

## Осторожно, DOI!

### “Библиографический детектив” эпохи цифровизации

В. С. Козякин

*Институт проблем передачи информации РАН, Москва, Россия*

Поступила в редколлегию 27.02.2019

**Аннотация**—Описывается пример несоответствия информации, предоставляемой популярными библиографическими сервисами, и обсуждаются причины такого несоответствия.

**КЛЮЧЕВЫЕ СЛОВА:** библиография, цифровой идентификатор объекта, DOI, BibTeX, RIS

*Гладко было на бумаге,  
да забыли про овраги,  
а по ним — ходить!*

Составление точных и полных библиографических ссылок является неотъемлемой частью любой научной работы. Ученые, заставшие “докомпьютерную” эпоху, помнят, каких трудов стоило правильно составить библиографические ссылки в статье в соответствии с требованиями оформления библиографии, принятыми в том или ином журнале. Ситуация кардинально изменилась с внедрением компьютерных методов подготовки публикаций — я буду говорить в основном о системе L<sup>A</sup>T<sub>E</sub>X и различных ее надстройках, хотя сказанное в большей или меньшей степени относится и к другим издательским системам, как проприетарным так и свободным.

Распространенным и, главное, удобным способом создания библиографических ссылок в L<sup>A</sup>T<sub>E</sub>X явилась программа BibTeX, созданная Ореном Паташником. Процесс создания библиографических ссылок с помощью BibTeX разбивается на два этапа — “ручное” создание базы данных необходимых публикаций, в которую для каждой публикации в определенном формате заносятся требуемые библиографические данные, а затем “автоматическое” (с помощью BibTeX) оформление списка публикаций в соответствии с теми или иными стилевыми файлами .bst, которые многие издательства создают в соответствии с их предпочтениями оформления библиографии и стилей цитирования.

Стоит отметить, что создание базы данных публикаций для последующей обработки с помощью BibTeX отнимает, конечно, некоторое время и требует внимания и определенных знаний о правилах оформления ее структурных элементов. Однако время и усилия, затраченные на составление баз данных публикаций, с лихвой окупаются простотой дальнейшего применения этих баз данных при оформлении библиографии в различных статьях и, главное, кардинальным сокращением ошибок в оформлении публикаций. Стоит отметить также, что на самом деле “ручное” составление базы данных BibTeX в большинстве случаев, как правило, даже не требуется — в нужном формате необходимые записи обычно формируются издательствами, а также многими библиографическими онлайн сервисами.

На рис. 1 приводится фрагмент страницы с выходными данными статьи [1]:



Рис. 1. Фрагмент страницы с выходными данными статьи [1]

Соответствующая статье [1] запись в формате BibTeX, полученная в библиографической системе MR LOOKUP Американского математического общества (American Mathematical Society, <https://mathscinet.ams.org/mrlookup>), имеет следующий вид:

```
@article {MR2999086,
  AUTHOR = {Kloeden, Peter E. and Kozyakin, Victor S.},
  TITLE = {Asymptotic behaviour of random tridiagonal {M}arkov chains in
    biological applications},
  JOURNAL = {Discrete Contin. Dyn. Syst. Ser. B},
  FJOURNAL = {Discrete and Continuous Dynamical Systems. Series B. A Journal
    Bridging Mathematics and Sciences},
  VOLUME = {18},
  YEAR = {2013},
  NUMBER = {2},
  PAGES = {453--465},
  ISSN = {1531-3492},
  MRCLASS = {60J10 (15B48 92C99)},
  MRNUMBER = {2999086},
  MRREVIEWER = {Ross S. McVinish},
  DOI = {10.3934/dcdsb.2013.18.453},
  URL = {https://doi.org/10.3934/dcdsb.2013.18.453}}
```

а аналогичная запись, выдаваемая системой ZMATH Европейского математического общества (The European Mathematical Society, <https://zbmath.org>), такова:

```
@Article{zbMATH06146721,
  Author = {Peter E. {Kloeden} and Victor {Kozyakin}},
  Title = {{Asymptotic behaviour of random tridiagonal Markov chains in biological applications.}},
  FJournal = {{Discrete and Continuous Dynamical Systems. Series B}},
  Journal = {{Discrete Contin. Dyn. Syst., Ser. B}},
  ISSN = {1531-3492; 1553-524X/e},
  Volume = {18},
  Number = {2},
  Pages = {453--465},
  Year = {2013},
  Publisher = {American Institute of Mathematical Sciences (AIMS), Springfield, MO},
  Language = {English},
  MSC2010 = {60J10 15B48 37H10},
  Zbl = {1277.60118}}
```

Как видно, вся существенная библиографическая информация (авторы, название журнала, выходные данные работы и т.п.) в обеих BibTeX-записях совпадает. В то же время оформление соответствующих записей немного различается и, кроме того, в них включаются и некоторые “индивидуальные” поля (например, учетные номера записей в соответствующих системах —

MRNUMBER и Zbl), отражающие предпочтения составителей данных записей. В частности, запись от MR LOOKUP содержит важное поле DOI — *цифровой идентификатор объекта*, по которому с помощью сервиса DIGITAL OBJECT IDENTIFIER (The International DOI Foundation (IDF), Digital Object Identifier, <http://www.doi.org>) можно перейти на WEB-страницу издательства, содержащую как минимум аннотацию и библиографические данные искомой публикации (а иногда и ее полный текст).

## НАКОНЕЦ МЫ ДОБРАЛИСЬ ДО DOI

Система DOI была создана по инициативе издательской индустрии, которая признала необходимость однозначно идентифицировать объекты контента, а не ссылаться на них по местоположению. В 1998 году для разработки системы был основан Международный Фонд DOI (International DOI Foundation); для внедрения системы DOI были разработаны необходимые технологии и стандарты [2, 3]. Первый сервис регистрации имен DOI начал работать в 2000 году; к началу 2009 года через восемь регистрационных сервисов было назначено уже около восьми миллионов имен DOI. Наиболее широко используемым применением системы DOI является служба перекрестных ссылок между издателями CROSSREF (<https://www.crossref.org>), которая позволяет связывать ссылки из цитаты непосредственно с цитируемым контентом на платформе другого издателя с учетом методов контроля доступа целевого издателя.

Оригинальные имена DOI могут представляться достаточно длинными строками символов, что бывает неудобно при организации ссылок. Чтобы избежать этого, Международный Фонд DOI (International DOI Foundation) создал сервис сокращенных имен DOI, названный SHORTDOI® (<http://shortdoi.org>). При обращении к SHORTDOI® по оригинальному DOI создается его укороченный псевдоним в формате 10/abcde (или возвращается ранее созданный укороченный псевдоним), с которым можно в дальнейшем работать так же, как с оригинальным DOI.

В настоящее время различные функции с использованием DOI выполняют многочисленные сервисы, проприетарные и свободные менеджеры библиографии, такие, например, как

- ENDNOTE компании Clarivate Analytics, <https://endnote.com>,
- MENDELEY компании Elsevier, <https://www.mendeley.com>,
- CITAVI компании Swiss Academic Software, <https://www.citavi.com>,
- ZOTERO, <https://www.zotero.org>,
- ZOTEROBIB, <https://zbib.org> (упрощенная версия ZOTERO),
- DOCEAR, <http://www.docear.org>,

а также свободные десктопные приложения

- JABREF, <http://www.jabref.org> (Windows, Linux, MacOS),
- BIBDESK, <https://bibdesk.sourceforge.io> (MacOS),
- KBIBTEX, <https://userbase.kde.org/KBibTeX> (Linux),

и множество других интернет-сервисов и десктопных программ, среди которых упомянем еще сервис DOI2BIB (<https://www.doi2bib.org>), осуществляющий конвертацию имен DOI в библиографические записи в формате BIBTEX.

Внедрение системы DOI кардинально изменило всю технологию использования библиографических данных — пользователи получили инструмент для “мгновенного” доступа к электронной версии публикации через сервис DIGITAL OBJECT IDENTIFIER, а также столь же мгновенного получения требуемой библиографической информации с помощью упомянутых выше сервисов CROSSREF, ENDNOTE, MENDELEY и множества других.

Казалось бы, для пользователей библиографических данных наступила долгожданная пора “всеобщего счастья”, когда все требуемые данные могут быть практически мгновенно получены, причем в верифицированном виде. Но оказалось, что, к сожалению, все не так хорошо, как хотелось бы (см. эпиграф).

1. Появились “мертвые” DOI, которые ничему не соответствуют. Причины тут могут быть разнообразные — ошибка в самом имени DOI, закрытие или изменение структуры сайта, на котором располагалась соответствующая публикация, перенос публикации на другой сайт и т.п.
2. Появились “полуживые” DOI, которые обслуживаются одними сервисами, но игнорируются другими. Например, DOI 10.1000/182 публикации [2] обслуживается, похоже, только сервисом DIGITAL OBJECT IDENTIFIER и не обслуживается сервисами CROSSREF, ENDNOTE, MENDELEY. Такая же ситуация с DOI, генерируемыми по запросу пользователей социальной сетью для ученых RESEARCHGATE (<https://www.researchgate.net>).
3. Появились ошибочные DOI, которые указывают на другие публикации.
4. Наконец, оказалось, что **библиографические данные, выдаваемые различными сервисами по запросу DOI, могут отличаться!** Например, запрос данных об упомянутой выше статье [1] по ее DOI 10.3934/dcdsb.2013.18.453 при обращении к сервису CROSSREF приводит к следующей цитате:

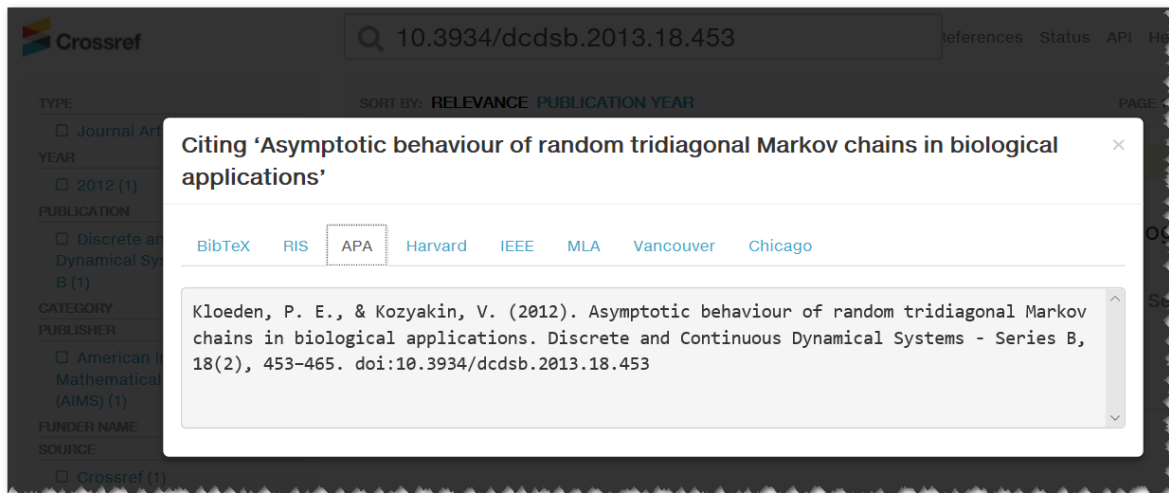


Рис. 2. Снимок экрана сервиса CROSSREF по запросу DOI 10.3934/dcdsb.2013.18.453

в которой год публикации 2012, отличается от истинного года (2013) журнальной (бумажной) публикации. Аналогичная ситуация возникает при обращении к сервисам ENDNOTE, MENDELEY, ZOTEROBIV, DOI2BIV, а также при попытке получения данных с помощью программы JAVREF — все они также выдают неверный год публикации статьи [1] по ее DOI.

Первые два указанных недостатка не являются столь уж критичными — здесь, по крайней мере, соответствующий сервис, с помощью которого запрашиваются библиографические данные по DOI, немедленно уведомляет о невозможности выдачи соответствующих данных. Третий недостаток досаден, но что поделать — ошибиться может каждый. К счастью, первые три недостатка имеют случайный характер.

Но вот последний недостаток уже представляется достаточно неприятным, поскольку он проявляется систематически и при этом ни один из указанных выше сервисов никоим образом

не предупреждает, что выдаваемые им данные требуют дополнительной верификации — это в значительной степени принижает сам смысл “цифрового идентификатора объекта”, DOI.

## РАССЛЕДОВАНИЕ

Возникает естественный вопрос:

*Каким образом могло получиться так, что по одному и тому же DOI различные сервисы выдают различную информацию?*

И тут следует обратить внимание на тот факт, что BibTeX является не единственным, и даже не самым распространенным форматом хранения библиографических данных. BibTeX получил распространение в среде научных публикаций, подготавливаемых по большей части с помощью системы L<sup>A</sup>T<sub>E</sub>X и различных ее надстроек. В издательской же индустрии распространены (появившиеся задолго до BibTeX'a) другие форматы хранения и обмена библиографическими данными, среди которых одним из наиболее распространенных является формат RIS, развиваемый компанией Research Information Systems, Incorporated и используемый в качестве основного формата такими цифровыми библиотеками как IEEE Xplore, SCOPUS, SCIENCE DIRECT, SPRINGERLINK и библиографическими сервисами ZOTERO, CITAVI, MENDELEY, ENDNOTE, CROSSREF. Оказалось, что в формате RIS запись для статьи [1], полученная при обращении к CROSSREF, имеет следующий вид:

```
TY - JOUR
DO - 10.3934/dcdsb.2013.18.453
UR - http://dx.doi.org/10.3934/dcdsb.2013.18.453
TI - Asymptotic behaviour of random tridiagonal Markov chains in
    biological applications
T2 - Discrete and Continuous Dynamical Systems - Series B
AU - Kloeden, Peter E.
AU - Kozyakin, Victor
PY - 2012
DA - 2012/11
PB - American Institute of Mathematical Sciences (AIMS)
SP - 453-465
IS - 2
VL - 18
SN - 1531-3492
ER -
```

В этой записи имеются два параметра, характеризующих дату — DA (Date) и PY (Publication year), причем в обоих этих параметрах указан год **2012** !

К сожалению, в описании формата RIS, доступном по ссылке [https://web.archive.org/web/20120526103719/http://refman.com/support/risformat\\_intro.asp](https://web.archive.org/web/20120526103719/http://refman.com/support/risformat_intro.asp) не содержится детального разъяснения смысла этих параметров. Однако значение 2012/11 параметра PY в RIS-записи статьи [1] совпадает с появлением онлайн версии (first online) этой статьи и, вероятно, именно таков смысл параметра PY — это дата **первого публичного** появления публикации. В то время как параметр Year в BibTeX описывается как **год в выходных данных** печатной публикации! По-видимому, именно это, в явном виде нигде не упоминаемое различие в трактовке понятия “год публикации”, и послужило причиной несоответствия в библиографических данных статьи [1], выдаваемых MR LOOKUP и ZBMATH с одной стороны и CROSSREF, MENDELEY, ZOTEROBIB, DOI2BIB, JAVREF с другой.

## ВЫВОД

Автор обратился с описанием несоответствия в данных, предоставляемых различными библиографическими сервисами, в форум одной из программ (ее название в данном случае несус-

щественно), недвусмысленно продвигающей идею преимущества получения библиографических данных из интернет-источников. К сожалению, достаточно ожидаемо это обращение ни к чему не привело: был получен ответ, “объясняющий”, что программа тут не при чем, — она “доверяет” тем данным, которые ей выдают сервисы, и эти данные не перепроверяет. Организовать диалог между двумя группами информационных сервисов с призывом согласовать/стандартизировать их способы трактовки библиографических данных обычному пользователю также явно не под силу. А учитывая, что ситуация, описанная выше, не единична, вывод из данной заметки таков:

***Используй DOI? Доверяй, но проверяй!***

#### СПИСОК ЛИТЕРАТУРЫ

1. *Kloeden P. E., Kozyakin V.* Asymptotic behaviour of random tridiagonal Markov chains in biological applications // *Discrete Contin. Dyn. Syst. Ser. B.* 2013. V. 18. No. 2. P. 453–466. DOI: 10.3934/dcdsb.2013.18.453.
2. DOI® Handbook. International DOI Foundation. [Online; updated August 16, 2018]. DOI: 10.1000/182.
3. *Paskin N.* Digital Object Identifier (DOI®) System // *Encyclopedia of Library and Information Sciences*, 3rd Edition / Ed. by M. J. Bates, M. N. Maack. Boca Raton: CRC Press, 2009. P. 1586–1592. URL: [http://www.doi.org/overview/DOI\\_article\\_ELIS3.pdf](http://www.doi.org/overview/DOI_article_ELIS3.pdf).

### Beware of DOI!

#### A “bibliographic detective story” of the era of digitalization

**Kozyakin V.**

An example of inconsistencies in information provided by popular bibliographic services is described, and the reasons for such inconsistencies are discussed.

**KEYWORDS:** bibliography, digital object identifier, DOI, BibTeX, RIS