OpenQTS: A Metadata-Rich Open Dataset for Tang Dynasty Poetry

Robin Young
University of Cambridge
Cambridge, UK

Abstract

Tang Dynasty poetry (618–907 CE) represents one of the most significant corpora in Chinese literature, yet no openly accessible, computationally-suitable resource exists with comprehensive author metadata. We present OpenQTS, a complete machine-readable edition of the *Quan Tang Shi* (全唐詩), containing 48,300 poems from 2,602 authors across 900 volumes. Unlike existing resources, OpenQTS provides systematic biographical metadata through algorithmic inference and authority file integration. Each poem receives a unique identifier, with proper handling of multi-part poems (15.5% of corpus) and collaborative works. All metadata includes provenance tracking, documenting whether information derives from Wikidata, biographical text inference, or historical markers.[1]

1 Introduction

Tang poetry (618–907 CE) stands as one of the most celebrated literary traditions in Chinese civilization, producing works by over 2,600 documented poets including Li Bai, Du Fu, and Wang Wei. The *Quan Tang Shi* ("Complete Tang Poems") (Peng et al., 1705), compiled during the Qing dynasty, represents the most comprehensive anthology of this period, containing approximately 48,900 poems across 900 volumes. Despite its cultural and historical significance, no openly accessible computational resource exists that combines the complete corpus with rich, structured metadata for modern research.

Existing digital resources face limitations. The Chinese Text Project (Sturgeon, 2021), while providing web access to classical Chinese texts, requires institutional subscription for bulk API access and offers minimal structured author metadata. Community-driven efforts such as the `chinese-poetry` GitHub repository provide downloadable data but lack systematic quality control, metadata, and scholarly validation.

This gap has significant implications. Without systematic author metadata, temporal analysis of stylistic evolution remains impractical. Without gender data for example, studies of female poets' contributions require manual identification. Without authority file integration, cross-corpus research and international collaboration are hindered. The lack of systematic identifiers prevents reliable citation and reproducibility across studies.

We present OpenQTS (**Open** **Q**uan **T**ang **S**hi), a complete, metadata-rich edition of the Tang poetry corpus designed for computational research. The dataset contains 48,300 poems from 2,602 authors, each with systematic unique identifiers and proper handling of complex structures including multi-part poems and collaborative works. We enrich 37.7% of authors with authority file links (Wikidata, CBDB, VIAF) and provide English names for 28.7%, enabling international accessibility and cross-corpus research. Through algorithmic inference, we classify 72% of poems by Tang dynasty period (Early/High/Middle/Late Tang) and infer gender for 68.3% of authors, identifying 164 female poets. All metadata includes provenance tracking, documenting whether information derives from Wikidata, biographical inference, or textual markers. The complete dataset, source code, and validation scripts are freely available, providing transparent, reproducible infrastructure for classical Chinese computational research.

OpenQTS enables systematic analysis of author representation across Tang periods, tem-

[1]The dataset is available on Zenodo with source code on GitHub.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1364–1376
November 6-10, 2026 ©2026 Association for Computational Linguistics

poral evolution of poetic forms, cross-lingual studies through authority file links, and large-scale training data for classical Chinese NLP models. By providing transparent, validated, reproducible infrastructure, we aim to accelerate computational research on one of world literature's most important corpora.

The complete dataset, source code, and documentation are available on Github with CC licensing. We release all extraction and inference scripts to enable verification, extension, and application to other classical Chinese corpora.

2 Related Work

2.1 Classical Chinese Text Resources

The Chinese Text Project (Sturgeon, 2021) represents the largest digital library of premodern Chinese texts, with over 7.1 billion characters spanning pre-Qin through Qing dynasty materials. The project employs domain-optimized OCR and wiki-based crowdsourcing to create transcriptions of historical texts (Sturgeon, 2020). While invaluable for web-based reading and research, the Chinese Text Project requires institutional subscription for full API access to bulk text retrieval, and provides minimal structured author metadata beyond basic bibliographic information. For Tang poetry specifically, texts are available for browsing but lack the systematic biographical enrichment necessary for large-scale computational analysis.

Community-driven digitization efforts such as the `chinese-poetry` GitHub repository (chinese-poetry contributors, 2016) provide openly downloadable Tang and Song poetry in JSON format, claiming approximately 55,000 Tang poems. However, these resources rely on contributions without systematic source control or validation, resulting in inconsistent data structures and variable quality across entries. Author metadata typically consists only of names and dynasty affiliations, with no authority file linking or biographical enrichment. While useful for poetry generation experiments, these datasets lack the metadata infrastructure required for scholarly research.

Specialized datasets for classical Chinese poetry generation include the Ancient Poems Dataset (Liu et al., 2018), and various collections from Tsinghua University's THUAIPoet project (Yi et al., 2020). These resources focus on specific poetic forms optimized for generation tasks rather than historical coverage, and similarly lack rich metadata integration.

2.2 Historical Biographical Databases

The China Biographical Database (CBDB) (Harvard University et al., 2019) provides comprehensive biographical information for historical Chinese figures, with particular strength in official biographies and social networks. However, CBDB is maintained independently from textual corpora, requiring manual effort to link individuals to their literary works. Wikidata (Erxleben et al., 2014) offers multilingual, linked open data for historical figures including classical Chinese poets, but coverage of Tang dynasty authors varies significantly and identifiers are not systematically integrated with poetry corpora.

The Virtual International Authority File (VIAF) (OCLC, 2003) and Library of Congress (1977) authority files provide standardized identifiers for authors across library systems worldwide. While some prominent Tang poets have VIAF and LC identifiers, systematic linking to poetry collections has not been previously undertaken. Our work bridges this gap by integrating multiple authority files directly with the complete Tang poetry corpus.

2.3 Computational Approaches to Classical Chinese

Poetry generation has been the primary focus of computational work on classical Chinese poetry, from early neural approaches (Zhang and Lapata, 2014) through planning-based (Wang et al., 2016) and memory-augmented methods (Zhang et al., 2017). These efforts focus on generation rather than analysis, typically operating on small subsets without leveraging author metadata.

2.4 Biographical Metadata and Provenance

Several projects address challenges in biographical metadata extraction and provenance that parallel our work. BiographyNet (Fokkens et al., 2014) manages provenance at multiple levels for Dutch biographical data,

tracking how information is derived and aggregated across sources. Magistry et al. (2019) mine the Biographical Dictionary of Republican China, extracting structured data from semi-structured biographical entries for network analysis, which is a task closely related to our extraction of occupations, degrees, and temporal markers from *Quan Tang Shi* biographies. These projects demonstrate the importance of provenance-aware metadata design for historical corpora, an approach we adopt through systematic `_source` fields on all inferred metadata.

2.5 Resource Design and Metadata Standards

Recent NLP resource papers demonstrate the value of provenance tracking for enabling diverse research. The BigScience ROOTS corpus (Laurençon et al., 2022) emphasizes transparency through detailed source documentation across 59 languages. We adopt similar provenance practices for classical Chinese poetry, documenting the derivation of each metadata field to support quality-based filtering by downstream users.

3 Dataset Overview

3.1 Source Material

The *Quan Tang Shi* (全唐詩, "Complete Tang Poems") is the comprehensive anthology of Tang Dynasty poetry, compiled during the Qing Dynasty (1705 CE) under imperial commission. The anthology spans the entire Tang Dynasty period (618–907 CE) and includes works from over 2,600 poets across 900 volumes. We extract our corpus from the Wikisource edition, which provides a vetted, community-maintained digital transcription with version-controlled editing history and open licensing (CC BY-SA).

3.2 Data Model

OpenQTS structures each poem as a JSON object with systematic identifiers and metadata. Each poem receives a unique identifier following the pattern `QTS_VVV_EEE_PP`, where `VVV` is the volume number (001–900), `EEE` is the entry number within that volume, and `PP` is the part index for multi-part poems. This UID scheme enables citation and reproducible reference across studies.

Poems contain fields including `uid`, `title`, `poem` (array of lines), `volume`, and a nested `author` object with biographical metadata. For multi-part poems (15.5% of corpus), we provide `part_index`, `total_parts`, and `part_label` fields to properly index sequences ranging from 2 to 303 parts.

Author metadata integrates data from multiple sources with provenance tracking. Fields include canonical and recorded names, biographical information (birth/death years, period, gender), professional data (occupations, academic degrees), authority file identifiers (Wikidata, CBDB, VIAF, Library of Congress), and English names for international accessibility. All inferred metadata includes a corresponding `_source` field (e.g. `gender_source`, `period_source`) documenting whether information derives from Wikidata, biographical text analysis, or rule-based inference. Complete examples of author and poem records are provided in Appendix.

3.3 Corpus Statistics

Table 1 presents comprehensive corpus statistics. OpenQTS contains 48,300 poems from 2,602 authors across 900 volumes. Period classification covers 72% of poems (34,887 poems), while gender inference achieves 68.3% author coverage, identifying 164 female poets. Authority file integration reaches 37.7% coverage (981 authors), with English names provided for 28.7% (747 authors). The detailed extraction pipeline, inference algorithms, and validation methods are described in Section 4.

4 Methodology

We describe our end-to-end pipeline for constructing OpenQTS, from data extraction through metadata enrichment. All inference methods are rule-based, using lookup tables of historically documented terms (official titles, era names, examination degrees) rather than learned models. The complete lookup tables are provided in Appendix B, and the full source code is released with the dataset.

Attribute	Value
Total poems	48,300
Total authors	2,602
Volumes	900
Poem characteristics	
Multi-part poems	7,508 (15.5%)
Longest sequence	303 parts
Avg. lines per poem	4.9
Author metadata coverage	
Period classified	1,871 (72%)
Gender inferred	1,778 (68.3%)
Authority file linked	981 (37.7%)
English names	747 (28.7%)
Temporal distribution	
Pre-Tang (–618)	298 (0.6%)
Early Tang (618–712)	2,795 (5.8%)
High Tang (713–765)	9,520 (19.7%)
Middle Tang (766–835)	14,018 (29.0%)
Late Tang (836–907)	7,491 (15.5%)
Five Dynasties/Song (907–)	763 (1.6%)
Unclassified	13,413 (27.8%)
Gender distribution	
Male authors	1,614 (62.0%)
Female authors	164 (6.3%)
Unknown	824 (31.7%)

Table 1: OpenQTS dataset statistics. Percentages for author metadata indicate coverage among 2,602 total authors; percentages for poems indicate distribution among 48,300 total poems.

4.1 Data Extraction Pipeline

4.1.1 Wikisource Scraping

We extract poem text and basic metadata from Wikisource's structured *Quan Tang Shi* pages. Each of the 900 volumes is organized as a separate wiki page with volume header, followed by author sections, each containing one or more poems with titles.

We preserve the original text as transcribed, including variant characters and annotations where possible. Each poem is assigned a unique identifier based on its position: volume number, entry number within volume, and part index for multi-part works.

For multi-part poems, we detect sequential titles with part markers (其一, 其二, 其三, etc.) and group them into sequences, recording both the part index and total parts. This handles complex cases including the 303-part sequence by Hanshan (寒山).

4.1.2 Author Name Resolution

The *Quan Tang Shi* presents author names with considerable variation such as multiple names per author (courtesy names, art names, posthumous names), variant character forms, and uncertain attributions. We establish a canonical name for each distinct author through a semi-automatic process.

We extract all author name strings from volume headers and poem attributions. We then apply rule-based clustering to group probable variants: detecting parenthetical aliases (e.g. "張九齡 (字子壽)"), comparing character similarity, and identifying common patterns like "某某一作某某" (indicating alternative attributions). Ambiguous cases are resolved through Wikidata cross-reference when available, or retained as distinct authors when insufficient evidence exists for merging.

For collaborative works (聯句), we record all participating authors while assigning primary authorship to the first-listed poet for organizational purposes.

4.2 Authority File Integration

We link authors to established authority files to enable cross-corpus research and provide internationally-recognized identifiers.

We query Wikidata for each author using their canonical Chinese name, filtering results with Tang Dynasty temporal constraints. When multiple candidates exist, we use biographical details from the *Quan Tang Shi* (birth/death years, known aliases) to disambiguate. Successful matches provide Wikidata QIDs and enable extraction of structured biographical data including birth/death years, occupations, and English names.

For authors linked to Wikidata, we extract China Biographical Database (CBDB) identifiers and Virtual International Authority File (VIAF) identifiers from Wikidata's external identifier properties. We additionally query CBDB directly using name matching for authors without Wikidata entries. Library of Congress identifiers are extracted from VIAF records when available. This multi-source approach achieves 37.7% authority file coverage (981 authors).

4.3 Biographical Metadata Inference

Many authors in the *Quan Tang Shi* lack entries in modern databases, particularly minor poets and those from marginal groups. We develop four inference algorithms to derive biographical metadata from available evidence.

4.3.1 Period Classification

We classify authors into Tang Dynasty subdivisions (Early: 618–712, High: 713–765, Middle: 766–835, Late: 836–907) using a hierarchical approach:

1. **Direct dating**: If both birth and death years are known, assign period based on the author's productive years (approximated as ages 20–60).

2. **Single date**: If only birth or death year is known, estimate the probable active period using historical average lifespan (50 years).

3. **Biographical markers**: If dates are unavailable, extract temporal references from biographical text in Wikidata or CBDB (e.g. "served under Emperor Xuanzong" then High Tang).

This approach classifies 72% of poems (34,887 poems) by period. The remaining 28% lack sufficient temporal information. We record the classification method in `period_source` to enable quality-based filtering.

4.3.2 Gender Inference

Gender identification in historical Chinese texts presents unique challenges. Conventions differ from modern systems, explicit gender marking is rare, and information for minor authors is sparse. We employ a multi-tiered inference strategy prioritizing reliable sources:

1. **Wikidata**: If the author has a Wikidata entry with gender property, use this value (highest confidence). Source: `wikidata`.

2. **Explicit markers in names**: Detect gender-indicating characters:

   - Female markers: 女 (woman), 妃 (imperial consort), 夫人 (lady), 宮人 (palace woman), 姬 (concubine), 娘 (miss), 妓 (courtesan), 嬪 (consort), 婦 (wife), 妻 (wife), 后 (empress)
   - Male markers: 僧 (Buddhist monk), 釋 (Buddhist title), 道士 (Daoist priest), 仙 (immortal), 真人 (perfected person), 沙門 (śramaṇa)

   Source: `inferred_female_marker:X` or `inferred_male_marker:X`

3. **Biographical text analysis**: Extract gender-indicating phrases from author biographies:

   - Male indicators: Official titles (進士 "jinshi scholar", 刺史 "regional inspector", 郎中 "court gentleman", 御史 "censor", 尚書 "minister"), academic degrees, military ranks
   - Female indicators: Biographical phrases containing 女, 妃, 后

   Source: `inferred_biography_male:TITLE` or `inferred_biography_female:MARKER`

4. **Volume context**: The *Quan Tang Shi* groups some poems in thematic volumes:

   - Volumes 798–800: Exclusively female authors
   - Volumes 806–808: Buddhist monks
   - Volume 856–857: Daoist priests and nuns

   We infer gender from volume placement only when corroborated by other evidence. Source: `inferred_volume_female`, `inferred_volume_buddhist`, `inferred_volume_daoist`

5. **Default historical assumption**: For authors with official titles but no explicit gender markers, infer male gender based on historical Tang Dynasty gender restrictions in civil service. Source: `inferred_default_historical`

This hierarchical approach achieves 68.3% coverage (1,778 authors), identifying 164 female poets (6.3% of identified authors). The tiered strategy with provenance tracking allows researchers to filter by confidence level or exclude inference methods they consider unreliable.

Table 2 shows the composition of gender classifications by source. The majority of inferences (87.1%) are either externally validated through Wikidata or deductive from well-established institutional facts: Tang Dynasty civil service examinations, official positions, and religious orders were restricted by gender as a matter of law and institutional

Source	Count	%	Type
Wikidata	731	41.1	Validated
Biography (titles, degrees)	450	25.3	Deductive
Volume placement	213	12.0	Deductive
Name markers	149	8.4	Deductive
Biography (female)	5	0.3	Deductive
Default historical	230	12.9	Inferential

Table 2: Gender inference source breakdown for 1,778 classified authors. Validated and deductive sources account for 87.1% of classifications.

practice, not assumption. Only 230 authors (12.9%) rely on the `default_historical` category, which applies to authors with some association with officialdom but no specific title recorded. Researchers concerned about this category can exclude it via the `gender_source` field.

4.3.3 Occupation and Degree Extraction

We extract occupational titles and academic degrees from biographical text using pattern matching on classical Chinese official titles. The Tang Dynasty civil service used standardized titles (進士 "jinshi scholar", 明經 "classicist", 刺史 "regional inspector", etc.) that appear consistently in biographical records.

Our extraction identifies 43.9% of authors (1,143) with occupation data and 25.7% (670) with academic degrees. Common occupations include government official (793 authors), poet (596), regional governor (253), and various court positions. The most common degree is 進士 (jinshi scholar), the highest level of the imperial examination system.

4.4 Quality Assurance

We implement multiple validation layers to ensure data quality and detect issues requiring manual attention. Automated checks verify that all poem UIDs are unique and follow the prescribed format (`QTS_VVV_EEE_PP`), that all 900 volumes are present, and that poem counts per volume fall within expected ranges based on the source edition.

Our validation pipeline includes heuristics checks for six categories of potential issues: (1) unsplit multi-part poems where titles indicate multiple parts (e.g. 三首, "three poems") but the poem wasn't properly split into constituent parts; (2) biography entries or misattributed poems where the title appears to be an author's name rather than a poem title; (3) collaborative poems (聯句) containing inline attribution markers requiring special handling; (4) suspicious content such as single-line poems with metadata-like markers (缺, 佚, "missing"); (5) abbreviated titles containing ellipsis or truncation markers (. 〔略〕) indicating incomplete text; and (6) content quality issues including non-Chinese characters, uncaught brackets, or encoding errors that suggest OCR problems.

Our automated validation pass identified approximately 50 volumes containing non-Chinese characters or encoding errors in the Wikisource source. Rather than working around these issues, we corrected them directly in Wikisource by consulting a *Quan Tang Shi* digitized print edition where necessary and re-ran our extraction pipeline after the corrections were accepted. This approach improved the source data for all Wikisource users while ensuring our dataset's accuracy.

We conducted manual validation on random samples across three layers. For extraction accuracy, we verified correct text, author attribution, and volume assignment on 250 randomly sampled poems, finding no errors. For authority file linking, we evaluated a random sample of 100 Wikidata-linked authors by verifying that each QID referred to the correct historical individual. Of 100 links, 99 were correct; the single error involved 柳氏 ("Lady Liu"), a generic surname-plus-feminine-marker pattern that matched a Wikidata surname disambiguation page rather than the specific individual. This class of error, affecting authors recorded without personal names, is detectable through automated entity-type checking.

For gender and period inference, the majority of classifications are deterministic mappings where accuracy is definitional rather than empirical as era names map to known date ranges, and examination titles or official positions map to gender based on documented Tang Dynasty institutional restrictions (see Table 2 and Appendix B). Empirical validation is meaningful only for the smaller inferential categories (e.g. the 230 authors classified via `default_historical`), where independent verification would require specialized expertise in Tang Dynasty biographical history not read-

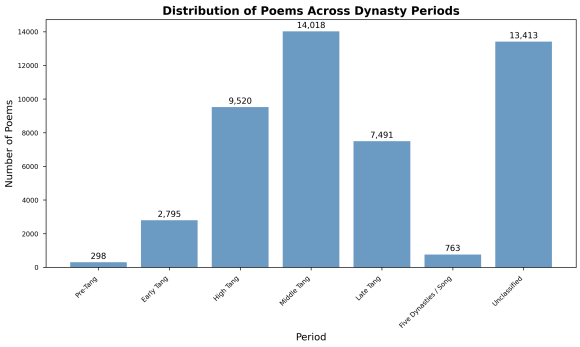


Figure 1: Distribution of poems across Tang Dynasty periods and adjacent eras. The four main Tang subdivisions account for 72% of poems, with Middle Tang showing highest concentration.

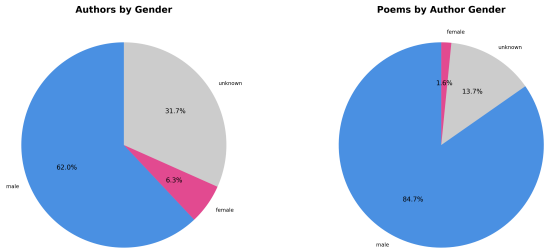


Figure 2: Gender distribution among authors (left) and poems (right). The productivity gap suggests female poets either wrote less prolifically or faced greater preservation losses.

ily available for annotation at scale.

5 Dataset Analysis

We demonstrate OpenQTS's utility through analyses that were previously impractical due to lack of systematic metadata. These examples illustrate research enabled by our dataset rather than presenting comprehensive literary conclusions.

5.1 Temporal Distribution and Period Classification

Figure 1 visualizes the temporal distribution summarized in Table 1. Middle Tang's dominance (29.0%) reflects both historical reality with the period's 70-year span being a flourishing era for poetry and preservation bias favoring established poets. The 27.8% unclassified portion represents authors lacking reliable biographical records, primarily minor poets. This transparency enables researchers to assess whether their analyses require complete temporal coverage or can proceed with the 72% classified subset.

5.2 Gender Representation Temporal Dynamics

While Table 1 shows overall gender distribution (164 female poets, 6.3% of identified authors), temporal analysis reveals additional patterns. Figure 2 shows that female poets wrote fewer poems on average: 6.3% of authors produce only 1.6% of poems, suggesting differential productivity or preservation rates.

More significantly, female representation varies dramatically across periods. Early Tang shows 3.22% female authorship (90 poems), while High Tang, the dynasty's literary golden age, shows only 0.12% (11 poems). Middle and Late Tang show modest recovery (0.71% and 0.72%). This temporal variation was invisible without systematic gender metadata and invites investigation into social dynamics, preservation bias, and women's changing access to literary production.

The most prolific identified female poets include Lady Huarui (花蕊夫人, 157 poems), Xue Tao (薛濤, 90 poems), Yu Xuanji (魚玄機, 50 poems), and Empress Wu Zetian (武則天, 49 poems), which were figures predominantly associated with imperial courts, suggesting social status mediated literary access.

6 Discussion

6.1 Comparison with Existing Resources

Table 3 compares OpenQTS with the two primary alternatives for Tang poetry research: the Chinese Text Project and the chinese-poetry GitHub repository. While both provide valuable access to Tang poetry, neither offers metadata infrastructure for large-scale computational research.

The Chinese Text Project (Sturgeon, 2021) primarily exists as a web-based reading interface with scholarly commentary and cross-references, serving a role for traditional philological research. However, its design prioritizes browsing over bulk analysis; full API access requires institutional subscription, Tang poetry is embedded within a multi-dynasty corpus without systematic extraction, and author metadata focuses on bibliographic rather than biographical information. For computa-

Feature	ctext.org	chinese-poetry	OpenQTS
Access model	Web/API†	Open	Open
Tang poems	Included*	∼55K	48.3K
Bulk download	No	Yes	Yes
Systematic UIDs	Yes	No	Yes
Author metadata	Some	Minimal	Rich
Period classification	No	No	72%
Gender data	No	No	68%
Authority files	No	No	38%
English names	No	No	29%
Provenance tracking	Yes	No	Yes

Table 3: Comparison of OpenQTS with existing Tang poetry resources. †API access requires institutional subscription. *Tang poetry included within larger corpus without separate export.

tional research requiring reproducible datasets with rich metadata, these limitations are significant.

The chinese-poetry GitHub repository provides open bulk access and has enabled numerous poetry generation projects. However, its crowdsourced nature results in doubtful provenance, variable quality, and minimal author metadata beyond names and dynasty affiliations. The lack of documented quality control limits its use for research requiring reproducible metadata.

OpenQTS addresses these gaps by combining open bulk access with systematic metadata, transparent methodology, and comprehensive provenance tracking. Our approach prioritizes reproducibility. Systematic UIDs enable citation, documented inference methods allow quality-based filtering, and released source code enables verification and extension.

6.2 Research Applications

OpenQTS enables several categories of research previously impractical at scale. Systematic gender metadata makes possible the first large-scale study of female Tang poets' participation patterns, thematic preferences, and social contexts. The temporal variation we document, with Early Tang's 3.22% female representation dropping to 0.12% in High Tang, invites investigation into how political changes, including Empress Wu Zetian's reign, affected women's literary access. Prior work on female Tang poets relied on manual identification and small samples; our systematic approach enables population-level analysis.

Period classification covering 72% of poems enables controlled temporal comparisons of stylistic evolution. We observe, for example, that average poem length declines from 5.64 lines in Early Tang to 4.47 lines in Late Tang. Researchers can investigate whether this trend reflects changing genre preferences, audience expectations, or compositional practices, though the 28% unclassified poems should be considered when interpreting such patterns.

Authority file integration enables cross-corpus research by linking Tang poets to other anthologies, biographical databases, and translation collections. The 37.7% coverage with English names (28.7%) facilitates international collaboration and comparative studies across languages and dynasties. Scholars studying the global reception of Tang poetry can use these links to connect Chinese originals with translations, critical editions, and secondary scholarship.

For NLP applications, OpenQTS provides structured data for several tasks beyond general language modeling. The 7,508 multi-part poems, with sequences up to 303 parts, provide natural data for studying thematic coherence and stylistic consistency across poem sequences, which is a harder generation problem than single-poem production that existing poetry datasets cannot support. Collaborative poems (聯句), where multiple poets contribute to a single work with inline attribution markers, provide ground-truth data for author attribution and style transfer, with labels for who wrote which part of a jointly authored text. Period, gender, and occupation metadata enable conditional generation tasks (e.g. generating in the style of a specific period or social group) and controlled evaluation of model behavior across demographic and temporal categories. Systematic UIDs enable reproducible train/test splits across studies, addressing a gap in classical Chinese NLP evaluation.

6.3 Design Decisions

Several design choices warrant discussion. We source from Wikisource rather than creating new transcriptions, prioritizing reproducibility and community validation over editorial control. While this introduces dependency on Wikisource's editorial choices, it ensures our work builds on established, openly-licensed transcriptions rather than duplicating effort or creating proprietary versions.

Our inference algorithms balance coverage with transparency. Rather than maximizing metadata coverage through aggressive inference, we implement conservative methods with explicit provenance tracking. This allows researchers to filter by confidence level (e.g. accepting only Wikidata-sourced gender vs. including inferred values) based on their research needs. We believe transparent limitations serve scholarship better than false completeness.

The JSON format prioritizes accessibility for computational researchers over space efficiency. While plaintext or binary formats would reduce file size, JSON's human-readability, wide tool support, and self-documenting structure align with our goal of lowering barriers to entry for digital humanities research.

7 Conclusion

We presented OpenQTS, a comprehensive, metadata-rich edition of the Tang Dynasty poetry anthology designed for computational research. The dataset contains 48,300 poems from 2,602 authors with systematic unique identifiers, authority file integration (37.7% coverage), and algorithmic metadata inference including period classification (72% coverage) and gender identification (68.3% coverage). All metadata includes provenance tracking, documenting whether information derives from Wikidata, biographical inference, or rule-based patterns.

OpenQTS addresses a gap in classical Chinese NLP resources. Existing alternatives either require institutional subscription for bulk access or lack the systematic metadata necessary for temporal analysis, gender studies, and cross-corpus integration. By combining open access with transparent methodology and comprehensive documentation, we provide infrastructure that enables research previously impractical at scale.

We hope OpenQTS contributes to computational research on one of world literature's most important traditions and serves as a model for transparent, reproducible resource creation in digital humanities. Future work includes expanding coverage to other dynasties, refining inference algorithms through community feedback, and developing additional metadata layers (poetic form classification, thematic tagging, intertextual references).

Limitations

We acknowledge several limitations in OpenQTS that researchers should consider when using the dataset.

7.1 Source Limitations

Our corpus derives from a single source, the Wikisource edition of *Quan Tang Shi*. While Wikisource provides community-validated transcriptions, it inherits the editorial choices and potential biases of the original Qing dynasty compilation. The *Quan Tang Shi* itself has known issues with poems misattributed to wrong authors, variant texts consolidated arbitrarily, and possible exclusion of works. We do not claim our dataset is definitive but rather provide a systematically structured version of one widely used edition.

Wikisource transcriptions may contain errors introduced during digitization or community editing. While we implement quality checks, we do not independently verify every character against print editions. Researchers requiring accuracy for specific poems should consult original sources.

7.2 Metadata Coverage and Inference Limitations

Our metadata coverage, while higher than existing alternatives, remains incomplete. Period classification covers 72% of poems, leaving 28% unclassified primarily from minor poets lacking reliable biographical records. Gender inference achieves 68.3% author coverage, with 31.7% unknown. Authority file integration reaches 37.7% of authors. This incomplete coverage may introduce selection bias if researchers analyze only metadata-complete subsets: well-documented poets (typically male, high-status officials) are overrepresented.

Our gender inference employs hierarchical methods reflecting historical Tang Dynasty contexts, including inferences based on official titles (which were restricted to men). While we document all inference sources through provenance tracking, researchers should evaluate whether our methods align with their

theoretical frameworks. In particular, downstream users should be aware that applying historical gender categories to computational analysis risks reifying labels that may not reflect the full complexity of individual identities. We recommend that users who work with inferred gender metadata report which provenance tiers they include and acknowledge the 12.9% of classifications that rely on default historical assumptions rather than direct evidence.

Period classification relies on birth/death years and biographical markers that may be uncertain, disputed, or inaccurate in historical sources. We classify based on productive years (ages 20–60), but actual creative periods vary. Poets with long lifespans spanning multiple periods are assigned to a single category, potentially obscuring stylistic evolution.

Occupation and degree extraction depends on pattern matching against classical Chinese titles, which may miss non-standard phrasing or misinterpret context. The 43.9% occupation coverage reflects both true data scarcity and limitations of our extraction methods.

Our inference algorithms are rule-based rather than using modern machine learning techniques. While this choice prioritizes transparency and interpretability, more sophisticated approaches might achieve higher coverage or accuracy. We provide conservative estimates rather than probabilistic confidence scores, which limits researchers' ability to weight metadata by reliability.

7.3 Scope Limitations

OpenQTS covers only Tang Dynasty poetry from the *Quan Tang Shi*, excluding other Tang literary genres (prose, drama) and poetry from other dynasties. Researchers studying cross-dynasty trends or multi-genre patterns should integrate OpenQTS with other resources.

We provide Chinese text only, without translations. While English names facilitate author identification, non-Chinese readers cannot directly engage with poem content. Future work might integrate existing translations, though copyright and attribution complexities present challenges.

The dataset includes limited linguistic annotation beyond basic structure (lines, characters). We do not provide phonological reconstruction, meter analysis, rhyme schemes, or grammatical parsing, which are all valuable for certain research questions but beyond our current scope.

7.4 Usage Considerations

Researchers should exercise caution when making strong claims based on incomplete metadata. For example, analyzing "all female Tang poets" actually means "all female poets we successfully identified," which may exclude certain groups (e.g. women without official connections, whose gender is harder to infer from historical records).

Temporal analyses using period classification should acknowledge the 28% unclassified poems and consider whether missingness is random or systematic. If unclassified poems disproportionately represent certain periods, genres, or author demographics, temporal comparisons may be biased.

Cross-corpus integration via authority files assumes these databases accurately represent historical figures, which is an assumption that should be validated for specific research contexts.

Despite these limitations, OpenQTS provides higher metadata coverage than existing open alternatives for Tang poetry. We document these limitations to help researchers assess fitness for their specific use cases, and welcome community contributions to extend coverage.

References

chinese-poetry contributors. 2016. chinese-poetry: The most comprehensive database of Chinese poetry. https://github.com/chinese-poetry/chinese-poetry. Accessed: 2026-01-13.

Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. Introducing wikidata to the linked data web. In *The Semantic Web – ISWC 2014*, volume 8796 of *Lecture Notes in Computer Science*, pages 50–65, Cham. Springer.

Antske Fokkens, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, and Guus Schreiber. 2014. BiographyNet: Methodological issues when NLP supports historical research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3728–3735, Reykjavik, Ice-

land. European Language Resources Association (ELRA).

Harvard University, Academia Sinica, and Peking University. 2019. China biographical database (CBDB). https://projects.iq.harvard.edu/cbdb. Accessed: 2026-01-13.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Benallal, Francesco De Toni, and 35 others. 2022. The bigscience roots corpus: a 1.6tb composite multilingual dataset. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Library of Congress. 1977. Library of Congress Name Authority File. https://id.loc.gov/authorities/names.html.

Dayiheng Liu, Quan Guo, Wubo Li, and Jiancheng Lv. 2018. A multi-modal chinese poetry generation model. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Rio de Janeiro, Brazil. IEEE.

Pierre Magistry, Cécile Armand, and Christian Henriot. 2019. Mining the biographical dictionary of republican china: From print to network exploration. In *BD2019 - 3rd Conference on Biographical Data in a Digital World 2019*, pages 1–7, Varna, Bulgaria. Institute of Information and Communication Technologies, Bulgarian Academy of Sciences.

OCLC. 2003. Virtual International Authority File (VIAF). https://viaf.org.

Dingqiu Peng, Sanzeng Shen, Zhongne Yang, Shihong Wang, Yi Wang, Mei Yu, Shuben Xu, Dingjin Che, Conglü Pan, and Sili Zha. 1705. *Complete Tang Poems*. Imperial Printing Office of the Qing Court, Beijing. Chinese title: 《全唐詩》. Compiled by imperial order under the Kangxi Emperor. Chinese compilers: 彭定求、沈三曾、楊中訥、汪士鋐、汪繹、俞梅、徐樹本、車鼎晉、潘從律、查嗣璨; edited and printed under imperial command by 曹寅.

Donald Sturgeon. 2020. Digitizing premodern text with the chinese text project. *Journal of Chinese History*, 4(2):486–498.

Donald Sturgeon. 2021. Chinese text project: A dynamic digital library of premodern chinese. *Digital Scholarship in the Humanities*, 36(Supplement_1):i101–i112.

Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 2016. Chinese poetry generation with planning based neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1051–1060, Osaka, Japan. The COLING 2016 Organizing Committee.

Xiaoyuan Yi, Ruoyu Li, Cheng Yang, Wenhao Li, and Maosong Sun. 2020. Mixpoet: Diverse poetry generation via learning controllable mixed latent space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9450–9457.

Jiyuan Zhang, Yang Feng, Dong Wang, Yang Wang, Andrew Abel, Shiyue Zhang, and Andi Zhang. 2017. Flexible and creative Chinese poetry generation using neural memory. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1364–1373, Vancouver, Canada. Association for Computational Linguistics.

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680, Doha, Qatar. Association for Computational Linguistics.

A Data Format Examples

A.1 Author Record Example

Figure 3 shows a complete author record for Emperor Taizong of Tang (李世民), demonstrating the structure and metadata of OpenQTS author entries. The record integrates multiple biographical sources from the volume preface of *Quan Tang Shi* and the Wikidata entry with explicit provenance tracking via the `source` field. Authority file integration spans four databases (Wikidata, CBDB, VIAF, Library of Congress), enabling cross-corpus research and international identification. Occupations derive from multiple sources as indicated by `occupations_source: "wikidata+inferred_biography"`, demonstrating the multi-source aggregation approach. External links to Wikipedia, Wikisource, and Wikimedia Commons images provide access to additional scholarly resources. Comprehensive temporal data includes both birth and death years (598–649 CE) and places (Wugong County and Hangfeng Hall), supporting historical and geographical analysis.

```
{
  "李世民": {
    "canonical": "李世民",
    "recorded_variants": ["李世民"],
    "biographies": [
      {
        "text": "帝姓李氏，諱世民，神堯次子，聰明英武...",
        "source": "volume_1"
      },
      {
        "text": "唐太宗李世民（598年1月28日—649年7月10日）
            ...",
        "source": "wikipedia"
      }
    ],
    "wikidata_id": "Q9701",
    "english_name": "Emperor Taizong of Tang",
    "birth_year": 598,
    "death_year": 649,
    "period": "Early Tang",
    "period_source": "wikidata",
    "gender": "male",
    "gender_source": "wikidata",
    "cbdb_id": "0013060",
    "viaf_id": "67810502",
    "loc_id": "n82074092",
    "birth_place": "Wugong County",
    "death_place": "Hangfeng Hall",
    "occupations": ["academician", "emperor", "poet", "
        scholar"],
    "occupations_source": "wikidata+inferred_biography",
    "wikipedia_url": "https://en.wikipedia.org/wiki/
        Emperor_Taizong_of_Tang",
    "wikisource_url": "https://zh.wikisource.org/wiki/Author
        :李世民",
    "image_url": "https://commons.wikimedia.org/.../Tang%20
        Taizong.jpg"
  }
}
```


Figure 3: Example author record for Emperor Taizong (李世民) showing metadata with provenance tracking and authority file integration.

A.2 Poem Record Examples

Figure 4 shows two consecutive poems from a multi-part sequence, demonstrating how OpenQTS handles complex poetic structures. The systematic UIDs (`QTS_001_001_01`, `QTS_001_001_02`) enable citation while the part indexing suffix differentiates components within sequences. Multi-part sequence tracking through `part_index` and `total_parts` fields allows researchers to programmatically identify and analyze complete poetic series. In this case, these are the first two parts from Emperor Taizong's 10-part "Imperial Capital" sequence. Author metadata is embedded directly within each poem record, enabling self-contained access without requiring joins across files. The second poem demonstrates textual variant tracking. Three characters have documented alternatives with sources (e.g. line 4's 鳳/鳥 variation). The first poem includes editorial notes and a complete preface, preserving scholarly apparatus from the source edition. Poem text is stored as an array of strings with one line per element, preserving the original structure and enabling line-level analysis while maintaining readability.


```
{
  "uid": "QTS_001_001_01",
  "volume": 1,
  "author": {
    "canonical": "李世民",
    "recorded": "李世民",
    "gender": "male",
    "period": "Early Tang",
    "english_name": "Emperor Taizong of Tang",
    "wikidata_id": "Q9701"
  },
  "title": "帝京篇十首 一",
  "poem": ["秦川雄帝宅", "函谷壯皇居", "綺殿千尋起",
           "離宮百雉餘", "連甍遙接漢", "飛觀迥凌虛",
           "雲日隱層闕", "風煙出綺疏"],
  "notes": [{
    "content": "並序",
    "context": "帝京篇十首",
    "original_title": "帝京篇十首（並序）"
  }],
  "preface": "予以萬幾之暇，游息藝文。觀列代之皇王...",
  "part_index": 1,
  "total_parts": 10
},
{
  "uid": "QTS_001_001_02",
  "volume": 1,
  "author": {
    "canonical": "李世民",
    "recorded": "李世民",
    "gender": "male",
    "period": "Early Tang",
    "english_name": "Emperor Taizong of Tang",
    "wikidata_id": "Q9701"
  },
  "title": "帝京篇十首 二",
  "poem": ["巖廊罷機務", "崇文聊駐輦", "玉匣啓龍圖",
           "金繩披鳳篆", "韋編斷仍續", "縹帙舒還卷",
           "對此乃淹留", "欹案觀墳典"],
  "part_index": 2,
  "total_parts": 10,
  "variants": [
    {"line": 4, "text": "鳳", "variant": "鳥", "source": "一
        作「鳥」"},
    {"line": 5, "text": "仍", "variant": "方", "source": "一
        作「方」"},
    {"line": 7, "text": "淹留", "variant": "忘憂", "source":
        "一作「忘憂」"}
  ]
}
```


Figure 4: Example poem records showing parts 1 and 2 of Emperor Taizong's 10-part sequence, demonstrating multi-part indexing, embedded author metadata, editorial notes, prefaces, and textual variants.

These examples demonstrate OpenQTS's schema design principles: systematic identifiers enable precise citation, embedded metadata allows self-contained access, provenance tracking documents all inference sources, and comprehensive structure preservation maintains scholarly fidelity to source materials.

B Inference Lookup Tables

Tables 4–7 present the lookup tables used in our inference pipeline. Each classification is a deterministic string match against these historically documented terms.

Term	Meaning
宮人	Palace woman
妻	Wife
夫人	Lady
妓	Courtesan/entertainer
女 / 女子	Woman/daughter
娘	Lady/maiden
姬	Concubine/lady
氏女	Daughter of (surname)
婦	Woman/wife
嬪	Imperial concubine
妃	Consort
后	Empress

Table 4: Female gender markers (13 terms). Matched against author names and biography text.

Category	Examples	Count
Religious clergy	僧, 釋, 道士, 沙門	7
Examinations	進士, 明經, 秀才	7
Central govt.	宰相, 尚書, 御史, 翰林	32
Regional/local	刺史, 太守, 縣令	12
Military	將軍, 都督, 參軍	10
Judicial	司法, 大理, 法曹	3
Nobility	王, 公, 侯, 帝	5

Table 5: Male gender markers by category (76 terms total). All represent roles restricted to men by Tang Dynasty law or institutional practice.

Volume	Gender	Rationale
797–805	Female	Dedicated female poet volumes
806–851	Male	Buddhist clergy volumes
852–862	Male	Daoist practitioner volumes

Table 6: Volume-based gender inference. Applied only when corroborated by other evidence.

Era name	Years	Period
武德	618–626	Early Tang
貞觀	627–649	Early Tang
開元	713–741	High Tang
天寶	742–756	High Tang
大曆	766–779	Middle Tang
貞元	785–805	Middle Tang
元和	806–820	Middle Tang
會昌	841–846	Late Tang
咸通	860–874	Late Tang
天祐	904–907	Late Tang

Table 7: Period inference era names (representative sample; full table contains 87 entries spanning Sui Dynasty through Five Dynasties). Each era name maps deterministically to a Tang period via its known date range.