УДК 681.3.06:686.1.03

© В. О. Кохановський, к.т.н., доцент, НТУУ «КПІ», Київ, Україна

ПАРАЛЕЛЬНА ПІДГОТОВКА ДРУКОВАНОГО НАУКОВО-ТЕХНІЧНОГО ДОКУМЕНТА ТА ЙОГО ЕЛЕКТРОННИХ ВЕРСІЙ

На даний час переважаюча частка інформації у Web представлена у форматі HTML (HyperText Markup Language). Використання HTML для публікацій матеріалів із великою кількістю математичних текстів на даний час досить обмежена. Класичний спосіб відтворення формул у вигляді графічного зображення має ряд суттєвих недоліків і суперечить сучасній концепції, згідно якої дані повинні бути відділені від форми. Для конвертації пропонується конвертор TeX4HT, який для будь-якої команди чи процедури Latex дає можливість визначити відповідний еквівалент мовою розмітки, що використовується вихідним документом (наприклад, HTML, XHTML, MathML). Для конвертора написані стильовий та конфігураційний файли, за допомогою яких можна отримати електроні версії у форматах PDF та HTML +MathML.

Ключові слова: науково-технічний документ; Latex; формат PDF; pdflatex; HTML сторінка; MathML; TeX4HT; плагін MathPlayer.

Постановка проблеми

Оформлення математичних текстів завжди було непростою справою. Велика кількість досить специфічних символів, використання кількох алфавітів (латинського, грецького, готичного та ін.), просторове розміщення різних частин тощо перетворює набір кожної більш менш складної формули на досить непросту та кропітку процедуру.

Одним з основних методів підготовки математичних текстів на даний час є широко розповсюджені процесори Microsoft Word та Math Type. Однак, форматам таких документів притаманний ряд суттєвих недоліків. Основні з них — це закритість та орієнтація, як правило, на поліграфічну розмітку всупереч логічній, яка слабко виражена. Текст настільки перевантажений різними форматуючими тегами, що не залишається місця для семантики та прагматики документа. Така характеристика ставить серйозні проблеми на шляху автоматичної обробки документів.

Існує ще одна проблема з документами Word — сумісність. Певні складнощі — від некоректного відображення деяких символів до повної нечитабельності формул, виникають навіть при використанні різних версій цьо-

ISSN 2077-7264.— Технологія і техніка друкарства.— 2015.— № 1(47)

ISSN 2077-7264.— Технологія і техніка друкарства.— 2015.— № 1(47)

го текстового процесора, а що вже говорити про інші комп'ютери і ОС.

На даний час, існує перевірений і розповсюджений (особливо на Заході) програмний продукт типу Tex для якісної підготовки математичних текстів. Проблема у тому, що робота з системою Tex схожа на програмування, розрахована на логічний, а не візуальний спосіб набору, використовує пакетну обробку файлів і численні конфігураційні файли. Розібратися в ній непросто, особливо користувачам, що звикли до графічних інтерфейсів та маніпулятора «мишки».

Видавнича система Tex несе на собі відбитки американського стилю оформлення видань, а тому виникла задача адаптації її до українських правил та традицій.

Втім, технології не стоять на місці, і на сьогодні є цілком можливим об'єднати універсальність і гнучкість цієї системи з простотою і інтуїтивністю сучасних текстових процесорів.

Аналіз попередніх досліджень

Визначальний крок у напрямку створення комп'ютеризованої видавничої системи для підготовки математичних видань належить видатному американському математику Дональду Кнуту [1]. Наприкінці 70-х років минулого століття він створив систему Tex, що не втратила свого значення дотепер і стала фактично стандартом для підготовки наукової літератури.

1) *Система Latex.* Розроблену Д. Кнутом видавничу систему багато зарубіжних спеціалістів відносять до одного з найбільших досягнень минулого століття, прирівнюючи її появу до створення друкарського верстата Гуттенбергом. Ось як сам Кнут характеризує своє дітище: «Це нова система набору, призначена для створення красивих книг і особливо таких, що містять багато математики. Підготувавши рукопис у форматі Tex, ви тим самим точно вкажете комп'ютеру, як перетворити рукопис у сторінки, друкарська якість яких порівнюється з роботою кращих друкарів» [2]. Керуючись благородними мотивами, Кнут надав своїй системі статус «public domain». Це означає, що вона не захищена від копіювання й вільно розповсюджується з навчальною та просвітницькою метою.

2) *Кириличні шрифти.* Сучасні реалізації Tex'а, зокрема, Miktex та TexLive містять достатню кількість ефективних засобів для роботи з кирилицею. Як приклад зазначимо PS-шрифт «PsCyr», що входить в обидва зазначені дистрибутиви. Це якісний та безкоштовний шрифт, що дозволяє оформляти наукові тексти, не піклуючись про узгодження шрифтів для формул та тексту, та представляти документ у форматі PDF.

Остання серйозна перешкода для повсюдного переходу від Latex до Pdflatex зникла в кінці 2001 року, коли В. Волович створив PostScript-версію кириличних LH-шрифтів. Пояснимо, що так звані mf-шрифти, технологія яких була розроблена Д. Кнутом спеціально для сис-

теми Tex, не придатні для впровадження в документ pdf — в Acrobat Reader вони виглядають жахливо. Згадані PS-шрифти на даний час включені в дистрибутив Miktex.

3) *Переноси для кириличного тексту.* Для переносів слів у системі Tex використовується алгоритм М. Лайєнга. Цей алгоритм використовує мало пам'яті, працює досить швидко й знаходить майже всі правильні точки переносів. Алгоритм відзначається гнучкістю, і його можна застосувати до будь-якої мови, а також для переносів у кількох мовах одночасно. Для застосування цього методу до конкретної мови потрібно побудувати для неї так звані таблиці зразків. Розробкою таких зразків для української мови займались Д. Вуліс, М. Поляков, А. Швайка та інші. При переносі слова система спочатку шукає його в словнику винятків, де зберігаються слова, переноси яких не охоплені зразками або протирічать їм. Якщо в ньому слова немає, то Tex звертається до таблиці зразків, на основі якої й знаходить можливі точки переносів. Підключення шрифтів і зразків переносів відбувається простим підключенням відповідних пакетів.

4) *Адаптація системи Latex до вітчизняної поліграфії.* Документи, підготовлені згідно американських та вітчизняних правил, суттєво різняться. Наведемо кілька прикладів таких розбіжностей, зазначивши при цьому, що їх перелік можна значно збільшити.

У нас при наборі чисел ціла частина від дробової відділяється десятковою комою, а числа при перерахуванні відділяються одне від одного крапка з комою. У американському наборі ціла частина від дробової відділяється десятковою крапкою, а числа розділяються комами.

Перенесення частини формули здійснюють на знаку операції (порівняння, арифметична операція, тощо). При цьому знак операції дублюється на початку наступного рядка згідно вітчизняних традицій і не дублюється згідно американських, що доцільно з програмістської точки зору.

У вітчизняному наборі кожний абзац починається відступом, чого не скажеш про американський, де перший абзац розділу відступу не має. В американських книгах кожний розділ починається з правої сторінки, у нас він може розпочинатися і з лівої.

У підписах до рисунків та таблиць ми використовуємо крапку після номера, а в американському наборі використовується двокрапка. Сказане стосується теорем, лем та інших математичних абзаців.

В американському наборі використовують інше накреслення та порядок розміщення зовнішніх та внутрішніх лапок, знаків тире, трикрапок та ін. Існує чимало розбіжностей при позначенні різних операцій. Наприклад, по-різному позначаються нестрогі нерівності, дійсна та уявна частини комплексного числа, знак порожньої множини та ін.

Зусиллями багатьох вчених та ентузіастів створений пакет

ISSN 2077-7264.— Технологія і техніка друкарства.— 2015.— № 1(47)

ISSN 2077-7264.— Технологія і техніка друкарства.— 2015.— № 1(47)

кирилізації T2, у якому є засоби для адаптації Tex'а до українських умов. Цей пакет продовжує розвиватися, і наразі його можливостей вистачає для якісної підготовки математичних текстів українською мовою [3–6].

Результати проведених досліджень

1) *Конвертація формату Latex у формат PDF.* Перетворення документа Latex у формат PDF здійснюється буквально в один дотик.

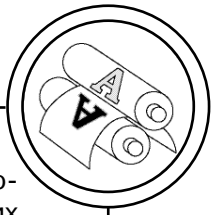
Автори книги «Latex по-русски» І. Котельникова і П. Чеботаєва [3] пишуть, що з недавніх пір традиційний сценарій компіляції tex-документів у формат DVI став не потрібним, оскільки тепер є все необхідне для прямого перетворення такого тексту у формат PDF. Простіше всього це зробити, замінивши компілятор latex на компілятор pdflatex. Програма pdflatex на виході створить файл у форматі PDF. Його можна проглянути на екрані або надрукувати за допомогою програми Acrobat Reader, безкоштовно поширюваною фірмою Adobe, або за допомогою іншої широко відомої програми Gsview, які поставляються окремо від системи. При друкуванні pdf-файл дає таку ж високу якість, як і dvi-файл, але його значно простіше переслати електронною поштою або експонувати в інтернеті, оскільки програма Acrobat Reader може вбудовуватися в найбільш поширені браузери.

До сказаного можна додати, що формат PDF став де-факто стандартом у видавничо-поліграфічній галузі. Редакції журналів присилають авторам статті на коректуру саме у форматі PDF. На сайтах видавництв опубліковані статті виставляються здебільшого у форматі PDF. Учені вважають також за краще обмінюватися вже опублікованими статтями у вигляді файлів PDF, хоча ще кілька років тому в подібних випадках доводилося відправляти електронною поштою вхідний текст (у розмітці Latex) і рисунки (у вигляді окремих файлів). Обмін друкованими матеріалами у форматі PDF гарантує, що при відтворенні матеріалу на будь-якому вихідному пристрої (принтері) матеріал буде відображений абсолютно однаково.

Обнадійливо виглядають pdf-формати в питаннях захисту інформації від несанкціонованого копіювання. Документ у форматі PDF може бути зашифрований і захищений паролем як на відкриття документа, так і на виконання таких типових операцій, як друкування документа або «запозичення» ілюстрацій та різних фрагментів.

PDF запізнився з виходом в Інтернет. Тільки з 1996 року стало можливим показувати документи PDF в інтернет-браузерах, коли у Web-просторі вже міцно вкоренився HTML.

Документи у форматі PDF, як правило, мають більший розмір, ніж HTML-документи. Це плата за точність відображення інформації. На зорі інтернету проблема пропускної спроможності мережі стояла дуже гостро, а типова наукова стаття у форматі PDF має розмір від 0,5 Mb і вище. Зараз гострота проблеми



зменшується, але якась частина користувачів інтернету свій вибір вже зробила.

Істотне значення відіграють і психологічні аспекти. Навіть людина без спеціальної освіти здатна швидко навчитися правити HTML-документи вручну, використовуючи простий текстовий редактор і кілька необхідних зразків. Однак, документи PDF неможливо редагувати без спеціальних інструментів.

2) *Конвертація формату Latex у формат HTML+MATHML.* Впровадження математики в Web — це не просто пошук способів відображення математичної інформації у вікні браузера. Всесвітня мережа представляє фундаментально новий підхід до зберігання знань, у якій взаємодія різних додатків, комп'ютерів та їхніх платформ відіграє центральну роль. Стає все більш важливим знайти способи взаємодії між різними прикладними програмами й документами, підготовленими на їхній основі. Це завдання значно ширше, ніж просте відображення математичних формул в Інтернеті.

Зрозуміло, що математика та її нотація не одне й те ж. Математичні ідеї (зміст) існують незалежно від способу їхнього представлення (форми). Однак, можливість маніпулювати ідеями в символьній формі — визначальна риса математичного апарату як інструменту опису й аналізу. Труднощі при впровадженні математики в Web полягають у тому, щоб зафіксувати форму й зміст так, аби в документах максимально використати високо розвинуту систему математичної нотації й потенціал взаємодії електронних засобів інформації.

На даний час переважаюча частка інформації в Web представлена у форматі HTML (HyperText Markup Language). Проте використання HTML для публікацій матеріалів із великою кількістю математичних текстів на даний час досить обмежена, незважаючи на наявність близько 20 програмних продуктів, призначених для цієї мети (див. http://w3c.com.org/Math#SoftWare). Класичний спосіб відтворення формул у вигляді графічного зображення суперечить сучасній концепції, згідно якої дані повинні бути відділені від форми, оскільки графічне зображення вже є формою.

Результатом зберігання формул у вигляді графічного зображення є неможливість оперувати формулами, як даними. Конкретніше, у формулах-рисунках неможливо здійснювати пошук та заміну, форматування й інші, звичні для нас, операції з текстом. Таке представлення формул утруднене для індексації з метою пошуку, не дає можливості використовувати їх за прямим призначенням: виконувати обчислення, будувати графіки, діаграми й тощо. З іншого боку, документи з великою кількістю графічних формул займають великий об'єм, вантажаться тривалий час і т. ін. Ці та інші проблеми викликали гостру необхідність у розробці ефективних засобів для повноцінного представлення математики в інтернеті.

Спробою розв'язати розглянуті проблеми стало створення

ISSN 2077-7264.— Технологія і техніка друкарства.— 2015.— № 1(47)

наприкінці минулого століття мови MathML. Буквально абревіатура MathML [2] розшифровується як Mathematical Markup Language (мова математичної розмітки). MathML — це заснована на принципах XML мова розмітки документів для запису математичних формул і виразів. Вона описує як зовнішній вигляд формул, так і їхній зміст. Основний принцип MathML полягає в тому, щоб будувати та вставляти математичні конструкції в HTML-документ досить простим способом. Мова пропонує гнучку й розширювану систему запису математичного матеріалу, дозволяє взаємодіяти із зовнішніми програмами, забезпечує можливість високоякісного відтворення в різних інформаційних середовищах. MathML забезпечує також повторне використання математичних виразів в інших застосуваннях або іншому контексті, а також дозволяє індексацію математичної інформації з метою її ефективного пошуку.

Важко не погодитися з тим, що вибір та тестування програмних продуктів — це досить важка частина дослідження. Якщо спочатку прийняти невірне рішення, то в кращому разі буде безповоротно загублений час, у гіршому — робота буде провалена. На даний час багато програмних продуктів та технології застаріли, чимало — розв'язують тільки дуже вузьке завдання, деякі — явно відстали через неадекватний маркетинг, інші — ще занадто молоді, щоб зрозуміти, чи буде з них користь. Важко не розгубитися в такім різноманітті. Єдине, що можна констатувати з упевненістю — так це те, що розв'язання проблеми відображення математичних текстів в Інтернеті далеке від ідеалу.

При написанні статті було проаналізовано майже 20 програм, з яких вибраний конвертор TeX4HT [2]. Він підтримує всі інструменти Latex, у тому числі перехресні посилання, автоматичну нумерацію і т. ін. У конверторі передбачені широкі можливості для настроювання й розширення його функцій. TeX4HT може створювати на виході гіпертекстові документи в різних форматах, у тому числі HTML+растр, XHTML+MathML, XML і т. ін.

Практично важливою властивістю системи TeX4HT є широкі можливості для її конфігурації. TeX4HT для читання вхідного тексту документа використовує систему Latex, тому не має обмежень на набір команд та процедур. Для будь-якої команди та процедури Latex можна визначити відповідний еквівалент мовою розмітки, що використовується вихідним документом (наприклад, HTML, XHTML, MathML). Замість HTML вихідний документ може бути записаний у розмітці XML або іншій, це залежить від того, як визначені еквіваленти команд і процедур Latex.

Автор системи TeX4HT Эйтан Гурарі розробив кілька варіантів конфігураційних і стильових файлів, які дозволяють одержувати на виході системи файли в розмітці HTML, XHTML, XML у сполученні з різними формами представлення математичних формул: від малюнків у форма-

ISSN 2077-7264.— Технологія і техніка друкарства.— 2015.— № 1(47)

тах GIF, PNG, JEPG до розмітки MathML. Перерахувати всі варіанти нелегко, оскільки при виборі розмітки можливі подальші варіанти: можна або орієнтуватися на представлення розмітки MathML у вікні браузера за допомогою плагіна MathPlayer, або за допомогою XSL стилів, розроблених Д. Карлайлом (David Carlisle) [2], орієнтованих на більш широкий спектр браузерів. Зокрема, стилі Карлайла дозволяють переглядати документи з розміткою MathML у вікні браузера Netscape 7, тоді як перший варіант дозволяє переглядати результати за допомогою браузера Microsoft Internet Explorer (див. http://www.oasis-open.org/cover/gurari-ml9808.html).

Широкі можливості пакета TeX4HT у його конфігуруванні й визначили наше рішення — використовувати його в якості базового для побудови системи перетворення наукових і навчальних документів із формату Latex у формат HTML+MathML. Ми зупинили свій вибір на відображенні розмітки MathML браузером Internet Explorer за допомогою плагіна MathPlayer. З появою інших, широко використовуваних браузерів, здатних відображати розмітку MathML, адаптація вихідного файла до цих браузерів не буде представляти особливих труднощів.

Процес трансляції вхідного документа в гіпертекст складається із трьох етапів: компіляції вхідного тексту (ми використали для цієї мети MiKTex) в DVI-код, обробки DVI-коду програмою TeX4HT і виконання заключних процедур, необхідних для завершення трансляції.

Висновки

1) Можна констатувати з упевненістю, що розв'язання проблеми відображення математичних текстів в Інтернеті далеке від ідеалу.

2) Пропонована технологія тестувалася й відпрацьовувалася при підготовці друкованої та електронних версій навчального посібника [7]. У результаті були написані стильовий та конфігураційний файли, що дало змогу уникнути помилок, неминучих при такого роду перетвореннях. У кінцевому результаті за допомогою систем Latex, TeX4HT та деяких інших допоміжних програм створена інтернет-версія видання.

Зазначений посібник — досить складне видання. Понад 50 % його об'єму складають формули, таблиці, графіки та діаграми. Успішна конвертація такого складного видання у різні електронні версії свідчить, що в основному технологія вибрана вірно, і нею можна користуватися при розв'язанні подібних задач.

3) Технологія дає можливість отримати паралельно з друкованим виданням, ще й електронні у форматах PDF та HTML + MathML. Останнє позначення означає, що для представлення формул використовується спеціальна мова математичної розмітки MathML, оскільки їхнє традиційне зображення у вигляді рисунків має цілий ряд принципових недоліків.

ISSN 2077-7264.— Технологія і техніка друкарства.— 2015.— № 1(47)

Список використаної літератури

1. Кнут Д. Все про TEX / Д. Кнут. — М. : Издательский дом Вильямс, 2003. — 687 с.
2. Гуссенс М. Путеводитель по пакету LaTeX и его Web-приложениям / М. Гуссенс, С. Ратц. — М. : Мир, 2001. — 606 с.
3. Котельников И. А. З. LaTeX по-русски / И. А. Котельников, П. Чеботаев. — Новосибирск : Сибирский хронограф, 2004. — 496 с.
4. Гуссенс М. Путеводитель по пакету LaTeX и его расширению LaTeX2e / М. Гуссенс, Ф. Миттельбах, А. Самарин. — М. : Мир, 1999. — 612 с.
5. Спивак М. Восхитительный TeX / М. Спивак : Руководство по комфортному изготовлению научных публикаций в пакете AMS-TeX. — М. : Мир, 1993. — 452 с.
6. Львовский С. М. Набор и верстка в пакете LaTeX / С. М. Львовский. — 3-е издание. — М. : Космоинформ, 2003. — 448 с.
7. Дорош А. К. Теорія ймовірностей та математична статистика / А. К. Дорош, О. П. Коханівський. — Київ : Київський політехнік, 2006. — 300 с.

References

1. Knut D. Vse pro TEH / D. Knut. — M. : Izdatel'skij dom Vil'jams, 2003. — 687 s.
2. Gussens M. Putevoditel' po paketu LaTeX i ego Web-prilozhenijam / M. Gussens, S. Ratc. — M. : Mir, 2001. — 606 s.
3. Kotel'nikov I. A. Z. LaTeX po-russki / I. A. Kotel'nikov, P. Chebotaev. — Novosibirsk : Sibirskij hronograf, 2004. — 496 c.
4. Gussens M. Putevoditel' po paketu LaTeX i ego rasshireniju LaTeX2e / M. Gussens, F. Mittel'bah, A. Samarin. — M. : Mir, 1999. — 612 s.
5. Spivak M. Voshititel'nyj TeH / M. Spivak : Rukovodstvo po komfortnomu izgotovleniju nauchnyh publikacij v pakete AMS-TeH. — M. : Mir, 1993. — 452 s.
6. L'vovskij S. M. Nabor i verstka v pakete LaTeX / S. M. L'vovskij. — 3-e izdanie. — M. : Kosmoinform, 2003. — 448 c.
7. Dorosh A. K. Teoriia ymovirnostei ta matematychna statystyka / A. K. Dorosh, O. P. Kokhanivskyi. — Kyiv : Kyivskyi politekhnik, 2006. — 300 s.

В работе рассмотрены вопросы использования формата tex для параллельной подготовки печатного научно-технического издания и его электронных версий. Для конвертации предлагается пакет TeX4HT, широкие возможности которого дают возможность конвертировать научно-технические документы из формата tex в электронные документы типа HTML+MathML.

Ключевые слова: научно-технический документ; Latex; формат PDF, pdflatex; HTML страница; MATHML; TeX4HT; плагин MathPlayer.

The article considered the questions to using Tex-format for a parallel preparation printing edition of scientific and technical publications and its electronical versions. To convert proposed package TeX4HT, opportunities which allow to convert scientific and technical documents from the Tex-format in electronic documents such as HTML + MathML.

Keywords: scientific and technical document; Latex; PDF format; pdflatex; HTML page; MATHML;TeX4HT; MathPlayer.

Рецензент — В. Ф. Морфлюк, д.т.н., професор, НТУУ «КПІ»

Надійшла до редакції 31.03.15