Живучесть научных публикаций при долговременном хранении в Интернет-среде

Березин Б.А., Ландэ Д.В.

Институт проблем регистрации информации НАН Украины

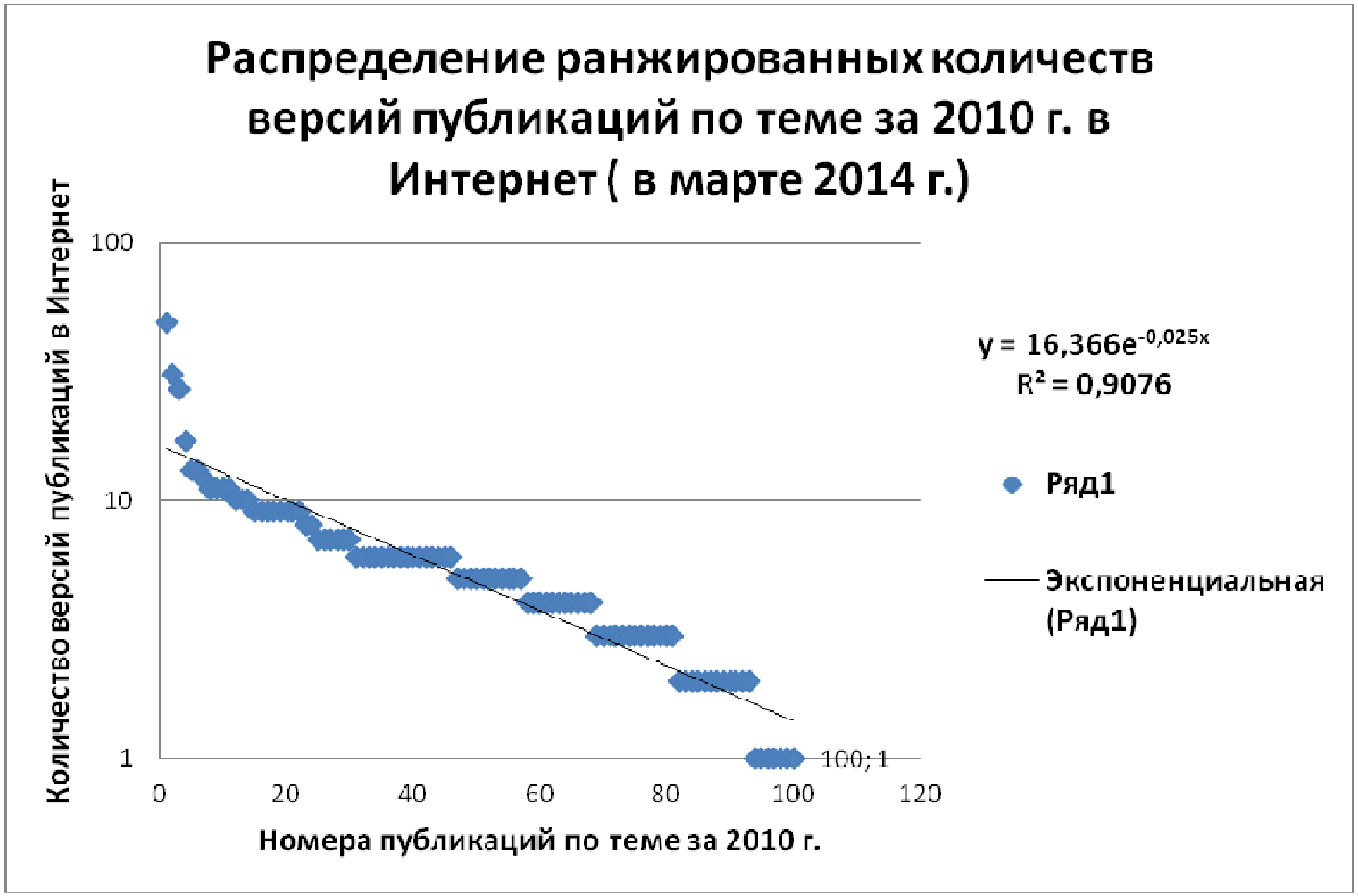
- **Развитие контента Интернет ведет к появлению информационных ресурсов, требующих долговременного хранения. Среди них правовая информация (например, в США принят "Типовой закон о правовых актах, публикуемых в электронном виде "), электронные журналы и научные публикации (Реестр хранителей е-журналов), профессиональные блоги и т.д.**

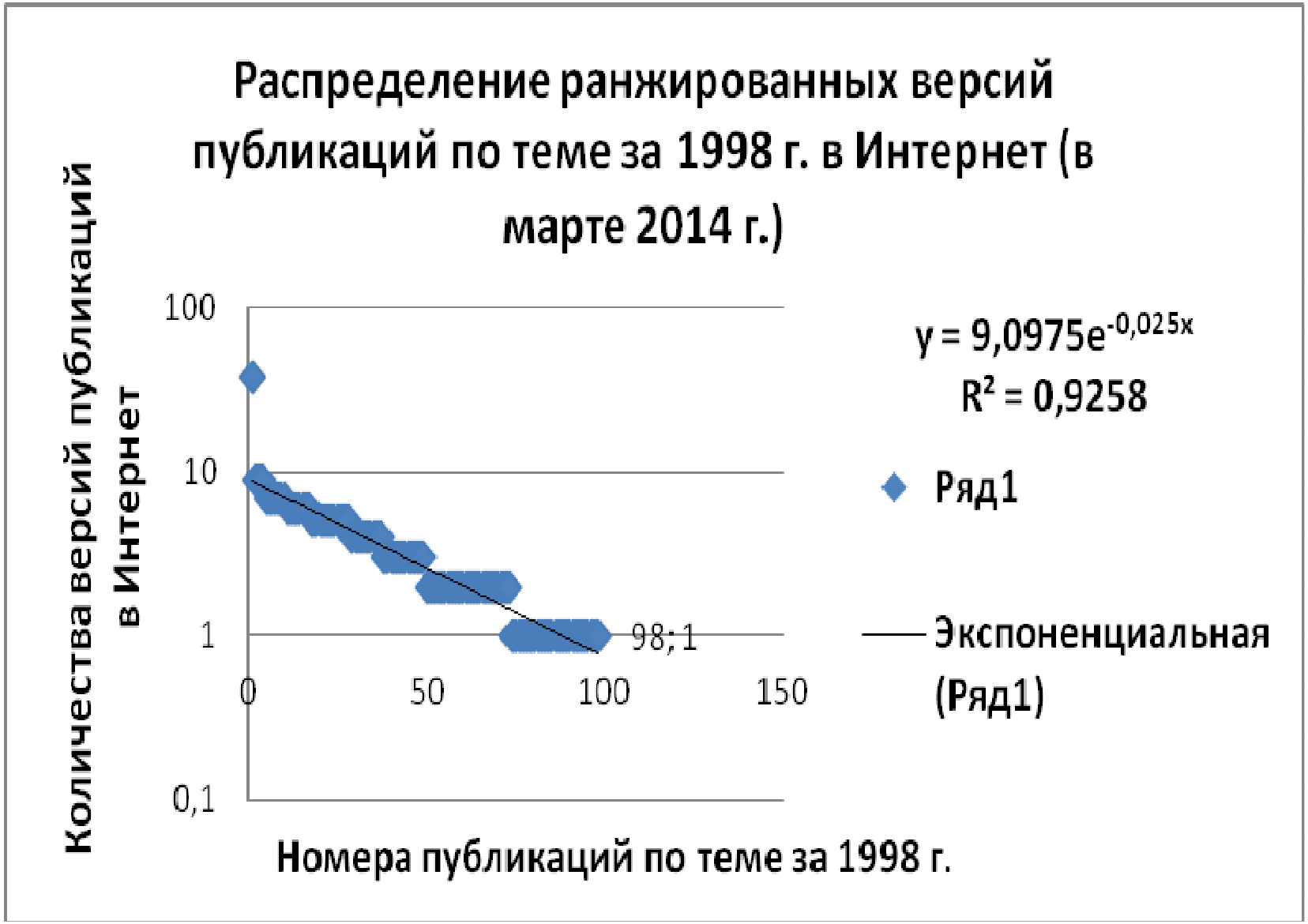
Учитывая актуальность договременного хранения ИО в Интернет, в работе исследовались характеристики Сети, влияющие на хранение научных публикаций (НП).

- **Время жизни ссылок на НП анализировалось с помощью запросов к поисковой системе Google Scholar. Результат на основе ограниченных выборок показан на графике. Для более точных результатов разрабатывается программа анализа профилей цитирования в Google Scholar.**

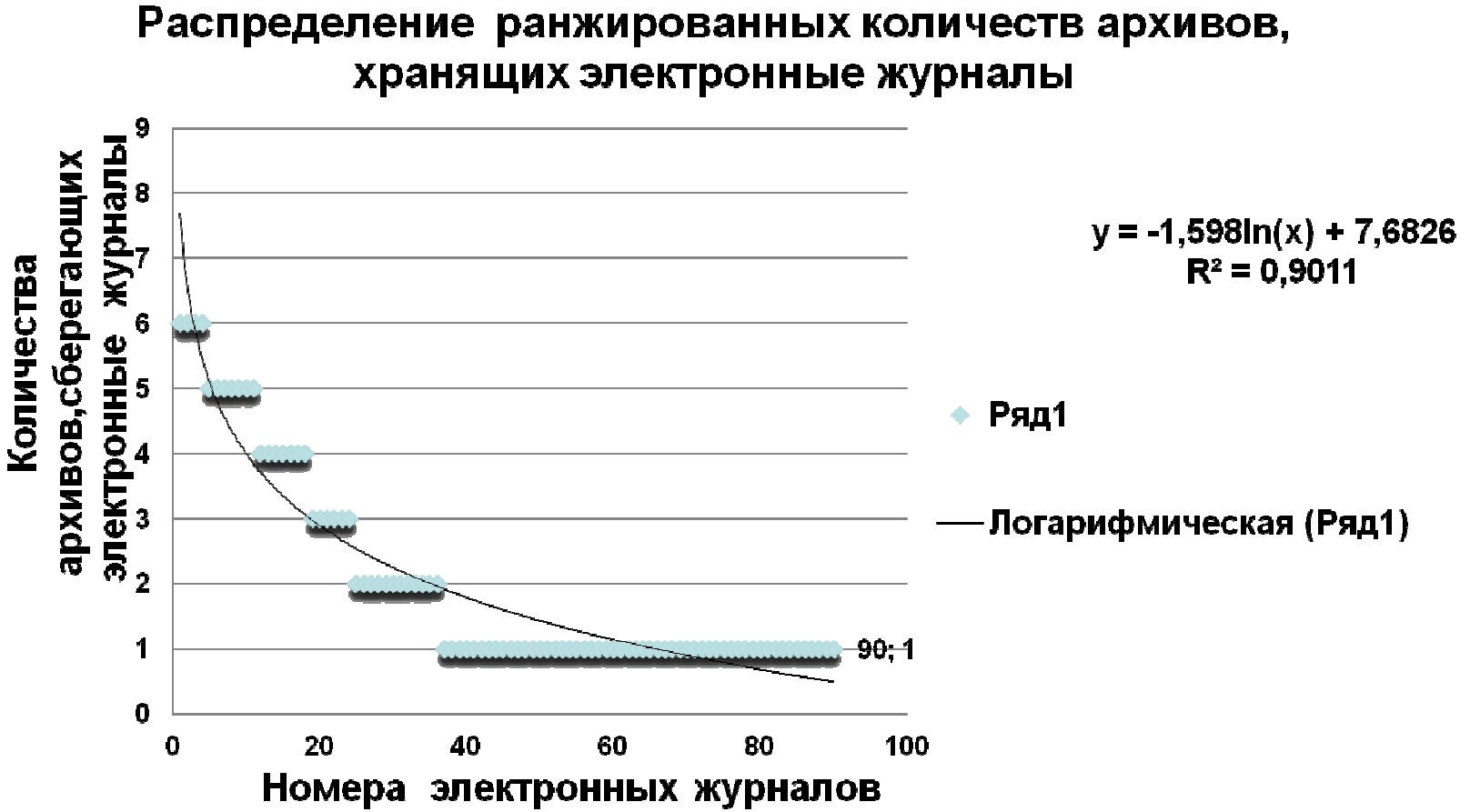

Доля доступных ссылок на Интернет-ресурсы в зависимости от года издания научной публикации
1,2
1
0,8
0,6
0,4
0,2
0
Доля доступных ссылок
2013
2010
2005
2000
Год публикации
Ряд1

- **Как и следовало ожидать, из-за влияния неблагоприятных воздействий (отказы и старение аппаратуры и ПО, ошибки операторов, экономические ошибки, компьютерные атаки и т.п.) с течением времени доля доступных ссылок в Интернет уменьшается. Но с другой стороны, из-за использования НП происходит увеличение числа их копий, версий, т.е. стихийное распространение в Интернет.**
- **С помощью запросов к Google Scholar было проанализировано изменение числа версий НП с течением времени.**

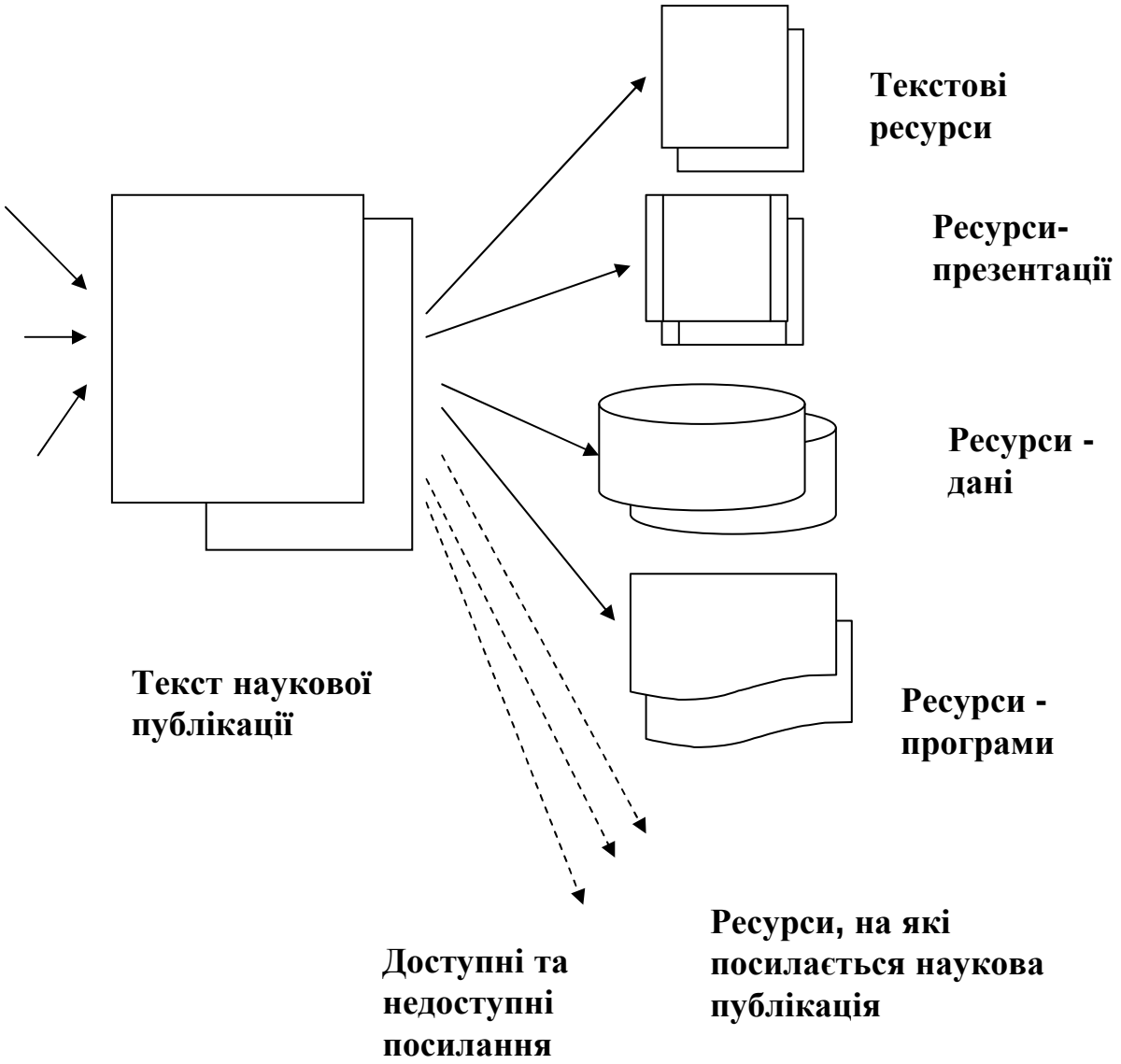

Распределение ранжированных количеств версий публикаций по теме за 2010 г. в Интернет (в марте 2014 г.)
Количество версий публикаций в Интернет
100
10
1
0
20
40
60
80
100
120
100; 1
Номера публикаций по теме за 2010 г.
y = 16,366e-0,025x
R² = 0,9076
Ряд1
Экспоненциальная (Ряд1)


Распределение ранжированных версий публикаций по теме за 1998 г. в Интернет (в марте 2014 г.)
Количества версий публикаций в Интернет
100
10
1
0,1
0
50
100
150
98;1
y = 9,0975e-0,025x
R² = 0,9258
Ряд1
Экспоненциальная (Ряд1)
Номера публикаций по теме за 1998 г.

Для сравнения, результат анализа хранения е-журналов в сетевых архивах



Структура ИО “научная публикация” при довговременном хранении в Интернет-среде



Будем считать, что живучесть научной публикации (НП) зависит от следующих основных факторов:

1) Живучести текста публикации, на которую влияют количество копий (версий) публикации в Интернет-среде; доступность серверов, на которых хранятся версии НП; индексированность текста публикации в универсальных и научных поисковых системах и т.д.;

2) Доли доступных Интернет-ссылок, используемых в научной публикации;

3) Живучести ИО, на которые есть Интернет-ссылки в научной публикации (тоже зависит от количества копий, версий, доступности серверов, индексированности и т.п.).

Тогда живучесть НП будем представлять двумя значениями: живучестью текста НП и живучестью ее ссылок.

Живучесть текста НП будем оценивать количеством версий НП с учетом доступности и индексированности:

$$ST = \sum_{i=1}^{VT} AT_i * IT_i$$

ST – живучесть текста НП;

AT_i – доступность *i*-й версии текста НП;

IT_i – индексированность в поисковых системах *i*- й версии текста НП;

Живучесть ссылок НП будем оценивать усредненным количеством версий ссылок с учетом доступности и индексации, а также с учетом общей доли доступных ссылок.

$$SR = \frac{RRL}{RRC} * \frac{\sum_{j=1}^{RRL} \sum_{i=1}^{VR_j} AR_{ij} * IR_{ij}}{RRL}$$

SR – **живучесть ресурсов, на которые ссылается НП;**

AR_{ij} – **доступность i-й версии j-го ресурса, на который ссылается НП;**

IR_{ij} – **индексированность в поисковой системе i - й-версии j - го ресурса, на который ссылается ЧП;**

VR_j – количество версий j-го ресурса, на который ссылается НП;

RRL – количество “живых” ссылок, на которые ссылается НП;

RRC – общее количество Интернет-ссылок, которые есть в НП.

- **В данных формулах, при оценке живучести, под доступностью текста НП (AT_i) понимаем долю, которую составляет время, когда к тексту НП можно обратиться через Интернет-среду, от общего времени существования НП в Интернет-среде.**
- **Под индексированностью текстов НП (IT_i) в поисковой системе понимаем долю, которую составляют адреса, представляемые поисковой системой на соответствующий запрос, к общему количеству текстов НП, отвечающих этому запросу в Интернет.**

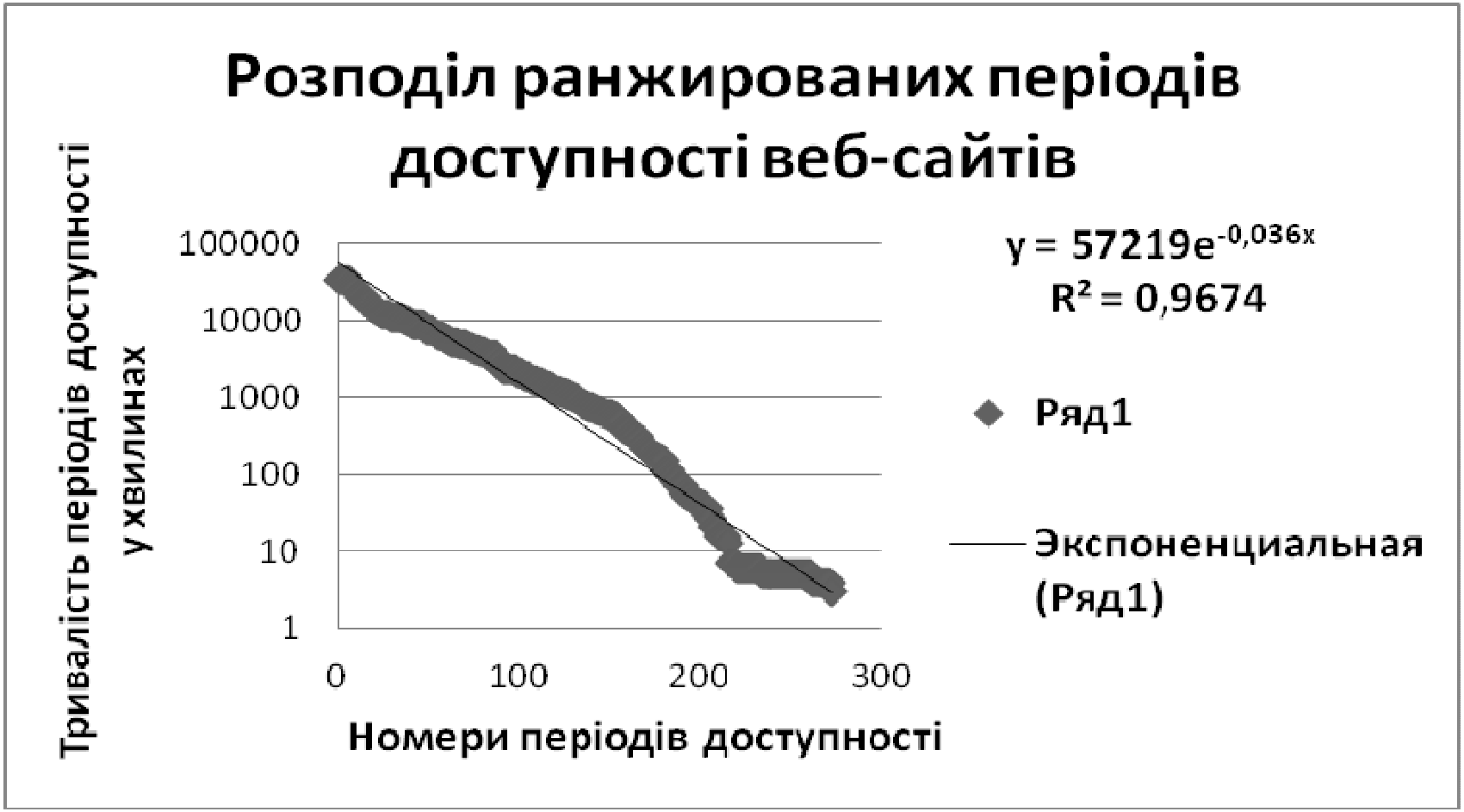
При оценке живучести НП из научной периодики Украины (на сайте Национальной б-ки им. В.Вернадского) необходимо учесть следующее.

- **Для отдельного научного издания среднее значение кол-ва версий НП в Интернет составляет около 2. Обычно это версии НП на сервере НБУВ и сервере института-издателя.**

(Для сравнения, среднее кол-во версий НП в Интернет для электронного архива arxiv.org составляет около 7).

- **Доступность серверов, на которых хранятся версии НП может оцениваться с помощью сервиса uptimerobot (для сервера НБУВ составляет около 0,99). Индексированность НП на сервере НБУВ не высокая.**

Результат анализа доступности веб-сайтов



- **Исходя из предложенной модели живучести НП, среди основных путей ее повышения:**
- **- увеличение количества версий НП в Интернет-среде;**
- **- повышение доступности серверов, где размещаются версии НП;**
- **- повышение уровня индексированности НП в поисковых системах и т.д.**

- Целесообразным направлением увеличения количества версий НП является использование архивных сервисов сети Интернет (webarchive.org, webcite.org и других) наряду с размещением дополнительных версий НП в открытых репозиториях, создаваемых в вузах и других научных учреждениях. Но, при использовании архивных сервисов для увеличения количества версий НП важно обеспечить их индексацию, т.е. представление их адресов в результатах обработки запросов поисковыми системами.

- Например, проведенный анализ результатов поиска Google Scholar показывает, что 10% - 20% найденных НП имеют копии в Internet Archive, но в общем списке версий публикаций Google Scholar эти копии не представляет. Решением этой проблемы может быть размещение адресов копий в полях метаданных НП (например, в международном стандарте метаданных для архивных материалов ISAD (G) предусмотрены данные про наличие и местонахождение копий.

Выводы

- **- Исследованы характеристики Интернет, влияющие на долговременное хранение информационных объектов (ИО), в частности научных публикаций (НП).**
- **- Предложена модель оценки живучести НП и пути ее повышения при долговременном хранении НП в Интернет.**

Спасибо за внимание! Пожалуйста, задавайте вопросы

boberezin@gmail.com

– Березин Борис Александрович