

Improving the Quality of Scientific Repositories Through Artificial Intelligence

Brian Bales

ORCID: 0000-0003-1452-2785

Abstract

Scientific repositories depend on accurate and consistent metadata, but maintaining that quality becomes increasingly difficult as collections grow and staff resources remain constrained. Large language models offer a means of extending automated quality assurance beyond conventional structural validation while still operating within explicitly defined rules.

This paper examines an artificial intelligence quality assurance agent developed for the International Nuclear Information System (INIS) of the International Atomic Energy Agency. The agent reviews newly registered records according to a defined set of bibliographic, indexing, and repository-quality rules. It corrects only clearly identifiable problems, such as improperly capitalized titles, incomplete affiliations, missing DOI identifiers in defined circumstances, or clearly unrelated subject descriptors. Other issues—including questionable resource types, misplaced publication information, indexing uncertainty, and unusual or historically sensitive content—are referred for human review.

The approach deliberately distinguishes correction from recommendation. The language model is instructed not to improve acceptable metadata or infer changes beyond its defined scope. Operational experience indicates that this constrained use of artificial intelligence can provide systematic quality control across large numbers of records while directing professional attention toward cases that genuinely require human judgement.

Keywords: artificial intelligence; scientific repositories; metadata quality; subject indexing; controlled vocabularies; large language models; quality assurance; information management

1. Introduction

Scientific repositories depend on metadata to make their collections discoverable, interpretable, and reusable. Errors in titles, affiliations, publication information, document types, identifiers, or subject descriptors can reduce retrieval quality and diminish confidence in the repository.

Repository systems already perform many forms of automated validation. Required fields can be checked for presence, values can be tested against controlled lists, and formats can be validated programmatically. Such checks are highly reliable where a rule

can be stated precisely. They are less effective when determining whether apparently valid metadata is nevertheless questionable in context.

Large language models provide another form of automated review. They can examine several elements of a record simultaneously, interpret natural-language metadata, and identify cases that warrant correction or professional attention. Their usefulness, however, depends heavily on how much authority they are given. A general instruction to “improve the metadata” would permit unnecessary changes and introduce an unacceptable level of uncertainty.

The INIS quality assurance agent therefore follows a deliberately restrictive approach. Its instruction is essentially to verify rather than improve. A field that is already acceptable is left unchanged. Corrections are permitted only for specifically defined conditions, while other concerns are returned as recommendations for human review.

The agent is used with newly registered records in the International Nuclear Information System (INIS), the IAEA's repository of scientific and technical literature concerning the peaceful uses of nuclear science and technology and related fields. Its role is not to replace cataloguing, indexing, or professional judgement. Instead, it provides an additional quality-control layer capable of reviewing every new record consistently and identifying a relatively small number of records requiring attention.

2. Quality Assurance in INIS

INIS records contain structured bibliographic and subject metadata including titles, personal and organizational authors, affiliations, abstracts, publication information, dates, identifiers, resource types, and controlled subject descriptors.

Many potential problems can be detected deterministically. A record identifier, for example, can be tested against a regular expression, and the presence or absence of required metadata can be established without artificial intelligence. These checks remain preferable whenever the rule is unambiguous.

Other checks require limited interpretation. An institutional affiliation may be recognizable but incomplete because it lacks a city or country. A descriptor may be a valid term from the INIS Thesaurus yet be clearly unrelated to the subject of the record. A technically valid resource type may appear inconsistent with the bibliographic characteristics of a publication. Older records may contain terminology or scientific ideas that are historically important but require contextual warning.

The QA agent brings these different checks together within a single review process. Importantly, it does not have unrestricted authority to modify records. The instructions enumerate the fields that may be examined, the conditions under which corrections may be proposed, and the issues that should instead be reported for review.

This produces three possible outcomes for a metadata element:

1. No action, when the value is already acceptable.
2. Correction, when the instructions define a clear error and the appropriate change is sufficiently certain.
3. Recommendation for human review, when an issue has been detected but should not be resolved automatically.

The distinction considerably reduces the risk of unnecessary intervention. An acceptable field is not rewritten simply because the model could produce an alternative version.

3. Design and Operation of the QA Agent

The QA agent operates as a post-processing quality assurance step after records have been registered and indexed by NADIA. Each day, the agent processes all newly eligible records added since the previous run. For each record, the associated metadata and document content are supplied to a large language model together with detailed instructions defining the checks it may perform and the issues it should report. The results are consolidated into a daily email report for human review.

The instructions emphasize restraint:

Do not add or change anything unless it is strictly necessary to fix a clear error. The goal is not to improve, but to verify correctness and identify problems.

3.1 Titles and Abstracts

Title correction is intentionally narrow. Titles written entirely in uppercase or lowercase are converted to appropriate English title case. Titles already in sentence case or title case remain unchanged. Bold HTML markup is removed, but other formatting is preserved.

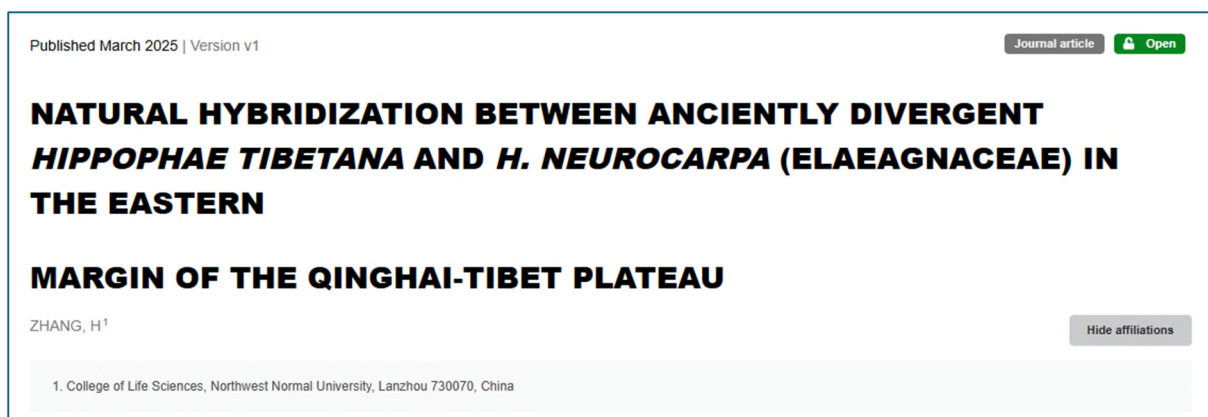


Figure 1: Title Provided by Contributor, in all-caps

In this example, the contributor provided a title in all-caps, bolded and with an odd line break. This is probably because of cut and paste from a PDF. The QA Agent has reformatted the title (as well as the author) in title case. Note that it left the species names in italics.

Published March 2025 | Version v1

Journal article

Open

Natural Hybridization Between Anciently Divergent *Hippophae tibetana* and *H. neurocarpa* (Elaeagnaceae) in the Eastern Margin of the Qinghai-Tibet Plateau

Zhang, H¹

Show affiliations

Figure 2: Title (and Author) Corrected by QA Agent

Abstract handling is similarly constrained. The agent does not generate a missing abstract from the document. A correction is made only when the English abstract is absent and an abstract in another language is already present in the metadata. In that situation, the existing abstract can be translated into English.

3.2 Identifiers

The agent checks for one specific identifier problem: a valid DOI appearing in the notes but missing from the structured related-identifiers field. When this occurs, the DOI can be added in the prescribed format.

This is a good example of a narrowly defined AI-supported correction. The agent is not asked to search externally for identifiers or infer a DOI from bibliographic information. It acts only on evidence already contained in the record.

3.3 Affiliations and Organizational Authors

Affiliations and organizational authors can be automatically fixed when they are clearly incomplete or poorly formatted, particularly when a city or country is missing.

For example:

Original affiliation:

Kernforschungsanlage Jülich

Recommended affiliation:

Kernforschungsanlage Jülich GmbH, Jülich, Germany

Country names are normalized using ISO 3166 English short names. Note the affiliations below. Each is missing a city name, and the QA Agent was able to identify the city of each of the universities given. It was even able to distinguish between differing locations of the same university – Chitkara University in Punjab and in Himachal Pradesh – providing the correct city or district for each.

```
"affiliation_corrections": [
  {
    "old_affiliation": "Saveetha Institute of Medical and Technical Sciences, Tamil Nadu, India",
    "recommended_affiliation": "Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India"
  },
  {
    "old_affiliation": "SRM Institute of Science and Technology, Tamil Nadu, India",
    "recommended_affiliation": "SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India"
  },
  {
    "old_affiliation": "Chitkara University, Punjab, India",
    "recommended_affiliation": "Chitkara University, Rajpura, Punjab, India"
  },
  {
    "old_affiliation": "Chitkara University, Himachal Pradesh, India",
    "recommended_affiliation": "Chitkara University, Solan, Himachal Pradesh, India"
  }
]
```

These corrections are useful because incomplete institutional forms may be perfectly valid strings from a technical perspective and therefore escape ordinary schema validation.

3.4 Dates

Date checking is intentionally limited to obvious anomalies. An absurd or missing publication year—for example, a value such as 500, 1900 in an obviously contemporary context, or a null value—can be flagged.

The agent does not attempt to resolve ordinary uncertainty over publication dates unless the problem falls within the defined rule.

3.5 Subject Descriptors

Automated subject indexing is performed by NADIA, a machine-learning system trained on patterns derived from approximately one million previously indexed records. In addition to predicting subject categories and indexing terms from these learned patterns, NADIA performs a grid search for semantic similarity across the title, abstract, subject classification and candidate indexing terms. This allows the system to identify relevant terms that might not otherwise be predicted from the training data alone, but it can also introduce false associations.

One recurring example involves isotopes. Semantic similarity can occasionally cause an isotope designation to be selected because of an incidental resemblance to other metadata rather than because the isotope is actually discussed in the document. A particularly persistent example occurs when subject category S36, Materials Science, is assigned: NADIA frequently also proposes the indexing term *sulfur-36*, despite the isotope having no relevance to the document.

Patterns such as these illustrate an important role for the QA agent. Rather than duplicating NADIA's indexing process, the agent examines the resulting metadata in the context of the document and identifies terms that appear unsupported or inconsistent.

3.6 Repository-Rule Recommendations

A separate group of checks produces recommendations rather than direct metadata corrections.

Examples include:

- a conference paper or book with page numbers stored in the journal field rather than the imprint field;
- a journal article containing an INIS lead-record identifier;
- a lead-record identifier that does not conform to the required identifier pattern;
- missing country-of-publication or country-of-input information;
- transliterated Russian appearing where a Cyrillic additional title is expected;
- a resource type that appears inconsistent with the publication;
- and records marked “Insufficient Index Confidence” by the NADIA automated indexing system.

In the last case, the agent reports:

“NADIA insufficient confidence, please index manually.”

These are deliberately surfaced for a person to evaluate rather than automatically resolved.

In the excerpt below from the daily email, the QA agent has made various suggestions and identified issues for further human checking. In the examples, most of the suggestions were implemented. However, the out-of-scope recommendation on the chemical rocket engine was found to be in scope and not removed.

- <https://inis.iaea.org/records/pzsb2-gfq19>
 - Record lacks country of publication (iaea:country); recommend adding "Brazil"
 - For conference papers, page numbers should be moved from journal:journal to iaea:imprint
- <https://inis.iaea.org/records/537we-9cq06>
 - Consider changing resource type to 'Publication/Conference paper'
 - For conference papers, page numbers should be under iaea:imprint rather than journal:journal
- <https://inis.iaea.org/records/jkeh7-6pk94>
 - Out of scope: Focuses solely on chemical rocket engine regenerative cooling with supercritical hydrogen, no nuclear context.
- <https://inis.iaea.org/records/3emcj-t2521>
 - Publication date is in the future: 2026-10
- <https://inis.iaea.org/records/tpww9-qkk14>
 - Out of scope: Focuses solely on insect mass rearing without nuclear techniques.
- <https://inis.iaea.org/records/1n6bg-95s39>
 - Consider changing resource type to 'Publication/Report'

Figure 3: Excerpt from QA email, identifying issues for human checking

The model identifies the issue, but the final bibliographic decision remains with the information professional.

3.7 Scope and Trustworthiness

The agent also performs a small number of higher-level quality checks.

It determines whether a record appears to fall within the defined scope of INIS. The scope is deliberately broad, encompassing nuclear science and technology together with areas including materials science, nanotechnology, medical applications, basic and theoretical physics, general energy policy and technology, climate change, and environmental science. Only clearly out-of-scope material is flagged.

The agent can also identify records containing apparent pseudoscience or fringe claims. Such content is not automatically corrected or rejected; instead, it is marked for human review.

- <https://inis.iaea.org/records/1eqt3-szd48>
 - Out of scope: Focuses solely on synthetic biology and biocomputation, not nuclear science.
 - The claim that genetically engineered bacteria can identify prime numbers and vowels appears to be pseudoscientific and requires verification.

Figure 4: Example of a record identified as possible pseudoscience

Similarly, older terminology, ideas, or methods can be identified as requiring historical context. This allows potentially important historical literature to remain represented accurately while alerting reviewers that contemporary interpretation may require qualification.

These functions illustrate a particularly appropriate role for an LLM: recognizing circumstances that warrant professional attention without giving the model authority to rewrite or suppress the record.

4. Conclusion

The INIS quality assurance agent demonstrates a practical use of large language models in scientific information management. It operates not as a replacement for established cataloguing and indexing processes, but as an additional layer of review applied after automated indexing has taken place.

This distinction is important. NADIA performs large-scale automated subject indexing using patterns learned from approximately one million previously indexed records, supplemented by semantic similarity techniques that can identify useful terms beyond those suggested directly by the training data. These methods increase coverage, but they can also introduce characteristic false associations. The QA agent addresses this different problem: it examines the resulting metadata in the context of the associated document and asks whether selected elements appear to be supported, consistent and correct.

The system therefore combines several complementary forms of quality control. Deterministic validation handles structural and rule-based checks. Machine learning and semantic methods support automated indexing. The large language model provides contextual assessment across metadata fields and document content, while information professionals retain responsibility for ambiguous, policy-dependent and specialist decisions.

The value of the QA agent lies in this constrained role. It is instructed not to improve records broadly or make unnecessary changes, but to identify clear problems, make only those corrections that can be justified with confidence, and raise uncertain cases for human review. Its usefulness derives not from unrestricted autonomy, but from carefully defined authority.

Because the agent runs each day across newly eligible records after NADIA indexing, systematic quality review can be applied continuously at a scale that would be difficult to sustain manually. Straightforward problems can be addressed quickly, while the daily exception report concentrates professional attention on records most likely to require judgement.

For scientific repositories, this layered approach may represent one of the more immediately useful applications of generative artificial intelligence. Rather than replacing cataloguers, indexers or existing machine-learning systems, large language models can provide a further layer of scrutiny around them: checking automated decisions against the evidence in the resource, identifying anomalies that other methods may miss, correcting only what can safely be corrected, and referring the remainder to a human.