# Reasoning models for adaptive information extraction in scientific documents

Rashid Turgunbaev
Kokand State University

**Abstract:** The exponential growth of scientific publishing has made accurate and efficient metadata extraction a crucial task for enabling search, retrieval, and knowledge management in scholarly communication. However, the diversity of journal formats and evolving publication practices pose significant challenges to traditional rule-based extraction systems. This article explores reasoning models as a foundation for adaptive metadata extraction in scientific documents. It examines the strengths and limitations of rule-based, case-based, probabilistic, and hybrid reasoning approaches, showing how they can be integrated to support robust and flexible extraction processes. An adaptive workflow is described in which annotated examples guide the generation of extraction rules that are refined through iterative reasoning strategies. The article argues that reasoning models not only improve the accuracy and scalability of metadata extraction but also provide interpretability, adaptability, and resilience to variations in document structures. Future directions point toward hybrid systems that combine reasoning with advances in machine learning and natural language processing, creating intelligent infrastructures for the dynamic landscape of scientific publishing.

**Keywords:** reasoning models, information extraction, scientific documents, metadata extraction, adaptive systems, scholarly communication

The rapid growth of scientific publishing in recent decades has created both opportunities and challenges for the research community. Vast quantities of articles are being produced daily across a wide range of disciplines, and while this expansion enriches the collective body of knowledge, it also introduces considerable complexity in terms of access, organization, and utilization of information. One of the fundamental tasks that enables effective navigation of this knowledge is the extraction of metadata from scientific documents. Metadata, which includes elements such as title, authorship, affiliation, abstract, keywords, and references, provides the structural backbone for indexing, retrieval, and interoperability across digital libraries, repositories, and citation databases. The process of metadata extraction, however, is far from straightforward, as scientific articles are published in a wide array of formats, templates, and styles. Traditional approaches that rely on rigid rules or heuristics often fail to accommodate the diversity of journal layouts and the evolving nature of publication practices. In this context, reasoning models provide a promising foundation for developing adaptive methods capable of generalizing across different document structures while preserving accuracy and robustness.

Reasoning, broadly conceived, is the ability to infer conclusions from available data, context, and prior knowledge. In artificial intelligence and computer science, reasoning models underpin many approaches to problem solving, ranging from deductive logic to probabilistic inference. When applied to information extraction, reasoning serves as the mechanism that bridges raw document content with structured metadata representation. The central challenge is to design reasoning models that are not only accurate in recognizing metadata components in a single article but also adaptive enough to generalize from one annotated example to other articles from the same or even different journals. This requires a synthesis of multiple reasoning paradigms, since no single model is sufficient to handle the complexity and variability of scientific documents.

Rule-based reasoning has historically been one of the most prominent approaches to information extraction. In this paradigm, extraction is guided by explicit rules that link document features to metadata categories. For instance, a rule might specify that the title corresponds to the first text block on the first page with the largest font size, or that the abstract is the segment of text following the word "Abstract" and preceding the section titled "Keywords." Such rules, when carefully crafted, can be highly effective in journals that adhere to strict templates. The strength of rule-based reasoning lies in its transparency and interpretability: each extraction decision can be traced back to a clear logical condition. However, the rigidity of this method often becomes a limitation when confronted with variations across journals, or even within a single journal over time as formatting guidelines evolve. Moreover, rule-based systems struggle with noisy data such as scanned PDFs, where text positioning and font features may be distorted during digitization.

Case-based reasoning offers an alternative perspective by emphasizing analogical thinking. In this paradigm, the system learns from a previously annotated document and applies similar patterns to new documents. For example, if the user marks the title, authors, and abstract in one sample article, the system can analyze the structural relationships and textual markers associated with those annotations, and then attempt to reproduce similar extractions in subsequent articles. This approach mirrors the way humans often learn to interpret new document formats: by recalling a previously encountered example and adapting its interpretation to the new case. The advantage of case-based reasoning is its adaptability; it does not require the manual specification of detailed rules but rather infers them from experience. Nevertheless, case-based reasoning can face challenges in generalization, as the similarity between cases may be superficial, and without additional mechanisms it may misinterpret documents with subtle structural differences.

Probabilistic reasoning and statistical approaches introduce another dimension of flexibility by modeling uncertainty directly. Instead of relying on strict rules or direct analogies, probabilistic models assign likelihoods to various hypotheses about document structure. For instance, a Bayesian model may infer that a certain text block is most likely the abstract because it is preceded by the word "Abstract" with high probability, appears near the top of the first page, and contains typical abstract-length sentences. Similarly, probabilistic models can handle ambiguities in identifying author affiliations, where multiple candidate segments of text may exist. The ability to quantify uncertainty makes probabilistic reasoning particularly valuable in noisy or heterogeneous document environments. However, probabilistic models typically require substantial training data to estimate distributions accurately, which may not always be available for specific journals or domains.

Non-monotonic reasoning also plays a role in adaptive extraction by allowing systems to revise conclusions when new evidence arises. For example, an initial hypothesis may identify a section of text as the abstract, but upon encountering a later segment labeled "Extended Abstract" or "Summary," the system may retract its initial conclusion and update the metadata assignment. This dynamic flexibility is critical in dealing with the evolving conventions of publishing, where new section labels and formatting variations are continually introduced.

Hybrid reasoning models emerge as a particularly powerful solution for adaptive information extraction, as they integrate multiple paradigms to leverage their complementary strengths. A hybrid system might combine rule-based reasoning for highly predictable sections such as references, case-based reasoning for flexible elements such as author affiliations, and probabilistic reasoning for ambiguous contexts like distinguishing acknowledgments from author contributions. By layering reasoning strategies, the system achieves greater robustness across a wide range of document types. Hybrid models also facilitate incremental learning, where rules derived from one journal can be adapted and extended to new formats with minimal user intervention.

An illustrative workflow for such an adaptive reasoning model begins with user annotation of a single article from a given journal. The system records both textual and structural features associated with each annotated metadata element, including position, font properties, lexical markers, and relational cues. Using these observations, the system generates an initial set of candidate rules and probabilistic models. These rules are not fixed but are stored in a flexible representation that allows refinement over time. When applied to a new article, the system tests multiple reasoning strategies: it first applies structural rules, then checks for analogical patterns from the annotated case, and finally employs probabilistic inference to resolve uncertainties. If the extraction results are inconsistent or ambiguous, non-monotonic reasoning allows the system to revise its conclusions, possibly requesting minimal user feedback for disambiguation. Over successive iterations, the system refines its reasoning models, improving accuracy while reducing the need for manual input.

The effectiveness of reasoning models for adaptive information extraction can be evaluated through both accuracy metrics and practical usability. Accuracy is typically measured by comparing the automatically extracted metadata with ground truth annotations, using metrics such as precision, recall, and F1-score. However, usability is equally important: a system that extracts metadata with moderate accuracy but requires extensive manual corrections may be less valuable than a system that achieves slightly lower accuracy but minimizes human intervention through adaptive learning. Furthermore, scalability is a crucial consideration, as scientific publishers release thousands of articles monthly, and any extraction system must operate efficiently across large datasets.

The implications of reasoning-based adaptive extraction extend far beyond metadata capture. Once robust metadata is available, it enables powerful applications in bibliometrics, knowledge graph construction, trend analysis, and scholarly communication. For example, reliable author affiliation data supports studies of institutional collaboration networks, while accurate keyword extraction facilitates topic modeling and thematic mapping of research landscapes. By grounding these processes in reasoning models, we can ensure that metadata is not only extracted accurately but also interpretable, traceable, and adaptable to new contexts.

Future directions in this domain point toward deeper integration of reasoning with machine learning and natural language processing. Large language models, for instance, demonstrate remarkable capabilities in parsing unstructured text, but they often lack transparency and control in specialized tasks such as metadata extraction. By embedding them within reasoning frameworks, it becomes possible to balance the generative power of such models with the rigor and adaptability of structured reasoning. Similarly, advances in document layout analysis and computer vision open opportunities for multi-modal reasoning, where textual, visual, and structural cues are combined to produce richer metadata extraction.

In conclusion, reasoning models provide a vital foundation for adaptive information extraction in scientific documents. Rule-based reasoning offers precision and interpretability, case-based reasoning contributes adaptability, probabilistic reasoning introduces uncertainty management, and hybrid approaches synthesize these strengths to deliver robust performance. The central challenge lies in designing systems that can learn from minimal annotation, generalize across diverse publication formats, and evolve alongside the ever-changing landscape of scholarly communication. By advancing reasoning models in this direction, the research community can not only enhance metadata extraction but also contribute to the broader vision of intelligent, adaptive, and accessible scientific knowledge infrastructures.

## References

1. Azimjonov, J., & Alikhanov, J. (2018). Rule based metadata extraction framework from academic articles. arXiv preprint arXiv:1807.09009.

2. Khankasikam, K. (2014). Thai Metadata Extraction by Using Case-based Reasoning. Naresuan University Engineering Journal, 5(2), 21–27.

3. Liu, P., Gao, W., Dong, W., Ai, L., Gong, Z., Huang, S., Li, Z., Hoque, E., Hirschberg, J., & Zhang, Y. (2024). A Survey on Open Information Extraction from Rule-based Model to Large Language Model. Findings of the Association for Computational Linguistics: EMNLP 2024.

4. Turgunbaev, R. (2021). Keysga asoslangan fikrlash va uni akademik metama'lumotlarni avtomatik ekstraksiya qilishda tadbiq qilinishi. Science and Education, 2(9), 129-144.

5. Boukhers, Z., & Yang, C. (2025). Comparison of Feature Learning Methods for Metadata Extraction from PDF Scholarly Documents. arXiv (January 2025).

6. Turgunbaev, R. (2021). Metadata: features, types and standards. Science and Education, 2(5), 167-175.

7. Atkinson, J., Gonzalez, A., Munoz, M., & Astudillo, H. (2014). Web Metadata Extraction and Semantic Indexing for Learning Objects Extraction. Applied Intelligence, 41, 649–664.

8. Turgunbaev, R., & Elov, B. (2021). The use of machine learning methods in the automatic extraction of metadata from academic articles. International Journal of Innovations in Engineering Research and Technology, 8(12), 72-79.