

RESEARCH ARTICLE • PEER-REVIEWED • OPEN ACCESS • **SERIES B** • 05.00.00 — TECHNICAL SCIENCES
• UDC: 025.3:004.65

A model of a unified bibliographic metadata layer with embedded quality control for a national scientific publications platform

Milliy ilmiy nashrlar platformasi uchun sifat nazorati o'rnatilgan yagona bibliografik metama'lumotlar qatlami modeli

Модель единого слоя библиографических метаданных со встроенным контролем качества для национальной платформы научных изданий

RAYIMJONOV Xurshidjon G'ayratjon o'g'li¹, MIRRAHIMOV Mirkomil Miromil o'g'li²

¹ Scientific and Technical Information Center under the Agency for Innovative Development of the Republic of Uzbekistan, Tashkent, Uzbekistan

² Turin Polytechnic University in Tashkent, Tashkent, Uzbekistan

ARTICLE INFO

ARTICLE TYPE

Research Article

LANGUAGE OF FULL TEXT

Русский (RU)

SCIENTIFIC FIELD / OAK

☒ 05.00.00 – Technical Sciences

Техника fanlari

Series B

UDC / UDK

025.3:004.65

ARTICLE HISTORY

Received: 2026-09-02

Revised: 2026-09-10

Accepted: 2026-09-16

Published: 2026-09-30

DOI

<https://doi.org/10.66073/ReFocus.V5I9.0907>

CONTACT

ReFocusJournal@gmail.com

LICENSE



Creative Commons Attribution 4.0 International (CC BY 4.0). This is an open access article distributed under the terms of the CC BY 4.0 licence.

AUTHOR, TITLE AND ABSTRACT METADATA (EN / UZ / RU)

ENGLISH

RAYIMJONOV Xurshidjon G'ayratjon o'g'li¹, MIRRAHIMOV Mirkomil Miromil o'g'li²

¹ Scientific and Technical Information Center under the Agency for Innovative Development of the Republic of Uzbekistan, Tashkent, Uzbekistan

² Turin Polytechnic University in Tashkent, Tashkent, Uzbekistan

Abstract. The article proposes a model of a unified bibliographic metadata layer for the National Scientific Publications Platform of Uzbekistan, with quality control embedded at every processing stage. Each field value is stored as an assertion with its source, method, time and confidence; the canonical value is selected by source rank. Quality is assessed along six dimensions; completeness, correctness and consistency are combined into a metadata quality index guiding recommendations to journals.

Keywords: bibliographic metadata; data quality; data provenance; assertion model; metadata quality index; national scientific publications platform

O'ZBEKCHA

RAYIMJONOV Xurshidjon G'ayratjon o'g'li¹, MIRRAHIMOV Mirkomil Miromil o'g'li²

¹ O'zbekiston Respublikasi Innovatsion rivojlanish agentligi huzuridagi Ilmiy-texnik axborot markazi, Toshkent, O'zbekiston

² Toshkent shahridagi Turin politexnika universiteti, Toshkent, O'zbekiston

Annotatsiya. Maqolada O'zbekiston Milliy ilmiy nashrlar platformasi uchun yagona bibliografik metama'lumotlar qatlami modeli taklif etilgan; ushbu modelda sifat nazorati yozuvga ishlov berishning barcha bosqichlariga o'rnatilgan. Har bir maydon qiymati manbasi, usuli, vaqti va ishonchiligi ko'rsatilgan tasdiq (assertion) sifatida saqlanadi, kanonik qiymat esa manbalar darajasi bo'yicha tanlanadi. Sifat oltita o'lchov bo'yicha baholanadi; to'liqlik, to'g'rilik va muvofiqlik ko'rsatkichlari jurnallarga manzilli tavsiyalar berishda qo'llaniladigan metama'lumotlar sifati indeksiga jamlanadi.

Kalit so'zlar: bibliografik metama'lumotlar; ma'lumotlar sifati; ma'lumotlar kelib chiqishi; tasdiqlar modeli; metama'lumotlar sifati indeksi; milliy ilmiy nashrlar platformasi

РУССКИЙ

РАЙИМЖОНОВ Хуршиджон Гайратжон угли¹, МИРРАХИМОВ Миркомил Миромил угли²

¹ Научно-технический информационный центр при Агентстве инновационного развития Республики Узбекистан, Ташкент, Узбекистан

² Туринский политехнический университет в городе Ташкенте, Ташкент, Узбекистан

How to cite / Iqtibos keltirish / Как цитировать: RAYIMJONOV, X. G., MIRRAHIMOV, M. M. (2026). A model of a unified bibliographic metadata layer with embedded quality control for a national scientific publications platform. Research Focus International Scientific Journal, 5(9), 73–77. <https://doi.org/10.66073/ReFocus.V5I9.0907>

© 2026 The Author(s). Published by Research Focus International Scientific Journal. Open access article under the CC BY 4.0 licence.

HAC / OAK: On the HAC list of national scientific publications (Academy of Sciences of the Republic of Uzbekistan) for 06.00.00 Agricultural Sciences and 16.00.00 Veterinary Sciences. В перечне ВАК при АН РУз: 06.00.00 – Сельскохозяйственные науки; 16.00.00 – Ветеринарные науки.

Аннотация. В статье предложена модель единого слоя библиографических метаданных для Национальной платформы научных изданий Узбекистана; в этой модели контроль качества встроен во все этапы обработки записи. Каждое значение поля хранится как утверждение с указанием источника, метода, времени и достоверности, а каноническое значение выбирается по рангам источников. Качество оценивается по шести измерениям; показатели полноты, корректности и согласованности сводятся в индекс качества метаданных, используемый для адресных рекомендаций журналам.

Ключевые слова: библиографические метаданные; качество данных; происхождение данных; модель утверждений; индекс качества метаданных; национальная платформа научных изданий

1. ВВЕДЕНИЕ

Introduction

В Узбекистане издается более 200 научных журналов, но сведения о статьях рассредоточены по сайтам редакций, а данные о публикациях ученых для аттестации и отчетности собираются вручную. Между тем национальные библиографические базы данных с полным охватом научной продукции страны могут лучше подходить для библиометрического анализа, чем коммерческие международные индексы, особенно в социально-гуманитарных науках [1].

Любой агрегатор неизбежно наследует дефекты источников: пропуски аффилиаций и списков литературы, разночтения в написании имен и значениях полей. Отклонение неполных записей сокращает охват, а их некритическое принятие снижает достоверность. Если качество данных – это пригодность для использования потребителями [2], контроль должен сопровождать запись на всем пути от источника до сервиса. Цель статьи – обосновать модель единого слоя библиографических метаданных, в которой происхождение значений и результаты проверок качества – свойства самих данных.

2. МАТЕРИАЛЫ И МЕТОДЫ

Materials and methods

Материалом исследования послужили концепция и техническое задание первого этапа Национальной платформы

научных изданий. Методами документального и системного анализа выделены механизмы контроля на каждом этапе жизненного цикла записи; они сопоставлены с общепринятыми измерениями качества данных и метаданных [2; 3].

3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Results and discussion

Сведения о статьях поступают только от журналов: автоматически (по протоколу OAI-PMH или сканированием страниц с метатегами citation_*) либо вручную в кабинете журнала с предзаполнением полей из PDF-файла и по DOI; внешние источники (Crossref, ORCID, ISSN Portal) записей не создают, а проверяют и дополняют их. Запись сохраняется в неизменном исходном слое (RAW) с хеш-суммой SHA-256, временем и адресом получения; запись, удаленная в источнике, не стирается, а помечается, как и в OAI-PMH [4].

Ключевой элемент модели – переход от записи как целого к утверждениям (assertions) о значениях отдельных полей. Утверждение фиксирует поле, значение, источник, метод, достоверность, время и ссылку на исходную запись (рис. 1) и тем самым связывает значение с породившим его действием и ответственным агентом в терминах онтологии происхождения PROV-O [5].

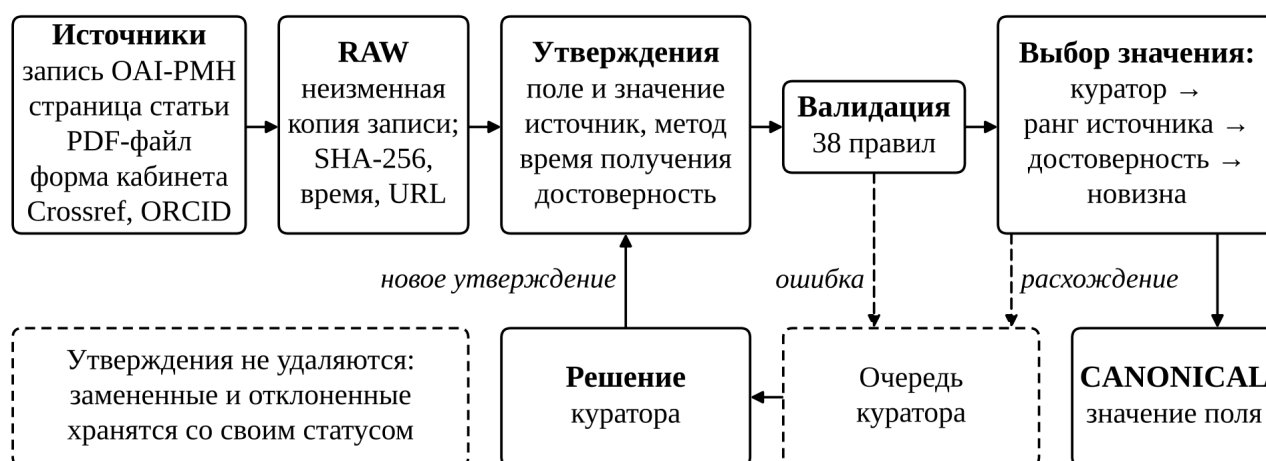


Рис. 1. Модель утверждений и выбора канонического значения поля / Fig. 1. Model of assertions and selection of the canonical field value

Каноническое значение (CANONICAL) выбирается среди утверждений, прошедших валидацию: приоритет имеет решение куратора, затем ранг источника, достоверность и новизна. Ранги задаются по полям: для DOI старший источник – Crossref, где идентификатор и метаданные регистрирует издатель [6], а для заглавия и аннотации – журнал, отвечающий за содержание. Если два старших утверждения по ключевому полю (DOI, заглавие, авторы, год)

существенно расходятся, запись получает статус «требуется рассмотрение».

Нарушения правил валидации (в техническом задании их 38) делятся в основном на ошибки и предупреждения. Ошибки – неверный контрольный символ ISSN или ORCID, недопустимый синтаксис DOI, совпадение DOI у разных работ, расхождение года публикации и года выпуска – в ручном

канале блокируют размещение записи, а в автоматическом переводят ее в статус «требуется рассмотрения»; предупреждения запись не блокируют, но учитываются в индексе качества метаданных (ИКМ). Для автоматически собранных записей так реализуется принцип «не отклонять, а маркировать»: дефект остается видимым, а журналу направляется запрос на исправление конкретного поля.

Таблица 1

Измерения качества метаданных и механизмы их контроля / **Table 1.** Metadata quality dimensions and their control mechanisms

Измерение	Механизм контроля	Показатель
Полнота	Профиль обязательных и рекомендуемых полей; данные со страницы статьи и из PDF	Составляющие C ₁ , C ₂ , R индекса (1)
Корректность	Правила валидации: контрольные символы ISSN и ORCID, синтаксис DOI, даты, страницы	Составляющая V индекса (1)
Согласованность	Сопоставление утверждений, сверка с Crossref, ORCID и ISSN Portal	Составляющая A индекса (1)
Уникальность	Правила дедупликации, очередь куратора, эталонные выборки	Точность ≥ 99,5 %; полнота ≥ 95 %
Своевременность	Ежедневный инкрементальный сбор, статусы источников	Задержка: автосбор ≤ 24 ч, кабинет ≤ 15 мин
Прослеживаемость	Утверждения со ссылкой на RAW, версии записей, журнал аудита	100 % значений выборки связаны с RAW

Полнота, корректность и согласованность сводятся в ИКМ (0–100):

$$\text{ИКМ} = 40 \cdot C_1 + 20 \cdot C_2 + 15 \cdot V + 15 \cdot A + 10 \cdot R, \quad (1)$$

где C₁ и C₂ – доли заполненных обязательных и рекомендуемых полей; V – доля полей без ошибок валидации; A – доля имеющихся идентификаторов (DOI, ORCID, ISSN), подтвержденных внешним источником; R – доля разобранных позиций списка литературы. Все составляющие вычисляются без экспертного просмотра записей, что соответствует подходам к автоматической оценке качества метаданных [7]. ИКМ не является рейтингом: он служит для адресных рекомендаций редакциям и упорядочения кураторских очередей.

Своевременность контролируется по источникам: статус сбора по протоколу OAI-PMH – «зеленый» при успешном последнем сеансе и доле ошибочных записей не более 5 %, «желтый» – при большей доле ошибок или частично выполненном сеансе, «красный» – после трех неудачных сеансов подряд или при отсутствии обновлений более 7 дней. При подключении журнала пробный сбор 100 записей дает отчет о соответствии – какие поля извлекаются и откуда дополнить недостающие; сбои следует отделять от пробелов в данных.

Уникальность обеспечивается консервативной дедупликацией: записи объединяются автоматически лишь в двух случаях – при совпадении DOI, если журнал и год не расходятся, и при совпадении журнала, года и первой страницы (номера статьи), если сходство заголовков не ниже 0,95, а совпадение авторов – не ниже 0,8. Пары с интегральной оценкой 0,80–0,97 передаются куратору, с меньшей – не связываются. Такое разбиение на зоны автоматического связывания, ручной проверки и отказа восходит к модели Феллеги – Сантера [8]; точность объединения (≥ 99,5 %) и полнота выявления дубликатов (≥ 95 %) ежеквартально проверяются на эталонных выборках.

4. ЗАКЛЮЧЕНИЕ

Conclusion

Таким образом, в предложенной модели качество метаданных – не итоговый фильтр, а свойство единого слоя данных: неизменный исходный слой, утверждения с указанием

качества оценивается по шести измерениям (табл. 1), включающим классические характеристики качества – полноту, корректность (точность) и своевременность [2]; набор близок к параметрам качества метаданных Т. Брюса и Д. Хиллманн, где происхождение – отдельный параметр [3].

происхождения, ранги источников, маркирующая валидация, индекс качества и консервативная дедупликация сохраняют охват и обеспечивают проверяемость каждого значения. Практическая ценность модели – в адресных рекомендациях журналам и проверяемом происхождении данных; хранение исходных записей и PDF-файлов вместе с версионированием алгоритмов позволяет пересчитывать показатели без повторных запросов к журналам. Дальнейшая работа – калибровка весов индекса и порогов сопоставления по данным пилотного подключения 20–30 журналов.

DECLARATIONS

Declarations

Author contributions CRediT

Conceptualization; Methodology; Investigation; Data curation; Writing – original draft; Writing – review & editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Ethics approval

The study is a documentary and system analysis of the design documents of an information platform; it involved no human participants or animals and required no ethics committee approval.

Informed consent

All authors have read and approved the final version of the manuscript and have given their consent to its submission and publication. The study involved no human participants and no personal data.

Data availability statement

The data analysed in the study are presented in the text, figure and table of this article; further details are available from the authors on reasonable request.

Conflict of interest

The authors declare no conflict of interest.

Generative AI disclosure

Generative AI (Claude, Anthropic) was used in the editorial preparation of this article: for language editing, translation of the title, abstract and keywords between Uzbek, Russian and English, preparation of the extended English summary from the authors' text, and formatting of the

reference list. No data, results or conclusions were generated by AI. The authors are responsible for the content of the publication.

EXTENDED SUMMARY IN ENGLISH

Kengaytirilgan inglizcha xulosa / Расширенный реферат на английском языке

Purpose and research problem

More than 200 scientific journals are published in Uzbekistan, but information about their articles is scattered across editorial websites, and data on researchers' publications for attestation and reporting are still collected manually. National bibliographic databases with full coverage of a country's scientific output may be better suited for bibliometric analysis than commercial international indexes, especially in the social sciences and humanities. Any aggregator inevitably inherits the defects of its sources — missing affiliations and reference lists, variant spellings of names and inconsistent field values. Rejecting incomplete records reduces coverage, whereas accepting them uncritically lowers reliability. If data quality is understood as fitness for use by data consumers, quality control has to accompany a record along its whole path from the source to the service. The aim of the article is to substantiate a model of a unified bibliographic metadata layer for the National Scientific Publications Platform of Uzbekistan in which the provenance of values and the results of quality checks are properties of the data themselves.

Materials and methods

The material of the study was the concept and the technical specification of the first stage of the National Scientific Publications Platform. Using documentary and system analysis, the authors identified the quality control mechanisms applied at each stage of the record life cycle and compared them with generally accepted dimensions of data quality (Wang and Strong) and metadata quality (Bruce and Hillmann). The model was described in terms of record intake channels, storage of source records, field-level assertions with provenance, selection of canonical values, validation rules, quality dimensions and indicators, monitoring of harvesting sources and deduplication.

Results and discussion

Article records come only from journals: automatically, via the OAI-PMH protocol or by crawling article pages with citation_* meta tags, or manually in the journal's account, with fields pre-filled from the PDF file and by DOI. External sources (Crossref, ORCID, ISSN Portal) do not create records but verify and enrich them. Each record is kept in an immutable source layer (RAW) with a SHA-256 hash and the time and address of retrieval, and a record deleted at the source is marked rather than erased, as in OAI-PMH. The key element of the model is the shift from the record as a whole to assertions about the values of individual fields: an assertion fixes the field, value, source, method, confidence, time and a link to the source record, which links each value to the activity and agent that produced it in terms of the PROV-O provenance ontology. The canonical value is selected among validated assertions by curator decision, then by source rank, confidence and recency; ranks are defined per field (Crossref for the DOI, the journal for the title and abstract), and a substantial disagreement between two senior assertions on a key field sends the record for review. The 38 validation rules of the technical specification are divided mainly into errors — invalid ISSN or ORCID check digits, invalid DOI syntax, the same DOI for different works, a mismatch between the publication year and the issue year — which block manually entered records and flag automatically harvested ones, and warnings, which do not block a record but are counted in the metadata quality index, following the principle "mark, do not reject". Quality is assessed in six dimensions — completeness, correctness, consistency, uniqueness, timeliness and traceability — each with its own control mechanism and indicator (Table 1). Completeness, correctness and consistency are combined into a metadata quality index (MQI, 0–100): $MQI = 40 \cdot C1 + 20 \cdot C2 + 15 \cdot V + 15 \cdot A + 10 \cdot R$, where C1 and C2 are the shares of filled mandatory and recommended fields, V is the share of fields without validation errors, A is the share of identifiers confirmed by an external source and R is the share of parsed reference list items. All components are calculated without expert review of records, and the index serves for targeted recommendations to editors

and for ordering curator queues rather than as a rating. Timeliness is monitored by source status (green, yellow or red depending on the success of the last harvesting session, the share of erroneous records and the time since the last update), and a trial harvest of 100 records when a journal is connected produces a compliance report. Uniqueness is ensured by conservative deduplication: records are merged automatically only when DOIs match and the journal and year agree, or when the journal, year and first page match with a title similarity of at least 0.95 and an author match of at least 0.8; pairs with a combined score of 0.80–0.97 are passed to a curator, following the Fellegi–Sunter model, and merge precision ($\geq 99.5\%$) and duplicate recall ($\geq 95\%$) are checked quarterly on reference samples.

Conclusion and contribution

In the proposed model, metadata quality is not a final filter but a property of the unified data layer: the immutable source layer, assertions with provenance, source ranks, marking validation, the quality index and conservative deduplication preserve coverage and make every value verifiable. The practical value of the model lies in targeted recommendations to journals and in verifiable data provenance; storing source records and PDF files together with versioned algorithms allows the indicators to be recalculated without repeated requests to journals. Further work will calibrate the index weights and matching thresholds using data from a pilot connection of 20–30 journals.

СПИСОК ЛИТЕРАТУРЫ

References (original language)

- [1] Sile L. [et al.] Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey // *Research Evaluation*. – 2018. – Vol. 27, No. 4. – P. 310–322. – DOI: 10.1093/reseval/rvy016.
- [2] Wang R. Y., Strong D. M. Beyond accuracy: What data quality means to data consumers // *Journal of Management Information Systems*. – 1996. – Vol. 12, No. 4. – P. 5–33. – DOI: 10.1080/07421222.1996.11518099.
- [3] Bruce T. R., Hillmann D. I. The continuum of metadata quality: defining, expressing, exploiting // *Metadata in Practice* / ed. by D. I. Hillmann, E. L. Westbrooks. – Chicago: American Library Association, 2004. – P. 238–256.
- [4] The Open Archives Initiative Protocol for Metadata Harvesting. Version 2.0. – URL: <https://www.openarchives.org/OAI/openarchivesprotocol.html> (дата обращения: 07.10.2026).
- [5] PROV-O: The PROV Ontology. W3C Recommendation, 30 April 2013. – URL: <https://www.w3.org/TR/prov-o/> (дата обращения: 07.10.2026).
- [6] Hendricks G., Tkaczyk D., Lin J., Feeney P. Crossref: The sustainable source of community-owned scholarly metadata // *Quantitative Science Studies*. – 2020. – Vol. 1, No. 1. – P. 414–427. – DOI: 10.1162/qss_a_00022.
- [7] Ochoa X., Duval E. Automatic evaluation of metadata quality in digital repositories // *International Journal on Digital Libraries*. – 2009. – Vol. 10, No. 2–3. – P. 67–91. – DOI: 10.1007/s00799-009-0054-4.
- [8] Fellegi I. P., Sunter A. B. A theory for record linkage // *Journal of the American Statistical Association*. – 1969. – Vol. 64, No. 328. – P. 1183–1210. – DOI: 10.1080/01621459.1969.10501049.

REFERENCES (ROMAN SCRIPT)

Transliterated list — required by Scopus and Web of Science

- [1] Sile L. [et al.] Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey // *Research Evaluation*. – 2018. – Vol. 27, No. 4. – P. 310–322. – DOI: 10.1093/reseval/rvy016.
- [2] Wang R. Y., Strong D. M. Beyond accuracy: What data quality means to data consumers // *Journal of Management Information Systems*. – 1996. – Vol. 12, No. 4. – P. 5–33. – DOI: 10.1080/07421222.1996.11518099.
- [3] Bruce T. R., Hillmann D. I. The continuum of metadata quality: defining, expressing, exploiting // *Metadata in Practice* / ed. by D. I.

- Hillmann, E. L. Westbrooks. – Chicago: American Library Association, 2004. – P. 238–256.
- [4] The Open Archives Initiative Protocol for Metadata Harvesting. Version 2.0. – URL: <https://www.openarchives.org/OAI/openarchivesprotocol.html> (accessed: 07.10.2026).
- [5] PROV-O: The PROV Ontology. W3C Recommendation, 30 April 2013. – URL: <https://www.w3.org/TR/prov-o/> (accessed: 07.10.2026).
- [6] Hendricks G., Tkaczyk D., Lin J., Feeney P. Crossref: The sustainable source of community-owned scholarly metadata // Quantitative Science Studies. – 2020. – Vol. 1, No. 1. – P. 414–427. – DOI: 10.1162/qss_a_00022.
- [7] Ochoa X., Duval E. Automatic evaluation of metadata quality in digital repositories // International Journal on Digital Libraries. – 2009. – Vol. 10, No. 2–3. – P. 67–91. – DOI: 10.1007/s00799-009-0054-4.
- [8] Fellegi I. P., Sunter A. B. A theory for record linkage // Journal of the American Statistical Association. – 1969. – Vol. 64, No. 328. – P. 1183–1210. – DOI: 10.1080/01621459.1969.10501049.