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 в мережі Інтернет, що впливають на їх живучість при довготерміновому зберіганні. Запропановано моделі для оцінки живучості наукових публікацій при зберіганні у мережі Інтернет та показано можливості їх 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 їх 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 в значній мірі створюються у цифровому вигляді, не мають паперової копії та потребують довготермінового зберігання є електронні журнали. На теперішній час в світі нараховується більше двадцяти тисяч найменувань електронних журналів.

В 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. В 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 аналізувалась доступність посилань в Інтернеті. В результаті виявилось, 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 доступінсть 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 та використовується їх читачами [3]. За допомогою засобів проекту LOCKSS [4], Інтернет-контент, до якого бібліотеки повинні надавати довготерміновий доступ своїм читачам, збирається з відповідних веб-сайтів та 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 р.р. Доступність посилань аналізувалась у вересні 2014 р.

Довготермінове зберігання е-журналів в Інтернет забезпечується електронними архівами. Реєстр зберігачів (The Keepers Registry, http://thekeepers.org), відображає стан із зберіганням та архівуванням е-журналів. У переліку зберігачів в 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 і т.і. Тобто, для довготермінового зберігання та доступу до ІО необхідно забезпечення живучості ІО. Живучість об’єкту - це властивість виконувати

основні функції в умовах негативних впливів (НВ), при необхідності тимчасово відмовляючись від виконання деяких другорядних функцій [6].

В 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 в мережі Інтернет, які впливають на їх живучість при довготерміновому зберіганні, а також відповідної моделі для оцінки живучості НП при довготерміновому зберіганні в Інтернет-середовищі.

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

Републікація. При користуванні інформаційними ресурсами в мережі Інтернет може відбуватися копіювання, розмноження їх версій, тобто републікація. Оцінка живучості ІО при довготерміновому зберіганні в Інтернет-середовищі може в значній мірі залежати від кількості версій НП. При визначенні характеристик такого процесу републікації контенту в Інтернет використовувалися запити до пошукової системи Google Scholar для отримання списку публікацій з заданої тематики (наприклад, тематики "довготермінове зберігання") на протязі періоду в один рік. В отриманому списку публікацій за допомогою пакету Excel аналізувалось поле "всі версії статті". Проранжировані кількості версій НП по вибраній тематиці за 1998 р. в Інтернет 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 р.р. також були апроксимовані за допомогою експоненціальної функції. Додатково до аналізу републікації контенту НП було проаналізовано републікацію контенту електронних журналів при їх зберіганні в електронних архівах. Дані для цього були отримані з Реєстру The Keepers Registry, *http:// thekeepers.org*. В результаті, дані про вибірку біля 100 електронних журналів, були ранжировані за кількістю електронних архівів, які зберігають ці журнали в Інтернет з апроксимацією логарифмічниою 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 їх доступності та недоступності.

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

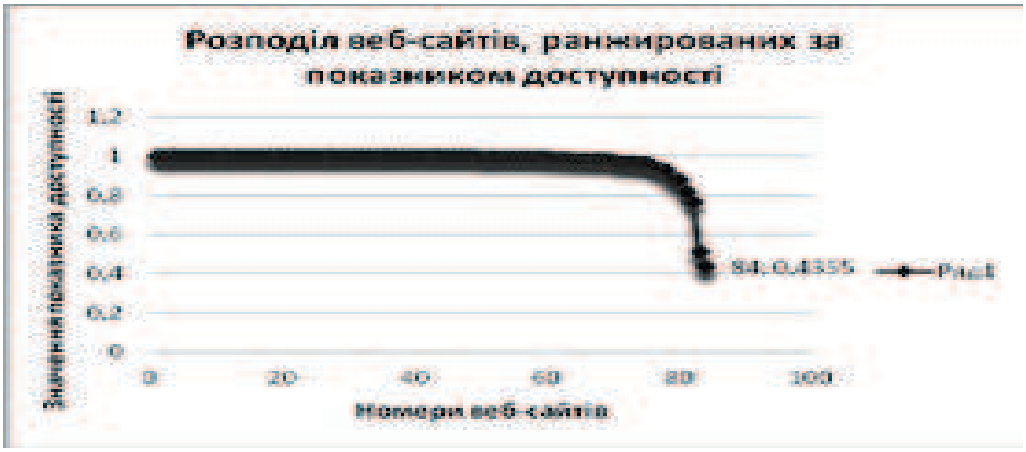

Рис. 3. Дані про вибірку біля 100 веб-сайтів, ранжированих за показником доступності.

При цьому період доступності сайту розглядається як наробіток між відмовами (наробіток об'єкта від завершення відмовлення його працездатного стану після відмови до виникнен-ня наступної відмови), а період недоступності – це час, що витрачається на відновлення працездатного стану сайта. 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:> . В 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 в Google, а нижній - в Scholar Google.

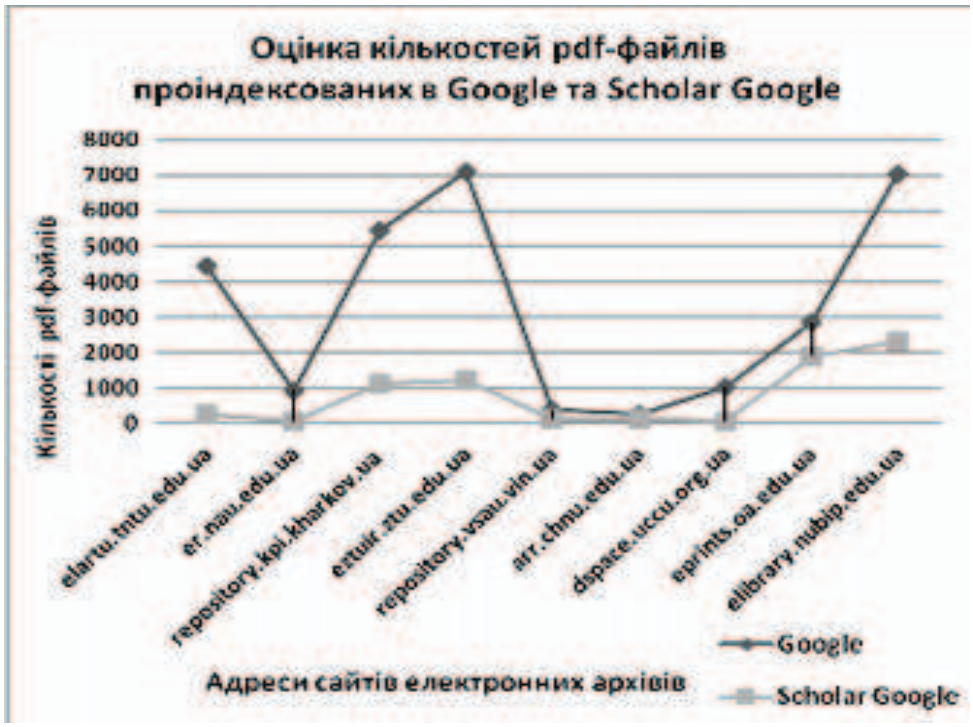

Рис. 4. Оцінка кількостей файлів формату pdf (в якому найчастіше зберігаються НП), проіндесованих в Google та Scholar Google на сайтах електронних архівів університетів України. (За даними запитів до Google та Scholar Google).

Рис. 4 показує, що доля файлів, проіндексованх в Scholar Google по відношенню до файлів, проіндексованх Google, може змінюватися від одиниц до десятків відсотків, в залежності 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, пов'язаних з відмовами сайтів (відсутність доступу до файлів), або неможливістю знайти НП через пошукові системи (із-за великого списку результатів пошуку або відсутності індексації) на їх живучість при

зберіганні в Інтернет може впливати старіння форматів, в 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 р.р. за допомогою запитів до Google були отримані відповідні дані. Результати, що стосуються світової мережі Інтернет, наведені в Таблиці 1. Аналіз даних показує, що серед основних форматів даних в Інтернет до числа найбільш поширених (тобто складаючих найбільшу частку) належать html (більш ніж 90%), pdf , в якому представлена більшість НП в Інтернет (2%-3%), doc (1%-2%), txt (1%-2%). За короткий проміжок часу спостереження 2013-2014 р.р. частка формату pdf в мережі Інтернет зменшилася в світі майже на 2%, а в домені України більш ніж на 3%.

Таблиця 1. Оцінка поширеності основних форматів в світовому Інтернет-середовищі у 2013 та 2014 р.р. в світі (за даними запитів до Google).

Фор-мат	Оцінка кількості файлів у 2013 р.	Оцінка кількості файлів у 2014 р.	Частка формату у 2013 р.	Частка формату у 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 2013р.).

Моделі живучості наукових публікацій при довготерміновому зберіганні в Інтернет. В загальному випадку інформаційний об'єкт "наукова публікація" має мережевий характер. В 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).

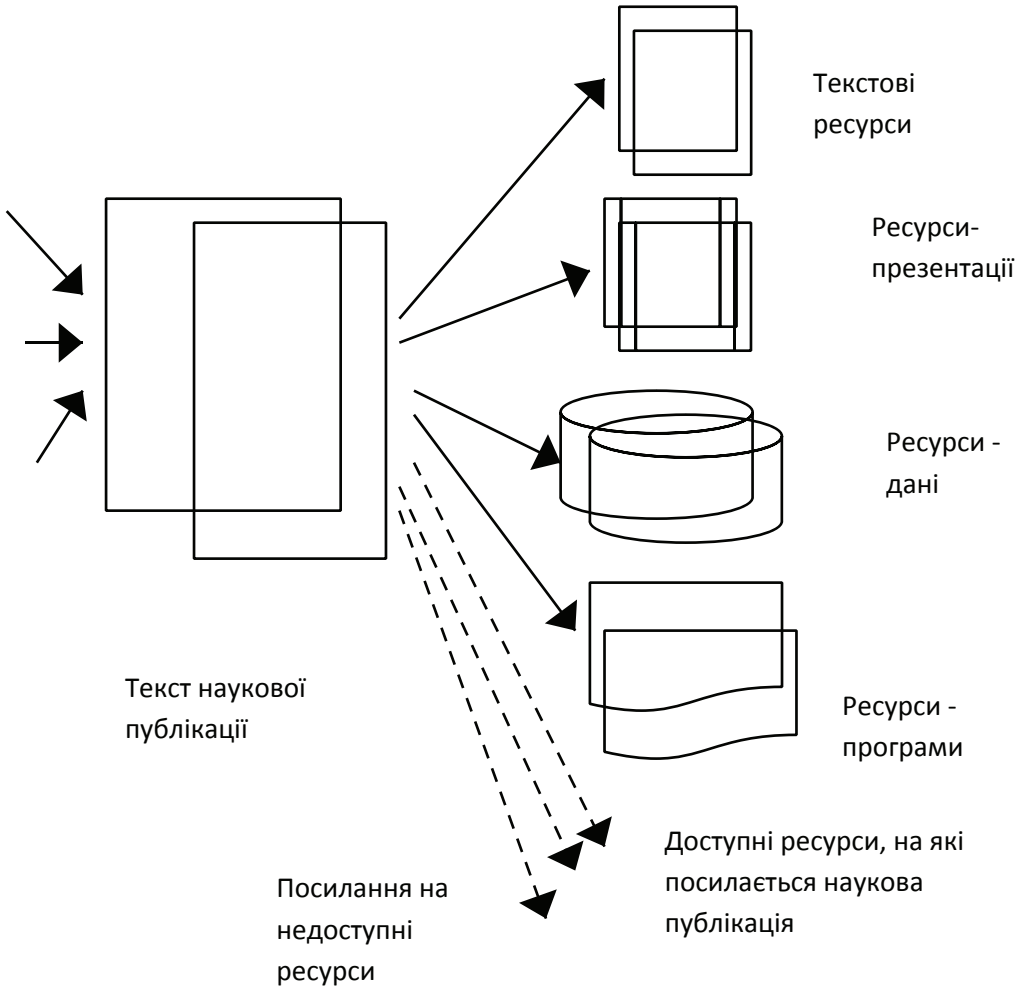

Рис. 6. Структура ІО "наукова публікація" при довготерміновому зберіганні в Інтернет-середовищі.

На основі наведених вище особливостей представлення наукових публікацій в мережі Інтернет, будемо вважати, що живучість НП залежить від слідуючих 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 вище наведеного, живучість ІО "наукова публікація" при довготерміновому зберіганні в Інтернет-середовищі можна оцінювати на основі кількості версій НП та її посилань в Інтернет. 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:

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

$ST -$ живучість тексту НП;

$AT_i -$ доступність *i*-ї версії тексту НП;

$IT_i -$ індексованість в пошуковій системі *i*- ї версії тексту НП;

$PFT_i -$ поширеність формату *i*-ї версії тексту НП в Інтернет.

В 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 в пошуковій системі будемо розуміти частку, яку складають адреси, що представляються пошуковою системою на відповідний запит, до загальної кількості текстів НП в даному масиві. Під поширеністю формату тексту будемо розуміти частку, яку дана версія формату складає на даний час в Інтернет. Тобто, доступність, індексованість текстів НП та поширеність їх 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.

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

SR – живучість ресурсів, на які посилається НП;

AR_{ij} – доступність *i*–ї версії *j*–го ресурсу, на який посилається НП;

IR_{ij} – індексованість в пошуковій системі *i* -ї-версії *j* – го ресурсу, на який посилається НП;

PFR_{ij} – поширеність формату *i* - ї версії *j* – го ресурсу НП в Інтернет.

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 запропанованої моделі може бути показана на прикладі наукової періодики України, представленої в Електронному архіві наукових періодичних видань на сайті Національної бібліотеки України імені В.І. Вернадського. 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 р.р.). 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 оцінювалась також середня кількість версій НП в Інтернет-середовищі для одного з найбільших електронних архівів наукових публікацій (http://arxiv.org). Для вибірки за 2008-2010 р.р. 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 запропанованої 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 в 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 їх індексацію, тобто представлення їх 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 в полях метаданих НП (наприклад, в міжнародному стандарті метаданих для архівних матеріалів ISAD (G) передбачені дані про наявність та 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 серверу. Враховуючи актуальність довготермінового зберігання ІО, зокрема НП в Інтернет-середовищі, представляється доцільним реалізація вище розглянутих функцій в 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. Таким чином, в 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 в пошукових системах; поширеність форматів даних. На базі цих особливостей запропановано моделі для оцінки живучості НП при довготерміновому зберіганні в Інтернет-середовищі, показано їх 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 по оцінкам 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 Information Archive's Examination of URL Stability // Law Library Journal. – 2010. – Vol. 102 – No. 33. – P. 581-597.
2.Sanderson R., Phillips M., Van de Sompel H. Analyzing the persistence of referenced web resources with Memento //arXiv preprint arXiv:1105.3459. – 2011.
3. Cornwall D., Jacobs J. R. Distributed Globally, Collected Locally: LOCKSS for Digital

Government Information //Against the Grain. – 2013. – Т. 21. – №. 1. – P. 42- 45.
4. Rosenthal D. S. H., Vargas D. L. Distributed Digital Preservation in the Cloud //International Journal of Digital Curation. – 2013. – Т. 8. – №. 1. – P. 107-119.
5. Kelly B., Guy M. Approaches to archiving professional blogs hosted in the cloud //7th International Conference on Preservation of Digital Objects (iPRES 2010). – University of Bath, 2010.
6. Додонов А.Г., Ландэ Д.В. Живучесть информационных систем. — К.: Наук. думка, 2011. — 256 с.
7. Березін Б.О., Ланде Д.В. Живучість наукових публікацій при довготерміновому зберіганні в інтернет-середовищі//Тези доповідей. Міжнародної науково-технічної конференції "Інтелектуальні технології лінгвістичного аналізу". - Київ: НАУ, 2014. - С. 10.

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