The Journal of Open Source Software

DigitalChild: Human Rights Data Pipeline for Child and LGBTQ+ Digital Protection

S. C. Vollmer [1] and **D. T. Vollmer** [2]

1 York University, Toronto, Canada **2** Resilient LLP, Toronto, Canada

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 17 February 2026
Published: unpublished

Summary

`DigitalChild` is an open-source Python pipeline for systematically scraping, processing, and analyzing human rights documents with a focus on child and LGBTQ+ digital protection. The software addresses the growing need for evidence-based analysis of how countries implement digital rights protections for vulnerable populations, particularly in the context of rapidly evolving technologies like artificial intelligence.

The pipeline automates the collection of policy documents from international organizations (African Union, UN Treaty Bodies, Universal Periodic Review mechanisms) and processes them through a comprehensive analysis workflow. It includes a scorecard system tracking 10 human rights indicators across 194 countries, with 2,543 validated authoritative source URLs. The software has been used to support research presented at international conferences on children's rights and published in peer-reviewed legal journals (S. Vollmer & Vollmer, 2022).

Statement of Need

While quantitative human rights research has grown substantially, a critical gap remains: **the lack of transparent, reproducible computational methods**. Existing research often presents results without sharing the underlying code, data processing steps, or validation procedures. This opacity creates several problems:

1. **Reproducibility crisis**: Published findings cannot be independently verified or extended
2. **Hidden methodological decisions**: Researchers cannot assess how data cleaning, categorization, or source selection impacts conclusions
3. **Barrier to entry**: New researchers must reinvent data collection and processing infrastructure
4. **Stale data**: Without automated pipelines, updating analyses as policies change is prohibitively expensive

Human rights organizations and researchers commonly rely on manual document review or proprietary tools that do not expose their methods. `DigitalChild` addresses this by providing complete documentation (25+ markdown files), tested modular code (124 automated tests), explicit version control for all tagging rules, comprehensive installation guides, and provenance tracking that maintains source URLs and processing history for every data point. This enables researchers to validate findings, adapt methods, or build upon the infrastructure.

State of the Field

Existing human rights data analysis tools fall into three categories: manual review processes, proprietary databases, and ad-hoc research scripts. UPR Info tracks Universal Periodic Review recommendations but focuses on aggregation rather than analysis pipelines. The Human

Rights Data Analysis Group publishes research but does not release underlying processing infrastructure. Academic researchers build custom scripts that rarely include documentation or reusability features.

Most critically, existing tools do not emphasize transparency and reproducibility as core design principles. `DigitalChild` uniquely combines automated collection, transparent methods, comprehensive testing, and complete documentation to enable reproducible research. The versioned configuration system allows researchers to compare how methodological choices impact results—a capability absent from existing tools.

Software Design

Build vs. Contribute Justification: We chose to build new software rather than contribute to existing tools (UPR Info, HRDAG) because those projects either (1) do not release their processing infrastructure as open-source code, precluding contributions, or (2) focus on data aggregation and presentation rather than providing reusable analysis pipelines. Contributing reproducibility features to closed-infrastructure projects would be impossible. Existing academic scripts lack the architectural foundation for transparent, versioned configurations and comprehensive testing. Our requirements—complete provenance tracking, versioned regex rules with comparison tools, modular extensibility, and documentation-first design—necessitated ground-up architecture. Building new software enables the research community to adopt transparent methods, whereas extending aggregation-focused tools would not address the reproducibility gap.

`DigitalChild` follows a modular pipeline architecture with four core stages: scraping, processing, tagging, and export. Each data source has dedicated scrapers (requests-based and Selenium variants) implementing rate limiting and respectful crawling. A fallback handler system attempts multiple processors (PDF, DOCX, HTML) sequentially, prioritizing reliability over speed.

Installation requires Python 3.12+ with standard dependencies (pandas, requests, BeautifulSoup4, pypdf, python-docx, selenium). Users run `python pipeline_runner.py --source au_policy --tags-version latest` to execute the full pipeline: scraping documents, converting to text, applying tags, and generating CSV exports. The scorecard workflow operates independently via `python pipeline_runner.py --mode scorecard --scorecard-action all`.

Regex-based rules defined in versioned JSON configurations enable transparent, auditable theme identification. While machine learning might offer higher accuracy, explicit regex patterns allow researchers to understand exactly why each tag was applied—critical for reproducibility and methodological transparency. All operations maintain complete provenance through JSON metadata files tracking processing history, tag versions, and timestamps. This design increases storage requirements but ensures full auditability. The architecture prioritizes extensibility through clear module boundaries, enabling researchers to add new scrapers or processors without modifying core pipeline logic.

Research Impact Statement

`DigitalChild` demonstrates both realized and credible research impact. The software supported research presented at the Second International Conference on Children's Rights (Stellenbosch, September 2025), examining digital rights protections for LGBTQ+ children across African states (D. Vollmer & Vollmer, 2025). The pipeline's predecessor methodology was published in a peer-reviewed legal journal (S. Vollmer & Vollmer, 2022).

The scorecard system provides 2,543 validated source URLs covering 194 countries across 10 human rights indicators, creating a publicly available dataset for comparative analysis. Active development continues with planned NLTK integration for sentiment analysis and expansion of SGBV monitoring to address pandemic-era violence patterns and wartime data preservation challenges—contexts where marginalized women and children face heightened rights violations (S. Vollmer & Vollmer, 2022). With comprehensive documentation and automated testing, the pipeline lowers barriers for researchers lacking extensive technical expertise.

Key Features

The pipeline includes seven specialized scrapers for international organizations (AU, OHCHR, UPR, UNICEF, regional mechanisms) with both requests-based and Selenium variants. Document processors convert PDF, DOCX, and HTML formats with fallback handlers. A regex-based tagging system with version control identifies themes (child rights, LGBTQ+ rights, AI policy, data protection) maintaining complete operation history.

The scorecard system tracks 10 indicators across 194 countries with validated sources from UNESCO, UNCTAD, ILGA, and UNICEF. Quality assurance includes 124 automated tests, parallel URL validation with retry logic, and diff monitoring to detect source changes. Timeline export enables temporal analysis at global, country, and regional levels. Pre-commit hooks enforce code quality, and CI/CD ensures tests pass before deployment.

AI Usage Disclosure

AI tools assisted with manuscript editing and formatting to JOSS specifications. All technical work is original.

Acknowledgements

This work was presented at the Second International Conference on Children's Rights (Stellenbosch, 2025). We thank the conference participants for their engagement with this research.

References

Vollmer, D., & Vollmer, S. (2025, September). Queer AI for the digital child: Examining the response to advanced digital technologies on the human rights of LGBTQ+ children in Africa. *Proceedings of the Second International Conference on Children's Rights*.

Vollmer, S., & Vollmer, D. (2022). Global perspectives of Africa: Harnessing the universal periodic review to address sexual and gender-based violence in SADC member states. *Stellenbosch Law Review*, *33*(1), 8–41. https://doi.org/10.47348/SLR/2022/i1a1