

СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ/SYSTEM ANALYSIS, MANAGEMENT AND PROCESSING OF INFORMATION

DOI: https://doi.org/10.60797/itech.2026.10.1 EDN: PTJNVY

ОБЗОР ДАТАСЕТА TAWOS ДЛЯ АНАЛИЗА ТРУДОЗАТРАТ И ПРОЦЕССОВ РАЗРАБОТКИ ПО

Обзор

Цыварев И.В.[1, *]

[1] Санкт-Петербургский государственный университет телекоммуникаций имени профессора М. А. Бонч-Бруевича, Санкт-Петербург, Российская Федерация

* Корреспондирующий автор (cyvarev.ilya156[at]gmail.com)

Аннотация

Настоящая статья представляет собой детальный обзор открытого набора данных TAWOS, предназначенного для анализа трудозатрат и процессов разработки программного обеспечения. Дается комплексное описание структуры и содержания датасета, который агрегирует информацию из 12 открытых репозиториев систем управления проектами и включает данные о 458 тысячах задач, полученных из 39 проектов с открытым исходным кодом от таких организаций, как Apache, MongoDB, Atlassian и других.

Основное внимание уделяется реляционной структуре набора данных. Подробно описываются ключевые сущности, такие как Issue (задачи с их атрибутами), Comment (комментарии к задачам), Change_Log (история изменений), Component (компоненты ПО), Version (версии), Sprint (спринты) и другие, а также связи между ними. Отмечается, что датасет предоставляет многогранные данные, охватывающие метаданные задач, временные затраты, активность участников и ход итераций.

В статье также рассматриваются потенциальные сценарии практического применения TAWOS, включая прогнозирование трудозатрат и сроков выполнения задач, анализ производительности команд и выявление закономерностей в процессах разработки. При этом подчеркиваются существующие ограничения набора данных, такие как его формирование исключительно на основе публичных проектов, неполнота и вариативность метаданных.

Ключевые слова: датасеты, набор данных, Jira, Agile, оценка трудозатрат.

AN OVERVIEW OF THE TAWOS DATASET FOR ANALYSING LABOUR COSTS AND SOFTWARE DEVELOPMENT PROCESSES

Review article

Tsyvarev I.V.[1, *]

[1] St. Petersburg State University of Telecommunications named after Professor M. A. Bonch-Bruevich, Saint-Petersburg, Russian Federation

* Corresponding author (cyvarev.ilya156[at]gmail.com)

Abstract

This article provides a detailed overview of the TAWOS open dataset, designed for analysing labour costs and software development processes. It provides a complex description of the structure and content of the dataset, which aggregates information from 12 open project management repositories and includes data on 458,000 tasks obtained from 39 open-source projects from organisations such as Apache, MongoDB, Atlassian and others.

The main focus is on the relational structure of the dataset. Key entities such as Issue (tasks and their attributes), Comment (comments on tasks), Change_Log (change history), Component (software components), Version, Sprint and others, as well as the relationships between them. It is noted that the dataset provides multifaceted data covering task metadata, time spent, participant activity and the progress of iterations.

The paper also examines potential scenarios for the practical application of TAWOS, including forecasting working hours and task completion times, analysing team productivity, and identifying patterns in development processes. It highlights existing limitations of the dataset, such as the fact that it is based exclusively on public projects, as well as the incompleteness and variability of the metadata.

Keywords: datasets, Jira, Agile, effort estimation.

Введение

Современная разработка программного обеспечения характеризуется высокой динамичностью процессов, разнообразием используемых методологий и возрастанием роли данных в управлении проектами. Для повышения эффективности планирования, оценки трудозатрат и анализа производительности команд все чаще применяются подходы, основанные на анализе реальных проектных данных. В этой связи особую ценность приобретают открытые датасеты, отражающие реальные процессы разработки, взаимодействие участников и изменение состояния задач в ходе жизненного цикла проекта.

Одним из таких источников является открытый набор данных TAWOS. Он объединяет данные о задачах, спринтах, коммитах и активности участников из систем управления проектами Jira, предоставляя исследователям возможность изучать закономерности распределения трудозатрат, прогнозировать сроки выполнения задач и оценивать эффективность командных процессов. Настоящая статья посвящена анализу структуры и возможностей применения

датасета TAWOS, а также рассмотрению его ограничений и научной значимости для исследований в области программной инженерии [1], [2], [3].

Методы и принципы исследования

В рамках данного обзора использовался метод структурного и сравнительного анализа открытых датасетов программной инженерии. Исследование включает:

- анализ схемы данных и реляционных связей датасета TAWOS;
- классификацию сущностей и их атрибутов с точки зрения задач анализа трудозатрат;
- сопоставление TAWOS с другими широко используемыми наборами данных;
- обобщение опубликованных и типовых практик использования аналогичных датасетов в прикладных и исследовательских проектах.

TAWOS рассматривается как база данных, ориентированная на анализ Agile-процессов. Основное внимание уделяется его применимости для количественных исследований: прогнозирования трудозатрат, анализа производительности команд и изучения динамики разработки.

Основные результаты

Датасет TAWOS основан на данных систем учета времени, использованных при разработке следующими командами и компаниями:

- Apache;
- Appcelerator;
- Atlassian;
- DNN Tracker;
- Hyperledger;
- Lsstcorp;
- Lyrasis;
- MongoDB;
- Moodle;
- Mulesoft;
- Sonatype;
- Spring.

Датасет включает в себя данные о 458 тысячах задач из 39 проектов с открытым исходным кодом, полученных из 12 открытых репозиториев с данными системы учета времени Jira. Набор данных TAWOS представлен в виде реляционной базы данных, которую можно загрузить и установить в систему управления базами данных MySQL.

На рисунке 1 представлена диаграмма сущностей датасета и отношений между ними.

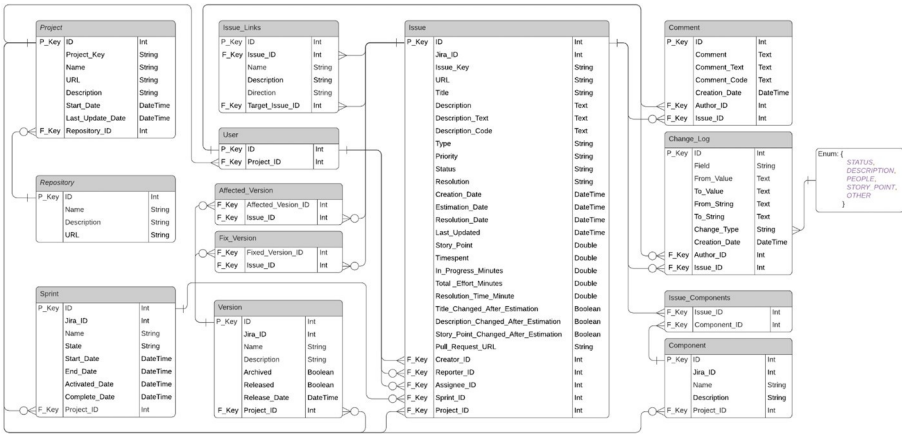

Рисунок 1 - Диаграмма сущностей датасета, их полей и отношений между ними

DOI: https://doi.org/10.60797/itech.2026.10.1.1

В таблице 1 перечислены сущности, составляющие датасет, и их назначения.

Таблица 1 - Сущности, составляющие датасет TAWOS

DOI: https://doi.org/10.60797/itech.2026.10.1.2

Название сущности	Назначение
Issue	Основная сущность, хранящая различные извлеченные, производные и вычисляемые характеристики каждой задачи.
Comment	Содержит каждый комментарий, написанный к

Название сущности	Назначение
	задаче, включая время создания и идентификатор автора комментария. Персональные данные заменены тегами.
Change_Log	Хронологически упорядоченные изменения атрибутов задач. Каждая запись содержит предыдущее и новое значение атрибута.
Issue_Components	Промежуточная таблица, связывающая задачи и компоненты (связь многие-ко-многим).
Component	Хранит информацию о компонентах, из которых состоит каждый программный продукт.
Issue_Links	Содержит связи между задачами, которые указывают на их взаимосвязь (например, дублирование, зависимость или блокировка).
User	Содержит уникальных пользователей, которые взаимодействовали с проектами в наборе данных. Идентификатор пользователя сгенерирован БД и не связан с ID в исходном репозитории.
Affected_Version	Промежуточная таблица, связывающая задачи с версиями, в которых была обнаружена ошибка или проблема.
Fix_Version	Промежуточная таблица, связывающая задачи с версиями, в которых функция была исправлена какая-либо ошибка.
Version	Хранит информацию о версиях разрабатываемого ПО (имя, описание, дата выпуска).
Project	Хранит информацию о проектах, включенных в базу данных.
Repository	Хранит информацию о репозиториях, включенных в базу данных.
Sprint	Хранит информацию о спринтах (итерациях) в процессе разработки, включая состояние, даты начала и окончания.

Стоит отметить, что значения категориальных полей сущности Issue могут варьироваться от проекта к проекту. Это означает, что для обозначения статусов, типов и категорий в одном проекте может использоваться один набор значений, а в другом проекте — иной.

Для лучшего понимания исследовательской ценности TAWOS целесообразно рассмотреть его в сравнении с другими известными датасетами:

1. PROMISE Repository — один из наиболее известных репозиториев датасетов программной инженерии. Содержит данные о дефектах, метриках кода и проектах, но практически не включает детальную информацию о процессах Agile, спринтах и трудозатратах [4].

2. GHTorrent — крупный датасет, агрегирующий данные GitHub (репозитории, issues, pull requests). Обладает масштабом, но не содержит структурированной информации о спринтах, оценках трудозатрат и Agile-итерациях. Поддержка остановлена в 2019 году [5].

Во многих работах TAWOS используется в качестве основного источника эмпирических данных для анализа Agile-проектов и разработки моделей машинного обучения. Например, в репозитории проекта перечислены исследования, опирающиеся на данные TAWOS, такие как:

1. Анализ эффективности использования story points для оценки трудозатрат, где использовались данные о задачах из TAWOS для изучения соответствия оценок фактическим трудозатратам в Agile-разработке [6].

2. Исследование методов кластеризации для оценки усилий по задачам, в котором TAWOS служил исходным набором задач для обучения и сравнения методов группировки [7].

3. Эксперименты с глубоким обучением для оценки усилий, в частности в работе по изучению эффективности Deep-SE и его репликации с использованием TAWOS (31 960 задач), где датасет использовался для оценки точности предсказаний ML-моделей внутри- и между-проектными сценариями [8].

Помимо академических исследований, TAWOS применяется и в современных задачах машинного обучения. Так, репликационный пакет исследования “Impact of Request Formats on Effort Estimation: Are LLMs Different than Humans?” включает код, который извлекает данные из MySQL-базы TAWOS (user stories и описания задач) и использует их для сравнения моделей оценки усилий, интегрируя их с LLM-моделями (GPT, Gemini, LLAMA) для генерации предсказаний трудозатрат [9].

Другое исследование, опубликованное в журнале Knowledge-Based Systems, использует TAWOS для классификации уровней серьёзности багов с помощью ансамблевых ML и NLP-методов, включая XGBoost, LightGBM и CatBoost. В этом случае данные TAWOS использовались для извлечения текстовых и структурных признаков из задач и комментариев с целью улучшения точности автоматической классификации [10].

Все эти примеры показывают, что TAWOS не ограничивается только описательной статистикой, но служит основой для обучения моделей, сравнительного анализа методов оценки усилий, автоматизации классификационных задач и разработки репликационных пакетов с реальными данными. Такая практика демонстрирует перспективность TAWOS как общего стандартизированного источника данных для анализа жизненного цикла Agile-проектов и применения методов Data-Driven Software Engineering.

Обсуждение

Благодаря многоаспектному охвату данных — от метаданных задач и времени их исполнения до активности участников и хода спринтов — TAWOS может использоваться для анализа производительности команд разработки. Сопоставление трудозатрат с параметрами задач позволяет выявлять закономерности в распределении нагрузки, оценивать эффективность распределения ролей и прогнозировать возможные отклонения от плановых сроков. Кроме того, наличие информации о связях между задачами и коммитами открывает возможности для анализа соответствия плановых и фактических результатов разработки.

Одним из перспективных направлений применения датасета является построение и обучение моделей прогнозирования сроков и трудозатрат. Используя текстовые описания задач, историю спринтов и метрики участников, можно создавать алгоритмы машинного обучения для автоматической оценки сложности задач и предсказания длительности их выполнения. Это особенно актуально в контексте современных подходов к оценке трудозатрат, где используются большие языковые модели (LLM) [11], [12], [13].

TAWOS также предоставляет основу для анализа качества процессов разработки и зрелости команд. Сопоставление данных об ошибках, комментариях и кодовых изменениях позволяет оценивать качество коммуникаций, выявлять «узкие места» в процессах ревью и интеграции, а также исследовать факторы, влияющие на устойчивость и успех проектов. Такие исследования могут способствовать построению моделей зрелости процессов и выявлению зависимостей между организационной культурой и эффективностью выполнения задач [14], [15].

Наконец, датасет может использоваться для изучения факторов успешности программных проектов в целом. Сравнение метрик различных команд и проектов — например, объёмов задач, средней продолжительности их выполнения и частоты релизов — позволяет выявлять ключевые детерминанты успешности и устойчивости процессов. Это делает TAWOS универсальной исследовательской платформой, обеспечивающей как количественный, так и качественный анализ разработки программного обеспечения [16], [17].

Несмотря на широкий спектр возможностей анализа, датасет TAWOS обладает рядом ограничений, которые необходимо учитывать. Прежде всего, он формируется на основе данных из публичных репозиториев и проектных трекеров, что может приводить к неполному охвату типов проектов и практик разработки. Такая выборка ограничивает репрезентативность данных и снижает возможность обобщения результатов на другие контексты и методологии.

Также существенным ограничением является неполнота и вариативность метаданных. В ряде случаев отсутствуют оценки трудозатрат, сроки или комментарии, что усложняет построение корректных выборок и требует дополнительной очистки данных. Кроме того, логи активности отражают лишь формализованные действия пользователей, не учитывая неформальные взаимодействия или внешние инструменты, используемые в процессе работы.

Значительную сложность создаёт и разнородность форматов данных. Поскольку информация собрана из различных проектов, использующих разные подходы к ведению трекеров, структура и семантика полей могут отличаться. Это осложняет сопоставление данных между проектами и требует нормализации, что, в свою очередь, может привести к потере контекстной информации.

Наконец, необходимо учитывать ограниченность интерпретации данных: длительность выполнения задачи или объём активности разработчиков не всегда напрямую отражают сложность работы или эффективность команды.

Заключение

Датасет TAWOS представляет собой значимый ресурс в области программной инженерии, предлагая структурированные и взаимосвязанные данные о задачах, спринтах, коммитах и активности разработчиков. Его комплексная структура позволяет исследовать широкий спектр вопросов — от оценки трудозатрат и прогнозирования сроков до анализа эффективности команд и зрелости процессов разработки.

Проведённый анализ показал, что TAWOS обладает высокой исследовательской ценностью благодаря охвату ключевых аспектов жизненного цикла проектов и возможности интеграции с методами машинного обучения и аналитики данных. Вместе с тем использование датасета требует внимательного подхода к предварительной обработке и интерпретации данных, учитывая их неполноту, вариативность и доменную специфику.

Таким образом, TAWOS может рассматриваться как основа для построения новых моделей прогнозирования трудозатрат, оценки командной производительности и выявления факторов успешности проектов. Его применение способствует развитию данных-ориентированных подходов в исследовании и управлении разработкой программного обеспечения, формируя базу для дальнейших научных и практических достижений в области инженерии ПО.

Конфликт интересов

Не указан.

Рецензия

Сообщество рецензентов журнала «Cifra. Информационные технологии и телекоммуникации».
DOI: https://doi.org/10.60797/itech.2026.10.1.3

Conflict of Interest

None declared.

Review

Community of Reviewers of the "Cifra. Information technology and telecommunications".
DOI: https://doi.org/10.60797/itech.2026.10.1.3

Список литературы на английском языке / References in English

1. SOLAR Research Group. The TAWOS Dataset: Tasks, Worklogs and Sprints // GitHub Repository. — 2024. — URL: https://github.com/SOLAR-group/TAWOS (accessed: 01.10.2025).

2. University College London (UCL) Research Data Repository. The TAWOS Dataset. — 2023. — URL: https://rdr.ucl.ac.uk/articles/dataset/The_TAWOS_dataset/21308124 (accessed: 01.10.2025).

3. Tawosi A. A Versatile Dataset of Agile Open-Source Software Projects / A. Tawosi, A. Al-Subaihin, M. Moussa et al. // Proceedings of the 19th International Conference on Mining Software Repositories (MSR 2022). — 2022. — DOI: 10.48550/arXiv.2202.00979.

4. University of Ottawa PROMISE Software Engineering Repository // PROMISE, University of Ottawa. — URL: http://promise.site.uottawa.ca/SERepository/ (accessed: 02.01.2026).

5. GHTorrent — GitHub Archive Dataset // GitHub Repository. — URL: https://github.com/ghtorrent (accessed: 02.01.2026).

6. Tawosi V. On the Relationship Between Story Points and Development Effort in Agile Open-Source Software / V. Tawosi, R. Moussa, F. Sarro // Proceedings of the 16th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '22). — 2022. — URL: https://solar.cs.ucl.ac.uk/pdf/tawosi2022esem.pdf (accessed: 02.01.2026).

7. Sarro F. Investigating the Effectiveness of Clustering for Story Point Estimation – Replication Package / F. Sarro [et al.] // Solar Research Group, UCL. — 2022. — URL: https://solar.cs.ucl.ac.uk/pdf/tawosi2022saner.pdf (accessed: 02.01.2026).

8. A Versatile Dataset of Agile Open Source Software Projects // Proceedings of the 19th International Conference on Mining Software Repositories (MSR'22). — 2022. — URL: https://arxiv.org/abs/2201.05401 (accessed: 02.01.2026).

9. ZENODO Dataset Record // Zenodo. — URL: https://zenodo.org/records/15205608 (accessed: 02.01.2026).

10. Labeling issues through semantic patterns in open-source Agile practices // Knowledge-Based Systems. — 2022. — URL: https://www.sciencedirect.com/science/article/pii/S0950705125012389 (accessed: 02.01.2026).

11. Rodríguez Sánchez E. Effort and Cost Estimation Using Decision Tree Techniques and Story Points in Agile Software Development / E. Rodríguez Sánchez, A. Cruz, I. Requena et al. // Mathematics. — 2023. — Vol. 11, № 6. — P. 1477. — DOI: 10.3390/math11061477.

12. Chen X. Leveraging Large Language Models for Predicting Cost and Duration in Software Engineering Projects / X. Chen, Y. Li, M. Zhou. — 2024. — URL: https://arxiv.org/html/2409.09617v1 (accessed: 18.10.2025).

13. Zhang S. Large Language Models for Software Engineering: A Systematic Literature Review / S. Zhang, A.E. Hassan, P. Devanbu. — 2023. — URL: https://ar5iv.labs.arxiv.org/html/2308.10620 (accessed: 18.10.2025).

14. Pasuksmit P. Towards Just-Enough Documentation for Agile Effort Estimation: What Information Should Be Documented? / P. Pasuksmit, P. Thongtanunam, S. Karunasekera. — 2021. — DOI: 10.48550/arXiv.2107.02420.

15. Kula R. Dynamic Prediction of Delays in Software Projects Using Delay Patterns and Bayesian Modeling / R. Kula, S. Greuter, A. van Deursen et al. — 2023. — DOI: 10.48550/arXiv.2309.12449.

16. Baratto G.J. Z-Se2: A Model for Software Effort Estimation Using LLM GPT-3.5 / G.J. Baratto, E.M. De Bortoli Fávero, D. Casanova et al. — SSRN, 2024. — URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5041669 (accessed: 18.10.2025).

17. Kumar R. Leveraging Transformer-Based Large Language Models for Parametric Estimation of Cost and Schedule in Agile Software Development Projects / R. Kumar, P. Singh, A. Sharma. — ResearchGate, 2024. — URL: https://www.researchgate.net/publication/392901590 (accessed: 18.10.2025).