Toward Research Knowledge Graphs for AI Agents: Automatic Metadata Extraction from Research Notes Using Japanese Open LLMs and the Model Context Protocol — A Proof of Concept on GakuNin RDM

Toshiyuki Kataoka[1,*], Yusuke Komiyama[1], Masaharu Hayashi[1], Hiroyuki Yasuda[2], Mikiko Tanifuji[1] and Kazutsuna Yamaji[1]

[1]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
[2]Chiba Institute of Technology, 2-17-1 Tsudanuma, Narashino, Chiba 275-0016, Japan

Abstract

Metadata generation for papers and research data is being automated by AI agents, and the generation and use of knowledge graphs has come within reach, so that support extends beyond organizing data to higher-level work such as search and analysis across papers and research data. Pre-publication research data, however, are confidential, and the demand is for metadata-generation support by locally executed AI agents; alongside graphs built from public information, closed knowledge graphs built from non-public information will matter. This paper examines the feasibility of such support, conforming to a metadata schema in operational use in a real research project, with AI agents driven by locally executed Japanese open large language models, and verifies the mapping from that schema to ontologies. The target is the pre-symptomatic disease schema operated by a project under Goal 2 of the Moonshot Research and Development Program: 171 fields comprising project metadata and dataset metadata. Thirty semi-structured Japanese research notes written for this verification were used as input, metadata under that schema was extracted from each note three times, and the JSON that GakuNin RDM registers was emitted. Judging 29,320 field-level decisions in total, llm-jp-4.1-32b-a3b-thinking reached a micro-averaged precision of 0.563, recall 0.640 and F1 0.599. F1 divides sharply — 0.800 for project metadata against 0.508 for dataset metadata — and the largest error category, 3,633 cases, is output to fields that should have been empty (over-extraction). As a step toward future work, the 171 fields were mapped to standard vocabularies and the mapping was formally validated. Candidate predicates or classes were assigned to 162 fields (94.7%), and the nine fields with no corresponding vocabulary — ethics review, anonymization, conflict of interest, acknowledgment — are concepts specific to Japanese research-data-management practice. Of the assignments, 115 fields carry a predicate the vocabulary defines and 16 do not. Shapes generated from the mapping were run over the extraction output and over the gold metadata: no instance conforms, the gold produces more violations (849) than the output (680), and only 282 of the output's 1,064 are attributable to the extraction. The shapes detect 46.0% of the project-metadata errors a human marked and 4.1% of the dataset-metadata ones, the gap following from a schema that declares no required and 5 enumerated dataset fields against 109 free-text ones.

Keywords

knowledge graphs, research data management, metadata extraction, large language models, AI agents, GakuNin RDM, Model Context Protocol, JSON-LD

1. Introduction

The management and publication of research data from publicly funded research projects has been institutionalized through domestic initiatives such as the research data ecosystem construction project [1]. NII Research Data Cloud (NII RDC) combines GakuNin RDM (GRDM), JAIRO Cloud (WEKO3), and CiNii Research, covering management, publication, and search [2]. GRDM provides research data management functions and is used by 231 institutions in Japan as of September 2026 [3], supporting closed registration and sharing of pre-publication data among project members. GRDM handles two metadata types: one for publicly funded research projects, and one for pre-symptomatic disease re-

IJCKG 2026 Posters, Demos and Workshops, November 19–21, 2026, Bangkok, Thailand
*Corresponding author.
toshikataoka@nii.ac.jp (T. Kataoka); komiyama@nii.ac.jp (Y. Komiyama); mhaya@nii.ac.jp (M. Hayashi); yasuda.hiroyuki@chibatech.ac.jp (H. Yasuda); tanifuji@nii.ac.jp (M. Tanifuji); yamaji@nii.ac.jp (K. Yamaji)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-06

CEUR-WS.org/Vol-4316/rdm-full-3.pdf

search corresponding to Goal 2, "Comprehensive Pre-symptomatic Disease Database," of the Moonshot Research and Development Program.

Metadata for pre-symptomatic disease research includes fields concerning data derived from human subjects, such as ethics review, informed consent, and anonymization, and must therefore remain under organizational control; this PoC executes all processing, including inference, in an environment under our own administration. Field designs premised on cross-project data sharing have been reported [4, 5]. This has been formalized as the NII RDC application profile (NII RDC AP), whose applicability to field-specific schemas has been verified [6].

Knowledge graphs linking research outputs and data are being advanced in academic fields [7]. What they are built from is information about published results, papers above all. Graphs over data that are still in progress and not yet published are useful in their own right, as a basis for exploration and reuse inside a project, but for such data no source paper exists from which information could be extracted. One way forward is to use the metadata a collaborative project has been maintaining in any case — for search and analysis among its members, and for eventual publication — which makes the registration and management of non-public metadata the more important. That metadata, however, contains many fields, including hierarchical structures, conditional branching, mandatory conditions and input rules, so manual registration imposes a substantial burden on researchers. Reducing the cost at which researchers produce it is therefore not only a question of metadata management but a preparation for building a closed knowledge graph. Non-public research information is moreover often confidential, and automating metadata generation normally presupposes commercial cloud services, whereas pre-symptomatic disease metadata requires an execution environment that does not transmit data externally.

This paper reports a PoC that extracts metadata from researchers' experimental research notes, using a Japanese open LLM, according to a metadata schema for pre-symptomatic disease research, and outputs it in the JSON format used for GRDM registration. The PoC comprises a schema retrieval tool and a metadata registration tool implemented as an MCP server, together with four AI agents. Contributions: (1) for a schema in actual use on a domestic public infrastructure, we evaluated at field level the extraction feasibility of a fully locally executed Japanese open LLM, over 30 research notes extracted three times each and 29,320 field-level judgements; (2) we report accuracy separately for project and dataset metadata and identify over-extraction — a key emitted for a field the note does not support — as the dominant error mode, and the one that overwrites a stored value on registration; (3) we compared the llm-jp-3 and llm-jp-4 generations on identical notes under a schema both generations can read, quantifying the difference between them; and (4) we mapped the schema's 171 fields to schema.org, DCMI Metadata Terms, DCAT, PROV-O, CSVW and the GA4GH Data Use Ontology, validated those assignments formally, and ran shapes derived from the mapping over both the extraction output and the gold metadata, separating the errors validation can catch from those it cannot. Graph construction is outside scope; requirements for connecting extracted metadata to a graph are discussed in Section 8.

2. Related Work and Positioning of This Study

Automatic metadata extraction progressed from machine-learning header classification [8] and CRF reference-string analysis [9] to GROBID [10], and LLM-based extraction with verification [11], all targeting formalized literature, not private research notes. Other methods—retrieval-augmented metadata generation [12]—center on CEDAR and BioSample and often rely on commercial cloud APIs. Experimental-record work—an LLM-equipped materials-science ELN [13]—remains confined to specific ELN products or domains, with little verification linking extraction to repository registration APIs. This study targets GRDM's operational schemas for publicly funded research projects and pre-symptomatic disease research, uses a Japanese open LLM run entirely within an environment under our own administration—important given human-subject data in pre-symptomatic disease research—and verifies schema acquisition and registration separated as MCP tools, upstream of research knowledge graph construction without increasing researcher burden.

3. Subjects and Methods

3.1. Moonshot Goal 2 and the Comprehensive Pre-symptomatic Disease Database Project

The schema originates in the Comprehensive Pre-symptomatic Disease Database Project under Goal 2 of the Moonshot Research and Development Program, spanning species, instruments, and analytical methods, which requires shared data infrastructure [5] and metadata processing design [4]. Two schemas—project-level information common to publicly funded research projects, and pre-symptomatic-disease-specific measurement content—are GRDM registration forms verified against the NII RDC application profile [6]. Their fields, hierarchical structures, and conditional branching burden manual registration.

3.2. Target Metadata Schema

The target schema is the project and dataset metadata for Moonshot Goal 2 (pre-symptomatic disease database) implemented in GRDM, comprising paired Japanese/English fields, with required fields marked by a trailing "*". Enumerated fields (e.g., dataset research field = "Shizen-kagaku ippan" [general natural science], "Raifu-saiensu" [life science], "Joho-tsushin" [information and communication], "Kankyo" [environment], "Nanoteku" [nanotechnology]) coexist with free-text fields, and fields such as keywords and creators use array or nested structures; some fields concatenate Japanese and English values with a pipe symbol (e.g., "Raifu-saiensu|Life Science"). Two schema forms were prepared: a JSON Schema (draft-07) describing hierarchy, conditional branching, required conditions, and input rules, and a Markdown format expressing the same content as plain text; AI-processing instructions were added to each field's description, alongside GRDM's existing descriptions. Metadata comprises project, dataset, extended, and file metadata; extended metadata ingests arbitrary files, and file metadata shares field composition with dataset metadata but targets individual data items rather than collections. Japanese label strings quoted in this paper are romanized; the schema stores them in Japanese script.

Mechanically aggregating the registration form definition files yields 345 field definitions, including array/nested parent elements, heading labels, and redeclared items. Excluding these, the substantive field count is 171 (57 project metadata, 114 dataset metadata). Of these, 120 have bilingual titles, 23 are enumerated (100 choice values total), and 14 use regular-expression validation. Biomedical-content fields number only 7 concepts: data type, measurement target, related organ name, similar data-type items, measurement device classification, measurement device name, and image type. Of these, only "analysis target data" is enumerated (10 values); the remaining 6 are free-text. Remaining fields cover administrative, rights, and technical metadata (license, access rights, informed consent, repository, date, contact information, folder structure, file naming conventions). All choice values are bilingual label strings; no fields take concept URIs from external vocabularies as values.

3.3. Research-Note Dataset

Research notes were typologized by (1) degree of structuring (structured / semi-structured / unstructured) and (2) schema conformance (high / medium / low). (1) distinguishes key–value, semi-structured, and free-text-only notes; (2) reflects how closely key/item strings match schema field names. The model comparison of Section 4.1 used a single "semi-structured × medium schema conformance" note from Moonshot program sample data, describing gene expression in high-fat-diet metabolic mice (C57BL/6) covering pre-symptomatic and post-onset metabolic syndrome across 13 organs, 16 time points, and 10 individuals.

For the evaluation in Section 4.5 a dataset of 30 research notes was constructed. The project's real research notes cannot be released, so the notes were built from 100 published papers by Moonshot Goal 2 researchers under a procedure designed so that no note can be linked back to a source paper, while the content still reflects actual research practice.

Generation. The corpus was indexed into a document-frequency (df) table over 50,494 terms. Each attribute of a paper—purpose and target, measurement method and device, data format and file organization, analysis procedure, and management and release conditions—was generalized until it could be expressed with vocabulary of $\mathrm{df} \geq 5$, giving an attribute pool per category; terms with $\mathrm{df} \leq 2$ were abstracted one level or replaced by coined words. The generator is given only this pool, a co-occurrence constraint table recording which measurement methods, target types, analysis methods, access rights and consent conditions are mutually realistic, and a pseudonym dictionary. It never sees the source text, so there is no path by which a rare term specific to one paper can leak: in this corpus 67% of the technical terms occur in a single paper, and mixing several papers into a prompt does not prevent such terms from surfacing. Person and institution names are made overtly fictitious, e-mail addresses and URLs use reserved domains, and DOIs, repository identifiers and grant numbers are placeholders. Device, software, gene, protein and cell-line names are coined rather than marked as fictitious, which would read as unnatural in a research note, and are checked for $\mathrm{df} = 0$ and for non-collision with product databases, HGNC/UniProt and Cellosaurus. Project names are not invented but drawn from the five enumerated values of `project-name`, since an invented name would violate the schema.

Linkage control. An attribute combination is admissible only if the number of corpus papers matching it is 0 or at least 5; the band 1–4 is forbidden, and at most two fields of a single note may originate from the same paper. Each note is then checked mechanically for fingerprint terms, pseudonym collisions, the k-anonymity condition and attribution. Because the notes are Japanese, the document-similarity part of the check is run on an English translation prepared for inspection only; run on the Japanese text it passes unconditionally. An LLM attribution test additionally supplies one note together with the titles and abstracts of the 100 papers and asks for the source paper, requiring the top-1 hit rate over the 30 notes to remain at chance level.

Freezing. Once a note and its gold metadata pass these checks they are frozen and their SHA-256 digests recorded, and neither is modified afterwards on the basis of any extraction result. This is a deliberate constraint rather than a formality. The note is the only freely adjustable variable in the pipeline, so admitting accuracy as a reason to revise it would drive the dataset towards notes written to be machine-readable, and it would cease to be a test bed for other models. Low accuracy on a particular note is therefore recorded as an observation about the extractor, never treated as a defect of the note.

Controlled variation. Varying the notes at random would make a drop in accuracy impossible to attribute, so variation follows a 3-factor × 3-level orthogonal design (Table 1): the naming of item keys, the proportion of prose, and the number of file-type groups. Twenty-seven notes realize all 27 combinations and three leading notes take the diagonal. The proportion of prose is defined as the character count of unlabelled text divided by the total, with the sentence count as a secondary measure. An earlier definition—the number of schema items recoverable from prose alone—proved unusable: Japanese notes do not formally separate field labels, section headings and running text, so that count moved between 0 and 10 for the same note depending on the labelling rule, whereas the character proportion shifts by only one to four points. At the highest level the bullet structure is retained, with at least 30 labelled lines, because turning numeric listings such as image width, height and resolution into prose would make the note diverge from real practice and confound "the note was unnatural" with "the model could not read it".

Gold metadata. A field is given a value only where the note actually states it; every other field is `null`, and a value supplied by the extractor for such a field counts as a spurious extraction and hence an error. English (`-en`) fields are filled only where the note itself carries an English rendering. Enumerated values are recorded in the concatenated bilingual form, numbers as written, and dates in any recognizable format, the scorer normalizing them before comparison. Two-dimensional structures expressed as text, such as a folder tree, are kept whole in a single value rather than decomposed. The gold metadata were produced by a model different from the generator and reviewed by a second person, so that one model's reading of the note is not fixed as the reference against which another model is judged.

Table 1
Orthogonal variation design of the 30 research notes. Each factor is realized at three levels; the 27 full combinations are assigned one note each and three leading notes take the diagonal.

Factor	Level 1	Level 2	Level 3
F1 Item keys	Almost identical to the schema titles	About one third replaced by synonyms	More than half omitted or paraphrased, so the target key must be inferred
F2 Proportion of prose	25–35% of characters, 4–6 sentences	40–50%, 7–9 sentences	55–65%, 10–13 sentences
F3 File-type groups	1 group (text, image or spreadsheet)	2 of the 4 groups, all 6 pairs	3 of the 4 groups, all 4 triples

3.4. Configuration of the PoC System

The PoC comprised an AI server (AWS p4d.24xlarge) and an API server (AWS t3.xlarge) (Figure 1). Users upload a research notes file and select the LLM and metadata type (project or dataset). An MCP client sequentially executes four agents: schema retrieval, metadata extraction, schema conformance validation, and metadata registration. Inference runs locally via llama.cpp or Ollama. The registration tool logs the JSON-formatted string; writing to GRDM was outside scope.

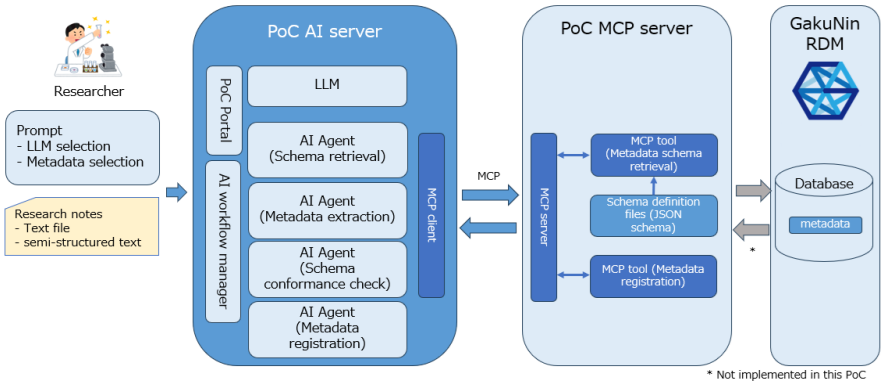


Figure 1: Configuration of the PoC system. On the PoC AI server, the AI workflow management unit sequentially executes four agents (schema acquisition, metadata extraction, schema matching/verification, and metadata registration), invoking tools on the MCP server side via an MCP client. Inference by the LLM is executed within the PoC AI server and does not go through an external API. The asterisk indicates a route not implemented in this PoC; registration to GRDM was limited to log output.

3.5. Models Evaluated

Nine models (Table 2) were deployed in the PoC environment. LLM-jp provides fully open Japanese LLMs [14]: llm-jp-3-13b-instruct3, llm-jp-3.1-8x13b-instruct4, and SIP-jmed-llm-3-8x13b (8x13B MoE, 8bit quantized); comparators were gpt-oss:20b, gpt-oss:120b, GPT-OSS Swallow 20B RL, glm-4.7-flash, qwen3.5:27b, and Nemotron-Nano-9B-JP. All 9 were evaluated in Section 4.1; the per-criterion comparison of project and dataset metadata there targeted SIP-jmed-llm-3-8x13b, gpt-oss:20b, and gpt-oss:120b. The evaluation of Section 4.5 used llm-jp-4.1-32b-a3b-thinking, released on 28 September 2026, the successor of LLM-jp-4 32B-A3B (April 2026; approximately 32 billion total and 3.8 billion active parameters) [15].

3.6. Evaluation Method

Extraction was assessed per field against expected values on a three-level scale: conforming (○), partially conforming (△), non-conforming (×). Partially conforming means the expected value is present

Table 2
Verification Models Implemented in the PoC Environment

Model	Parameters	Architecture	Quantization	Size	Inference Engine
llm-jp-3-13b-instruct3	13B	Dense	Q8_0	14GB	llama.cpp
gpt-oss:20b	20B	MoE	MXFP4	14GB	Ollama
SIP-jmed-llm-3-8x13b	104B (MoE)	8x13B MoE	Q8_0	73GB	llama.cpp
gpt-oss:120b	116.8B	MoE	MXFP4	65GB	Ollama
GPT-OSS Swallow 20B RL	20B	MoE	MXFP4	12GB	Ollama
glm-4.7-flash	Unknown	Dense + reasoning	Default	19GB	Ollama
qwen3.5:27b	27.8B	Dense + reasoning	Q4_K_M	17GB	Ollama
Nemotron-Nano-9B-JP	9B	Mamba2 + Attention (reasoning)	Q8_0	8.9GB	Ollama (Modelfile)
llm-jp-3.1-8x13b-instruct4	104B (MoE)	8x13B MoE	Q4_K_M	45GB	Ollama (Modelfile)

but carries surplus or imprecision; non-conforming means the requirement itself is violated. Initial verification assessed extraction items; detailed verification defined assessment criteria: date format, suppression of guessing for English fields, retention of required-field marks, and JSON output stability. No statistical testing was performed; each condition was tried once.

4. Results

4.1. Selection of Model and Schema Description Format

An initial verification with SIP-jmed-llm-3-8x13b on 8 fields produced 6 exact matches. Free-text fields (dataset name, data type, measurement device category) matched. For ”number of files of the same type,” the model output 16, but the schema requires the product of measurement time points and individuals, i.e., 16×10=160: the numeral was transcribed correctly, but the arithmetic was not performed. For the enumerated field ”dataset field,” ”Raifu-saien” was output, missing the choice’s final character. Table 3 shows the same 8 fields given to 9 models. Both errors were specific to SIP-jmed-llm-3-8x13b; truncation did not occur in the other 7 models, and 6 models from the gpt-oss, glm, qwen, and Nemotron families output 160. The 9B Nemotron-Nano-9B-JP matched all 8 fields, while the 104B SIP-jmed-llm-3-8x13b matched 6, indicating these errors reflect model-family characteristics rather than scale. llm-jp-3-13b-instruct3 and llm-jp-3.1-8x13b-instruct4 produced no JSON output, so field-level evaluation could not be established.

Table 3
Number of exact matches for 9 models in initial validation (8 items)

Model	Optimal configuration	Match count	Non-matching items
glm-4.7-flash	v1+v1	8/8	—
gpt-oss:20b	v1+v2	8/8	—
Nemotron-Nano-9B-JP	v1+v2	8/8	—
gpt-oss:120b	v1+v2	7/8	Dataset name (abbreviated notation)
qwen3.5:27b	v1+v2	7/8	Measurement target (half-width space intrusion)
GPT-OSS Swallow 20B RL	v1+v2	7/8	Related organ name ((R,L) omitted)
SIP-jmed-llm-3-8x13b	v1+v1	6/8	Dataset field (truncated text), number of same-type files (calculation rule not applied)
llm-jp-3-13b-instruct3	—	0/8	JSON structural collapse (all items unassessable)
llm-jp-3.1-8x13b-instruct4	—	0/8	JSON output failure (generated schema description)

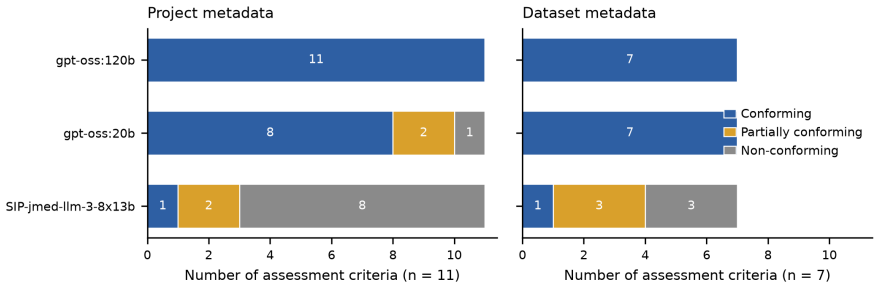


Figure 2: Breakdown of assessments by assessment criterion. Left: project metadata (assessment criteria n=11); right: dataset metadata (assessment criteria n=7). Conforming, partially conforming and non-conforming counts are shown. Assessments are based on a single-trial visual evaluation of one sample research note, and quantization conditions differ across models (Table 2).

Comparing Markdown- and JSON Schema-format schemas: Markdown yields Japanese field names as keys, whereas JSON Schema yields English kebab-case keys more easily mapped to GRDM's internal field IDs; keywords were flat arrays in Markdown but arrays of objects in JSON Schema, and nested structure for creator and administrator was preserved only in JSON Schema. Given the hierarchical and conditional structure of the metadata, JSON Schema was judged appropriate. Writing instructions into the field descriptions changed adding, to field descriptions, a prohibition on guessing together with the output format (JSON Schema v2). Under v1, gpt-oss:20b guessed 4 English fields, guessed data-id as "12abcdE", and transcribed another field's value into repository URL; all became empty under v2. For gpt-oss:120b, ic-is-no, empty under v1, was correctly extracted under v2. However, "time-series data" continued to be appended to kaiseki-taisho deta (analysis-target data) across both models and all formats, and regressions occurred: gpt-oss:20b dropped the pipe-delimited English portion of the license and left ic-is-no empty, while gpt-oss:120b shifted the date from ISO format to the as-written notation. Undocumented combinations are marked "—".

Figure 2 shows assessment breakdowns for project metadata (11 criteria) and dataset metadata (7 criteria). gpt-oss:120b was conforming on all criteria. gpt-oss:20b was conforming on 8, partially conforming on 2, and non-conforming on 1 project criterion, and conforming on all 7 dataset criteria; the non-conforming case was omission of the "*" required-field suffix, and partial conformance involved persistent temporal information and redundant description of the experimental objective, suggesting excessive suppression of guessing for dataset metadata. SIP-jmed-llm-3-8x13b was conforming on 1, partially conforming on 2, and non-conforming on 8 project criteria, and conforming on 1, partially conforming on 3, and non-conforming on 3 dataset criteria, showing name-spacing errors, date confusion or ISO deviation, misspelled English field names, email typos, free-text enumerated values, structural collapse, consolidation of 5 sections, and guessed folder names. gpt-oss:120b is expected to support registration with further tuning, whereas the other two models remain difficult to use at their current performance; gpt-oss:20b's high per-criterion conformance differs in granularity from the overall schema-wide assessment (over 100 fields for dataset metadata).

4.2. Validation of MCP Tools

The Metadata Schema Retrieval Tool correctly returned schema definitions. The Metadata Registration Tool logged received JSON, but AI-generated JSON sometimes contained incorrect field IDs or hierarchical structures, requiring registration-side validation with caller feedback. Per Section 4.5, JSON with an unfounded Key overwrites existing values with empty ones; validation must cover value formats and the Key set.

4.3. Tuning of the Extraction Pipeline

Before the evaluation the extraction pipeline was tuned along six lines (Table 4). Neither fine-tuning nor constrained decoding against the JSON Schema was used: every change lies in the model version, the inference parameters, the prompt, the schema descriptions, post-processing, or the agent workflow. Each change was scored at least three times against the gold list, because at temperature 0 the output still varies between runs and a single run cannot be separated from that variance.

For llm-jp-4.0 the decisive changes were removing the 4096-token output cap, which reasoning alone could exhaust so that no JSON was emitted at all, together with `reasoning_effort=medium`, treating `finish_reason=length` as an error so that a JSON fragment produced mid-reasoning is never accepted as a result, and adding to the prompt an item list—key, title, description and enumerated values—derived from the schema. Of the parsed candidates the largest syntactically complete JSON is taken, rather than a fragment appearing inside the reasoning text.

llm-jp-4.1 follows instructions such as "omit the field when there is no value" more literally than 4.0, which made imprecise instructions actively harmful. Project-level content—ethics review, informed consent, licence— migrated into the general-purpose remarks fields of the dataset metadata; creator and manager, separately labelled in the note, were both filled with both names; and the English (`-en`) fields, which 4.0 had filled by translating on its own initiative, came back empty. These were addressed by making the instruction explicit and by worked examples, with post-processing restricted to format-level repair. Moving values between fields by rule was rejected as a remedy because it does not generalize: where the model chooses the wrong destination, the schema or the prompt is the place to fix it.

Tuning the field descriptions helped where it generalized them to how notes are actually written—recognizing row and column descriptions stated without a heading, splitting array elements per file, normalizing "1,201 rows" to 1201, and keeping a folder tree verbatim in a single value—and hurt where it added prohibitions. "Do not supply a position if none is written" did not stop the model from numbering columns in sequence, and lengthening the descriptions made the output less stable elsewhere; using phrases from a particular note as an example made matters worse still.

The largest single gain came from a two-stage flow. In the first stage the note is passed with line numbers and the model returns only the line ranges of each region—measurement, folder, per-file, and lines shared by files of the same format. The text itself is then cut from the original by program, since asking the model to copy it introduces summarization and loss. In the second stage each region, and each file within it, is extracted with only the relevant part of the schema and the relevant lines; the results are merged and pass through the same post-processing, validation and registration path as before. On dataset metadata this raised accuracy from 84% to 95–98% (Table 5), at the cost of six to seven LLM calls per note instead of one and about 1.5 times the processing time. Confusion of spreadsheet rows with columns, and leakage of one file's description into another, both frequent under single-pass extraction, did not occur once. The two-stage path is taken only for schemas that have measurement, folder and file-group regions; others are extracted in a single pass as before, and both join the same post-processing.

The broader lesson is that for a model of this size, narrowing the input is more effective than adding instructions. Prohibitions were the least reliable device of all, and the gains that survived came from giving the model less to read at once.

4.4. Controlled Comparison of the llm-jp-3 and llm-jp-4 Generations

The observations of Sections 4.1 and 4.3 were made on different inputs, so they support no generational claim. We therefore ran the two generations against the same material: research notes 01–05, the same five files for both models, three extraction runs each, project metadata only. The models are SIP-jmed-llm-3-8x13b and llm-jp-4.1-32b-a3b-thinking.

The comparison required a change of schema. Against the operational project- and dataset-metadata schemas, llm-jp-3 extraction fails outright rather than degrading, so there is nothing to score; a simpli-

Table 4
Tuning applied to the extraction pipeline.

Line of tuning	Content
Model version	llm-jp-4.0 → llm-jp-4.1
Inference parameters	Output cap removed, `reasoning_effort=medium`, temperature 0, truncation detected via `finish_reason`
Prompt	System/user separation, Japanese instructions, note placed last, explicit extraction rules, worked examples for creator and manager, item list derived from the schema
Schema descriptions	Field descriptions aligned with how notes are actually written: headingless statements, row/column assignment, numeric normalization
Post-processing	Final JSON parsing, folder tree kept verbatim, nested arrays expanded for CSV
Agent workflow	LangGraph nodes (schema selection, extraction, validation, registration), batch processing of multiple files, two-stage segment-then-extract flow

Table 5
Effect of the two-stage flow on dataset metadata (two notes, three runs each, 70 gold items per run).

Method	Correct	Accuracy	LLM calls / note
Single-pass extraction	176/210	84%	1
Two-stage flow, prototype	205/210	98%	6–7
Two-stage flow, with file-name instruction	199/210	95%	6–7

fied project-metadata schema of 58 fields was prepared and both models were run against that same schema. This is itself the first part of the answer — the earlier observation that many-field schemas break the llm-jp-3 generation is reproduced here as an inability to score it at all — and it means the figures below are not comparable with those of Section 4.5, which uses the full schema and is reported next.

Over 870 item-level judgements per model, llm-jp-3 reaches F1 0.553 and llm-jp-4.1 0.822 (Table 6). The gap is not uniform across the confusion matrix: llm-jp-3's recall, 0.687, is within 0.16 of llm-jp-4.1's, whereas its precision, 0.463, is little more than half. llm-jp-3 finds a comparable share of the values the note supports and surrounds them with far more that it does not — 396 false positives against 107 — so over-extraction is the generational difference, not blindness.

The difference concentrates in one place. Of the 237 judgements whose gold value is an enumerated option, the schema states those options as bilingual label strings of the form `Japanese|English`; llm-jp-3 reproduced the exact option string in 159 (67.1%) and in 60 (25.3%) returned a value that does not carry the form at all, against 218 (92.0%) and 2 (0.8%) for llm-jp-4.1. The errors are of the kind that defeats a controlled vocabulary while reading plausibly: `Sonota|Other` returned as the repository's actual name, `Ka|Yes` as `Kano` [possible], `CC BY 4.0 Hyoji Kokusai|CC BY 4.0 Attribution International` truncated to `CC BY 4.0`. A smaller group picks the wrong option in the right form — `Raifu-saiensu|Life Science` returned as `Shizen-kagaku ippan|Natural Science` — which validation against the enumeration cannot catch.

Both models reproduce themselves closely at temperature 0: the second and third runs are identical field for field for each model, and the first differs on 20 of 290 fields for llm-jp-3 and 23 for llm-jp-4.1, so the ordering is not an artefact of sampling. Taken with Section 4.2, the comparison locates the generational gain precisely: what improved is the reproduction of strings fixed by the schema, and that is exactly the property a registration-side validator can check mechanically.

4.5. Evaluation on 30 Research Notes with llm-jp-4.1

The tuned pipeline was evaluated with llm-jp-4.1-32b-a3b-thinking, released on 28 September 2026, over the 30-note dataset of Section 3.3. Each note was extracted three times and every field of every extraction was matched against the frozen gold metadata in schema order and judged as a true or false positive or negative, giving 29,320 item-level judgements (Table 7). Dates are accepted in any

Table 6
The two generations on identical input — research notes 01–05, three runs each, project metadata under the simplified 58-field schema (micro-average over 870 item-level judgements per model).

Model	Items	TP	FP	FN	TN	P	R	F1
SIP-jmed-llm-3-8x13b	870	342	396	156	109	0.463	0.687	0.553
llm-jp-4.1-32b-a3b-thinking	870	422	107	76	311	0.798	0.847	0.822

format, row and column positions are accepted in any notation whose order is right, and procedures and free-text fields are accepted whether written as one block or split across several. Over the three runs precision is 0.563, recall 0.640 and F1 0.599.

The aggregate hides a split that matters more than the aggregate itself. Project metadata reaches F1 0.800, dataset metadata 0.508. The difference is not a matter of degree: for project metadata the extractor produces slightly fewer values than the note supports (528 false positives against 753 false negatives), whereas for dataset metadata it produces far more (4,236 against 2,702), and it is the dataset side that carries the file-level structural descriptions of Section 3.2. Across all three runs false positives outnumber false negatives 4,764 to 3,455, which is the item-level form of precision falling below recall. Of those, 1,131 are value mismatches that count on both sides; the remainder divides into 3,633 keys emitted for fields the note does not support and 2,324 supported fields left out.

Both sides were run under the tuned configuration of Section 4.3, the dataset side using the two-stage flow. The distance to Table 5 is therefore one of sample and of metric, not of configuration. Table 5 counts gold-valued dataset fields recovered on the two notes the flow was developed against, 70 per run; the same quantity over these 30 notes is recall, 0.570 — 57% against 95–98%. Two effects are mixed in that drop and are not separated here: the two notes were in-sample, and the 30 carry the controlled variation of Section 3.3. F1 is lower again at 0.508 because it also charges keys emitted on fields the note does not support, which the Table 5 metric cannot see.

Run-to-run reproducibility. Inference runs at temperature 0, and on the project side the effect is visible: the first two runs are identical field for field across all 1,939 judgements, and the third differs on 159 of the 1,936 fields the two have in common (8.2%). Extraction of project metadata is therefore close to deterministic, and a single run is nearly as informative as three. Dataset metadata behaves differently. The number of items to be judged itself changes between runs — 7,552, 7,768 and 8,183 — because the model emits different numbers of array elements for the per-file descriptions, and 9.0% and 15.2% of the values shared with the first run differ in the second and third. Temperature 0 fixes the sampling, not the structure of the output; where the schema leaves the number of elements to the model, the output remains unstable, and there a single run cannot be distinguished from the variance.

Scoring treats a field whose gold value is empty as correct only when the key is *absent* from the JSON, not when it is present with an empty value. The distinction is operational rather than formal: in GRDM registration an absent key leaves the stored field unchanged, whereas a key carrying an empty value overwrites it. A key emitted for a field the note does not support is therefore a false positive, and these are the errors that damage an existing record rather than merely leaving it incomplete. Suppressing key output for unsupported fields remains the single most valuable improvement, consistent with the effect of the explicit prohibition on guessing in Section 4.1 and with the need for registration-side validation in Section 4.2.

5. Attempted Mapping to Standard Vocabularies

5.1. Scope and Coverage

This chapter reports an exploratory trial estimating what would be required to use extracted metadata as a source for knowledge-graph generation, rather than presenting PoC results proper. For the 171 items described in Section 3.2, mapping to predicates and classes of schema.org [16], DCMI Metadata Terms [17], DCAT [18], PROV-O [19], CSVW and the GA4GH Data Use Ontology [20] was performed

Table 7
Item-level evaluation of llm-jp-4.1-32b-a3b-thinking over the 30 research notes, three runs per note, scored against the frozen gold metadata (micro-average).

Scope	Items	TP	FP	FN	TN	P	R	F1
Project metadata	5,817	2,568	528	753	2,301	0.829	0.773	0.800
Dataset metadata	23,503	3,580	4,236	2,702	13,783	0.458	0.570	0.508
All, three runs	29,320	6,148	4,764	3,455	16,084	0.563	0.640	0.599
run 1	9,491	2,024	1,477	1,177	5,168	0.578	0.632	0.604
run 2	9,707	2,046	1,588	1,155	5,301	0.563	0.639	0.599
run 3	10,122	2,078	1,699	1,123	5,615	0.550	0.649	0.596

by having Claude (Opus 5) enumerate candidates, finalized by the authors against each vocabulary's official definitions. The mapping targets predicates, not values: the schema's options are bilingual label strings without concept URIs, so grounding destinations for values were appended as proposals.

A predicate or class was assigned to 162 of 171 fields (94.7%), all of them satisfied with non-medical vocabularies; medical ontologies (NCBI Taxonomy, Cell Line Ontology, UBERON, OBI, EDAM) were needed only to ground values. The 9 fields for which no standard vocabulary was found—ethics review board approval, anonymization presence, conflict of interest, acknowledgment name and other-database metadata-file presence—are all on the project-metadata side: international vocabularies cover dataset description, creators, rights and provenance, but not concepts specific to domestic research-data-management operations.

Whether declaring the mapping as an `@context` converts key-value JSON into triples directly was examined empirically. Key-to-predicate mapping alone yields 7 triples whose subject is an unreferenceable blank node, with no IRI object and no language tag, against 12 forming a usable graph once subject IRI, type, object IRI and language tags are configured as well; JSON-LD expansion also silently discards undefined keys, so any omission loses extracted metadata without warning. Triplification therefore needs conversion processing beyond a declaration—subject-IRI assignment, bilingual-value decomposition and date normalization—which is deterministic and need not be entrusted to an LLM, since a single-character IRI error redirects a node. This mapping remains exploratory in that its semantic validity rests on the authors' judgment: that "data manager" belongs under `dcat:contactPoint` rather than `dcterms:contributor` is an editorial decision no tool settles. What a tool can settle—whether the terms exist and whether they accept the values—is reported next.

5.2. Formal Validation of the Mapping

We retrieved the six vocabularies in full and checked the 171 assignments against them. Of the 171, 115 carry a term the declaring vocabulary defines as a property and none is deprecated; 16 do not — eight row- and column-position fields assigned `csvw:number`, which CSVW does not define, and eight consent, provision, use and permission fields assigned DUO IRIs declared `owl:Class`. Thirty-one fields carry no predicate in the mapping table and nine have no corresponding term, so neither group could be checked. Range violations appeared only while nested fields were listed flat; with the containers represented as intermediate nodes, none remains.

Shapes were generated from the mapping table against the two schemas the extraction itself was given — 60 fields for the first registration page and 127 for the second — giving 3 node shapes with 25 property constraints for project metadata, 15 of them required, and 7 node shapes with 25 for dataset metadata; the required marks were taken from the GRDM form definition, the extraction schemas declaring none. The shapes were run with pySHACL over the extraction output of Section 4.5 — 30 notes, three runs, 90 instances per metadata type — and over the same instances built from the gold metadata, so that violations caused by the schema could be separated from those caused by the extraction (Table 8). No instance conforms, in any of the four sets. The gold produces more project-metadata violations than the output, 849 against 680, because the gold dates are not ISO 8601 and the

gold leaves concept-typed fields as label strings exactly as the extraction does. Of the 1,064 violations the output produces, 282 are attributable to the extraction — missing required fields, values outside an enumeration, keys the schema does not declare and ungrounded concept values — and the remaining 782 follow from the schema and from a conversion step that is not implemented.

Matched against the human judgements of Section 4.5, the shapes detect 436 of the 948 project-metadata errors (46.0%) and 249 of the 6,140 dataset-metadata errors (4.1%), 685 of 7,088 in total (9.7%). The gap between the two follows from the schema: of the 114 dataset-metadata leaves, none is required and 5 are enumerated against 109 free-text. Over-extraction is not covered at all, since a key emitted for a field the note does not support carries a declared predicate and is admitted. What limits detection further is the coverage of the conversion rather than the choice of constraints: a first conversion matched predicates on the leading part of each key and left the nested data-creator, data-manager and keyword groups out of the graph, and re-doing it on leaf names, with those containers raised to intermediate nodes, took the converted keys from 29 to 54 of 72, the triples per instance from 13.4 to 26.2 and the detection rate from 38.0% to 46.0%. Two limits are disclosed: the `dcat:contactPoint` node meets its declared range `vcard:Kind` at the class level but carries schema.org leaf predicates, mixing the vocabularies one level down; and the creator and contact nodes are blank nodes, so using researchers as a joining axis between datasets would require ORCID or GRDM account IRIs.

Table 8
SHACL violations over the extraction output and over the gold metadata, shapes derived from the two extraction schemas, 90 instances each (30 notes, three runs).

Constraint	Project		Dataset	
	output	gold	output	gold
Required field missing (`sh:minCount`)	109	39	0	0
Concept left as a label (`sh:nodeKind`)	495	624	384	246
Value outside the enumeration (`sh:in`)	65	0	0	0
Key outside the schema (`sh:closed`)	9	0	0	0
Date not ISO 8601 (`sh:datatype`)	2	186	0	0
Total	680	849	384	246
Conforming instances	0/90	0/90	0/90	0/90

6. Discussion

Results depended more on schema description than on model capability. JSON Schema suited hierarchically structured, conditionally branching metadata, and stating instructions and output format in field descriptions improved accuracy. The effect was model-family dependent: description-text instructions improved the gpt-oss family, SIP-jmed-llm-3-8x13b showed no change, the two llm-jp-3-generation models produced no JSON output at all, and glm-4.7-flash already matched all 8 fields (Table 3). Schema refinement thus benefits only models with stable JSON output. Efforts to formalize the NII RDC AP via ontology techniques [6] are significant from this perspective too.

llm-jp-3 generation errors concerned schema conformance—missing enumerated-option characters, non-ISO dates, omitted required-field marks, misspelled field names, structural collapse—not value content. The llm-jp-4 generation resolved these, and the controlled comparison of Section 4.4 quantifies the difference on identical input: under a simplified schema both generations can read, F1 rises from 0.553 to 0.822, and three quarters of the gain is the reproduction of the schema's own enumerated option strings. Schema conformance, however, is not what now limits extraction. Over 30 notes llm-jp-4.1 reaches F1 0.599, and the binding difficulty is no longer the shape of the JSON but the decision of whether a field is supported by the note at all — a decision the model makes far better for project metadata (F1 0.800) than for the file-level descriptions of dataset metadata (0.508). The tuning of Section 4.3 points the same way: of the six lines tried, the one that moved accuracy most was not a better instruction but a narrower input—the two-stage flow, which gives the model one region of

the schema together with the corresponding lines of the note—while prohibitions such as "do not supply a position that is not written" were never reliably obeyed, and lengthening the field descriptions destabilized fields that had previously been correct. Field-ID and hierarchy errors still occurred at registration, so registration tools should implement JSON Schema validation returning violation locations to prompt regeneration [11]. Since pre-symptomatic disease metadata may be adjacent to research-plan and human-subject information, processing entirely within an environment under our own administration broadens applicable domains; local models' performance limits make schema development and validation a realistic compensating design.

7. Limitations of This Study

The 30 notes are constructed data. They derive from 100 published papers of the programme and follow the operational schema, but a note written to be read differs from one accumulated as a by-product of an experiment, and the generalization that removes linkage also removes part of the specificity a real note would carry. The gold metadata were authored by a model and reviewed by hand, so a misreading shared by the model and the reviewer would not be caught. The notes populate the "semi-structured × medium schema conformance" cell of the typology at three levels of key naming and prose proportion; unstructured and low-conformance notes remain unverified.

Only one model was evaluated at this scale. The nine-model comparison of Section 4.1 is still a single trial over 8 fields with a three-level visual assessment (Table 3, Figure 2), quantization differs across those models (Q8_0, MXFP4, Q4_K_M; Table 2), and their generation parameters and prompts are not stated. The figures of Section 4.5 are broken down by metadata type and by run but not by variation factor or by error category, so the orthogonal design of Table 1 has not yet been used to separate the main effects it was built to expose, and the comparison with llm-jp-3 in Section 4.4 rests on five notes under a simplified schema, which the llm-jp-3 generation could read and the operational one it could not. The two notes of Table 5 were the ones the two-stage flow was developed against, so that figure is in-sample; the 30-note result is the out-of-sample one, and the gap between them is not decomposed into the in-sample effect and the effect of the controlled variation. Neither fine-tuning nor constrained decoding against the JSON Schema was attempted.

The registration tool is limited to log output; registration into GRDM and transmission to and publication on JAIRO Cloud are unverified. The mapping of Section 5 was enumerated with an LLM and finalized by the authors against the official vocabulary definitions, so its semantic validity still rests on that judgment; the formal validation of Section 5.2 bounds how far it has been checked rather than establishing that it is right. 31 fields carry no predicate and were not checked at all, the 120 fields derived by rule have not been confirmed individually, and 8 of the 13 container predicates needed for the dataset-metadata conversion were assigned by the authors rather than taken from the field table, and 45 leaves have no predicate at all. Of 171 field assessments, 52 fields common across file types were classified automatically by a prefix rule without individual scrutiny.

8. Outlook: From Metadata to Knowledge Graphs

Of the 171 fields, 162 map to standard-vocabulary predicates and classes, so the generated JSON can be interpreted as triples by deterministic transformation [21]: the project name references a project entity, and the object of measurement, the classification of the measuring device and the related organ reference, respectively, taxonomic and cell-line vocabularies (NCBI Taxonomy, Cell Line Ontology), an instrument vocabulary (OBI) and an anatomical vocabulary (UBERON). Grounding remains the challenge: the six concepts with the largest effect on graph construction are free-text (Section 5), and linking them to ontology terms requires language models [22].

Which fields do the joining, however, depends on whether the graph is public or internal. The one internal entity with strong reuse across the notes is the project name — 9 values over 29 notes, each shared by at least two — against 1.03 for device names, yet both are declared literals, the mapping

having been made against external vocabularies. Grounding the fields that have external counterparts is vocabulary work; a graph closed to the laboratory additionally needs the platform to issue IRIs for projects, researchers and devices, and a vocabulary for ethics review approval, which has none and is what access authorization must reason over. Graph generation is not implemented, so its accuracy is unevaluated.

9. Conclusion

Our PoC extracted metadata from research notes into JSON in GRDM's registration format, following the pre-symptomatic disease research metadata schema. Under the llm-jp-4 generation the output conforms to the schema, including the bilingual Japanese and English value format, and tuning the pipeline—most of all a two-stage flow that segments a note into regions before extracting each region against the corresponding part of the schema—raised dataset-metadata accuracy from 84% to 95–98%. Evaluated over 30 purpose-built research notes, each extracted three times, llm-jp-4.1-32b-a3b-thinking reached precision 0.563, recall 0.640 and F1 0.599 over 29,320 item-level judgements, splitting into F1 0.800 for project metadata and 0.508 for the file-level descriptions of dataset metadata. Precision below recall identifies over-extraction—keys emitted for fields the note does not support—as the dominant error mode, and these are precisely the errors that overwrite existing values with empty ones on registration. Mapping the 171 fields to six standard vocabularies assigned a predicate or class to 162 of them, 115 of which carry a term the declaring vocabulary defines; shapes derived from that mapping detect 46.0% of the project-metadata errors a human marked against 4.1% of the dataset-metadata ones and none of the over-extraction. Future work: suppressing over-extraction, evaluation on notes from actual operation, ELN ingestion, registration-side JSON validation, MCP-based access authorization control linking the GRDM API and the GakuNin IdP, and platform-side issuance of IRIs for the internal entities on which a closed graph joins.

Acknowledgments

This research was supported by JST Moonshot R&D Grant Number JPMJMS2021.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Declaration on Generative AI

During the preparation of this work, the authors used a large language model (Claude, Anthropic) in order to: enumerate candidate predicates and classes for the mapping of schema fields to standard vocabularies reported in Section 5, and to translate the manuscript from Japanese into English. All mapping candidates were subsequently checked by the authors against the official vocabulary definitions, and the translation was reviewed by the authors. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

A. Sample Research Note and Extracted Metadata

A.1. Research note (translated from the Japanese original)

The listing below comes from `note_01` of the 30-note dataset of Section 3.3, which realizes level 1 of all three variation factors. The note is Japanese; it is translated here for the reader, with layout, punctuation, indentation and field order exactly as in the original and only the Japanese strings replaced.

Shown are the opening, from which project metadata is read, and the file-related part of the dataset-metadata section, where the notes carry most of their structure; elided passages are marked. All names and identifiers are pseudonyms or coined words satisfying the $k \geq 5$ condition.

```
Infectious Disease Epidemic Forecasting Survey Dataset  November 15, 2024
  Created：  Kari-A University  Karioka Nao  karioka@example.ac.jp (Kari-A University, Karioka
     Nao)
  Managed：  Kari-B Medical Center  Kariyama Nako  kariyama@example.ac.jp (Kari-B Medical Cent
    er, Kariyama Nako)

Project 001：Mathematical model analysis of social risk factors and clinical outcomes in patie
    nts with systemic infectious disease

Last updated：2024/11/18

[...]

Folder structure：
├── ReadMe.txt # this file
├── data/ # questionnaire response data
□ └── LSQ20_responses.txt # all LSQ-20 response data
└── analysis/ # analysis results and configuration files
  ├── risk_index_summary.txt # aggregate values of risk indicators
  └── model_parameters.txt # dynamical-model parameters

Text files：
3 files in all; aggregate tables of questionnaire response scores and analysis results, not ti
    me-series measurements.
Delimiter：tab character
Character encoding：UTF-8

【LSQ20_responses.txt】
1,201 rows (including the header row); each row corresponds to the data of one patient.
21 columns, each column：
  column 0：Patient_ID (patient identification code)
  columns 1–20：LSQ20_Item01–LSQ20_Item20 (score of each questionnaire item, 1–5)
[...]
```

A.2. Extracted metadata (Japanese values likewise translated)

Below is what llm-jp-4.1-32b-a3b-thinking returned for that note: the opening of the project metadata and the file-related part of the dataset metadata, with the Japanese values translated on the same basis. Fields the note does not state are returned as empty strings, whereas the scoring rule of Section 4.5 requires the key to be absent. Column positions are returned as the note writes them, and one of the three file groups without a file name.

```
{
 "title-of-dataset": "Infectious Disease Epidemic Forecasting Survey Dataset  November 15, 202
    4",
 "title-of-dataset-en": "",
 "project-name": "Project 001",
 "date-registered-in-metadata": "2024-11-15",
  ... 39 further fields, of which 13 are returned as empty strings
}
{
 "d-fol-Structure-or-descriptions-of-folders-jp": [
  {"description-of-folder": "├── ReadMe.txt # this file\n├── data/ ... "}],
 "d-txt-group": [
  {"d-txt-file-name-convention-file-extension": "LSQ20_responses.txt",
   "d-txt-description-of-column": [
    {"Position-of-column": "column 0", "Name-of-term": "Patient_ID",
     "Description-of-term": "patient identification code"},
    ... 1 further column],
   "d-txt-number-of-rows": "1201", "t-txt-delimiter": "tab character"},
  ... 2 further file groups, one of them returned without a file name
 ]
}
```

References

[1] K. NAKANO, The Current Status of the Research Data Ecosystem Development Project, Joho Chishiki Gakkaishi (2024). doi:10.2964/jsik_2024_026, in Japanese.

[2] M. TANIFUJI, Research Data Cloud NII RDC and Open Science –NII's Initiatives in the Open Access Acceleration Program, The Journal of Information Science and Technology Association (2025). doi:10.18919/jkg.75.9_447, in Japanese.

[3] N. I. of Informatics, About GakuNin RDM, GakuNin RDM Support Portal (accessed 10 September 2026) (2026). URL: https://support.rdm.nii.ac.jp/about/, in Japanese.

[4] M. HAYASHI, Y. KOMIYAMA, I. FUJIWARA, M. ASAOKA, Y. MINAMIYAMA, R. IKEYA, H. YASUDA, M. TODORIKI, F. FUJII, S. KIMOTO, K. FUJIWARA, K. YAMAJI, A Metadata Handling for Research Data in the Comprehensive Pre-symptomatic Database Project, Joho Chishiki Gakkaishi (2024). doi:10.2964/jsik_2024_033, in Japanese.

[5] Y. KOMIYAMA, Comprehensive Pre-Disease Database and Moonshot Goal 2: Realizing Cross-Project Data Sharing and Utilization, in: Proceedings for Annual Meeting of The Japanese Pharmacological Society, 2023. doi:10.1254/jpssuppl.97.0_2-B-S22-3, in Japanese.

[6] M. ASAOKA, Y. MINAMIYAMA, M. HAYASHI, T. HIRAKI, Y. KOMIYAMA, T. SHIMOYAMA, S. YOKOYAMA, I. FUJIWARA, H. TAKEDA, K. YAMAJI, Evaluating the Applicability of the NII RDC Application Profile to Domain-Specific Metadata Schemas, Joho Chishiki Gakkaishi (2025). doi:10.2964/jsik_2025_023, in Japanese.

[7] S. Auer, V. Kovtun, M. Prinz, A. Kasprzik, M. Stocker, M.-E. Vidal, Towards a Knowledge Graph for Science, in: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics (WIMS), 2018. doi:10.1145/3227609.3227689.

[8] H. H. C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, Automatic document metadata extraction using support vector machines, in: Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL), 2003. doi:10.1109/jcdl.2003.1204842.

[9] I. G. Councill, C. L. Giles, M. Kan, ParsCit: an Open-source CRF Reference String Parsing Package, 2008. doi:10.63317/3pp9nrrhs7x9.

[10] P. Lopez, GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications, Lecture notes in computer science (2009). doi:10.1007/978-3-642-04346-8_62.

[11] Z. Alyafeai, M. S. Al-shaibani, B. Ghanem, MOLE: Metadata Extraction and Validation in Scientific Papers Using LLMs, Findings of the Association for Computational Linguistics: EMNLP 2025 (2025). doi:10.18653/v1/2025.findings-emnlp.655.

[12] P. Tinn, S. Sørbø, S. Jiang, K. Voutetakis, Pre-Meta: Priors-augmented Retrieval for LLM-based Metadata Generation, Bioinformatics (2025). doi:10.1093/bioinformatics/btaf519.

[13] M. Jalali, Y. Luo, L. Caulfield, E. Sauter, Large language models in electronic laboratory notebooks: Transforming materials science research workflows, Materials Today Communications (2024). doi:10.1016/j.mtcomm.2024.109801.

[14] LLM-jp, LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs, arXiv (2024). doi:10.48550/arXiv.2407.03963.

[15] Research, N. I. o. I. Development Center for Large Language Models, Release of new fully open Japanese LLMs 'LLM-jp-4 8B' and 'LLM-jp-4 32B-A3B' trained on a corpus of approximately 12 trillion tokens, News release (2026). URL: https://www.nii.ac.jp/news/release/2026/0403.html, in Japanese.

[16] R. V. Guha, D. Brickley, S. Macbeth, Schema.org: Evolution of Structured Data on the Web, Queue (2015). doi:10.1145/2857274.2857276.

[17] D. C. M. Initiative, DCMI: DCMI Metadata Terms, DCMI Recommendation (2020). URL: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

[18] W. D. E. W. Group, Data Catalog Vocabulary (DCAT) - Version 3, W3C Recommendation (2024). URL: https://www.w3.org/TR/vocab-dcat-3/.

[19] T. Lebo, S. Sahoo, D. M. (eds.), PROV-O: The PROV Ontology, W3C Recommendation (2013). URL: https://www.w3.org/TR/prov-o/.

[20] J. Lawson, M. N. Cabili, G. Kerry, T. Boughtwood, A. Thorogood, P. Alper, S. R. Bowers, R. Boyles, The Data Use Ontology to streamline responsible access to human biomedical datasets, Cell Genomics (2021). doi:10.1016/j.xgen.2021.100028.

[21] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutiérrez, Knowledge Graphs, ACM Computing Surveys (2021). doi:10.1145/3447772.

[22] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, From Local to Global: A Graph RAG Approach to Query-Focused Summarization, arXiv (2024). URL: http://arxiv.org/abs/2404.16130v2.