

О внедрении метаданных в PDF-файлы на основе XMP-схемы PRISM

Научный сервис в сети Интернет – 2019
(23–28 сентября 2019, Абрау-Дюрсо)

М. Н. Саушкин, А. Д. Зубкова
saushkin.mn@samgtu.ru

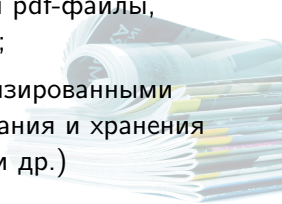
Самарский государственный технический университет

25 сентября 2019 г.



В настоящее время...

- тысячи научных статей по разным направлениям публикуются еженедельно;
- количество публикаций растет так быстро, что ученые физически не успевают обратить внимание на новые исследования;
- у ученых в личных архивах накапливаются pdf-файлы, которые так и не были ни разу прочитаны;
- ученые вынуждены пользоваться специализированными сервисами для управления, каталогизирования и хранения pdf-файлов (Kopernio, EndNote, Mendeley и др.)



... метаданные в pdf-файлах...

- облегчают видимость и индексацию публикаций в сети;
- облегчают импорт в специализированные сервисы для управления, каталогизирования и хранения pdf-файлов;
- облегчают поиск необходимых публикаций в сети и в личных архивах;
- повышают вероятность прочтения и возможно цитирования публикации;



... присутствуют у зарубежных издателей...

преимущественно для современных публикаций

- платформа ScienceDirect содержит публикации с метаданными (Dublin Core, Prism — не везде), использует приложение Elsevier;
- платформа SpringerLink содержит публикации с метаданными (Dublin Core), использует приложение Springer;
- платформа Arpha содержит публикации с метаданными (Dublin Core), использует метаданные от издателей (?)

... и отсутствуют у российских издателей и агрегаторов...

«мусор»¹ в метаданных присутствует

- на сайтах вузовских издательств;
- в таких агрегаторах как eLibrary, КиберЛенинка и ОЭК — системе учета обязательных электронных экземпляров;
- в изданиях на платформах EIPub и Эко-Вектор, оказывающих и издательские услуги

¹метаданные, не соответствующие публикации

... и отсутствуют у российских издателей и агрегаторов...

«мусор»¹ в метаданных присутствует

- на сайтах вузовских издательств;
- в таких агрегаторах как eLibrary, КиберЛенинка и ОЭК — системе учета обязательных электронных экземпляров;
- в изданиях на платформах EIPub и Эко-Вектор, оказывающих и издательские услуги;
- в трудах конференции «Научный сервис в сети Интернет»

¹метаданные, не соответствующие публикации

... портал MathNet.ru — приятное исключение

Портал MathNet.ru перед тем как отдать pdf-файл читателю с помощью утилиты pdftk

- «затирает» метаданные от издателя;
- внедряет титульную страницу с выходными данными публикации;
- в метаданные pdf-файла добавляет название статьи и авторов.

К сожалению, другие метаданные игнорируются и не внедряются, хотя они в полном объеме присутствуют в БД портала.



Цель (желание)

Иметь утилиту, которая работала бы как pdftk (консольное приложение, работа в *nix-системах, исходные данные в текстовом файле), но с более широким набором метаданных, например из схемы Prism.



Рассматривается только ПО, которое способно внедрять метаданные в уже существующие pdf-файлы.
Список не претендует на абсолютную полноту.



Adobe Acrobat Pro

Возможности (плюсы)

- поддерживает русский язык метаданных (Unicode);
- поддерживает все существующие схемы метаданных;
- позволяет изменять метаданные как интерактивно, так и с помощью импорта в формате XMP

Недостатки (минусы)

- отсутствие консольного режима;
- проблематична поддержка/запуск в *nix-системах;
- платное программное обеспечение

PDF Metadata Editor

Возможности (плюсы)

- поддерживает русский язык метаданных (Unicode);
- позволяет изменять метаданные как интерактивно, так и с помощью импорта в формате XMP;
- кроссплатформенное приложение (наличие Java VM);

Недостатки (минусы)

- поддержка только стандартных схем (Basic, XMP Pdf, XMP Dublin Core, XMP Rights)
- консольный (пакетный) режим только в платной версии



Pdftk

Возможности (плюсы)

- поддерживает русский язык метаданных (Unicode);
- позволяет импортировать метаданные, подготовленные в текстовом формате определенной структуры;
- консольное приложение;
- кроссплатформенное приложение (основан на iText);
- свободное программное обеспечение

Недостатки (минусы)

- поддержка только стандартных схем (Basic, Dublin Core)

Adobe PDF Library

Возможности (плюсы)

- development library (C++, .NET, Java);
- позволяет производить любые манипуляции с pdf-файлами;
- поддержка использования любых метаданных (?)

Недостатки (минусы)

- проприетарная лицензия;
- необходимость написания собственного ПО;
- отсутствие знакомства с библиотекой



iText

Возможности (плюсы)

- development library (Java, C#, and other .NET languages);
- позволяет производить любые манипуляции с pdf-файлами;
- поддержка использования любых метаданных;
- двойная лицензия (проприетарная и AGPL);
- наличие опыта взаимодействия с библиотекой

Недостатки (минусы)

- необходимость написания собственного ПО

`metimpdf` ([met]adata [im]port into a [pdf] file) — консольная утилита для импорта метаданных в pdf-файлы на основе XMP-схемы PRISM

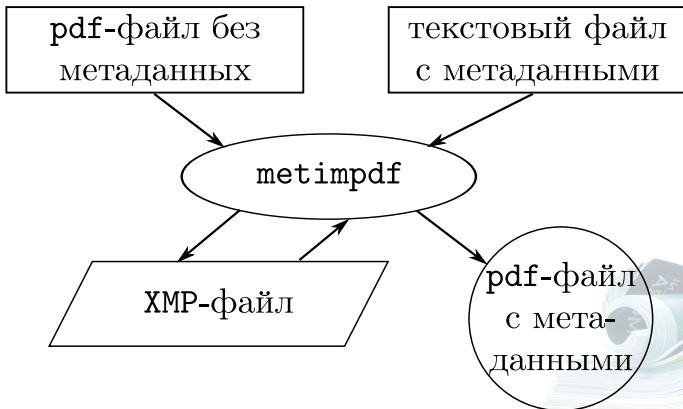


Используемое программное обеспечение

- библиотека iText v.5.5.13
- язык программирования C#
- среда разработки SharpDevelop v.5.1



Схема работы утилиты metimpdf



Запуск утилиты: `metimpdf source.pdf meta.data putput.pdf`

Структура файла с метаданными

```
meta.data — Блокнот
Файл Правка Формат Вид Справка
Title = {0 скорости стабилизации решений задачи Коши для уравнения Карлемана
с периодическими начальными данными}
Author = {Сергей Анатольевич Духновский}
Keywords = {кинетическое уравнение ; уравнение Карлемана ; Фурье-решение ;
состояние равновесия ; секулярные члены ; обобщенное решение}
Doi = {10.14498/vsgtu1529}DoiUrl = {http://dx.doi.org/10.14498/vsgtu1529}
Publisher = {Самарский государственный технический университет}
Rights = {@ Автор, 2017; Самарский государственный технический университет,
2017 (составление); Creative Commons Attribution 4.0 International (CC BY
4.0) (распространение)}
RightsUrl = {https://creativecommons.org/licenses/by/4.0/deed.ru}
Description = {Исследуется одномерная система уравнений для дискретной модели
газа (система уравнений Карлемана). Система Карлемана является кинетическим
уравнением Больцмана модельного одномерного газа, состоящего из двух частиц.
Для этой модели не сохраняются импульс и энергия. На примере модели Карлемана
хорошо видна суть уравнения Больцмана, которое описывает смесь
«конкурирующих» процессов: релаксацию и свободное движение. Доказывается
существование глобального решения задачи Коши для возмущения состояния
равновесия с периодическими начальными данными. Впервые устанавливается
скорость стабилизации к состоянию равновесия (экспоненциальная
стабилизация).}
Issn = {2310-7081 (online), 1991-8615 (print)}
Volume = {21}
Number = {1}
CoverDisplayDate = {Март, 2017}
CoverDate = {2017-06-22}
IssueName = {Вестн. Сам. гос. техн. ун-та. Сер. Физ.-мат. науки}
PageRange = {7-41}
StartingPage = {7}
EndingPage = {41}
CreatorTool = {Math-Net.Ru Meta Data Creator}
Producer = {Math-Net.Ru}
```



Элементы описания

Элемент	XMP-схемы	Описание элемента
Title	Dublin Core, Prism	Название статьи (документа)
Author	Dublin Core, Prism	Автор(ы) статьи (документа); может добавляться XMP-схему в виде коллекции
Keywords	Dublin Core, Prism, pdf	Ключевые слова статьи (документа); добавляется XMP-схему в виде коллекции
Doi	Dublin Core, Prism	Идентификатор DOI статьи (документа)
DoiUrl	Prism	Стандартная ссылка (URL) для доступа к статье (документу) по идентификатору DOI
Publisher	Dublin Core	Издатель статьи (документа)

Элементы описания

Элемент	XMP-схемы	Описание элемента
Rights	Dublin Core, xmpRights	Текстовое поле с описанием исключительных прав, прав доступа и использования (распространения) статьи (документа)
RightsUrl	xmpRights	Ссылка (URL) на web-документ, в котором приводится описание прав доступа и использования (распространения) статьи (документа)
Description	Dublin Core, Prism	Описание (аннотация) статьи (документа)
Issn	Prism	Номер ISSN издания, в котором опубликована статья (документ)
elssn	Prism	Номер eISSN издания, в котором опубликована статья (документ)



Элементы описания

Элемент	XMP-схемы	Описание элемента
Volume	Prism	Том издания, в котором опубликована статья (документ)
Number	Prism	Номер издания, в котором опубликована статья (документ)
Cover-Display-Date	Prism	Текстовый формат описания даты выхода издания, в котором опубликована статья (документ), например, Март, 2017
CoverDate	Prism	Машиночитаемый формат описания даты выхода издания, в котором опубликована статья (документ), например, 2017-06-22
IssueName	Prism	Название издания, в котором опубликована статья (документ)
PageRange	Prism	интервал страниц на которых опубликована статья (документ) в издании

Свойства документа без метаданных (XMP Dublin Core)

Свойства документа ×

Описание | Защита | Шрифты | Вид при открытии | Заказные | Дополнительные

Описание

Файл: dukhnovsky.pdf

Заголовок:

Автор:

Тема:

Ключевые слова:

Создан: 08.06.2017 7:16:04 Дополнительные метаданные...

Изменен: 08.06.2017 7:16:04

Приложение: LaTeX with hyperref package

Дополнительно

Производитель PDF: pdfTeX-1.40.14

Версия PDF: 1.5 (Acrobat 6.x)

Местонахождение: D:\Dropbox\metimpdf\metimpdf\bin\Debug\

Размер файла: 825,20 КБ (845 001 байт)

Размер страницы: 146 x 240 мм Количество страниц: 35

PDF с тегами: Нет Быстрый просмотр в Web: Нет

Свойства документа без метаданных (описание)

dukhnovsky.pdf

Описание
Дополнительно

Описание

Заголовок документа:

Автор:

Заголовок автора:

Описание:

Автор описания:

Ключевые слова:

Для разделения ключевых слов используются запятые

Наличие авторского права:

Уведомление об авторских правах:

URL владельца авторских прав:

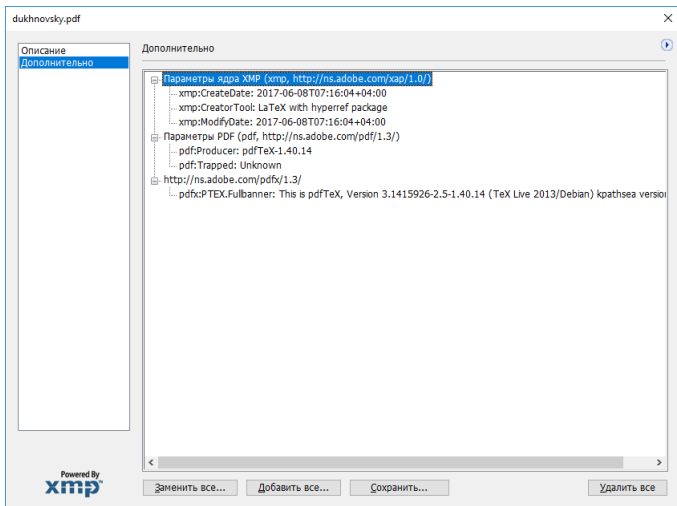
Перейти к URL...

Создано: 08.06.2017 7:16:04
Изменено: 08.06.2017 7:16:04
Приложение: LaTeX with hyperref package
Формат:

Powered By
xmp



Свойства документа без метаданных (дополнительно)



dukhnovsky.pdf

Описание
Дополнительно

Дополнительно

- Параметры ядра XMP (xmp, <http://ns.adobe.com/xap/1.0/>)
 - xmp:CreateDate: 2017-06-08T07:16:04+04:00
 - xmp:CreatorTool: LaTeX with hyperref package
 - xmp:ModifyDate: 2017-06-08T07:16:04+04:00
- Параметры PDF (pdf, <http://ns.adobe.com/pdf/1.3/>)
 - pdf:Producer: pdfTeX-1.40.14
 - pdf:Trapped: Unknown
- <http://ns.adobe.com/pdfx/1.3/>
 - pdfx:PTEX.Fullbanner: This is pdfTeX, Version 3.1415926-2.5-1.40.14 (TeX Live 2013/Debian) kpathsea version

Powered By **xmp**

Заменить все... Добавить все... Сохранить... Удалить все

Свойства документа с метаданными (описание)

output.pdf

Описание
Дополнительно

Описание

Заголовок документа: О скорости стабилизации решений задачи Коши для уравнения

Автор: Сергей Анатольевич Духновский

Заголовок автора:

Описание: Исследуется одномерная система уравнений для дискретной модели газа (система уравнений Карлемана). Система Карлемана является кинетическим уравнением

Автор описания:

Ключевые слова: кинетическое уравнение; уравнение Карлемана; Фурье-решение; состояние равновесия; секулярные члены; обобщенное решение

Для разделения ключевых слов используются запятые

Наличие авторского права: Защищен авторскими правами

Уведомление об авторских правах: © Автор, 2017; Самарский государственный технический университет, 2017 (составление); Creative Commons Attribution 4.0 International (CC BY 4.0) (распространение)

URL владельца авторских прав: <https://creativecommons.org/licenses/by/4.0/deed.ru>

Перейти к URL...

Создано:
Изменено: 21.05.2018 11:15:40
Приложение: Math-Net.Ru Meta Data Creator
Формат: application/pdf

Powered By
xmp

Свойства документа с метаданными (доп.: XMP DC)

The screenshot shows a software window titled 'output.pdf' with a tree view of document metadata. The left sidebar has 'Описание' and 'Дополнительно' tabs, with 'Дополнительно' selected. The main area displays a tree structure of metadata elements:

- Параметры Дублинского ядра (dc, <http://purl.org/dc/elements/1.1/>)
 - dc:format: application/pdf
 - dc:identifier: 10.14498/vsgtu1529
 - dc:title (alt container)
 - [x-default]: О скорости стабилизации решений задачи Коши для уравнения Карлемана с периодическим
 - dc:creator (seq container)
 - [1]: Сепрей Анатолевич Духновский
 - dc:publisher (bag container)
 - [1]: Самарский государственный технический университет
 - dc:rights (alt container)
 - [x-default]: © Автор, 2017; Самарский государственный технический университет, 2017 (составление
 - dc:description (alt container)
 - [x-default]: Исследуется одномерная система уравнений для дискретной модели газа (система урав
 - dc:subject (bag container)
 - [1]: кинетическое уравнение
 - [2]: уравнение Карлемана
 - [3]: Фурье-решение
 - [4]: состояние равновесия
 - [5]: секулярные члены
 - [6]: обобщенное решение
- Параметры PDF (pdf, <http://ns.adobe.com/pdf/1.3/>)
- <http://prismstandard.org/namespaces/basic/2.2/>
- Параметры управления правами XMP (xmpRights, <http://ns.adobe.com/xap/1.0/rights/>)
- Параметры ядра XMP (xmp, <http://ns.adobe.com/xap/1.0/>)

At the bottom of the window, there is a 'Powered By xmp' logo and four buttons: 'Заменить все...', 'Добавить все...', 'Сохранить...', and 'Удалить все'.

Свойства документа с метаданными (доп.: Prism)

output.pdf

Описание
Дополнительно

Дополнительно

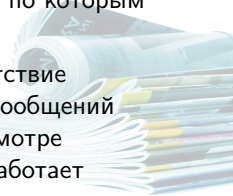
- [-] Параметры Дублинского ядра (dc, <http://purl.org/dc/elements/1.1/>)
- [-] Параметры PDF (pdf, <http://ns.adobe.com/pdf/1.3/>)
- [-] <http://prismstandard.org/namespaces/basic/2.2/>
 - prism:doi: 10.14498/vsgtu1529
 - prism:url: <http://dx.doi.org/10.14498/vsgtu1529>
 - prism:issn: 2310-7081 (online), 1991-8615 (print)
 - prism:volume: 21
 - prism:number: 1
 - prism:coverDisplayDate: Март, 2017
 - prism:coverDate: 2017-06-22
 - prism:issueName: Вестн. Сам. гос. техн. ун-та. Сер. Физ.-мат. науки
 - prism:pageRange: 7-41
 - prism:startingPage: 7
 - prism:endingPage: 41
- [-] Параметры управления правами XMP (xmpRights, <http://ns.adobe.com/xap/1.0/rights/>)
- [-] Параметры ядра XMP (xmp, <http://ns.adobe.com/xap/1.0/>)

Powered By **xmp**

Заменить все... Добавить все... Сохранить... Удалить все

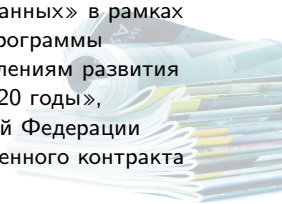
Результат

- Разработанная утилита metimpdf может использоваться на терминальном сервере для внедрения метаданных в pdf-файлы при их запросе с сервера (по аналогии как это сделано на портале MathNet.ru с утилитой pdftk).
- В отличие от утилиты pdftk, которая поддерживает только XMP Dublin Core, утилита metimpdf поддерживает XMP-схему Prism, которая может содержать полное библиографическое описание статьи, включая аннотацию и другие метаданные, по которым может производиться поисковый запрос.
- Корректное отображение метаданных и их соответствие исходным данным, а также отсутствие ошибок и сообщений о нарушении целостности pdf-файла при его просмотре позволяет сделать вывод, что утилита metimpdf работает корректно и соответствует поставленной цели.



Финансирование

Работа выполнена в рамках договоров МОН 2018/20 от 20 апреля 2018 г. и МОН 2019/40 от 29 мая 2019 г. на реализацию программы развития научного журнала «Вестник Самарского государственного технического университета. Серия «Физико-математические науки» между ФГБОУ ВО «СамГТУ» и НП «НЭИКОН», действующего в рамках Государственного контракта на выполнение работ (оказание услуг) для государственных нужд от 28 августа 2017 г. № 14.597.11.0035 по проекту «Продолжение конкурсной поддержки программ развития научных журналов с целью их вхождения в международные наукометрические базы данных» в рамках реализации мероприятия 3.3.1 федеральной целевой программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014–2020 годы», утвержденной постановлением Правительства Российской Федерации от 21 мая 2013 г. № 426 (Реестровый номер Государственного контракта 1771053913517000299).



Спасибо за внимание!

