УДК 004.7

Д. В. Ланде, Б. О. Березін
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна

Підхід до оцінки живучості наукових публікацій при довготерміновому зберіганні в інтернет-середовищі

Визначено особливості представлення наукових публікацій у мережі Інтернет, що впливають на їхню живучість при довготерміновому зберіганні. Запропановано моделі для оцінки живучості наукових публікацій при зберіганні у мережі Інтернет і показано можливості їхнього використання.

Ключові слова: *оцінка живучості, наукові публікації, довготермінове зберігання, републікація, доступність, формати даних.*

Постановка проблеми, її актуальність та аналіз публікацій

Розвиток інформаційних ресурсів мережі Інтернет веде до появи в їхньому складі якісно нових ресурсів, що вимагають довготермінового зберігання. Прикладами таких ресурсів є правова, урядова інформація, електронні журнали (які створюються переважно у цифровому вигляді, без паперових копій), професійні блоги та інші. Все частіше у США важливі правові матеріали більше не публікуються в друкованому вигляді і доступні тільки в глобальній мережі, у зв'язку з чим було прийнято «Типовий закон про правові акти, що публікуються в електронному вигляді» («Uniform Electronic Legal Material Act» — UELMA). Іншим видом інтернет-ресурсів, які значною мірою створюються у цифровому вигляді, не мають паперової копії та потребують довготермінового зберігання, є електронні журнали. На теперішній час у світі нараховується більше двадцяти тисяч найменувань електронних журналів. У той же час, дослідження показують обмеженість часу зберігання інформаційних ресурсів на веб-серверах мережі Інтернет. Аналіз доступності посилань на наукові публікації (НП) з плином часу було проведено за допомогою запитів до пошукової системи Google Scholar. Отримані результати (рис. 1) показують, що через рік після видання НП доступними є більше 90 % посилань, через 4 роки біля 70 %, а через 14 років лише біля 30 % посилань. У роботі [1], присвяченій стабільності URL (доступності посилань на веб-ресурси з правової інформації з плином часу) показано, як для набору з близько 600 веб-ресурсів аналізувалася доступність посилань в Інтернеті. В результаті виявилося, що протягом

першого року стали недоступними більш як 8 % URL, за другий рік кількість недоступних URL зросла до більш як 14 %, на третьому році недоступних посилань стало понад 28 %. Дані про доступність посилань наведені також в [2]. Серед існуючих прикладів подолання ризиків втрати правової інформації, що створюється у вигляді веб-сторінок — контент сайту http://public.resource.org та деяких інших, де розміщено сотні гбайт правової інформації. Ця інформація довготерміново зберігається в мережі USDocs Private LOCKSS Network, побудованої на базі майже двадцяти університетських бібліотек і використовується їхніми читачами [3]. За допомогою засобів проекту LOCKSS [4], інтернет-контент, до якого бібліотеки повинні надавати довготерміновий доступ своїм читачам, збирається з відповідних веб-сайтів і довготерміново зберігається у вигляді кількох копій на вузлах мережі P2P, що створюється на базі бібліотек.

Рис. 1. Результати аналізу частки доступних посилань на інтернет-ресурси в НП 2013–2000 рр. Доступність посилань аналізувалась у вересні 2014 р.

Довготермінове зберігання е-журналів в Інтернеті забезпечується електронними архівами. Реєстр зберігачів (The Keepers Registry, http://thekeepers.org), відображає стан зі зберіганням та архівуванням е-журналів. У переліку зберігачів у Реєстрі наведено біля 10 проектів та організацій: Global LOCKSS Network**,** CLOCKSS Archive**,** National Science Library (Chinese Academy of Sciences**)** та інші. Наведена статистика показує, що більше 22 тис. найменувань електронних серійних видань зберігаються принаймі одним зберігачем, а більш як 8 тис. — трьома і більше зберігачами. Нарешті, в роботі [5] наведені підходи для забезпечення зберігання та довготермінового доступу (long term access) до контенту професійних блогів (присвячених виконанню проектів, або відображаючих будь-які події тощо).

Довготермінове зберігання та доступ до інформаційних об’єктів (ІО) передбачає подолання різних видів загроз: пошкодження носіїв інформації, старіння

носіїв/обладнання, старіння програмного забезпечення/форматів, помилки операторів, атаки, природні катастрофи, економічні помилки тощо. Тобто, для довготермінового зберігання та доступу до ІО необхідне забезпечення живучості ІО. Живучість об'єкта — це властивість виконувати основні функції в умовах негативних впливів (НВ), за необхідності тимчасово відмовляючись від виконання деяких другорядних функцій [6].

У наведених вище прикладах, довготермінове зберігання різних видів контенту Інтернет забезпечується за допомогою електронних архівів, сервісів постійного зберігання, які використовують методи реплікації тощо. В той же час, проведений аналіз [7] показує, що не тільки всередині електронних архівів, але й в інтернет-середовищі контент зберігається, як правило, у вигляді кількох копій, версій тощо. Ця та інші особливісті представлення ІО в інтернет-середовищі, які впливають на живучість зберігання ІО на сьогодні детально не досліджені.

Метою роботи є визначення особливостей представлення наукових публікацій у мережі Інтернет, які впливають на їхню живучість при довготерміновому зберіганні, а також відповідної моделі для оцінки живучості НП при довготерміновому зберіганні в інтернет-середовищі.

Особливості представлення НП при зберіганні в Інтернеті

Аналіз показує, що серед основних особливостей представлення наукових публікацій у мережі Інтернет, які впливають на живучість при довготерміновому зберіганні, можна виділити републікацію, доступність, індексованість і поширеність форматів даних.

Републікація. При користуванні інформаційними ресурсами в мережі Інтернет може відбуватися копіювання, розмноження їхніх версій, тобто републікація. Оцінка живучості ІО при довготерміновому зберіганні в інтернет-середовищі може значною мірою залежати від кількості версій НП. При визначенні характеристик такого процесу републікації контенту в Інтернеті використовувалися запити до пошукової системи Google Scholar для отримання списку публікацій із заданої тематики (наприклад, тематики «довготермінове зберігання») протягом періоду в один рік. В отриманому списку публікацій за допомогою пакету Excel аналізувалося поле «всі версії статті». Проранжировані кількості версій НП за вибраною тематикою за 1998 р. в Інтернеті представлені на рис. 2 (ордината представлена в логарифмічних координатах). Отримані таким чином розподіли для виборок розміром біля 100 НП в усі роки інтервалу 1998–2013 рр. також були апроксимовані за допомогою експоненціальної функції. Додатково до аналізу републікації контенту НП було проаналізовано републікацію контенту електронних журналів при їхньому зберіганні в електронних архівах. Дані для цього було отримано з Реєстру The Keepers Registry, http://thekeepers.org. У результаті, дані про вибірку біля 100 електронних журналів були ранжировані за кількістю електронних архівів, які зберігають ці журнали в Інтернеті з апроксимацією логарифмічною функцією.

Доступність. Оцінка живучості ІО при довготерміновому зберіганні в інтернет-середовищі також залежить від доступності веб-серверів, на яких розміщені НП. Для визначення характеру цієї залежності було налагоджено моніторинг стану сайтів, на яких може бути розміщено контент довготермінового зберігання.

Для збору даних про стан сайтів було використано можливості ресурсу http://uptimerobot.com. На основі накопиченого матеріалу було розраховано дані й побудовано розподіли випадкових значень показника доступності (рис. 3) й недоступності сайтів, а також періодів їхньої доступності та недоступності.

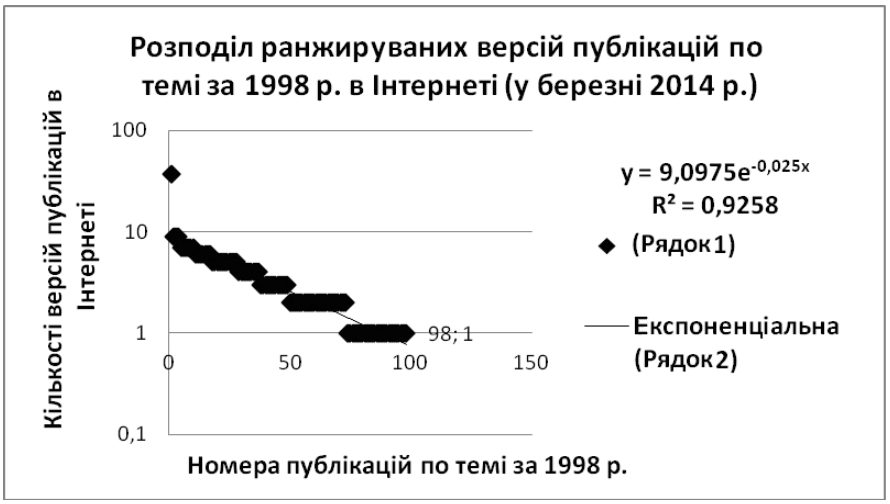

Рис. 2. Дані про вибірку біля 100 НП за 1998 р. з тематики зберігання даних, ранжирувані за кількістю версій публікацій в Інтернеті з апроксимацією експоненціальною функцією

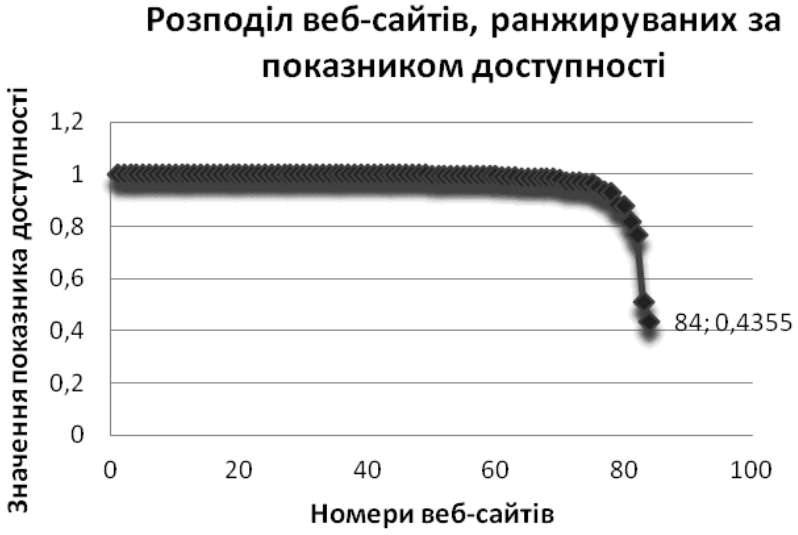

Рис. 3. Дані про вибірку біля 100 веб-сайтів, ранжируваних за показником доступності

При цьому період доступності сайту розглядається як наробіток між відмовами (наробіток об'єкта від завершення відмовлення його працездатного стану після відмови до виникнення наступної відмови), а період недоступності — це час, що витрачається на відновлення працездатного стану сайту. Показник недоступності сайту розраховується як протилежний показнику доступності сайту (відношенню

загального часу, коли сайт знаходився в робочому стані до загального часу існування сайту), який доповнює показник доступності до одиниці.

Індексованість. Оцінка живучості ІО при довготерміновому зберіганні в інтернет-середовищі також залежить від індексованості веб-серверів, на яких розміщені НП. Для визначення характеру цієї залежності було використано запити до пошукових систем Google та Scholar Google у формі <site: filetype:>. У запитах використовувались адреси сайтів, на яких звичайно зберігаються файли з НП (сайти відкритих репозиторієв, електронних архівів тощо), а також тип файлу pdf, найбільш характерний для зберігання НП. Результати запитів до сайтів електронних архівів університетів України для оцінки кількостей файлів формату pdf наведено на рис 4. Верхній графік відповідає кількостям файлів, проіндексованим у Google, а нижній — у Scholar Google.

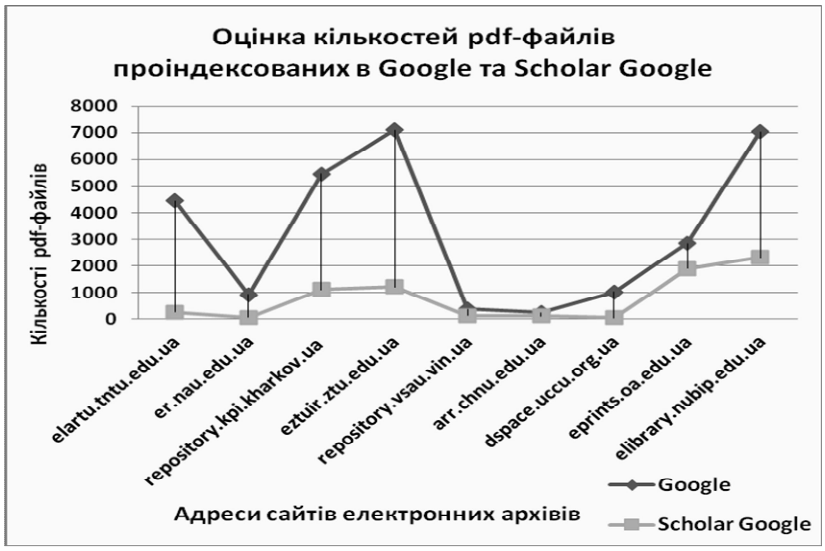

Рис. 4. Оцінка кількостей файлів формату pdf (в якому найчастіше зберігаються НП), проіндесованих у Google та Scholar Google на сайтах електронних архівів університетів України (за даними запитів до Google та Scholar Google)

Рис. 4 показує, що доля файлів, проіндексованх у Scholar Google відносно до файлів, проіндексованх у Google, може змінюватися від одиниць до десятків відсотків залежно від конкретного електронного архіву. Відповідно, може змінюватися властивість НП виконувати свої функції, тобто живучість.

Поширеність форматів. Оцінка живучості НП при довготерміновому зберіганні в інтернет-середовищі також може залежати від поширеності форматів, використаних для представлення НП та її посилань. Тобто, крім загроз, що пов'язані з відмовами сайтів (відсутність доступу до файлів), або неможливістю знайти НП через пошукові системи (із-за великого списку результатів пошуку або відсутності індексації) на їхню живучість при зберіганні в Інтернеті може впливати старіння форматів, у яких представлені ці НП (поява нових форматів і ПЗ, несумісного зі старими форматами представлення даних). Для оцінки поширеності основних форматів в інтернет-середовищі у світі та на сайтах домену України UA, в 2013 та

2014 рр. за допомогою запитів до Google були отримані відповідні дані. Результати, що стосуються світової мережі Інтернет, наведені в таблиці. Аналіз даних показує, що серед основних форматів даних в Інтернеті до числа найбільш поширених (тобто тих, що складають найбільшу частку) належать html (більш ніж 90 %); pdf, в якому представлена більшість НП в Інтернеті (2–3 %); doc (1–2 %); txt (1–2 %). За короткий проміжок часу спостереження 2013–2014 рр. частка формату pdf у мережі Інтернет зменшилась у світі майже на 2 %, а в домені України більш ніж на 3 %.

Оцінка поширеності основних форматів у світовому інтернет-середовищі в 2013 та 2014 рр. у світі (за даними запитів до Google).

Формат	Оцінка кількості файлів у 2013 р.	Оцінка кількості файлів у 2014 р.	Частка формату у 2013 р.	Частка формату в 2014 р.	Зміна частки формату у %
html	4,58E+09	2,53E+10	0,93	0,95	0,8
pdf	2,64E+08	9,72E+08	0,054	0,036	–1,8
doc	1,15E+07	2,86E+08	0,002	0,01	0,8
txt	6,28E+06	7,76E+07	0,001	0,003	0,2
rtf	8,79E+05	3,44E+07	0,0002	0,001	0,1
docx	2,77E+06	1.58E+07	0,0006	0,00059	0,002
xls	3,04E+06	1.81E+07	0,0006	0,0007	0,005
xlsx	311000	464000	6,4E–05	1,7E–05	–0,005
ppt	3,97E+06	2,94E+07	0,0008	0,001	0,03
odt	3,14E+06	1,95E+06	0,0006	7,3E–05	–0,06
pptx	6,09E+06	3,58E+06	0,001	0,0001	–0,1

Для оцінки зміни версій формату pdf було проаналізовано близько 230 тис. файлів міжнародних ресурсів з кешу пошукової системи. Було отримано розподіл контенту інтернет-середовища за різними версіями формату pdf, який представлено на рис. 5.

Рис. 5. Дані про розподіл версій формату pdf у світовому інтернет-середовищі (біля 230 тис. файлів проаналізовано в грудні 2013 р.)

Моделі живучості наукових публікацій при довготерміновому зберіганні в Інтернеті

У загальному випадку інформаційний об'єкт «наукова публікація» має мережевий характер. У ньому використовують інтернет-посилання на ресурси різних видів: документи різних форматів, презентації, дані, програми розроблені під різними ОС тощо, які складно зібрати на одному комп'ютері. Крім того, як показано вище, наукова публікація, а також усі ресурси, на які вона посилається, звичайно представлені в інтернет-середовищі кількома версіями (рис. 6).

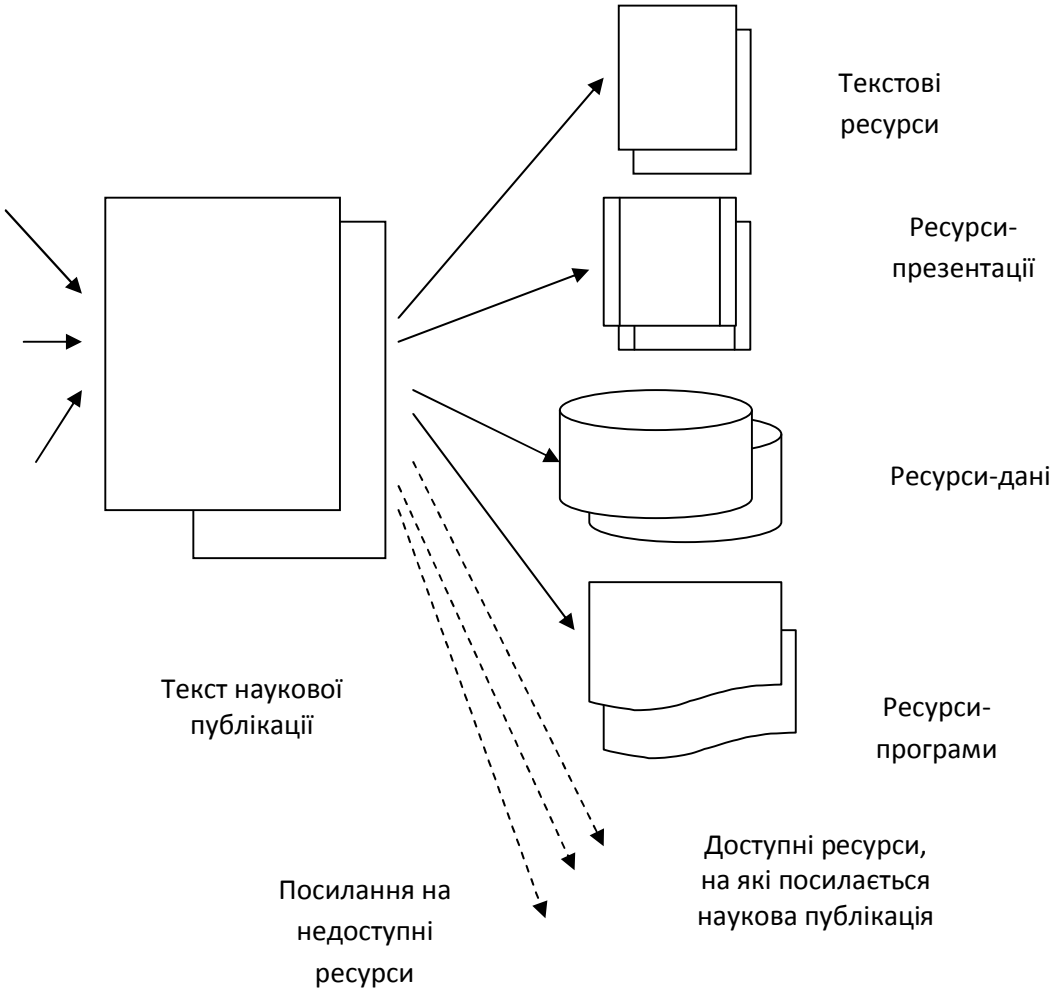

Рис. 6. Структура ІО «наукова публікація» при довготерміновому зберіганні в інтернет-середовищі

На основі наведених вище особливостей представлення наукових публікацій у мережі Інтернет, будемо вважати, що живучість НП залежить від наступних основних факторів:

— живучості тексту публікації, на яку впливають кількість копій публікації в інтернет-середовищі (в загальному випадку будемо розглядати кількість версій публікацій, тобто не лише повнотекстові копії, а й анотації статей); доступності

серверів, на яких зберігаються копії та версії НП; індексованості тексту публікації в універсальних та наукових пошукових системах; поширеності формату, в якому представлено текст тощо;

— частки доступних інтернет-посилань, що використовуються в науковій публікації;

— живучості ІО, на які є інтернет-посилання в науковій публікації (теж залежить від кількості копій, версій, доступності серверів, індексованості, поширеності форматів тощо).

З урахуванням вищенаведеного, живучість ІО «наукова публікація» при довготерміновому зберіганні в інтернет-середовищі можна оцінювати на основі кількості версій НП та її посилань в Інтернеті. Живучість НП будемо представляти двома значеннями (наприклад, координатами точки на площині): живучістю тексту НП та живучістю її посилань.

Живучість тексту НП будемо оцінювати кількістю версій НП з урахуванням доступності, індексації та поширеності формату

$$ST = \sum_{i=1}^{VT} AT_i * IT_i * PFT_i, \qquad (1)$$

де ST — живучість тексту НП; AT_i — доступність i-ої версії тексту НП; IT_i — індексованість у пошуковій системі i-ої версії тексту НП; PFT_i — поширеність формату i-ої версії тексту НП в Інтернеті.

У даних формулах, при оцінці живучості, під доступністю тексту НП будемо розуміти частку, яку складає час, коли до тексту НП можна звернутися через інтернет-середовище, від загального часу існування НП в інтернет-середовищі. Під індексованістю текстів НП у пошуковій системі розуміємо частку, яку складають адреси, що представляються пошуковою системою на відповідний запит, до загальної кількості текстів НП у даному масиві. Під поширеністю формату тексту будемо розуміти частку, яку дана версія формату складає на даний час в Інтернеті. Тобто, доступність, індексованість текстів НП і поширеність їхніх форматів при оцінці живучості входять у формули (1) та (2) як частки одиниці.

Живучість посилань НП будемо оцінювати усередненою кількістю версій посилань з урахуванням доступності та індексації, а також з урахуванням загальної частки доступних посилань:

$$SR = \frac{RRL}{RRC} * \frac{\sum_{j=1}^{RRL} \sum_{i=1}^{VR_j} AR_{ij} * IR_{ij} * PFR_{ij}}{RRL}, \qquad (2)$$

де SR — живучість ресурсів, на які посилається НП; AR_{ij} — доступність i-ої версії j-го ресурсу, на який посилається НП; IR_{ij} — індексованість у пошуковій системі i-ої версії j-го ресурсу, на який посилається НП; PFR_{ij} — поширеність формату

i-ої версії *j*-го ресурсу НП в Інтернеті; VR_{ij} — кількість версій *j*-го ресурсу, на який посилається НП; *RRL* — кількість «живих» посилань, на який посилається НП; *RRC* — загальна кількість інтернет-посилань, які є в НП.

Особливості оцінки живучості НП з використанням запропонованої моделі може бути показана на прикладі наукової періодики України, що представлена в Електронному архіві наукових періодичних видань на сайті Національної бібліотеки України імені В.І. Вернадського (НБУВ). Для вибірки НП окремого наукового видання з цього електронного архіву середнє значення кількості версій НП в інтернет-середовищі складає приблизно 2 (на основі даних пошукових систем Google Scholar та Google, аналізувалися 2008–2010 рр.). Зазвичай це версії НП на сервері НБУВ та на сервері наукової установи — видавця. Доступність серверів, на яких зберігаються версії НП видання наукової періодики України може оцінюватися за допомогою сервісу http://uptimerobot.com. Індексованість НП в електронному архіві наукових публікацій є незначною. Для порівняння оцінювалася також середня кількість версій НП в інтернет-середовищі для одного з найбільших електронних архівів наукових публікацій (http://arxiv.org). Для вибірки за 2008–2010 рр. вона складає біля 7 (на основі даних пошукових систем Google Scholar та Google). Враховуючи те, що кількість версій НП є однією з основних складових живучості, це показує порівняно низький рівень живучості НП періодичних видань України при довготерміновому зберіганні в інтернет-середовищі. Виходячи із запропонованої структури живучості НП, серед основних шляхів її підвищення: збільшення кількості версій НП в інтернет-середовищі; підвищення доступності серверів, де розміщуються версії НП; підвищення рівня індексованості НП у пошукових системах; використання більш нових форматів тощо.

Доцільним напрямком збільшення кількості версій НП є використання архівних сервісів мережі Інтернет (webarchive.org, webcite.org та деяких інших) поряд з розміщенням додаткових версій НП у відкритих репозиторіях, що створюються в університетах та інших наукових установах. Але при використанні архівних сервісів для збільшення кількості версій НП важливо забезпечити їхню індексацію, тобто представлення їхніх адрес у результатах обробки запитів пошуковими системами. Наприклад, проведений аналіз результатів пошуку Google Scholar показує, що 10–20 % знайдених НП мають копії в Internet Archive, але у загальному списку версій публікацій Google Scholar ці копії не представляє. Рішенням цієї проблеми може бути розміщення адрес копій у полях метаданих НП (наприклад, у міжнародному стандарті метаданих для архівних матеріалів ISAD (G) передбачені дані про наявність і місцезнаходження копій [8]). Іншим рішенням може бути створення загального реєстру адрес зберігання версій НП, подібного до Реєстру проектів-зберігачів електронних видань). І, нарешті, розміщення копій НП в архівних сервісах, як правило, підвищує показник доступності сервера. Враховуючи актуальність довготермінового зберігання ІО, зокрема НП в інтернет-середовищі, представляється доцільним реалізація вищерозглянутих функцій у рамках інтернет-сервісу, який дозволяє оцінювати стан живучості НП українських періодичних видань, розміщувати додаткові версії НП у сховищах з високою доступністю, а також забезпечувати індексованість усіх версій НП.

Висновки та подальші роботи

У роботі визначено особливості представлення наукових публікацій в мережі Інтернет, які впливають на живучість НП при довготерміновому зберіганні: републікація; доступність серверів, на яких зберігаються НП; індексація НП у пошукових системах; поширеність форматів даних. На базі цих особливостей запропановано моделі для оцінки живучості НП при довготерміновому зберіганні в інтернет-середовищі, показано їхні можливості на прикладі електронного архіву української наукової переодіки.

Використання цих моделей до оцінки живучості НП дозволяє сформулювати рекомендації по підвищенню живучості НП на основі збільшення кількості версій НП в інтернет-середовищі, підвищення доступності та індексованості цих версій, впровадження сучасних форматів даних або конвертації форматів. Подальші роботи передбачають накопичення статистики за оцінками живучості різних видів ІО при довготерміновому зберіганні в інтернет-середовищі та реалізацію елементів відповідного інтернет-сервісу.

1. *Rhodes S.* Breaking down link rot: the chesapeake project legal snformation archive's examination of URL stability / S. Rhodes // Law Library Journal. — 2010. — Vol. 102, N 33. — P. 581–597.

2. *Sanderson R.* Analyzing the persistence of referenced web resources with Memento / R. Sanderson, M. Phillips, H. Van de Sompel //arXiv preprint arXiv:1105.3459. — 2011.

3. *Cornwall D.* Distributed globally, collected locally: LOCKSS for digital government information / D. Cornwall, J.R. Jacobs // Against the Grain. — 2013. — Vol. 21, N 1. — P. 42–45.

4. *Rosenthal D.S.H.* Distributed digital preservation in the cloud / D.S.H. Rosenthal, D.L. Vargas // International Journal of Digital Curation. — 2013. — Vol. 8, N 1. — P. 107–119.

5. *Kelly B.* Approaches to archiving professional blogs hosted in the cloud / B. Kelly, M. Guy // 7th Internation. Conf. on Preservation of Digital Objects (iPRES 2010). — University of Bath. — 2010.

6. *Додонов А.Г.* Живучесть информационных систем / А.Г. Додонов, Д.В. Ландэ. — К.: Наук. думка, 2011. — 256 с.

7. *Березін Б.О.* Живучість наукових публікацій при довготерміновому зберіганні в інтернет-середовищі / Б.О. Березін, Д.В. Ланде // Тези доповідей Міжнар. наук.-техн. конф. «Інтелектуальні технології лінгвістичного аналізу». — К.: НАУ, 2014. — С. 10.

8. *Селиванова Ю.* Международные стандарты метаданных для описания библиотечных, архивных материалов и музейных объектов / Ю. Селиванова, Т. Масхулия // Бібліотечний вісник. — 2012. — № 4. — С. 18–29.

Надійшла до редакції 18.11.2014