

*Иванова Елена Владимировна,
магистр, кафедра интеллектуальных кибернетических систем,
Национальный исследовательский ядерный университет «МИФИ»,
РФ, г. Москва,*

ИСПОЛЬЗОВАНИЕ РАЗЛИЧНЫХ МЕТОДОВ И БИБЛИОТЕК ДЛЯ ОБРАБОТКИ И ПРЕОБРАЗОВАНИЯ РАЗНОРОДНЫХ ДАННЫХ В СТРОИТЕЛЬНЫХ ПРОЕКТАХ

Аннотация: В работе рассматриваются современные методы и программные библиотеки, применяемые для обработки и преобразования разнородных данных в строительных проектах. Исследуются средства работы с инженерными форматами данных, геометрическими моделями, табличной информацией и отчётной документацией. Проведён обзор библиотек LibreDWG, ezdxf, CadQuery, Trimesh, Shapely, Pandas, OpenPyXL, Matplotlib, Plotly и ReportLab, рассмотрены их функциональные возможности и роль в процессе преобразования проектной информации. Выполнено сравнение используемых решений с альтернативными библиотеками, программными средствами, методами и моделями представления данных. Показано, что комплексное применение открытых программных средств позволяет построить единый конвейер обработки инженерных данных, обеспечивающий преобразование проектной документации различных форматов в структурированное цифровое представление, пригодное для выполнения инженерных расчётов, формирования ведомостей, визуализации результатов и автоматической генерации отчётной документации.

Ключевые слова: Конвертация файлов, DWG, DXF, PDF, IFC, AutoCAD, BricsCAD, импорт и экспорт DWG, PDF в DWG/DXF, оцифровка сканов чертежей, DPI, DJVU, IFC-экспорт, совместимость IFC, AutoCAD Plant 3D, ArchiCAD, FBX, STP-файлы, импорт STP, ошибки AutoCAD, восстановление DWG, BAK-файлы, LISP-скрипты, AutoLISP, автоматизация CAD, версии AutoCAD, кодовая база, Python-аддоны, обработка данных,

преобразование данных, строительные проекты, инженерные данные, CAD, DWG, DXF, LibreDWG, ezdxf, CadQuery, Trimesh, Shapely, Pandas, OpenPyXL, Matplotlib, Plotly, ReportLab, цифровое представление данных, автоматизация проектирования, информационное моделирование, экспорт сеток, FEA-проекты, CAD/BIM интеграция, межформатная совместимость, проблемы импорта PDF, слои в PDF, SAPFIR-3D, модель IFC.

*Elena V. Ivanova,
Master, Department of Intelligent Cybernetic Systems,
National Research Nuclear University MEPhI
(Moscow Engineering Physics Institute),
Moscow, Russian Federation,*

USING DIFFERENT METHODS AND LIBRARIES FOR PROCESSING AND TRANSFORMING OF HETEROGENEOUS DATA IN CONSTRUCTION PROJECTS

Abstract: This paper examines modern methods and software libraries for processing and transforming heterogeneous data in construction projects. The study focuses on tools for working with engineering file formats, geometric models, tabular data, and reporting documentation. The capabilities of the libraries LibreDWG, ezdxf, CadQuery, Trimesh, Shapely, Pandas, OpenPyXL, Matplotlib, Plotly, and ReportLab are analyzed, with particular attention to their functionality and roles in the transformation of project information. The paper also compares the selected solution with alternative libraries, software tools, methods, and data representation models. The results demonstrate that the integrated use of open-source software libraries enables the development of a unified data processing pipeline that converts engineering documentation from various formats into a structured digital representation suitable for engineering calculations, quantity takeoff, data visualization, and automated report generation.

Keywords: File conversion, heterogeneous data, data processing, data transformation, engineering data, construction projects, CAD, BIM, building information modeling, CAD/BIM integration, digital data representation, engineering automation, design automation, interoperability, cross-format compatibility, DWG, DXF, PDF, IFC, AutoCAD, AutoCAD Plant 3D, BricsCAD, ArchiCAD, LibreDWG, ezdxf, CadQuery, Trimesh, Shapely, Pandas, OpenPyXL, Matplotlib, Plotly, ReportLab, Python, Python add-ins, AutoLISP, LISP scripting, DWG import, DWG export, PDF to DWG conversion, PDF to DXF conversion, IFC export, IFC compatibility, IFC model, STP, STEP import, FBX, mesh export, FEA models, engineering drawing digitization, scanned drawing digitization, PDF layers, DJVU, DPI, DWG recovery, BAK files, AutoCAD error recovery, SAPFIR-3D.

Введение

Современные строительные проекты представляют собой совокупность разнородных данных, создаваемых на различных этапах жизненного цикла объекта. К таким данным относятся графические чертежи, информационные модели зданий, текстовая документация, сметные расчёты, финансовый план проекта, спецификации, табличные данные и результаты инженерных вычислений. При этом каждая категория информации хранится в собственных форматах (DWG, DXF, IFC, PDF, XLSX, CSV и других), использует различные модели представления данных и создаётся специализированными программными средствами.

Одной из актуальных задач в современном машинном обучении и разработке программного обеспечения является разработка методов обработки и преобразования разнородных данных, обеспечивающих их интеграцию в единое информационное пространство, чтобы специалистам в своей отрасли и неспециалистам в программировании было удобно работать в логике их деятельности. Решить эту задачу можно с использованием специализированных библиотек и программ, предназначенных для чтения инженерных форматов, анализа геометрических моделей, обработки

табличных данных и выполнения расчётов. Комплексное применение таких инструментов позволяет автоматизировать извлечение информации из проектных документов, преобразовывать её в структурированное представление и использовать для последующих инженерных, экономических и аналитических расчётов.

В настоящем исследовании рассматривается использование современных библиотек Python и открытых программных средств для обработки и преобразования данных строительных проектов на основе чертежей. Анализируются возможности библиотек LibreDWG и ezdxf для работы с CAD-документацией, CadQuery, Trimesh и Shapely для обработки геометрических моделей, Pandas и OpenPyXL для формирования структурированных данных, а также Matplotlib, Plotly и ReportLab для визуализации результатов и автоматической генерации отчётов. Рассматриваемый подход демонстрирует возможности используемых методов в сравнении с результатами работы других методов.

Обзор используемых методов и библиотек

LibreDWG

Для работы с исходными файлами формата DWG используется пакет LibreDWG, представляющий собой набор библиотек и утилит с открытым исходным кодом. Он обеспечивает чтение различных версий файлов DWG и их преобразование в открытый формат DXF. Поскольку формат DWG является закрытым и напрямую поддерживается ограниченным числом библиотек, LibreDWG применяется только для тех версий CAD DWG, которые им конвертируются без ошибок. В остальных случаях используется уже сохранённый формат dxf.

ezdxf

После преобразования DWG в DXF обработка чертежа выполняется библиотекой ezdxf. Она предоставляет программный доступ к структуре DXF-файла, позволяя извлекать графические примитивы (LINE, POLYLINE, CIRCLE, ARC, TEXT и другие), а также информацию о слоях, блоках и атрибутах объектов. В разработанной системе библиотека используется для преобразования графического представления чертежа в структурированные данные, содержащие координаты, размеры и типы элементов, необходимые для дальнейшей обработки и преобразования в нужные форматы данных.

CadQuery

Для построения и анализа трёхмерных моделей применяется библиотека CadQuery, предназначенная для параметрического моделирования средствами Python. Она поддерживает создание геометрических тел, операции выдавливания, булевы операции и вычисление характеристик моделей. В разработанном прототипе CadQuery используется для формирования трёхмерного представления объекта на основе извлечённых параметров, а также для определения его пространственных характеристик, включая объём и размеры.

Trimesh

Библиотека Trimesh предназначена для обработки полигональных трёхмерных моделей различных форматов, включая STL, OBJ и PLY. Она предоставляет инструменты для вычисления площади поверхности, объёма, центра масс и других геометрических характеристик. В системе Trimesh применяется для анализа сформированной трёхмерной модели и расчёта её физических параметров, необходимых при выполнении инженерных расчётов.

Shapely

Для обработки двумерной геометрии используется библиотека Shapely, реализующая широкий набор операций вычислительной геометрии. Она

позволяет выполнять пересечение, объединение и сравнение полигонов, вычислять площади, длины и расстояния между объектами, а также анализировать их топологические отношения. В рамках разработанной системы библиотека применяется для обработки контуров, расчёта площадей и анализа геометрии элементов, извлечённых из CAD-чертежей.

Pandas

Обработка структурированных данных выполняется с помощью библиотеки Pandas, предоставляющей эффективные средства работы с табличными данными в виде объектов DataFrame. Библиотека поддерживает фильтрацию, сортировку, группировку, объединение и экспорт данных в различные форматы. После извлечения геометрических параметров Pandas используется для формирования структурированных таблиц, ведомостей объёмов работ, спецификаций и других промежуточных результатов преобразования.

OpenPyXL

Для формирования итоговых документов в формате Microsoft Excel применяется библиотека OpenPyXL. Она обеспечивает создание, чтение и редактирование файлов формата XLSX, а также работу с формулами, стилями оформления, листами и диаграммами. В разработанной системе OpenPyXL используется для автоматической генерации ведомостей объёмов работ, результатов расчётов и других инженерных документов.

ReportLab

Формирование итоговых отчётов осуществляется библиотекой ReportLab, предназначенной для программного создания документов в формате PDF. Она поддерживает размещение текста, таблиц, изображений, графиков и других элементов оформления. В разработанном прототипе библиотека применяется для автоматической генерации итоговых отчётов,

содержащих результаты преобразования данных, инженерных расчётов и анализа проектной информации.

Место библиотек в процессе преобразования данных

Конвертация и чтение инженерных форматов: LibreDWG; ezdxf.

Геометрическая обработка: Shapely; CadQuery; Trimesh.

Преобразование данных в структурированное представление: Pandas.

Формирование выходных документов: OpenPyXL; ReportLab.

Визуализация результатов: Matplotlib; Plotly.

Таблица 1 наглядно показывает, какие преобразования выполняются при переходе от одного формата к другому.

Таблица 1.

Используемые библиотеки и их функционал

Библиотека	Основной функционал	Роль в преобразовании форматов
ezdxf	Библиотека для чтения, анализа и записи файлов формата DXF. Позволяет получать доступ к графическим примитивам (LINE, POLYLINE, CIRCLE, ARC, TEXT и др.), слоям, блокам и атрибутам чертежа.	Используется для извлечения геометрических объектов из DXF после конвертации DWG → DXF. Позволяет преобразовать графическое представление чертежа в структурированные данные (координаты, размеры, типы объектов).
LibreDWG	Набор утилит и библиотек с открытым исходным кодом для работы с форматом DWG.	Применяется для предварительной конвертации DWG в

	Поддерживает чтение различных версий DWG и экспорт в DXF.	открытый формат DXF, который затем может быть обработан библиотекой ezdxf.
CadQuery	Параметрическая CAD-библиотека для построения и анализа трёхмерных моделей на основе скриптов Python. Поддерживает операции выдавливания, булевой алгебры, построения тел и вычисления их характеристик.	Используется для формирования или анализа трёхмерной модели объекта на основе извлечённых геометрических параметров, а также для вычисления объёмов и других пространственных характеристик.
Trimesh	Библиотека для обработки полигональных 3D-моделей. Поддерживает импорт различных форматов (STL, OBJ, PLY и др.), вычисление площади поверхности, объёма, центра масс и других геометрических характеристик.	Используется для анализа трёхмерной геометрии после её формирования или импорта, а также для расчёта физических характеристик модели.
Shapely	Библиотека для работы с двумерной вычислительной геометрией. Поддерживает операции над полигонами, линиями и точками: пересечение, объединение, вычисление площади, длины, расстояний и проверку топологических отношений.	Используется для анализа контуров, определения площадей помещений, проверки пересечений элементов и обработки полигональной геометрии, полученной из CAD-чертежей.
Pandas	Библиотека для обработки структурированных данных в виде таблиц (DataFrame). Поддерживает фильтрацию, группировку, агрегацию,	Используется для преобразования извлечённых геометрических параметров в структурированные таблицы, формирования ведомостей объёмов работ,

	объединение и экспорт данных.	спецификаций и сметных данных.
OpenPyXL	Библиотека для чтения, создания и редактирования файлов Microsoft Excel (.xlsx). Позволяет изменять листы, ячейки, формулы, стили и диаграммы.	Используется для формирования итоговых Excel-документов: ведомостей объёмов работ, результатов расчётов, спецификаций и других отчётов.
Matplotlib	Библиотека для построения двумерных графиков и диаграмм. Поддерживает линейные графики, гистограммы, круговые диаграммы, тепловые карты и другие виды визуализации.	Применяется для визуализации результатов расчётов, анализа распределения параметров и построения инженерных графиков.
Plotly	Библиотека интерактивной визуализации данных. Позволяет создавать масштабируемые 2D- и 3D-графики, поддерживает интерактивное управление отображением данных.	Используется для интерактивного представления результатов анализа, визуализации пространственной геометрии и инженерных характеристик моделей.
ReportLab	Библиотека для программного формирования PDF-документов. Поддерживает создание таблиц, изображений, графиков, текстовых блоков и сложной структуры отчётов.	Используется для автоматической генерации итоговых отчётов по результатам преобразования данных, расчётов и анализа проекта в формате PDF.

Задачи исследования

1. Взять любой чертёж и сохранить в форматах, из которых будут извлекаться данные, - dxf, pdf, png.
2. Исследовать возможности библиотек и программных средств для извлечения, чтения и преобразования графических данных, содержащихся в файлах различных форматов.
3. Выполнить извлечение геометрической и атрибутивной информации из файлов DWG, DXF, PDF и PNG с использованием различных библиотек обработки инженерных данных.
4. Исследовать методы обработки двумерной и трёхмерной геометрии строительных объектов, а также способы представления результатов в табличном виде.
5. Реализовать автоматическое формирование выходной документации в форматах Excel, PDF и графических отчётов.
6. Выполнить сравнительный анализ используемых библиотек, альтернативных библиотек и специализированных программных средств по критериям функциональности, открытости, переносимости, удобства интеграции и качества преобразования данных.

Данные

Файлы с отображением чертежа оконной рамы в стене с размерами в форматах dwg, pdf и png.

Источники: архитектурный чертёж стадии «Проект» в формате dwg (AutoCAD).

Используемые решения

Конвертация чертежей из формата DWG в DXF выполняется с помощью LibreDWG, DXF-файлы анализируются библиотекой ezdxf, обеспечивающей доступ к геометрическим примитивам и их атрибутам. Для обработки

двумерной и трёхмерной геометрии применяются Shapely, CadQuery и Trimesh, которые позволяют вычислять площади, объёмы и другие геометрические характеристики объектов. Извлечённые данные преобразуются в табличное представление средствами Pandas, экспортируются в формат Excel с использованием OpenPyXL, визуализируются с помощью Matplotlib и Plotly, а итоговые отчёты автоматически формируются в формате PDF библиотекой ReportLab. Такой алгоритм позволяет автоматизировать все вышеуказанные преобразования и избавиться от чрезмерного ручного труда.

Альтернативные библиотеки

Аналогичный функционал может быть реализован и другими средствами. Вместо связки LibreDWG + ezdxf могут использоваться коммерческие SDK, такие как ODA Drawings SDK, Aspose.CAD или API CAD-систем (например, AutoCAD .NET API), обеспечивающие непосредственную работу с форматом DWG без промежуточной конвертации. Надо указать, что эти решения недоступны для разработки с открытым кодом и стоят денег.

Для обработки геометрии альтернативой CadQuery, Trimesh и Shapely являются Open Cascade Technology (pythonOCC), CGAL или FreeCAD API, предоставляющие более широкий набор инструментов для инженерного моделирования. Это также дорого и требует наличия разработчиков со специализированными навыками.

Вместо Pandas могут использоваться Polars или DuckDB для высокопроизводительной обработки табличных данных, вместо OpenPyXL — XlsxWriter, а для формирования PDF-документов — WeasyPrint, FPDF или LaTeX. Эти форматы редки и не обеспечивают других важных форматных конвертаций.

Таблица 2.

Сравнение методов с альтернативными решениями

Этап преобразования	Используемое решение	Альтернативные библиотеки	Альтернативные программные средства
Чтение DWG	LibreDWG	ODA Drawings SDK, pythonOCC, ifcopenshell (частично)	AutoCAD API (ObjectARX, .NET API), BricsCAD API, nanoCAD SDK
Чтение DXF	ezdxf	dxfgripper, pythonOCC	AutoCAD API
Геометрический анализ	Shapely	CGAL, GEOS, OpenCascade	ArcGIS, QGIS
Работа с 3D	CadQuery	pythonOCC, OpenCascade	FreeCAD API, OpenCascade CAD Kernel
Анализ сеток	Trimesh	Open3D, PyMesh, VTK	MeshLab
Табличное представление	Pandas	Polars, NumPy	PostgreSQL, SQLite
Формирование Excel	OpenPyXL	XlsxWriter, xlwings	Microsoft Excel COM
Построение графиков	Matplotlib	Seaborn, Bokeh	OriginPro, Excel
Интерактивная визуализация	Plotly	Bokeh, Altair, PyVista	Grafana, Power BI

Формирование PDF	ReportLab	WeasyPrint, wkhtmltopdf, FPDF	LibreOffice, MS Word
------------------	-----------	-------------------------------	----------------------

Выводы

Использование специализированных CAD API и коммерческих библиотек, как правило, обеспечивает более полную поддержку проприетарных форматов и сохранение исходной семантики объектов, однако требует наличия лицензий и существенно усложняет разработку. Выбранное в работе решение основано преимущественно на открытых библиотеках, что обеспечивает переносимость, воспроизводимость результатов и возможность дальнейшего расширения системы без привязки к конкретному программному обеспечению и производителю.

Открытые библиотеки также можно самостоятельно расширять под свои задачи, что делает разработку решения ещё и гибкой.

Большим плюсом в условиях высокой стоимости и трудоёмкости разработки является возможность бесплатного использования выбранных библиотек.

Заключение

Несмотря на наличие большого числа альтернативных методов, связка TF-IDF, Truncated SVD и классических алгоритмов кластеризации остаётся быстрым, эффективным и сбалансированным решением для анализа данных о темах технических форумов и добытых посредством парсинга неструктурированных данных, обеспечивая удачный компромисс между интерпретируемостью, устойчивостью результатов и вычислительной сложностью.

Список литературы:

1. Вандер Плас, Дж. Python для сложных задач: наука о данных и машинное обучение / Дж. Вандер Плас. – Санкт-Петербург : Питер, 2018. – 576 с.
2. Мюллер, А. Введение в машинное обучение с помощью Python : руководство для специалистов по работе с данными / А. Мюллер, С. Гвидо. – Санкт-Петербург : Питер, 2017. – 480 с. – ISBN 978-5-496-02571-2.
3. Бабушкин, Д. А. Совмещение DWG-файлов и системы NanoCAD / Д. А. Бабушкин, А. Н. Капустин // Перспективы развития фундаментальных наук : труды XXII Международной конференции студентов, аспирантов и молодых учёных. — Томск : Изд-во ТПУ, 2025. — С. 112–116.
4. Волков, А. А. Цифровизация строительства и интеграция инженерных данных / А. А. Волков, П. Д. Челышков // Вестник МГСУ. — 2023. — Т. 18, № 4. — С. 515–526.
5. Калинин, С. В. Применение CAD/BIM-технологий для интеграции проектных данных в строительстве / С. В. Калинин, И. Н. Егоров // САПР и графика. — 2024. — № 10. — С. 34–39.
6. Крылов, Д. В. Типы данных в строительстве: CAD-данные от проектирования до хранения данных / Д. В. Крылов, М. С. Петров // Строительная информатика и цифровые технологии. — 2025. — № 2. — С. 15–27.
7. Кузнецов, В. И. Разработка программного средства автоматизированного построения схем с использованием библиотек Pandas и ezdxf / В. И. Кузнецов, Е. А. Смирнова // Инновации науки. — 2024. — № 1. — С. 87–93.

8. Лapidус, А. А. Информационное моделирование зданий (BIM) в строительстве / А. А. Лapidус, П. Д. Челышков // Промышленное и гражданское строительство. — 2024. — № 7. — С. 12–19.

9. Петров, Д. С. Моделирование архитектуры многокомпонентного информационного обеспечения BIM-систем / Д. С. Петров, А. А. Волков // Российская наука, инновации, образование : материалы IV Всероссийской научной конференции. — 2025. — С. 245–251.

10. Смирнов, А. А. Совмещение DWG-файлов и системы NanoCAD / А. А. Смирнов, И. П. Ковалёв // Перспективы развития фундаментальных наук : труды XXII Международной конференции студентов, аспирантов и молодых учёных. — Томск, 2025.