були створені (наприклад, за допомогою перетворень XSLT). OOXML не накладає обмежень на використання сервісів безпеки, таких як XML Digital Signature та XML Encryption.

В рамках OOXML розрізняють відповідність стандарту документа та відповідність стандарту застосунка. Споживаючий застосунок відповідає стандарту якщо він не відкидає документи очікуваного формату, що відповідають стандарту. Створюючий застосунок відповідає стандарту, якщо він здатен створювати документи, що відповідають стандарту. В стандарті введено поняття суворого (Strict) та перехідного (Transitional) форматів, які мають різні простори імен, щоб розрізняти документи в цих форматах поміж собою та відрізняти від документів у стандарті ECMA 376 першого видання, які фактично відповідають формату Transitional. Формат Transitional додано для сумісності з документами та застосунками, що вже існували на той час. Нові документи мають створюватися в форматі Strict, якщо вони проголошують відповідність до стандарту IS29500.

Додатково підтримуються такі мови розмітки: DrawingML (для подання графічних об’єктів у документі), VML (подання векторної графіки, залишено для зворотної сумісності та буде замінено на DrawingML), мови розмітки Math, Metadata, Custom XML, Bybliography, а загалом модель документів формату OOXML показана на рис. 3.

В OOXML використовуються схеми XSD, схеми RelaxNG мають інформативний характер.

Стандарт ISO\IEC 29500 «Office Open XML File Formats» складається з чотирьох частин:

- базові поняття та довідкове керівництво з мови розмітки. Містить всю необхідну інформацію щодо створення нових документів формату Strict [35];
- погодження щодо упаковки частин документів. Описано вимоги до формування файлу-контейнера та його структура;
- сумісність та розширюваність розмітки. Сформульовані правила визначення сумісності, описано підходи до забезпечення сумісності розмітки з майбутніми версіями стандарту, подано інструкції щодо роботи з невідомими (цій версії специфікації) просторами імен або їх елементами;
- особливості міграції у перехідний формат. Надано вказівки щодо високоякісної міграції з бінарних форматів документів Microsoft до документів формату OOXML Transitional.

Відповідність документа стандарту ODF:

- документи можуть містити елементи з зовнішніх просторів імен;
- споживаючий застосунок має прочитати (або записати) документ який був би відповідним, якщо з нього видалити всю зовнішню розмітку;
- споживаючі застосунки мають зберігати зовнішню розмітку.

В ODF схеми описуються за допомогою RelaxNG.

Стандартом ODF передбачена можливість зберігання документів як у вигляді

Рис. 3. Загальна модель документів OOXML

єдиного файлу XML, так і у вигляді декількох XML-файлів в контейнері (ZIP-архів). У випадку використання контейнера документ розбивається на структурні частини з власними кореневими елементами для подання змістовної частини, метаданих, стилів та налаштувань [37]. Для порівняння моделей документів ODF та OOXML розглянемо саме цей спосіб зберігання документа ODF (рис. 4).

В документах ODF розрізняють зумовлені метадані та метадані користувача (входять до елемента «метадані»). Так само як і в OOXML зумовлені метадані використовують напрацювання Dublin Core Metadata Initiative у вигляді множини елементів для опису документа (ресурсу) із наперед заданою семантикою [38], що також відображено у стандарті ISO 15836:2009 «The Dublin Core metadata element set». Метадані користувача можуть описувати будь-які додаткові властивості документа. Починаючи з версії 1.2 в стандарт ODF також інтегровано підтримку опису елементів метаданих за допомогою RDF, що дозволило використовувати елементи метаданих у змістовній частині документа та в частині, яка містить стилі.

Налаштування документа подані у вигляді множини, елементами якої можуть бути:

- поодинокі елементи – для зберігання конкретних налаштувань;
- множини елементів – перелік елементів у довільній послідовності;
- індексовані масиви елементів – доступ до елементів за індексом;
- асоціативні масиви елементів – доступ до елементів за ім’ям.

Останні три форми подання налаштувань є контейнерами та застосовуються для групування однотипних налаштувань. Для контейнерів, які належать кореневому елементу налаштувань, має бути вказано назву простору імен, що однозначно визначає належність налаштувань тому чи іншому застосунку.

З рис. 3 та 4 видно, що моделі документів у форматах ODF та OOXML суттєво відрізняються. В стандарті OOXML існує декілька типів документів: текстовий, електронна таблиця, презентація із власними мовами розмітки. Вони розроблялися в різний час і мають небагато спільного. В ODF визначено єдину мову розмітки в якій описано всі елементи документа та атрибути цих елементів. Документи ODF мають єдину загальну структуру. Елементи тіла документа, що визначають тип документа та відображають його специфіку, обов’язково мають власні дочірні елементи “атрибути”, “пролог”, “основна частина” та “епілог” з відповідними префіксами, їх наповнення залежить від типу документа. Типові конструкції основної частини тіла документа (текст, таблиці,

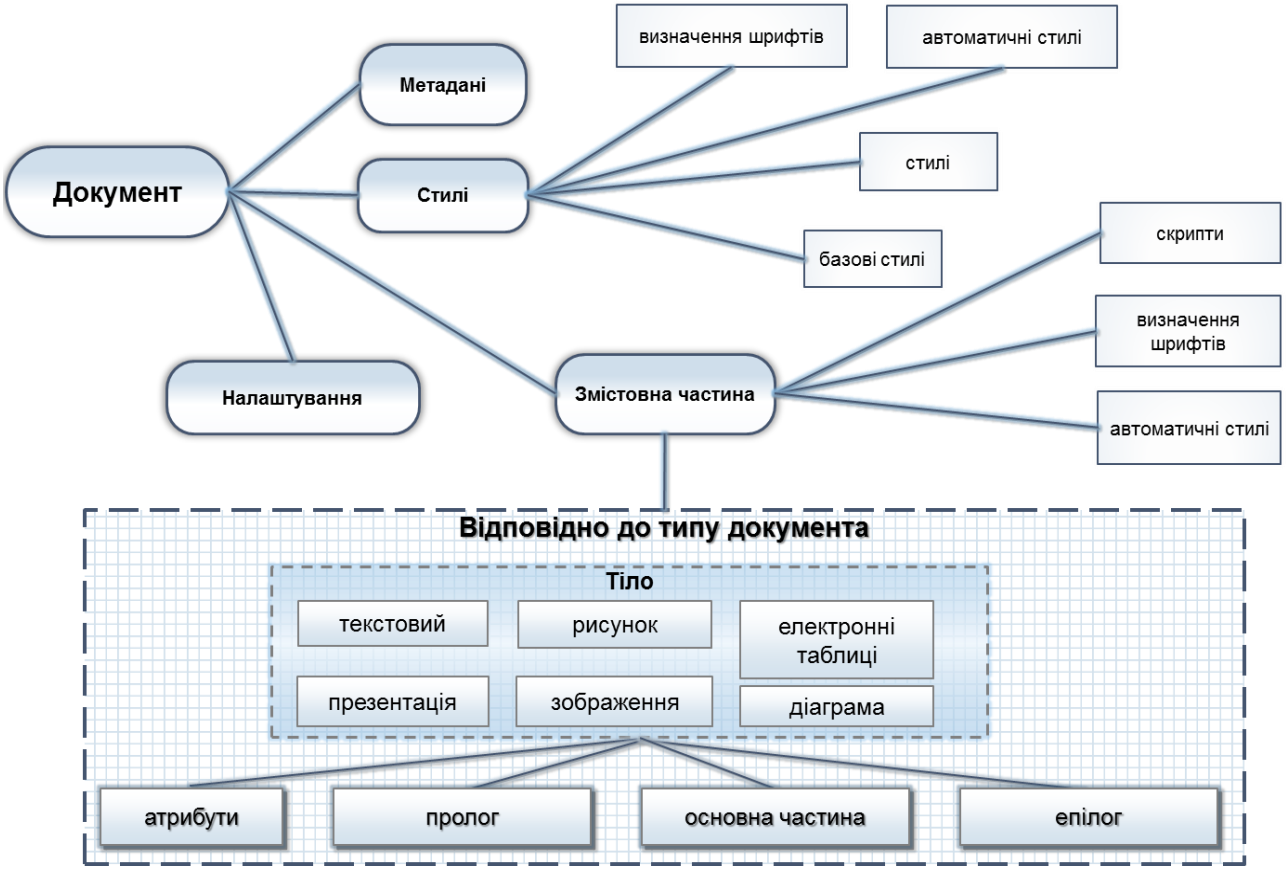

Рис. 4. Загальна модель документів стандарту ODF

рисунки тощо) формалізовані для спільного використання у будь-якому з документів, але від типу документа залежить, які елементи можуть у ньому з'явитися і в яких комбінаціях.

Документи обох стандартів підтримуються провідними офісними пакетами: Microsoft Office, OpenOffice, LibreOffice, KOffice, Google Docs, IBM Lotus Symphony, Corel WordPerfect Office, деякими іншими. Документи у форматі OOXML Strict можуть читати лише LibreOffice з версії 4.3.2 та Microsoft Office починаючи з версії 2010. Записувати у форматі OOXML Strict може лише Microsoft Office починаючи з версії 2013. Застосунки інших виробників використовують формат OOXML Transitional навіть тоді, коли документ не містить несумісних із форматом OOXML Strict частин [39].

Одним з важливих напрямків роботи над розглянутими стандартами є їх гармонізація, яка має полегшити конвертацію з одного формату до іншого. В першу чергу це прискорить та здешевить процес розробки програмного забезпечення, оскільки покращить сумісність застосунків, які мають справу з обміном документами у різних форматах, експортом чи імпортом даних з них. В стандарті ISO/IEC TR 29166:2011 "Керівні вказівки щодо переміщення між форматами документів ISO/IEC 26300 та ISO/IEC 29500" [40] проводиться аналіз основоположних концепцій в обох стандартах документів. На підґрунті ретельного вивчення кожного стандарту визначається наявність спільних рис та розбіжностей документів, побудовано таксономію властивостей документа взагалі. Шляхом порівняння оцінюється ступінь точності переміщення між документами в кожному зі стандартів та у разі можливості детально розглядається, як важлива функціональність одного формату документів може бути подана в іншому. Вивчаються саме стандарти, а не їх реалізації в застосунках, оскільки реалізації можуть породжувати додаткові питання сумісності [41].

Структуровані формати для обміну даними

Основними критеріями використання спеціалізованих мов опису даних у задачах обміну структурованими даними без необхідності їх візуального подання є простота та зручність перегляду. Проаналізуємо на відповідність цим критеріям мов опису структурованих даних JSON та YAML. Обидві мови не містять спеціалізованих тегів розмітки для позначення елементів структурованої інформації. Інформаційна модель цих мов гранично проста. Документ YAML являє собою спрямований граф, вершинами якого можуть бути колекції, списки, скалярні значення, порожні значення, що визначають типи даних в мові. Синтаксис для опису вузлів графа документа YAML мінімальний – «:» розділяє пари «ключ-значення», «%» вказує, що далі йде опис деякого структурованого типу («колекції»), який складається з неупорядкованих пар «ключ-значення» з унікальним ключем, «@» означає початок списку вузлів. Елементами списку може бути будь-який з дозволених типів даних. Входження пар «ключ-значення» в колекції, а також входження елементів в список визначається відступом всередині рядка (кількістю пробілів перед елементом). В JSON визначені 2 типи елементів документа. Колекція – складається з одної чи більше пар виду «ключ-значення». Синтаксично задається за допомогою «{» та «}», ключ і значення в парах відокремлюються «:», самі пари відокремлюються «,». Список (масив) впорядковано містить будь-який з дозволених типів (список, колекція, рядок, число). Початок та кінець масиву позначається «[» та «]», перелік ведеться через «,».

Вочевидь інформаційні моделі цих мов дуже схожі, однак модель збереження даних в YAML дозволяє визначати посилання на елементи документа («&») та в подальшому звертатися до цих елементів за посиланням («*»), що сприяє суттєвому скороченню обсягу інформації для передачі та зниженню ймовірності виникнення помилки при створенні часто повторюваних даних. Крім цього в YAML існує так званий «скорочений запис», який збігається з синтаксисом JSON. Таким чином в окремих випадках документ YAML при використанні скороченого запису може ставати допустимим документом JSON. На користь вибору JSON як спрощеного формату обміну структурованою інформацією говорить і те, що він пройшов стандартизацію у ISO та підтримується у бібліотеках багатьох популярних середовищ розробки.

Висновок

В результаті проведеного аналізу встановлено, що провідне положення серед способів подання структурованих електронних документів (СЕД) займає мова XML 1.0. Основними критеріями такого вибору є простота сприйняття людиною текстової інформації, чітка формалізація мови, а також спирання цієї мови на міжнародний стандарт SGML. До переваг XML також можна віднести наявність формалізованих програмних інтерфейсів DOM, SAX та StAX [42] – механізмів, що дозволяють будь-який XML-документ представити у вигляді програмного об'єкта, через який із структурою та/або даними документа може працювати будь-який програмний застосунок. Будь-яке сучасне програмне середовище має у своєму складі спеціальні програмні компоненти для роботи з XML-документами. XML досить добре пристосований для зберігання посилань, ієрархічних, табличних і бінарних даних. Це дозволяє використовувати його для зберігання найрізноманітніших типів інформації. Найважливішою особливістю XML є його розширюваність: у разі необхідності обміну якимись специфічними документами можливо розробити власний діалект мови XML, створити файл з його описом (XSD-документ) та поширити його між усіма учасниками документообігу. Цього буде достатньо для правильної обробки документів такого типу. Є величезна різноманітність діалектів XML для опису документів у самих різних предметних областях і каталоги подібного роду описів (наприклад, XBRL [43], ті ж самі OOXML та ODF). Ще однією важливою особливістю

XML саме в контексті ЕДО є можливість легкого перетворення будь-якого XML-документа в звичайний, легкий для читання HTML-документ за допомогою XSLT-перетворення.

В корпоративних інформаційних системах використання відкритих форматів офісних документів, заснованих на мові XML, яка є стандартом «де факто», дозволяє вирішити задачі:

- довгострокового зберігання документів – гарантується можливість коректного відображення документів не тільки поточними версіями офісних пакетів, а й майбутніми;
- обміну документами;
- однакового відображення документа на різних платформах, в тому числі і на мобільних приладах;
- обробки СЕД засобами текстового редактора чи табличного процесора з метою автоматичного отримання потрібної інформації для подальшого внесення її до бази даних;
- автоматичне формування СЕД з результатів запиту до бази даних;
- захисту документа за допомогою цифрового підпису;
- збереження разом із основною інформацією додаткової інформації, що відображає семантику документа.

Стандарти на обидва розглянутих відкритих формати (OOXML та ODF) представлення офісних документів мають детальний опис їх структури, тому вони можуть бути використані як внутрішній формат подання даних у корпоративній інформаційній системі. Це дозволить скоротити витрати часу на розробку та підтримку власного формату даних та повністю зосередитися на архітектурі майбутньої системи. Використання стандартизованих форматів дає ще одну перевагу: оскільки вони спираються на інші міжнародні стандарти, з'являється можливість залучити до роботи найкращі з існуючих їхніх реалізацій.

Розглянуті структуровані формати подання даних у вигляді тексту в корпоративній інформаційній системі доцільно використовувати для внутрішніх комунікацій між частинами системи.

Література

1. Goldfarb C.F. The Roots of SGML – A Personal Recollection [Електронний ресурс]. 1996. Режим доступу: http://www.sgmlsource.com/history/roots.htm. – Назва з екрана.
2. Goldfarb C.F. Design Considerations for Integrated Text Processing Systems. [Електронний ресурс]. IBM. Cambridge Scientific Center Technical Report G320–2094. 1973. Режим доступу: https://web.archive.org/web/20040627050036/http://www.sgmlsource.com:80/history/G320-2094/G320-2094.htm. – Назва з екрана.
3. Information processing – Text and office systems – Standard Generalized Markup Language (SGML) : ISO 8879:1986 [Електронний ресурс]. International Organization for Standardization. 1986. Режим доступу: https://www.iso.org/standard/16387.html. Назва з екрана.
4. Berners–Lee T. Information Management: A Proposal [Електронний ресурс]. CERN. 1989. Режим доступу: http://www.w3.org/History/1989/proposal.html. Назва з екрана.
5. Berners–Lee T., Connolly D. Hypertext Markup Language – 2.0 : RFC1866 [Електронний ресурс]. November 1995. Режим доступу: http://www.rfc-editor.org/rfc/rfc1866.txt. Назва з екрана.
6. Clark J. Comparison of SGML and XML : World Wide Web Consortium Note 15–December–1997 [Електронний ресурс]. W3C. 1997. Режим доступу: http://www.w3.org/TR/NOTE–sgml–xml–971215. Назва з екрана.
7. Berjon R. HTML5 : A vocabulary and associated APIs for HTML and XHTML : W3C Recommendation 28 October 2014 [Електронний ресурс]. W3C. 2014. Режим доступу: http://www.w3.org/TR/html5/. Назва з екрана.
8. Bray T. Extensible Markup Language (XML) 1.1 (Second Edition): W3C Recommendation 16 August 2006, edited in place 29 September 2006 [Електронний ресурс]. W3C. 2006. Режим доступу: http://www.w3.org/TR/2006/REC-xml11-20060816/. Назва з екрана.

9. Pemberton S. XHTML™ 1.0 The Extensible HyperText Markup Language (Second Edition) : A Reformulation of HTML 4 in XML 1.0 : W3C Recommendation 26 January 2000, revised 1 August 2002 [Електронний ресурс]. W3C. 2002. Режим доступу: http://www.w3.org/TR/xhtml1/. Назва з екрана.
10. Pemberton S. XHTML2 Working Group Home Page [Електронний ресурс]. W3C. 2007. Режим доступу: http://www.w3.org/MarkUp/. Назва з екрана.
11. HTML : Living Standard — Last Updated 22 November 2017 [Електронний ресурс]. – [WHATWG]. [2006?]. Режим доступу: https://html.spec.whatwg.org/multipage/. Назва з екрана.
12. Adobe Acrobat 1.0 - Product brochure [Електронний ресурс]. Adobe Systems. 1993. Режим доступу: http://www.planetpdf.com/planetpdf/pdfs/adobe_acrobat1_broch.pdf. Назва з екрана.
13. Warnock J. The Camelot Project [Електронний ресурс]. 1990. Режим доступу: http://www.planetpdf.com/planetpdf/pdfs/warnock_camelot.pdf. Назва з екрана.
14. Wootton A.T. How was the PDF format created? [Електронний ресурс]. Quora. 2014. Режим доступу: https://www.quora.com/PDF-file-format/How-was-the-PDF-format-created. Назва з екрана.
15. Johnson D. Is PDF an Open Standard? [Електронний ресурс]. 2010. Режим доступу: https://talkingpdf.org/is-pdf-an-open-standard/. Назва з екрана.
16. Document management – Portable document format. Part 1: PDF 1.7 : ISO 32000-1:2008 [Електронний ресурс]. International Organization for Standardization. 2008. Режим доступу: https://www.iso.org/standard/51502.html. Назва з екрана.
17. Document management – Portable document format. Part 2: PDF 2.0 : ISO 32000-2:2017 [Електронний ресурс]. International Organization for Standardization. 2017. Режим доступу: https://www.iso.org/standard/63534.html. Назва з екрана.
18. Document management – Electronic document file format for long-term preservation. Part 1: Use of PDF 1.4 (PDF/A-1) : ISO 19005-1:2005 [Електронний ресурс]. International Organization for Standardization. 2005. Режим доступу: https://www.iso.org/standard/38920.html. Назва з екрана.
19. Johnson D. White Paper: How to Implement PDF/A [Електронний ресурс]. 2010. Режим доступу:https://talkingpdf.org/white-paper-how-to-implement-pdfa/. Назва з екрана. 20. Jones B. Open XML timeline [Електронний ресурс]. Brian Jones. 2007. Режим доступу: https://blogs.msdn.microsoft.com/brian_jones/2007/07/09/open-xml-timeline/. Назва з екрана.
20. TAC approval on conclusions and recommendations on open document formats [Електронний ресурс]. IDABC: [Interoperable Delivery of European eGovernment Services to public Administrations, Businesses and Citizens]. 2006. Режим доступу: http://web.archive.org/web/20060720005118/http://ec.europa.eu/idabc/en/document/2592/5588. Назва з екрана.
21. Sayer P. “ISO Rejects Microsoft's OOXML as Standard” [Електронний ресурс]. PCWorld: IDG News Service. 2007: [04.09.2007]. Режим доступу: http://www.pcworld.com/article/136711/iso_rejects_microsofts_ooxml_as_standard.html. Назва з екрана.
22. Standard ECMA-376 : Office Open XML File Formats [Електронний ресурс]. ECMA International. 2006. Режим доступу: http://www.ecma-international.org/publications/standards/Ecma-376.htm. Назва з екрана.
23. History of OpenDocument [Електронний ресурс]. 2006. Режим доступу: http://opendocument.xml.org/milestones. Назва з екрана.
24. Cover R. OpenOffice.org XML File Format [Електронний ресурс]. OASIS. 2006. Режим доступу: http://xml.coverpages.org/starOfficeXML.html. Назва з екрана.
25. Durusau P. Developing an XML-based file format specification for office applications [Електронний ресурс]. OASIS Open Document Format for Office Applications (OpenDocument) TC. [2005?]. Режим доступу: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office. Назва з екрана.
26. Updegrove A. Another Open Document Format – From China [Електронний ресурс]. The Standards Blog. 2006. Режим доступу: http://www.consortiuminfo.org/standardsblog/article.php?story=2006110806164573. Назва з екрана.
27. Zhong Chen, Liyong Tang, Huiping Sun et al. ODF–UOF Converter [Електронний

ресурс]. 2006. Режим доступу: http://odf-to-uof.sourceforge.net/overview.html. Назва з екрана.

28. Updegrove A. Don't Forget UOF: Here Comes EIOffice 2009 (Updated 2X [Електронний ресурс]. The Standards Blog. 2008: [21.07.2008]. Режим доступу: http://www.consortiuminfo.org/standardsblog/article.php?story=20080721140512962. Назва з екрана.
29. Ingerson B., Clark Evans & Oren Ben-Kiki. Yet Another Markup Language (YAML) 1.0 : Working Draft 01 Aug 2001 [Електронний ресурс]. 2001. Режим доступу: http://web.archive.org/web/20070217091403/http://yaml.org:80/spec/history/2001-08-01.html. Назва з екрана.
30. Standard ECMA–262 : ECMAScript 2017 Language Specification [Електронний ресурс]. ECMA International. 2017. Режим доступу: http://www.ecma-international.org/publications/standards/Ecma-262.htm. Назва з екрана.
31. Standard ECMA-404: The JSON Data Interchange Format [Електронний ресурс]. ECMA International. 2013. Режим доступу: http://www.ecma-international.org/publications/standards/Ecma-404.htm. Назва з екрана.
32. Information technology – The JSON data interchange syntax: ISO/IEC 21778:2017 [Електронний ресурс]. International Organization for Standardization. 2017. Режим доступу: https://www.iso.org/standard/71616.html. Назва з екрана.
33. Ngo T. Office Open XML overview [Електронний ресурс]. ECMA International. [2006?]. Режим доступу: http://www.ecma-international.org/news/TC45_current_work/OpenXML%20White%20Paper.pdf. – Назва з екрана.
34. Information technology – Document description and processing languages – Office Open XML File Formats – Part 1: Fundamentals and Markup Language Reference: ISO/IEC 29500–1:2016 [Електронний ресурс]. International Organization for Standardization. – 2016. – Режим доступу: https://www.iso.org/standard/71691.html. Назва з екрана.
35. Weir R. A technical comparison: ISO/IEC 26300 vs. Microsoft Office Open XML [Електронний ресурс]. OpenOffice.org Conference (OOoCon 2006): September 11 – 13 2006. Lyon, France. 2006. (Доклад). 2006. Режим доступу: http://www.openoffice.org/marketing/ooocon2006/presentations/wednesday_o3.pdf. Назва з екрана.
36. ISO/IEC 26300-1:2015: Information technology – Open Document Format for Office Applications (OpenDocument) v1.2. Part 1: OpenDocument Schema [Електронний ресурс]. International Organization for Standardization. 2015. Режим доступу: https://www.iso.org/standard/66363.html. Назва з екрана.
37. Kunze J., Baker T. The Dublin Core Metadata Element Set [Електронний ресурс Request for Comments 5013]. The IETF Trust. 2007. Режим доступу: http://www.ietf.org/rfc/rfc5013.txt. Назва з екрана.
38. Sustainability of Digital Formats: Planning for Library of Congress Collections [Електронний ресурс]. [Library of Congress, U.S.?]. [2017?]. Режим доступу: https://www.loc.gov/preservation/digital/formats/fdd/fdd000401.shtml. Назва з екрана.
39. ISO/IEC TR 29166:2011 : Information technology – Document description and processing languages – Guidelines for translation between ISO/IEC 26300 and ISO/IEC 29500 document formats [Електронний ресурс]. International Organization for Standardization. 2011. Режим доступу: https://www.iso.org/standard/45245.html. Назва з екрана.
40. Eckert Klaus Peter, Goluchowicz Kerstin, Gauch Stephan, Kirchhoff Björn. Feature Based Document Profiling – a Key For Document Interoperability? [Електронний ресурс]. Fraunhofer FOKUS, Berlin. 2012. Режим доступу: https://cdn1.scrvt.com/fokus/403d3c76ef3c1e76/ea5963383889/Feature_Based_Document_Profiling.pdf. Назва з екрана. – ISBN 978-3-00-038675-6
41. Sanaulla M. Parsing XML using DOM, SAX and StAX Parser in Java [Електронний ресурс]. 2013. Режим доступу: https://sanaulla.info/2013/05/23/parsing-xml-using-dom-sax-and-stax-parser-in-java/. Назва з екрана.
42. Engel P., Hamscher Walter et al. Extensible Business Reporting Language (XBRL) 2.1 : Recommendation 31 December 2003 with errata corrections to 20 February 2013 [Електронний ресурс]. XBRL International. 2013. Режим доступу: http://www.xbrl.org/Specification/XBRL-2.1/REC-2003-12-31/XBRL-2.1-REC-2003-12-31+corrected-errata-2013-02-20.html. Назва з екрана.

References

1. Goldfarb C. F. The Roots of SGML – A Personal Recollection. 1996. [Electronic resource]. Mode of access: http://www.sgmlsource.com/history/roots.htm
2. Goldfarb C. F. Design Considerations for Integrated Text Processing Systems. IBM. Cambridge Scientific Center Technical Report G320–2094. 1973. [Electronic resource]. Mode of access: https://web.archive.org/web/20040627050036/http://www.sgmlsource.com:80/history/G320-2094/G320-2094.htm
3. Information processing -- Text and office systems -- Standard Generalized Markup Language (SGML) : ISO 8879:1986. International Organization for Standardization. 1986. [Electronic resource]. Mode of access: https://www.iso.org/standard/16387.html
4. Berners–Lee T. Information Management: A Proposal. CERN. 1989. [Electronic resource]. Mode of access: http://www.w3.org/History/1989/proposal.html
5. Berners–Lee T. Hypertext Markup Language – 2.0 : RFC1866. November 1995. [Electronic resource]. Mode of access: http://www.rfc-editor.org/rfc/rfc1866.txt
6. Clark J.Comparison of SGML and XML : World Wide Web Consortium Note 15–December–1997. W3C. 1997. [Electronic resource]. Mode of access: http://www.w3.org/TR/NOTE–sgml–xml–971215
7. Berjon R. HTML5 : A vocabulary and associated APIs for HTML and XHTML : W3C Recommendation 28 October 2014. W3C. 2014. [Electronic resource]. Mode of access: http://www.w3.org/TR/html5/
8. Bray T. Extensible Markup Language (XML) 1.1 (Second Edition) : W3C Recommendation 16 August 2006, edited in place 29 September 2006. W3C. 2006. [Electronic resource]. Mode of access: http://www.w3.org/TR/2006/REC-xml11-20060816/
9. Pemberton S. XHTML™ 1.0 The Extensible HyperText Markup Language (Second Edition) : A Reformulation of HTML 4 in XML 1.0 : W3C Recommendation 26 January 2000, revised 1 August 2002. W3C. 2002. [Electronic resource]. Mode of access: http://www.w3.org/TR/xhtml1/
10. Pemberton S. XHTML2 Working Group Home Page. W3C. 2007. [Electronic resource]. Mode of access: http://www.w3.org/MarkUp/
11. HTML : Living Standard — Last Updated 22 November 2017. [WHATWG]. [2006?]. [Electronic resource]. Mode of access: https://html.spec.whatwg.org/multipage/
12. Adobe Acrobat 1.0 - Product brochure. Adobe Systems. 1993. [Electronic resource]. Mode of access: http://www.planetpdf.com/planetpdf/pdfs/adobe_acrobat1_broch.pdf
13. Warnock J. The Camelot Project. 1990. [Electronic resource]. Mode of access: http://www.planetpdf.com/planetpdf/pdfs/warnock_camelot.pdf
14. Wootton A.T. How was the PDF format created?. Quora. 2014. [Electronic resource]. Mode of access: https://www.quora.com/PDF-file-format/How-was-the-PDF-format-created
15. Johnson D. Is PDF an Open Standard?. 2010. [Electronic resource]. Mode of access: https://talkingpdf.org/is-pdf-an-open-standard/
16. Document management – Portable document format -- Part 1: PDF 1.7 : ISO 32000-1:2008 [Електронний ресурс]. – International Organization for Standardization. – 2008. – Режим доступу: https://www.iso.org/standard/51502.html. – Назва з екрана.
17. Document management – Portable document format -- Part 2: PDF 2.0 : ISO 32000-2:2017. International Organization for Standardization. 2017. [Electronic resource]. Mode of access: https://www.iso.org/standard/63534.html
18. Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1) : ISO 19005-1:2005. International Organization for Standardization. 2005. [Electronic resource]. Mode of access: https://www.iso.org/standard/38920.html
19. Johnson D. White Paper: How to Implement PDF/A. 2010. [Electronic resource]. Mode of access: https://talkingpdf.org/white-paper-how-to-implement-pdfa/. – Назва з екрана.
20. Jones B. Open XML timeline. 2007. [Electronic resource]. Mode of access: https://blogs.msdn.microsoft.com/brian_jones/2007/07/09/open-xml-timeline/
21. TAC approval on conclusions and recommendations on open document formats. IDABC : [Interoperable Delivery of

European eGovernment Services to public Administrations, Businesses and Citizens]. 2006. [Electronic resource]. Mode of access: http://web.archive.org/web/20060720005118/http://ec.europa.eu/idabc/en/document/2592/5588
22. Sayer P. "ISO Rejects Microsoft's OOXML as Standard". PCWorld : IDG News Service. 2007. [Electronic resource]. Mode of access: http://www.pcworld.com/article/136711/iso_rejects_microsofts_ooxml_as_standard.html
23. Standard ECMA-376 : Office Open XML File Formats. ECMA International. 2006. [Electronic resource]. Mode of access: http://www.ecma-international.org/publications/standards/Ecma-376.htm
24. History of OpenDocument. 2006. [Electronic resource]. Mode of access: http://opendocument.xml.org/milestones
25. Cover R. OpenOffice.org XML File Format. OASIS. 2006. [Electronic resource]. Mode of access: http://xml.coverpages.org/starOfficeXML.html
26. Durusau P. Developing an XML-based file format specification for office applications. OASIS Open Document Format for Office Applications (OpenDocument) TC. [2005?]. [Electronic resource]. Mode of access: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office
27. Updegrove A. Another Open Document Format – From China. The Standards Blog. 2006. [Electronic resource]. Mode of access: http://www.consortiuminfo.org/standardsblog/article.php?story=2006110806164573
28. Zhong C. ODF–UOF Converter. 2006. [Electronic resource]. Mode of access: http://odf-to-uof.sourceforge.net/overview.html
29. Updegrove A. Don't Forget UOF: Here Comes EIOffice 2009 (Updated 2X. The Standards Blog. 2008. [Electronic resource]. Mode of access: http://www.consortiuminfo.org/standardsblog/article.php?story=20080721140512962
30. Ingerson B. Yet Another Markup Language (YAML) 1.0 : Working Draft 01 Aug 2001. 2001. [Electronic resource]. Mode of access: http://web.archive.org/web/20070217091403/http://yaml.org:80/spec/history/2001-08-01.html
31. Standard ECMA–262 : ECMAScript 2017 Language Specification. ECMA International. 2017. [Electronic resource]. Mode of access: http://www.ecma-international.org/publications/standards/Ecma-262.htm
32. Standard ECMA-404 : The JSON Data Interchange Format. ECMA International. 2013. [Electronic resource]. Mode of access: http://www.ecma-international.org/publications/standards/Ecma-404.htm
33. Information technology -- The JSON data interchange syntax : ISO/IEC 21778:2017. International Organization for Standardization. 2017. [Electronic resource]. Mode of access: https://www.iso.org/standard/71616.html
34. Ngo T. Office Open XML overview. ECMA International. [2006?]. [Electronic resource]. Mode of access: http://www.ecma-international.org/news/TC45_current_work/OpenXML%20White%20Paper.pdf
35. Information technology -- Document description and processing languages -- Office Open XML File Formats -- Part 1: Fundamentals and Markup Language Reference : ISO/IEC 29500–1:2016. International Organization for Standardization. 2016. [Electronic resource]. Mode of access: https://www.iso.org/standard/71691.html
36. Weir R. A technical comparison: ISO/IEC 26300 vs. Microsoft Office Open XML // OpenOffice.org Conference (OOoCon 2006) : September 11 - 13 2006 – Lyon, France. 2006. [Electronic resource]. Mode of access: http://www.openoffice.org/marketing/ooocon2006/presentations/wednesday_o3.pdf
37. ISO/IEC 26300-1:2015 : Information technology -- Open Document Format for Office Applications (OpenDocument) v1.2 -- Part 1: OpenDocument Schema. International Organization for Standardization. 2015. [Electronic resource]. Mode of access: https://www.iso.org/standard/66363.html
38. Kunze J. The Dublin Core Metadata Element Set Request for Comments 5013. The IETF Trust. 2007. [Electronic resource]. Mode of access: http://www.ietf.org/rfc/rfc5013.txt
39. Sustainability of Digital Formats: Planning for Library of Congress Collections. [Library of Congress, U.S.?]. [2017?]. [Electronic resource]. Mode of access: https://www.loc.gov/preservation/digital/formats/fdd/fdd000401.shtml
40. ISO/IEC TR 29166:2011 : Information technology -- Document description and processing languages -- Guidelines for translation between ISO/IEC 26300 and

ISO/IEC 29500 document formats. International Organization for Standardization. 2011. [Electronic resource]. Mode of access: https://www.iso.org/standard/45245.html

41. Eckert K.-P. Feature Based Document Profiling - a Key For Document Interoperability?. Fraunhofer FOKUS, Berlin. 2012. [Electronic resource]. Mode of access: https://cdn1.scrvt.com/fokus/403d3c76ef3c1e76/ea5963383889/Feature_Based_Document_Profiling.pdf. ISBN 978-3-00-038675-6
42. Sanaulla M. Parsing XML using DOM, SAX and StAX Parser in Java. 2013. [Electronic resource]. Mode of access: https://sanaulla.info/2013/05/23/parsing-xml-using-dom-sax-and-stax-parser-in-java/
43. Engel P. Extensible Business Reporting Language (XBRL) 2.1 : Recommendation 31 December 2003 with errata corrections to 20 February 2013. XBRL International. 2013. [Electronic resource]. Mode of access: http://www.xbrl.org/Specification/XBRL-2.1/REC-2003-12-31/XBRL-2.1-REC-2003-12-31+corrected-errata-2013-02-20.html

Одержано 09.01.2018

Про авторів:

Шарипанов Антон Веніамінович,
молодший науковий співробітник.
Кількість наукових публікацій в українських виданнях – 7.
Кількість наукових публікацій в зарубіжних виданнях – 5,
Індекс Хірша – 2.
https://orcid.org/0000-0001-6804-0533.

Щетинін Ігор Євгенович,
кандидат технічних наук,
старший науковий співробітник,
т.в.о. завідувача відділу.
Кількість наукових публікацій в українських виданнях – більше 30.
https://orcid.org/0000-0002-9131-4405.

Іванов Сергій Миколайович,
молодший науковий співробітник.
Кількість наукових публікацій в українських виданнях - 7.
https://orcid.org/0000-0002-2003-8538.

Місце роботи авторів:

Інститут кібернетики
імені В.М. Глушкова НАН України,
03187, м. Київ-187, проспект Глушкова, 40.
Тел.: +38(044) 526 4568.
E-mail: _sha_@ukr.net,
shchet@nas.gov.ua