УДК 004.942

Д.В.Косенко
Л.И.Воронова
В.И.Воронов
Москва, Россия

РАЗРАБОТКА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ ОБРАБОТКИ СЛОЖНОСТРУКТУРИРОВАННЫХ ДАННЫХ НАУЧНОГО ЭКСПЕРИМЕНТА

Аннотация. На сегодняшний день компьютерный эксперимент является наиболее значимым инструментом в областях, где имеется большой разрыв между возможностями теории и эксперимента. Сфера физической химии и металлургии является наиболее ярким примером из списка данных областей. Моделирование химических процессов, извлечение свойств и результатов, полученных в ходе компьютерного эксперимента, а также предоставление удаленного доступа к ним — основная цель проекта ИИС «MD_SLAGMELT».

Авторами статьи разработана подсистема, обеспечивающая конвертирование данных для базового Legacy Application ИИС «MD_SLAGMELT» из текстового в реляционный формат и автоматического переноса данных из файлового хранилища в реляционную базу данных.

В подсистеме реализованы следующие функциональные возможности:

– перенос термодинамических характеристик, энергетических параметров, кинетических коэффициентов компьютерного эксперимента в базу данных;
– проверка *.DAT файлов на наличие ошибок в генерации;
– перенос в базу данных результатов предыдущих экспериментов;
– формирование отчетности в формате *.xls;
– конфигурирование программы под текущие настройки базы данных.

Разработанное программное обеспечение позволяет переносить сложноструктурированные данные результатов компьютерных экспериментов в области физической химии из текстового в реляционный формат, строить представления для разных измерений гиперкуба свойств, формировать SQL-запросы, обеспечивающие выборки данных из нескольких таблиц базы данных ИИС «MD_SLAGMELT».

Рассмотрены существующие подходы считывания данных в текстовом формате, из которых выбран оптимальный, исходя из условий поставленной задачи, спецификации программного комплекса, особенности формата хранения результатов.

Разработана методика ведения отладочной информации в целях проверки целостности структуры итоговых результатов эксперимента.

В процессе внедрения программного обеспечения «Программа обработки сложноструктурированных данных для научного эксперимента в ИИС “MD_ SLAGMELT”» функциональность программы была протестирована на основном сервере проекта.

D.V.Kosenko
L.I.Voronova
V.I.Voronov
Moscow, Russia

DEVELOPING SOFTWARE FOR PROCESSING COMPLEX STRUCTURED DATA OF SCIENTIFIC EXPERIMENT

Abstract: Computer experiment is one of the most significant tools today, especially in areas with a large gap between the theory and experiment. The field of physical chemistry and metallurgy is the most striking example among such areas. The goals of IMS «MD_SLAGMELT» project are to construct chemical processes, to extract properties and results, obtained during a computer experiment, and to provide a remote access to such results.

The main purpose is the development of the methods of converting data from a text into a relational format during computer experiments, and the development of algorithms and tools they correspond for IMS «MD_SLAGMELT» project.

As a result of the project conduction the following functional tasks have been implemented:

– transferring of thermodynamic characteristics, energy parameters, kinetic coefficients of a computer experiment into the database;
– checking *.DAT files for generation errors;
– transferring the results of previous experiments into the database;
– developing a report in *.xls;
– forming a program for the current settings of the database.

The developed software allows transferring complex structured data results of computer experiments in physical chemistry from the text format to the relational one for IMS «MD_SLAGMELT». The existing approaches to reading data from text files were presented, the optimal of which was selected, based on the conditions of the task specification software system, especially the storage format of the results. In addition, some other techniques were presented: transfer characteristics approach in a relational database,and a technique of debugging information in order to verify an integrity of the structure of final results of an experiment.

During the application of “Processing program of complex-structured data for scientific experiment in IMS «MD_SLAGMELT»” software its functionality was tested on the main server of the project.

The obtained results provide a higher level of IMS «MD_SLAGMELT» project automation, intermediate inspections of the output files’ formation standards and reports in the specified range of structures for the given temperature (dependence of the structure on the property).

Полученные результаты обеспечивают повышение степени автоматизации проекта ИИС «MD_SLAG-MELT», промежуточной проверки стандартов формирования результирующих файлов, построение отчетов в заданном диапазоне составов при заданных температурах: исследование зависимости «состав — свойство».

Ключевые слова: компьютерный эксперимент; физическая химия; конвертирование данных; реляционные базы данных; ИИС «MD_SLAGMELT».

Key words: computer experiment; physical chemistry; data processing; reporting; IRS «MD_SLAGMELT».

Сведения об авторах: Дмитрий Владимирович Косенко[1], студент 4 курса факультета программной инженерии; Лилия Ивановна Воронова[2], заведующая кафедрой информационных систем и моделирования; Воронов Вячеслав Игоревич[3], доцент кафедры информационных систем и моделирования.

Место работы: [1] Национальный исследовательский университет «Высшая школа экономики»; [2, 3] Российский государственный гуманитарный университет.

About the authors: Dmitriy Vladimirovich Kosenko[1], 4th grade student at Program Engineering Faculty; Lilia Ivanovna Voronova[2], Head of the Department of Information Systems and modeling; Vyacheslav Igorevich Voronov[3], Associate Professor, Department of Information Systems and Modeling.

Place of employment: [1] National Research University «Higher School of Economics»; [2, 3] Russian State University for the Humanities.

Контактная информация: 111672, г. Москва, ул. Новокосинская, д. 40, кв. 204; тел.: 9160686838.
E-mail: voronova2001 @ mail.ru

Компьютерный эксперимент — это исследование математической модели объекта изучения на ЭВМ, состоящее в том, что по известным параметрам вычисляются искомые и на этой основе делаются выводы о свойствах объекта. Значение компьютерного эксперимента особенно велико в тех областях, где имеется большой разрыв между возможностями теории и эксперимента, к ним относятся физическая химия и металлургия. Как правило, в физической химии предметом исследования является взаимосвязь структурных характеристик и физико-химических свойств.

Для проведения КЭ создаются автоматизированные информационные системы (АИС), главной целью которых является расширение границы исследований, оптимизация научной работы и ускорение проведения исследований. Одной из таких систем является ИИС «Шлаковые расплавы» [6].

Преобразование сложноструктурированных данных, полученных в результате компьютерного эксперимента, в реляционный формат и обеспечение удаленного доступа к ним является одной из первоочередных задач в рамках проекта ИИС «Шлаковые расплавы».

Основной целью работы стала разработка методов конвертирования данных из текстового в реляционный формат при проведении компьютерных экспериментов, и реализующих их алгоритмов и инструментальных средств. Эта проблема связана с разработкой программного обеспечения для «Legacy application» (унаследованных приложений) ИИС «MD_SLAGMELT» [3], с переходом от локальных приложений, рассчитанных на моделирование нескольких тысяч частиц к системе с удаленным доступом, обеспечивающей компьютерный эксперимент для «больших данных» с количеством частиц порядка сотен тысяч, что серьезно осложняет задачу обработки и передачи данных между подсистемами.

ИИС позволяет вести моделирование в нескольких «*режимах*» с широким набором получаемых свойств:

– моделирование комплекса свойств определенного состава многокомпонентной системы вблизи выбранной температуры;

– моделирование многокомпонентной системы в заданном диапазоне составов при заданных температурах: исследование зависимости состав—свойство;

– моделирование состава по ряду температурных точек (плавление/затвердевание): исследование температурных зависимостей свойств;

– комплексное моделирование многокомпонентной системы (набор температурных зависимостей свойств состава для заданного диапазона составов): получение многомерных зависимостей состав—температура—свойство—структура.

Рис. 1. Архитектура ИИС «Шлаковые расплавы» [1—3]

Для исследования многомерных зависимостей состав—температура—свойство—структура разработана информационная модель оксидного расплава [2].

На рис. 1 приведена архитектура ИИС «MD_SLAGMELT», ядром которой является база данных для хранения результатов моделирования. Однако наполнение базы данных происходит после обработки файлового хранилища, куда в текстовых форматах записываются данные достаточно больших объемов.

Таким образом, предметом автоматизации является создание компонента «адаптер», предназначенного для преобразования сложноструктурированных данных, полученных в результате компьютерного эксперимента в реляционный формат.

Результаты проведенных экспериментов записываются в текстовом формате в набор файлов во внутреннем серверном файловом хранилище, структура которого формируется динамически в зависимости от входных данных. Для примера на рис. 2 приведена структура директории, сформированной после компьютерного эксперимента с системой SiO2-Na2O (0,5—0,5). Подобная директория создается для каждого из проведенных экспериментов, ее название и уровень вложенности формируется на основании названий элементов, входящих в название химической системы и мольных долей. Например, *SiONa → 0505 (мольные доли) → 806 (идентификационный номер математической модели)*

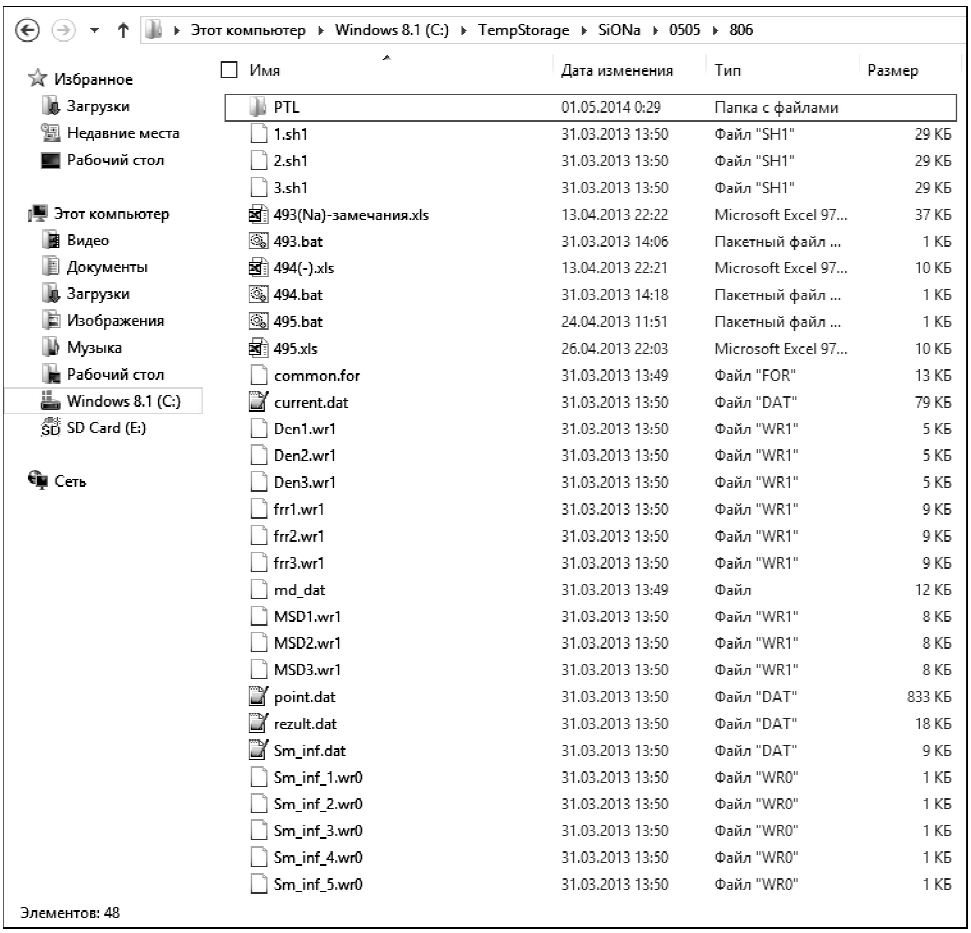


Рис. 2. Организация хранимых файлов по завершении проводимого эксперимента

Все результаты делятся на определенные группы. Они представляют собой документы формата «*.DAT». В них содержится информация о начальных условиях эксперимента, о каждой из групп характеристик (средних значений параметров, давления, длины связей и т.д.), полученных в результате исследования (рис. 3).

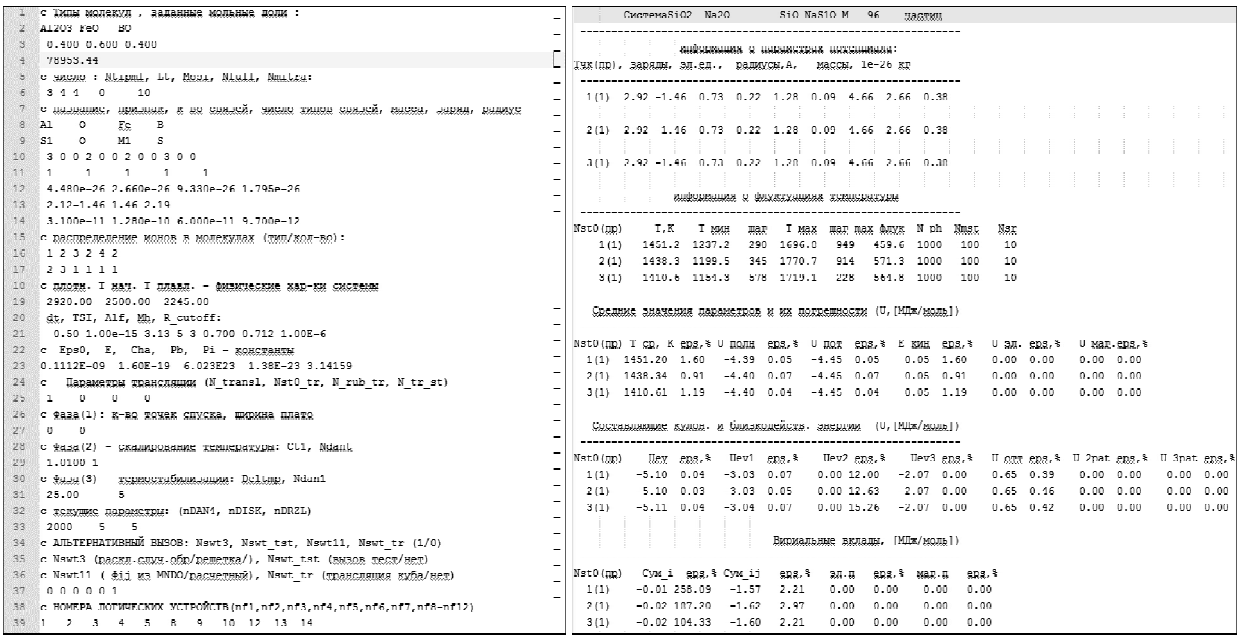


Рис. 3. Структура MD_DAT файла и SM_INF.DAT файла соответственно

Каждый из документов формата «*.DAT» представляет собой текстовый файл с позиционным расположением характеристик эксперимента. Между файлами существуют корреляции. Так, MD_DAT файл содержит информацию о количестве элементов, количестве запусков и температурных точках, что определяет размер файла SM_INF.DAT (рис. 3).

При разработке программного обеспечения было выбрано построчное считывание данных ввиду хранения результатов с разделителями «\t» и « » (рис. 4).

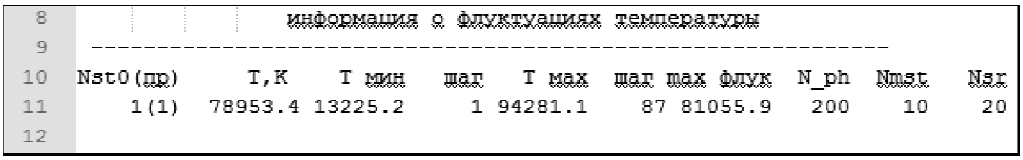


Рис. 4. Позиционное хранение на примере файла SM_INF.DAT

В целях распознавания элементов, разделяющихся одним или несколькими знаками « », была использована стандартная библиотека, входящая в состав Java Development Kit java.lang.Object String. В частности были использованы методы split() и разработанный метод класса StringUtils skip(), назначение которого заключается в пропуске «пустых» строк [7].

MD_DAT:

- Каждая строка определена для описания нижерасположенных свойств или самих свойств (рис. 4);
- Каждая строка содержит «заключительный» символ «_», позволяющий использовать его как точку останова при считывании массива данных (рис. 4);
- Содержит параметры, определяющие структуру файла SM_INF.DAT.

SM_INF.DAT:

- Количество данных зависит от количества температурных точек, указанных в MD_DAT;
- Каждая группа характеристик отделена одной из строк-разделителей (табл. 1, знач. 1—4);
- Каждый из значений эпсилонов результатов эксперимента может быть мал (10^{-7}), ввиду чего его значение не записывается в файл и имеет обозначение (табл. 1, знач. 5).

Таблица 1

Символьные обозначения [10]

№	Символьное обозначение	Кол-во в файле	Даты экспериментов
1	“----------“	1—2	От 07.2012
2	“ ----------”	16—20	От 07.2012
3	“ ----------“	1—2	От 07.2012
4	“ ---------”	1—4	До 07.2012
5	“******”	Зависит от кол-ва температурных точек	

При определении типа элементов были использованы стандартные классы JDK java.lang.Number: Integer, Float, Double (табл. 3).

Таблица 2

Числовые типы [4; 5; 8; 9]

№	Наименование	Разрядность	Диапазон значений
1	Integer	32	–2, 147, 483, 648 .... 2, 147, 483, 647
2	Float	32	3.4e-038 .... 3.4e+ 038
3	Double	62	1.7e-308 .... 1.7e+ 308

При попытке считывания очередного числового элемента при подаче нечислового значения (в случае, если файл был поврежден или числовое значение не было выведено) возникает обрабатываемое исключение NumberFormatException.

Для администратора проекта ИИС «MD_SLAGMELT» разработан вид логирования — ведения отладочной информации в целях проверки целостности структуры суммарных результатов эксперимента. После завершения выполнения программы, в случае возникновения исключительных ситуаций, администратор проекта, имеющий доступ к файловой системе, содержащей результаты проведенных экспериментов, имеет возможность ознакомиться с причиной, вызвавшей «некорректную» работу программы.

Завершение программы при возникновении исключительной ситуации ведет к построению и записи актуального информационного сообщения администратору в файле error.log.0 в директории с результатами связанного эксперимента.

Для удобства обеспечения отладки и поиска ошибок в формате сформированных результатов осуществляется хранение вплоть до пяти логов с соответствующей нумерацией *.0—*.4.

Представление формирования логов для файла SM_INF.DAT и запуска программы с аргументом “-makexls” — соответственно (рис. 5—6).

```
1  мая 25, 2014 4:18:16 PM Main main
2  SEVERE:
3  args{ "-insert" "E:/DataExamples/6/MD_DAT" "E:/DataExamples/6/SM_INF.dat" }
4  file: SM_INF.dat  row: 28  element: 2
5  Exception:
6  java.lang.NumberFormatException: For input string: "*********"
7      at sun.misc.FloatingDecimal.readJavaFormatString(Unknown Source)
8      at sun.misc.FloatingDecimal.parseFloat(Unknown Source)
9      at java.lang.Float.parseFloat(Unknown Source)
10     at model.MvweAtrbs.<init>(MvweAtrbs.java:58)
11     at controller.DatParser.Parser(DatParser.java:249)
12     at Main.main(Main.java:170)
```

Рис. 5. Сгенерированный лог для исключительной ситуации, возникшей при считывании SM_INF.DAT

```
1  мая 25, 2014 10:00:25 PM Main main
2  SEVERE:
3  args{ "-makexls" "2000" "700" "Na" }
4
5  Exception:
6  java.lang.Exception: Нет данных соответствующих запросу.
7      at Main.main(Main.java:271)
```

Рис. 6. Сгенерированный лог для исключительной ситуации, возникшей при формировании output.xls

При разработке логера была использована стандартная библиотека, входящая в состав Java Development Kit java.util.logging.Logger, некоторые из параметров управления логером могут быть изменены по желанию администратора в файле logging.propeties (рис. 7—8).

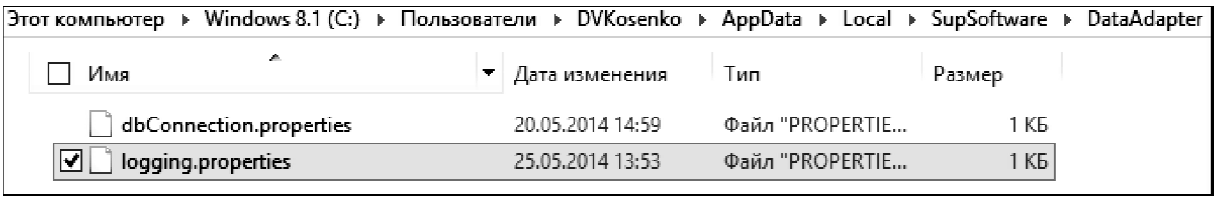
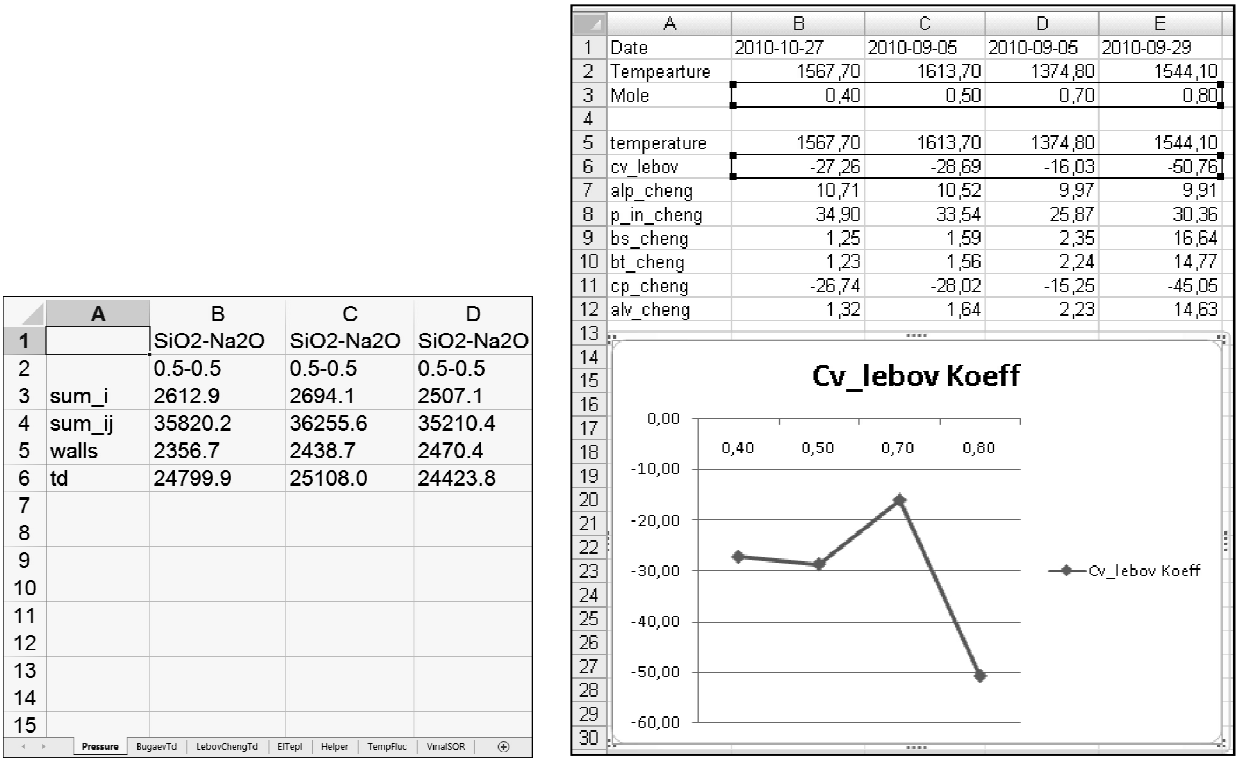


Рис. 7. Директория пути к файлу logging.properties

```
handlers = java.util.logging.FileHandler
java.util.logging.FileHandler.pattern = error.log
java.util.logging.FileHandler.limit = 1000000
java.util.logging.FileHandler.count = 5
java.util.logging.FileHandler.formatter = java.util.logging.SimpleFormatter
```

Рис. 8. Параметры логера в файле logging.properties

При запуске программы с входными параметрами {“-makexls” “filename.xls” “температурная_точка” “отклонение” “элемент”} происходит извлечение результатов экспериментов с температурными точками, находящимися в диапазоне от [температурная_точка – отклонение; температурная_точка + отклонение], содержащих “элемент”. Извлеченные из базы данных характеристики, соответствующие запросу, хранятся в оперативной памяти, после чего с помощью интерфейса прикладного программирования (API) jXLS формируется файл *.xls, содержащий искомые данные (рис. 9).



	A	B	C	D
1		SiO2-Na2O	SiO2-Na2O	SiO2-Na2O
2		0.5-0.5	0.5-0.5	0.5-0.5
3	sum_i	2612.9	2694.1	2507.1
4	sum_ij	35820.2	36255.6	35210.4
5	walls	2356.7	2438.7	2470.4
6	td	24799.9	25108.0	24423.8

	A	B	C	D	E
1	Date	2010-10-27	2010-09-05	2010-09-05	2010-09-29
2	Tempearture	1567,70	1613,70	1374,80	1544,10
3	Mole	0,40	0,50	0,70	0,80
4					
5	temperature	1567,70	1613,70	1374,80	1544,10
6	cv_lebov	-27,26	-28,69	-16,03	-50,76
7	alp_cheng	10,71	10,52	9,97	9,91
8	p_in_cheng	34,90	33,54	25,87	30,36
9	bs_cheng	1,25	1,59	2,35	16,64
10	bt_cheng	1,23	1,56	2,24	14,77
11	cp_cheng	-26,74	-28,02	-15,25	-45,05
12	alv_cheng	1,32	1,64	2,23	14,63

Рис. 9. Сформированный *.xls файл и практическая значимость отчетности

По каждой из групп характеристик, относящихся к определенному временному промежутку и пользователю, в сформированном программой «*.xls» файле оператор с помощью стороннего ПО может сформировать графики их изменения относительно мольных долей и температур.

Таким образом, решена задача переноса сложноструктурируемых данных, генерируемых Legacy application ИИС «MD_SLAGMELT» из набора текстовых файлов в реляционный формат. Реализовано программное обеспечение, позволяющее проводить необходимые преобразования данных и обеспечивающее запись в реляционную базу данных.

При этом разработана схема базы данных, архитектура программного обеспечения, обеспечен информационный обмен компонентами ИИС «MD_SLAGMELT», осуществлена реализация переноса всех групп выходных данных (термодинамических характеристик, энергетических параметров, кинетических коэффициентов) компьютерного эксперимента из текстовых файлов в реляционную базу данных; выстроена система проверки *.DAT файлов на наличие ошибок в генерации и проведено тестирование; разработана логика формирования отчетности в формате *.xls; осуществлено конфигурирование программы под текущие настройки базы данных.

В процессе внедрения программного обеспечения «Программа обработки сложноструктурированных данных для научного эксперимента в ИИС “MD_SLAGMELT”» функциональность программы была протестирована на основном сервере проекта.

ЛИТЕРАТУРА

1. Буч Г., Рамбо Д., Джекобсон А. UML. Руководство пользователя. М., 2001.
2. Воронова Л.И., Григорьева М.А., Воронов В.И., Трунов А.С. Программный комплекс «MD-SLAG-MELT» для моделирования наноструктуры и свойств многокомпонентных расплавов // Расплавы. 2013. № 4.
3. Воронова Л.И., Трунов А.С. Оптимизация параллельного алгоритма подсистемы распределенного молекулярно-динамического моделирования // Межотраслевая информационная служба. 2011. № 3.
4. Дейт К. Дж. Введение в системы баз данных. 8-е изд. М., 2005.
5. Дюбуа П. MySQL. Полное руководство. 3-е изд. М., 2006.
6. ИИС «MD-SLAG-MELT». URL: http://nano-md-simulation.com
7. Марка Д., МакГоуэн К. Методология структурного анализа и проектирования. М., 1993.
8. Моримото Р. Microsoft Windows Server 2012. Полное руководство / М.Ноэл, Г.Ярдени, О.Драуби, Э.Аббейт, К.Амарис. 2-е изд. М., 2013.
9. Рамбо Дж., Блаха М. UML 2.0. Объектно-ориентированное моделирование и разработка. 2-е изд. СПб., 2007.
10. Уорсли Дж. PostgreSQL. Для профессионалов. 3-е изд. СПб., 2003.

REFERENCES

1. Booch G., Rumbaugh J., Jacobson I. UML. User Guide. Moscow, 2001.
2. Voronov L.I., Grigoriev M.A., Voronov V.I., Trunov A.S. Software complex «MD-SLAGMELT» for modeling nanostructures and properties of multicomponent liquid alloys // Liquid Alloys. 2013. № 4.
3. Voronov L.I., Trunov A.S. Optimizing parallel algorithm of distributed molecular dynamics simulation subsystem // Interdisciplinary Information Service. 2011. № 3.
4. Date C.J. An Introduction to Database Systems. 8th Ed. Moscow, 2005.
5. DuBois P. MySQL. Complete Guide. 3rd Edition. Moscow, 2006.
6. IMS «MD-SLAG-MELT». URL: http://nano-md-simulation.com
7. Mark D., McGowan K. Methodology of structural analysis and design. Moscow, 1993.
8. Morimoto R. Microsoft Windows Server 2012. Complete Guide / M.Noel, G.Yardeni, O.Draubi, E.Abbeyt, K.Amaris. 2nd Edition. Moscow, 2013.
9. Rumbaugh J., Blaha M. UML 2.0. Object oriented modeling and design. 2nd Edition. St. Petersburg, 2007.
10. Worsley J. PostgreSQL. For professionals. 3rd edition. St. Petersburg, 2003.