УДК 004.021, 004.457

АНАЛИЗ МЕТОДОВ АВТОМАТИЗИРОВАННОЙ ГЕНЕРАЦИИ ДОКУМЕНТОВ ДЛЯ РЕШЕНИЯ ЗАДАЧИ СОЗДАНИЯ ГЕНЕРАТОРА КНИГИ АБСТРАКТОВ ДЛЯ КОНФЕРЕНЦИЙ ЛИТ ОИЯИ

Ильина Анна Владимировна[1], Кореньков Владимир Васильевич[2]

[1]Студент;
ГБОУ ВО МО «Университет «Дубна»,
Институт системного анализа и управления;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: ttrnkv75@yandex.ru.

[2]Директор, доктор технических наук, профессор;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6;
ГБОУ ВО МО «Университет «Дубна»,
Институт системного анализа и управления;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: korenkov@cv.jinr.ru.

В работе рассматриваются методы автоматизированной генерации документов, выделены достоинства и недостатки каждого из них, приведено обоснование выбора метода генерации документов на основе шаблонов формата DOCX для решения задачи автоматизированной генерации книги абстрактов для конференций ЛИТ ОИЯИ.

Ключевые слова: генерация документов, конференция, абстракт, программирование, *DOCX*, *LaTeX*.

ANALYSIS OF METHODS OF AUTOMATED GENERATION OF DOCUMENTS FOR THE CREATION OF A GENERATOR OF THE BOOK OF ABSTRACTS FOR LIT JINR CONFERENCES

Ilina Anna[1], Korenkov Vladimir[2]

[1]Student;
Dubna State University,
Institute of the system analysis and management;
141980, Dubna, Moscow reg., Universitetskaya str., 19;
e-mail: ttrnkv75@yandex.ru.

[2]Director, Doctor of Technical Science, professor;
Joint institute for nuclear researches,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, Joliot-Curie, 6;
Dubna State University,
Institute of the system analysis and management;
141980, Dubna, Moscow reg., Universitetskaya str., 19;
e-mail: korenkov@cv.jinr.ru.

The paper deals with the methods of automated document generation, highlighted the advantages and disadvantages of each of them, the rationale for the choice of the method of generating documents based on docx format templates to solve the problem of automated generation of abstracts for conferences LIT JINR.

Keywords: document generation, conference, abstract, programming, DOCX, LaTeX.

Введение

Проблема автоматизированной генерации документов особенно актуальна для систем, обрабатывающих огромное количество данных, которое в результате должно представлять из себя некоторые конкретизированные отчеты, оформленные специальным образом. Зачастую количество данных настолько большое, а формат выходного файла столь нетривиален, что вручную решить такую задачу не представляется возможным либо решение займет колоссальное количество времени.

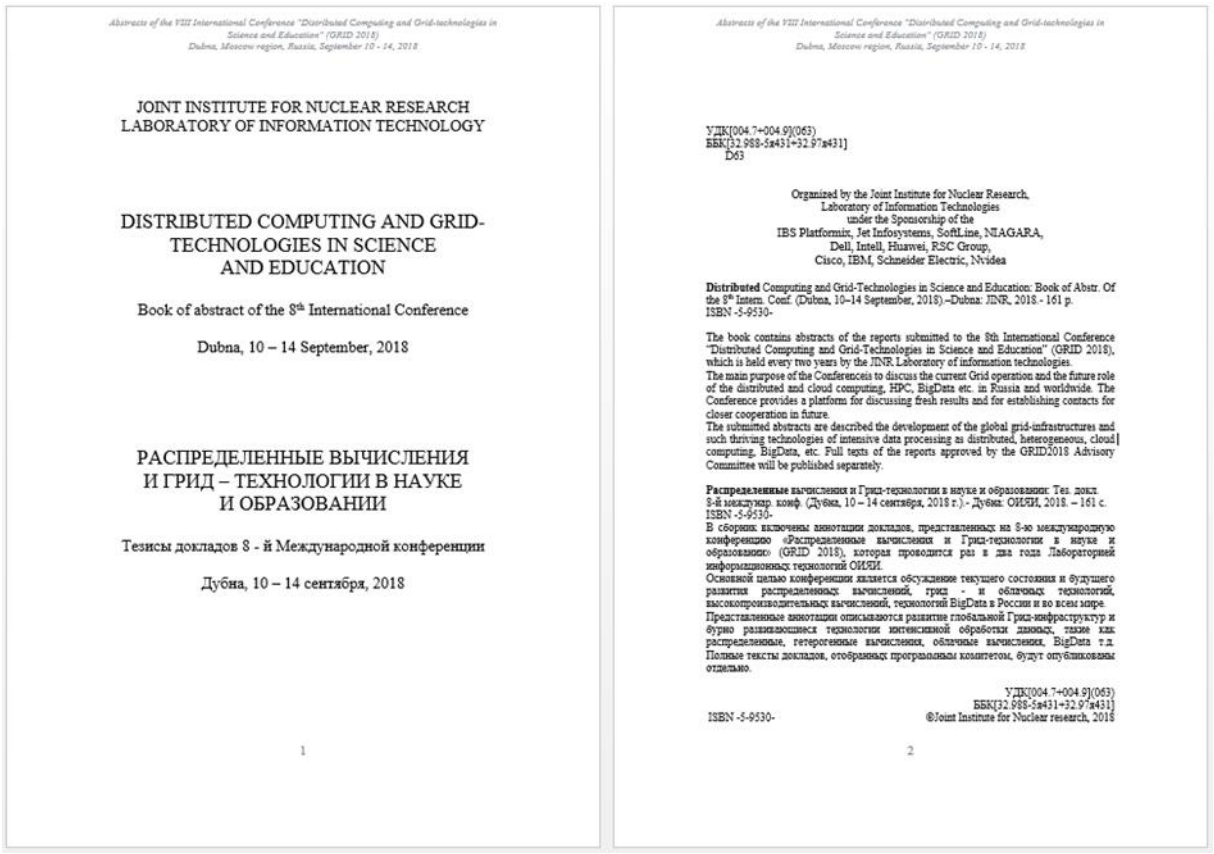
Задачу автоматизированной генерации документов нельзя считать тривиальной и имеющей некое общее решение. Это связано в первую очередь с особенностями стилей конкретных генерируемых документов, а также с форматом входных данных, содержащих ключевую для генерации информацию. Однако существуют некоторые общие методы и подходы, составляющие основу алгоритмизации данного процесса.

Задача автоматизированной генерации документов актуальна и для ОИЯИ – международной межправительственной научно-исследовательской организации, перед которой задача автоматизации генерации различного рода документов стоит особенно остро.

1. Структура книги абстрактов

Одной из задач является задача генерации некоторого документа для конференций, которые постоянно организуются в институте. Рассмотрим этот документ более подробно.

Организация каждой конференции включает в себя создание книги абстрактов, определенной и стилизованной специальным образом (см. рис. 1). Книга содержит большое количество страниц, содержащих в себе данные о конференции и выступающих (см. рис. 2), представленные в исходном файле формата *XML*, который генерируется в информационной системе организации конференций *Indico* (см. рис. 3). Существует проблема автоматизированной генерации такой книги, подходы к решению которой будут рассмотрены в данной работе.



Abstracts of the VIII International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2018)
Dubna, Moscow region, Russia, September 10 - 14, 2018

JOINT INSTITUTE FOR NUCLEAR RESEARCH
LABORATORY OF INFORMATION TECHNOLOGY

DISTRIBUTED COMPUTING AND GRID-TECHNOLOGIES IN SCIENCE AND EDUCATION

Book of abstract of the 8th International Conference

Dubna, 10 – 14 September, 2018

РАСПРЕДЕЛЕННЫЕ ВЫЧИСЛЕНИЯ И ГРИД – ТЕХНОЛОГИИ В НАУКЕ И ОБРАЗОВАНИИ

Тезисы докладов 8 - й Международной конференции

Дубна, 10 – 14 сентября, 2018

1

Abstracts of the VIII International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2018)
Dubna, Moscow region, Russia, September 10 - 14, 2018

УДК[004.7+004.9](063)
ББК[32.988-5я431+32.97я431]
D63

Organized by the Joint Institute for Nuclear Research,
Laboratory of Information Technologies
under the Sponsorship of the
IBS Platformix, Jet Infosystems, SoftLine, NIAGARA,
Dell, Intell, Huawei, RSC Group,
Cisco, IBM, Schneider Electric, Nvidea

Distributed Computing and Grid-Technologies in Science and Education: Book of Abstr. Of the 8th Intern. Conf. (Dubna, 10–14 September, 2018).–Dubna: JINR, 2018.- 161 p.
ISBN -5-9530-

The book contains abstracts of the reports submitted to the 8th International Conference "Distributed Computing and Grid-Technologies in Science and Education" (GRID 2018), which is held every two years by the JINR Laboratory of information technologies.
The main purpose of the Conferenceis to discuss the current Grid operation and the future role of the distributed and cloud computing, HPC, BigData etc. in Russia and worldwide. The Conference provides a platform for discussing fresh results and for establishing contacts for closer cooperation in future.
The submitted abstracts are described the development of the global grid-infrastructures and such thriving technologies of intensive data processing as distributed, heterogeneous, cloud computing, BigData, etc. Full texts of the reports approved by the GRID2018 Advisory Committee will be published separately.

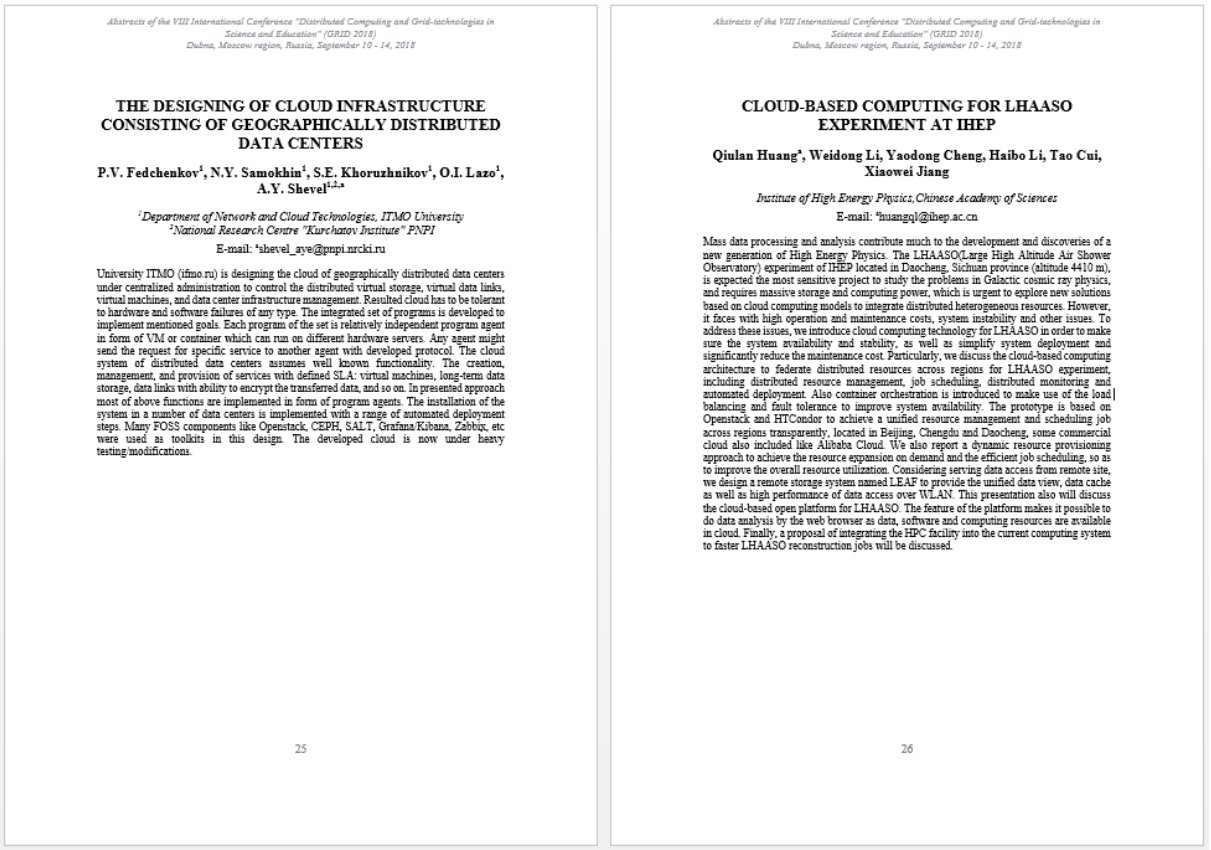
Распределенные вычисления и Грид-технологии в науке и образовании: Тез. докл. 8-й междунар. конф. (Дубна, 10 – 14 сентября, 2018 г.).- Дубна: ОИЯИ, 2018. – 161 с.
ISBN -5-9530-

В сборник включены аннотации докладов, представленных на 8-ю международную конференцию «Распределенные вычисления и Грид-технологии в науке и образовании» (GRID 2018), которая проводится раз в два года Лабораторией информационных технологий ОИЯИ.
Основной целью конференции является обсуждение текущего состояния и будущего развития распределенных вычислений, грид - и облачных технологий, высокопроизводительных вычислений, технологий BigData в России и во всем мире.
Представленные аннотации описывают развитие глобальной Грид-инфраструктур и бурно развивающиеся технологии интенсивной обработки данных, такие как распределенные, гетерогенные вычисления, облачные вычисления, BigData т.д.
Полные тексты докладов, отобранных программным комитетом, будут опубликованы отдельно.

УДК[004.7+004.9](063)
ББК[32.988-5я431+32.97я431]
ISBN -5-9530-

2

Рис. 1. Титульные страницы книги абстрактов



THE DESIGNING OF CLOUD INFRASTRUCTURE CONSISTING OF GEOGRAPHICALLY DISTRIBUTED DATA CENTERS

P.V. Fedchenkov[1], N.Y. Samokhin[1], S.E. Khoruzhnikov[1], O.I. Lazo[1], A.Y. Shevel[1,2,*]

[1]Department of Network and Cloud Technologies, ITMO University
[2]National Research Centre "Kurchatov Institute" PNPI

E-mail: *shevel_aye@pnpi.nrcki.ru

University ITMO (ifmo.ru) is designing the cloud of geographically distributed data centers under centralized administration to control the distributed virtual storage, virtual data links, virtual machines, and data center infrastructure management. Resulted cloud has to be tolerant to hardware and software failures of any type. The integrated set of programs is developed to implement mentioned goals. Each program of the set is relatively independent program agent in form of VM or container which can run on different hardware servers. Any agent might send the request for specific service to another agent with developed protocol. The cloud system of distributed data centers assumes well known functionality. The creation, management, and provision of services with defined SLA: virtual machines, long-term data storage, data links with ability to encrypt the transferred data, and so on. In presented approach most of above functions are implemented in form of program agents. The installation of the system in a number of data centers is implemented with a range of automated deployment steps. Many FOSS components like Openstack, CEPH, SALT, Grafana/Kibana, Zabbix, etc were used as toolkits in this design. The developed cloud is now under heavy testing/modifications.

CLOUD-BASED COMPUTING FOR LHAASO EXPERIMENT AT IHEP

Qiulan Huang*, Weidong Li, Yaodong Cheng, Haibo Li, Tao Cui, Xiaowei Jiang

Institute of High Energy Physics, Chinese Academy of Sciences

E-mail: *huangql@ihep.ac.cn

Mass data processing and analysis contribute much to the development and discoveries of a new generation of High Energy Physics. The LHAASO(Large High Altitude Air Shower Observatory) experiment of IHEP located in Daocheng, Sichuan province (altitude 4410 m), is expected the most sensitive project to study the problems in Galactic cosmic ray physics, and requires massive storage and computing power, which is urgent to explore new solutions based on cloud computing models to integrate distributed heterogeneous resources. However, it faces with high operation and maintenance costs, system instability and other issues. To address these issues, we introduce cloud computing technology for LHAASO in order to make sure the system availability and stability, as well as simplify system deployment and significantly reduce the maintenance cost. Particularly, we discuss the cloud-based computing architecture to federate distributed resources across regions for LHAASO experiment, including distributed resource management, job scheduling, distributed monitoring and automated deployment. Also container orchestration is introduced to make use of the load balancing and fault tolerance to improve system availability. The prototype is based on Openstack and HTCondor to achieve a unified resource management and scheduling job across regions transparently, located in Beijing, Chengdu and Daocheng, some commercial cloud also included like Alibaba Cloud. We also report a dynamic resource provisioning approach to achieve the resource expansion on demand and the efficient job scheduling, so as to improve the overall resource utilization. Considering serving data access from remote site, we design a remote storage system named LEAF to provide the unified data view, data cache as well as high performance of data access over WLAN. This presentation also will discuss the cloud-based open platform for LHAASO. The feature of the platform makes it possible to do data analysis by the web browser as data, software and computing resources are available in cloud. Finally, a proposal of integrating the HPC facility into the current computing system to faster LHAASO reconstruction jobs will be discussed.

Рис. 2. Содержимое книги абстрактов

```
▼<AbstractBook>
  ▼<Conference>
     The 8th International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2018)
   </Conference>
  ▼<abstract>
     <Id>180</Id>
    ▼<Title>
       Грид и облачная инфраструктура дата-центра Института Физики НАН Азербайджана
     </Title>
    ▼<Content>
       Основными направлениями развития дата-центра Института физики НАН Азербайджана являются грид и облачные технологии.
       Пользуясь тем что грид-сегмент дата-центра интегрирован в инфраструктуру EGI/WLCG в качестве грид-сайта уровня Tier3,
       пользователи дата-центра получают возможность участвовать в таких международных проектах как ATLAS (CERN). Сотрудничество с
       международными научными центрами ОИЯИ, ЦЕРН в области информационных технологий способствует эффективному развитию дата
       центра, а также помогает решать задачи пользователей в таких научных областях, как физика высоких энергий, физика твердого
       тела и т. д.
     </Content>
    ▼<field id="content">
       Основными направлениями развития дата-центра Института физики НАН Азербайджана являются грид и облачные технологии.
       Пользуясь тем что грид-сегмент дата-центра интегрирован в инфраструктуру EGI/WLCG в качестве грид-сайта уровня Tier3,
       пользователи дата-центра получают возможность участвовать в таких международных проектах как ATLAS (CERN). Сотрудничество с
       международными научными центрами ОИЯИ, ЦЕРН в области информационных технологий способствует эффективному развитию дата
       центра, а также помогает решать задачи пользователей в таких научных областях, как физика высоких энергий, физика твердого
       тела и т. д.
     </field>
     <field id="summary"/>
    ▼<PrimaryAuthor>
       <FirstName>Aleksey</FirstName>
       <FamilyName>Bondyakov</FamilyName>
       <Email>aleksey@jinr.ru</Email>
       <Affiliation>JINR (Joint Institute For Nuclear Research)</Affiliation>
     </PrimaryAuthor>
    ▼<Speaker>
       <FirstName>Aleksey</FirstName>
       <FamilyName>Bondyakov</FamilyName>
       <Email>aleksey@jinr.ru</Email>
       <Affiliation>JINR (Joint Institute For Nuclear Research)</Affiliation>
     </Speaker>
     <ContributionType>Sectional reports</ContributionType>
     <Track/>
   </abstract>
  ▼<abstract>
```

Рис. 3. Входной файл формата XML

2. Постановка задачи

Цель: автоматизировать генерацию документа абстрактов для конференций, проходящих в ЛИТ ОИЯИ.

Исходные данные:

1. Входной документ формата *XML*, содержащий данные абстрактов и конференции.
2. Пример выходного файла в формате *DOCX*.
3. Справочная литература по языкам объектно-ориентированного программирования высокого уровня.

Модельные представления:

1. Автоматический генератор документов должен представлять из себя программное решение, реализованное под ОС *Linux* и *Windows* (далее – программный продукт (ПП)).
2. ПП должен принимать на вход документ в формате *XML*, автоматический сгенерированный веб-сервисом для организации конференций *Indico*.
3. ПП должен генерировать конечный документ в формате *DOCX* (преимущественно) или *PDF*.

Ожидаемый результат: ПП, позволяющий автоматически сгенерировать книгу абстрактов для конференций, проходящих в ЛИТ ОИЯИ, в формате *DOCX* или *PDF* из документа формата *XML*.

Критерий оценки результата: стиль и форматирование автоматически сгенерированного документа должны быть идентичны стилю и форматированию документа, содержащего пример выходного файла.

Рассмотрим некоторые общие подходы к автоматизированной генерации документов.

3. Метод генерации документов с помощью системы LaTeX

LaTeX – это издательская система на базе *TeX* для верстки текстов с формулами. Сам по себе *TeX* представляет собой специализированный язык программирования, на котором пишутся издательские системы, используемые на практике. Точнее говоря, каждая издательская система на базе *TeX* представляет собой пакет макроопределений (макропакет) этого языка (см. рис. 4) [1].

Достоинства метода:

1. Напечатанный документ выглядит как в книге.
2. Указав с помощью простых средств логическую структуру текста, автор может не вникать в детали оформления, причем эти детали при необходимости нетрудно изменить.
3. Высокое качество и гибкость верстки абзацев и математических формул.
4. Система позволяет генерировать один большой документ из многих маленьких.

Недостатки метода:

1. Наличие определенных специальных навыков верстки и работы в системе *LaTeX*.
2. Работа с исходным текстом и просмотр того, как текст будет выглядеть на печати, – разные операции (не поддерживается свойство *WYSIWYG* (*What you see is what you get*)) [2].
3. *TeX* плохо приспособлен для верстки страниц со сложным взаимодействием текста и графики.
4. Чтобы изменить какую-либо информацию в документе, необходимо править исходный *.tex*-файл и осуществлять генерацию конечного документа заново.

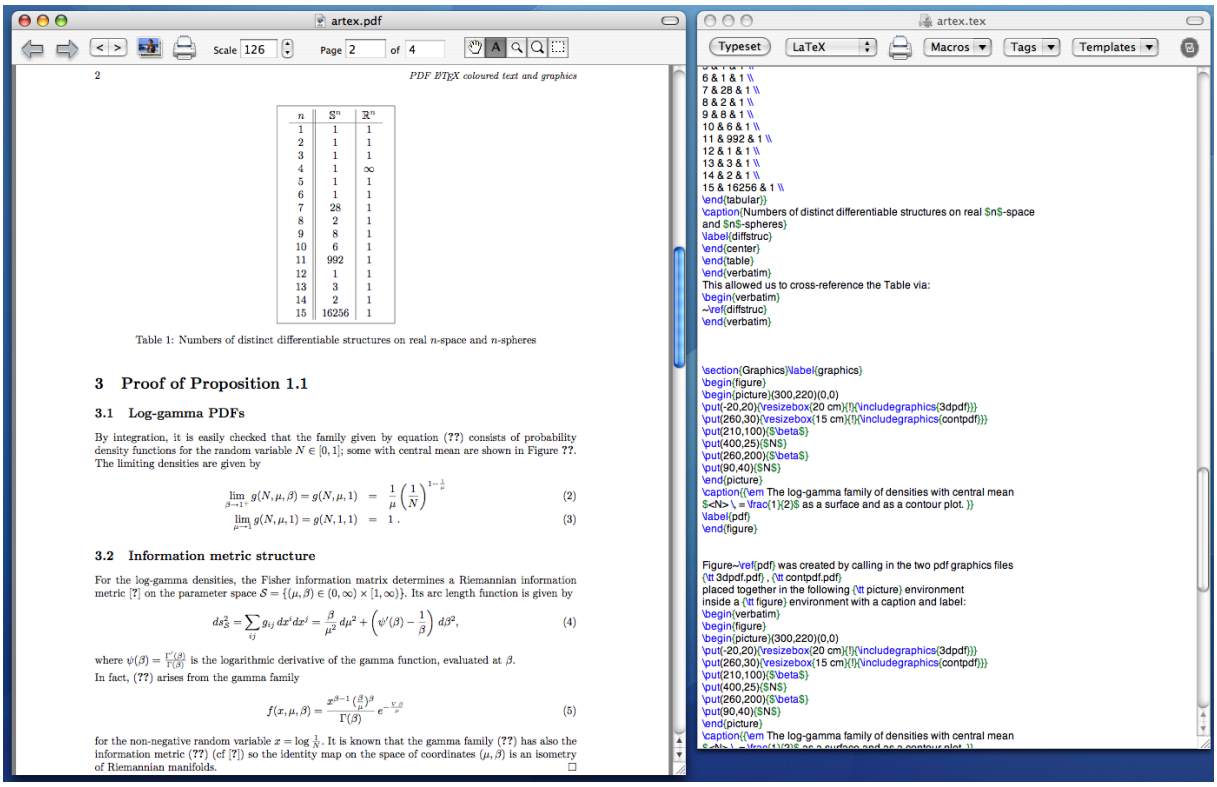


Рис. 4. Пример верстки документа в системе LaTeX

4. Метод генерации документов на основе шаблонов формата DOCX

В основе метода лежит достаточно быстрый процесс создания текстового шаблона, содержащего некоторые переменные (или поля) (см. рис. 5), которые будут автоматически заменены необходимыми конкретными данными с помощью специальных библиотек по генерации документов из шаблонов. Недостатком этого метода можно считать необходимое написание программного продукта с подключением соответствующей библиотеки. Такой продукт должен уметь считать информацию из входного файла и с помощью библиотеки заменить ей соответствующие переменные. К недостаткам также следует отнести и тот факт, что при некорректной работе соответствующей библиотеки конечный документ будет так же некорректно сформирован (существует зависимость от разработчиков стороннего ПО) [3]. Данный подход является наиболее практичным для генерации документов с большими объемами информации, которые должны быть оформлены специальным образом, поскольку:

1. Поддерживает свойство *WYSIWYG* (*What you see is what you get*) [2].
2. Не требует написания *tex*- или *html*-представлений.
3. Конечный документ имеет редактируемый формат *DOCX*.

В работе были рассмотрены некоторые общие подходы к автоматизированной генерации документов, а также выделены достоинства и недостатки каждого из них.

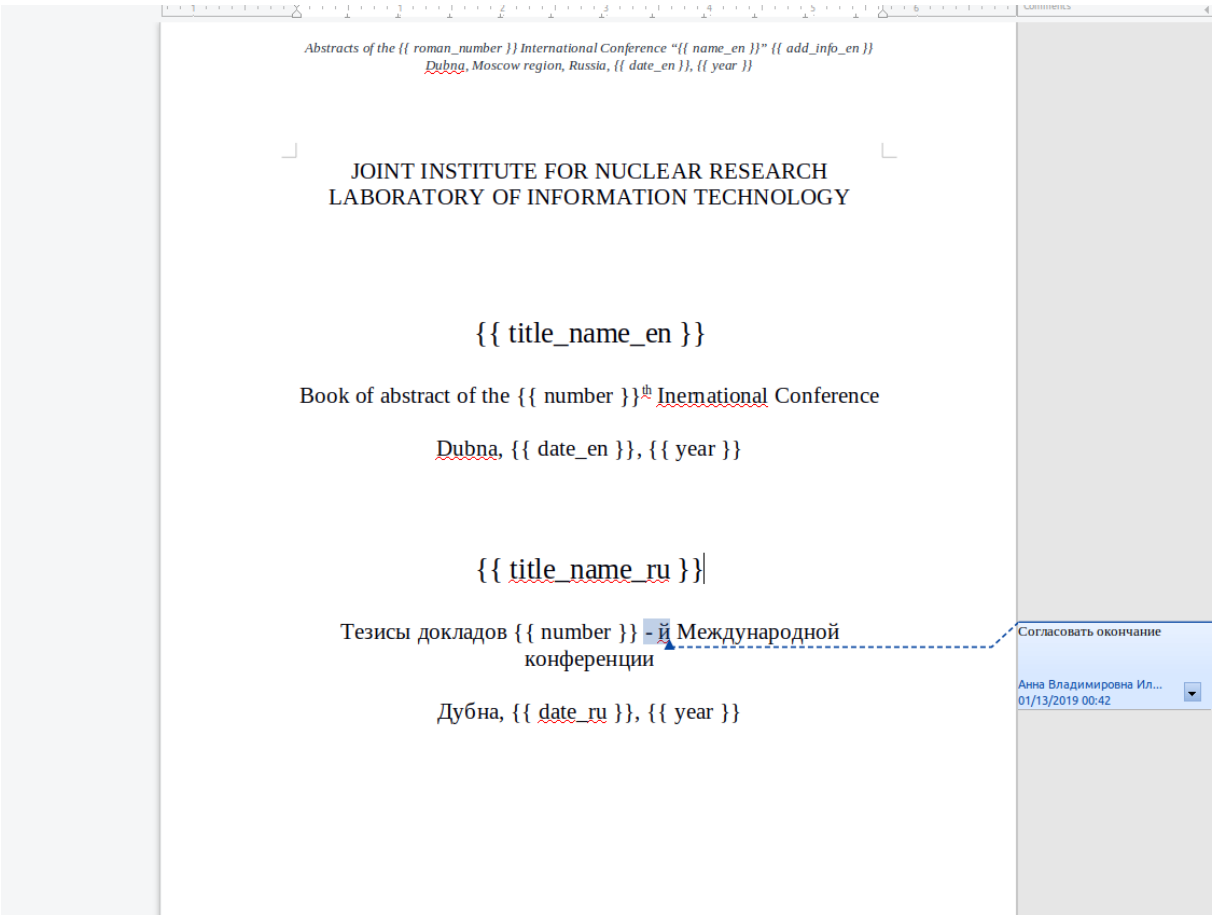


Рис. 5. Пример шаблона для генерации документа

Заключение

Наиболее практичным подходом для генерации книги абстрактов выбран метод генерации документа на основе шаблона формата *DOCX* с написанием соответствующего ПП, поскольку:

1. Позволяет достичь поставленной цели.
2. Не требует специальных навыков верстки и программирования для конечных пользователей.
3. Конечный документ имеет редактируемый формат *DOCX*, что гарантирует возможность правки информации в случае её изменения.
4. Документ формата *DOCX* легко экспортировать в документ формата *PDF* и тут же отправить в печать.

Список литературы

1. TeX – система для вёрстки текстов с формулами. – [Электронный ресурс]. URL: http://www.wikiznanie.ru/ru-wz/index.php/TeX_-_система_для_вёрстки_текстов_с_формулами.
2. WYSIWYG. – [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/WYSIWYG.
3. Автоматизируем генерацию docx документов по шаблону с помощью python и docxtpl. – [Электронный ресурс]. URL: http://tut-admin.ru/2017/05/31/автоматизируем-генерацию-docx-документо/.