Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Utilization of Japanese Public Educational Data by Retrieval Augmented Generation for Policy Research

Kyosuke TAKAMI
Education Data Science Center, National Institute for Educational Policy Research, Japan
takami@nier.go.jp

Abstract: Public educational data, including government-conducted national surveys and research cases, are widely available to the public and intended for use in municipal policymaking. However, some of this data has been published in PDF format and remains underutilized. Therefore, this study leverages new tools in the era of generative AI, such as Large Language Model (LLM) and Retrieval Augmented Generation (RAG), to process 705 public educational document PDF files in Japanese. This process involves extracting text, vectorizing it, and generating responses, thereby presenting a case study of methods for effectively utilizing public educational data. This study revealed that without using the RAG, the outputs from GPT-3.5 and GPT-4 were verbose, while the use of the RAG led to more specific answers based on the retrieval results. Furthermore, GPT-4 can be used to evaluate the quality of retrieval results. These results demonstrate that LLMs can be applied to local educational knowledge in countries with local languages, such as Japanese, and suggest that previously underutilized educational data can be leveraged to aid in formulating educational policies.

Keywords: Public Education Data Platform, GPT-3.5, GPT-4, Retrieval Augmented Generation, Educational Policy Making

1. Introduction

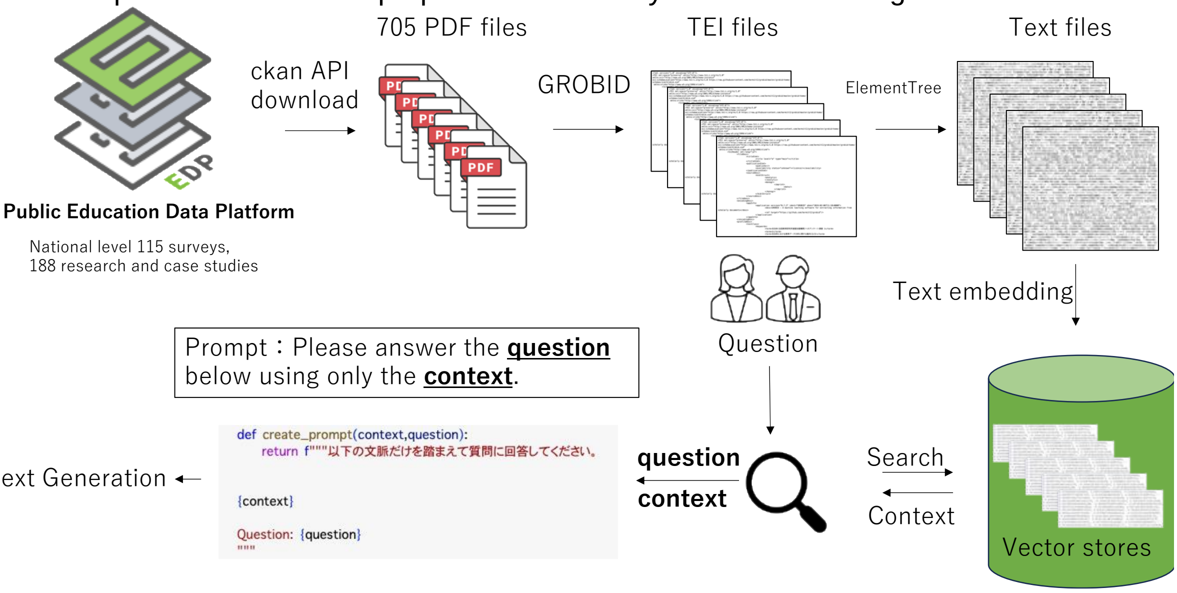
Public educational data, including government-conducted national surveys and research cases on topics such as academic achievement, bullying, and absenteeism, and school ICT, are broadly available to the public and are intended for use in municipal policymaking. Additionally, the availability of open data can increase public sector transparency, encourage citizen participation, and support the creation of new services (Attard et al., 2015). However, some of this data has been published in PDF format and remains underutilized.

On the other hand, the advancements in generative AI since the release of ChatGPT (OpenAI, 2022) in November 2022 have not been significant. In addition to OpenAI, other technology giants, such as Meta and Google, have introduced their own large language models (LLM), such as Meta's LLaMA-2 and Google's Gemini, stirring a sensation across society. In particular, it has been noted that LLMs perform worse in local languages other than English because LLMs are trained primarily in English, and in educational contexts, performance in Japanese has been reported to be worse than in English (Takami, 2023). One possible solution to these issues is the use of Retrieval-Augmented Generation (RAG). The RAG utilizes external information sources to retrieve relevant information and enhance the LLM response quality (Gao et al., 2023).

In this study, we explore whether applying RAG to local Japanese public education data can compensate for the decrease in LLM performance caused by a lack of training in local language knowledge. Therefore, we applied the RAG to Japanese education data.

2. Methods

The whole process of the RAG proposed in this study is illustrated in Figure 1.



Prompt：Please answer the **question** below using only the **context**.

```
def create_prompt(context,question):
    return f"""以下の文脈だけを踏まえて質問に回答してください。


{context}

Question: {question}
"""
```

Figure 1. Flow diagram of the RAG process.

2.1 Datasets

We used educational data provided by the Public Education Data Platform of Japan (National Institute for Educational Policy Research (NIER), 2023). This data platform provides open educational data from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the National Institute for Educational Policy Research (NIER). This web site is largely divided into a "Data Catalog" and a "Research and Case Studies." As of September 2023, approximately 300 research results and case studies including national surveys on academic achievement, bullying, absenteeism, and school ICT, are available on the Public Education Data Platform. This platform is built in CKAN (CKAN, n.d.), and the data can be retrieved using the CKAN API. From this platform, we obtained 705 PDF files for Japanese education.

2.2 PDF Information Extraction

We employed GROBID (GeneRation Of Bibliographic Data)(“Grobit”, 2008-2023), a machine learning library specifically tailored for extracting and processing data from scientific literature in PDF format. This library transforms unstructured PDF data into structured data in the form of TEI (a text-encoding initiative) (TEI Consortium, 2023) format, efficiently managing large volumes of files. We obtained structured TEI files and extracted text from them using ElementTree (Python). The text was further split into chunks using langchain.text_splitter so that chunk_size=1000 by langchan library (“LangChain”, 2023).

2.3 Text Embedding

The chunked text was vectorized using OpenAI's "text-embedding-ada-002" model and stored in the Chroma database. From this vector database, the top five texts with the highest similarity to the vector of question strings were used as the retrieval results.

2.4 Prompt (RAG)

The input question string was converted into a vector, and the five most similar to the vector were retrieved from the Chroma database and inserted into the prompt as a context. The prompt was

> “以下の文章だけを踏まえて質問に回答してください。(Please answer the questions below using only the context.)
>
> {context}
>
> Questions{question}”

The five texts of the search results were inserted into the context, and the input question was entered into the question. To compare and evaluate the texts generated by the LLM models, gpt-3.5-turbo and gpt-4 (OpenAI API) were generated using the RAG prompt.

3. Evaluation

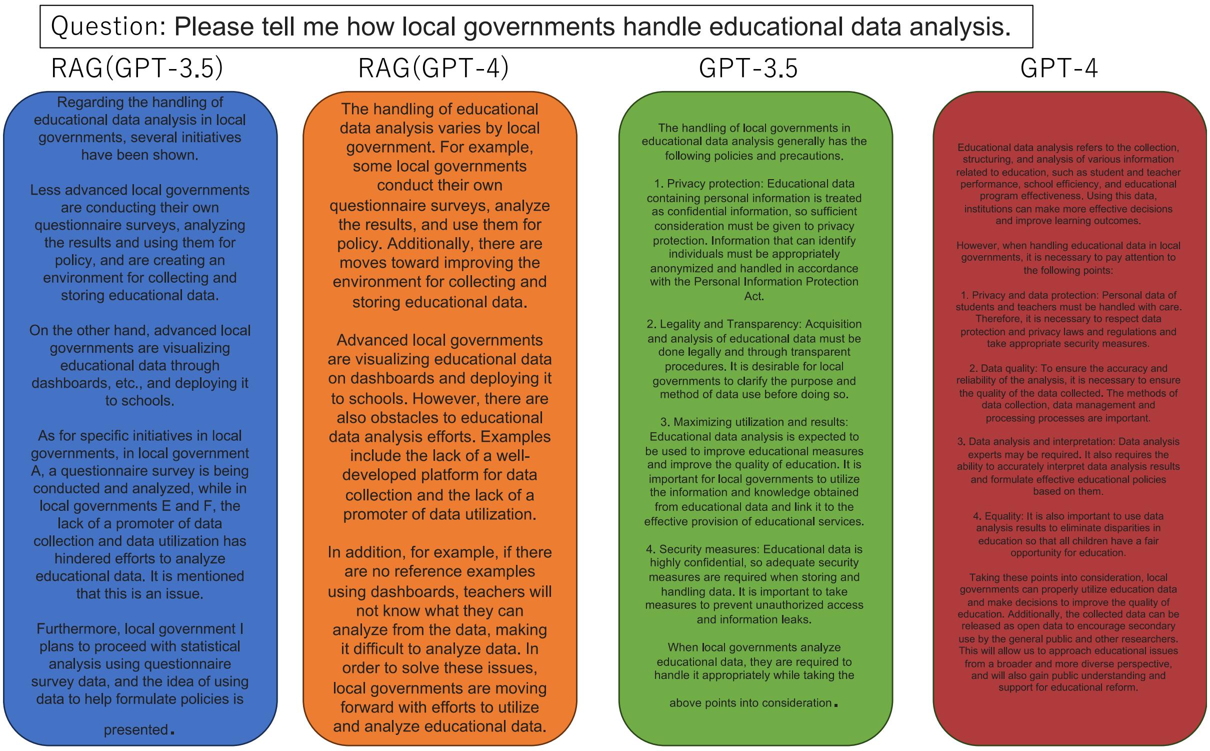
The evaluation of RAG models revolves around two main components: retrieval and generation modules (Gao et al., 2023). These evaluations employ established metrics, such as question-answering evaluations, fact-checking tasks, and similarity-based assessments of task-specific metrics. In this study, it was difficult to use automatic RAG evaluation tools (Ragas, 2023) that have been developed and are commonly used in English-speaking countries to evaluate the performance of RAGs in local Japanese education data. Therefore, we designed the original set of ten questions shown in Table 1 to examine whether LLMs could answer based on the knowledge contained in the educational dataset specific to this study. This list was created based on Japanese keywords specifically included in the Public Education Data Platform dataset. These questions were posed to the GPT-3.5 (gpt-3.5-turbo) and GPT-4 (gpt-4) models using the RAG, and the responses were evaluated. Similarly, we experimented by simply asking questions to the GPT-3.5, and GPT-4 models without RAG.

Table 1. *Questions for Japanese public educational data*

Questions (Japanese)	(Translation to English)
教育デジタル化の課題を教えてください。	Please tell me about the issues related to educational digitalization.
全国学力・学習状況調査について教えてください。	Please tell us about the national academic ability and learning status survey.
教育デジタル化の課題を教えてください。	Please tell me about the issues related to educational digitalization.
ICTリテラシーと資質・能力いついて教えてください。	Please tell me about ICT literacy, qualities and abilities.
学力向上するにはどうすれば良いですか？	How do I improve my academic ability?
非認知能力についての課題について教えて下さい。	Please tell us about the tasks concerning non-cognitive abilities.
いじめなどの問題行動について教えてください。	Please tell me about problems such as bullying.

不登校児童生徒にはどのような対応をすれば良いですか？	What should I do to school refusal students?
教員養成の課題について教えてください。	Please tell us about the task of teaching teachers.
地方教育行政について教えてください。	Please tell me about local education administration.

4. Preliminary Results



Question: Please tell me how local governments handle educational data analysis.

RAG(GPT-3.5)

Regarding the handling of educational data analysis in local governments, several initiatives have been shown.

Less advanced local governments are conducting their own questionnaire surveys, analyzing the results and using them for policy, and are creating an environment for collecting and storing educational data.

On the other hand, advanced local governments are visualizing educational data through dashboards, etc., and deploying it to schools.

As for specific initiatives in local governments, in local government A, a questionnaire survey is being conducted and analyzed, while in local governments E and F, the lack of a promoter of data collection and data utilization has hindered efforts to analyze educational data. It is mentioned that this is an issue.

Furthermore, local government I plans to proceed with statistical analysis using questionnaire survey data, and the idea of using data to help formulate policies is presented.

RAG(GPT-4)

The handling of educational data analysis varies by local government. For example, some local governments conduct their own questionnaire surveys, analyze the results, and use them for policy. Additionally, there are moves toward improving the environment for collecting and storing educational data.

Advanced local governments are visualizing educational data on dashboards and deploying it to schools. However, there are also obstacles to educational data analysis efforts. Examples include the lack of a well-developed platform for data collection and the lack of a promoter of data utilization.

In addition, for example, if there are no reference examples using dashboards, teachers will not know what they can analyze from the data, making it difficult to analyze data. In order to solve these issues, local governments are moving forward with efforts to utilize and analyze educational data.

GPT-3.5

The handling of local governments in educational data analysis generally has the following policies and precautions.

1. Privacy protection: Educational data containing personal information is treated as confidential information, so sufficient consideration must be given to privacy protection. Information that can identify individuals must be appropriately anonymized and handled in accordance with the Personal Information Protection Act.

2. Legality and Transparency: Acquisition and analysis of educational data must be done legally and through transparent procedures. It is desirable for local governments to clarify the purpose and method of data use before doing so.

3. Maximizing utilization and results: Educational data analysis is expected to be used to improve educational measures and improve the quality of education. It is important for local governments to utilize the information and knowledge obtained from educational data and link it to the effective provision of educational services.

4. Security measures: Educational data is highly confidential, so adequate security measures are required when storing and handling data. It is important to take measures to prevent unauthorized access and information leaks.

When local governments analyze educational data, they are required to handle it appropriately while taking the above points into consideration.

GPT-4

Educational data analysis refers to the collection, structuring, and analysis of various information related to education, such as student and teacher performance, school efficiency, and educational program effectiveness. Using this data, institutions can make more effective decisions and improve learning outcomes.

However, when handling educational data in local governments, it is necessary to pay attention to the following points:

1. Privacy and data protection: Personal data of students and teachers must be handled with care. Therefore, it is necessary to respect data protection and privacy laws and regulations and take appropriate security measures.

2. Data quality: To ensure the accuracy and reliability of the analysis, it is necessary to ensure the quality of the data collected. The methods of data collection, data management and processing processes are important.

3. Data analysis and interpretation: Data analysis experts may be required. It also requires the ability to accurately interpret data analysis results and formulate effective educational policies based on them.

4. Equality: It is also important to use data analysis results to eliminate disparities in education so that all children have a fair opportunity for education.

Taking these points into consideration, local governments can properly utilize education data and make decisions to improve the quality of education. Additionally, the collected data can be released as open data to encourage secondary use by the general public and other researchers. This will allow us to approach educational issues from a broader and more diverse perspective, and will also gain public understanding and support for educational reform.

Figure 2. Examples of generated text from each model

Examining the specific responses from each model (Figure 2), it appears that the models using RAG, namely RAG (GPT-3.5) and RAG (GPT-4), provide more specific and content-focused answers based on the retrieved information. In contrast, models without RAG, GPT-3.5 and GPT-4, tend to produce verbose, bullet-pointed, and generic responses. The use of the RAG allowed for responses based on search results, such as unique content found in the dataset; for example, initiatives by the Fukuoka Prefectural Board of Education confirmed that it provided more specific answers about local knowledge in Japan. Interestingly, in RAG (4.0), when the retrieval results are insufficient, it can start its response with "Unfortunately, the provided text does not explicitly contain information about specific issues related to non-cognitive abilities," suggesting it can judge the quality of retrieval results. This response was not observed for RAG (GPT-3.5).

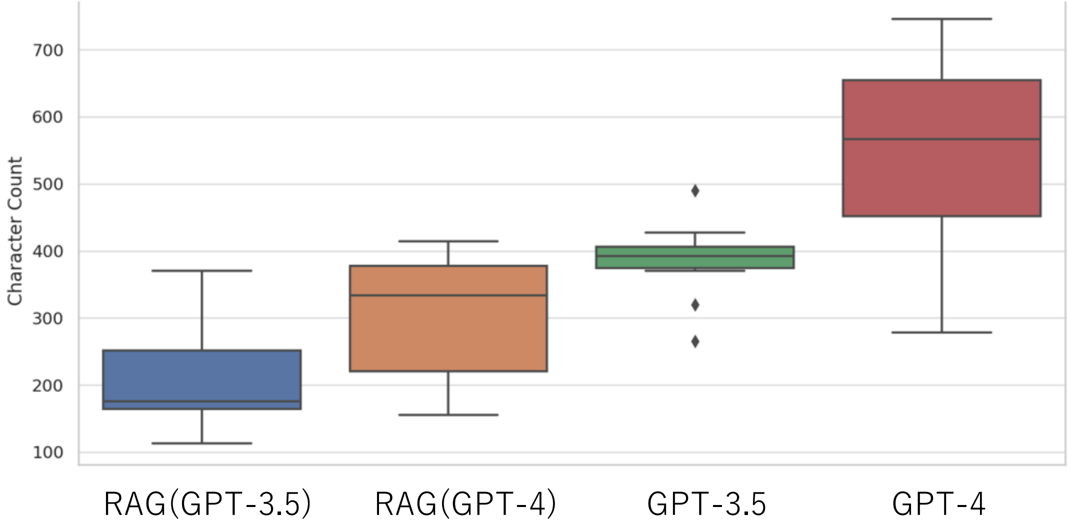


Figure 3. The length of the generated text in each model.

Next, we examined the lengths of the strings of text generated from a set of 10 questions (Figure 3). As you can see, the two graphs on the right, which do not use RAG, show that the length of the strings is longer compared to when RAG is applied. These results suggest that using RAG may allow for more concise and locally informed responses, which could be more accurate and easier for users to understand than the traditional longer replies. Furthermore, the similarity of texts across the models was investigated. Figure 4 shows the cosine similarity scores for each pair of models. This illustrates the degree of similarity between the text content of these models. While the similarity is high between RAG (GPT-3.5) and RAG (GPT-4.0), it is thought that the text was generated from similar retrieved text results by the RAG, implying the consistency and reproducibility of the retrieval results. In addition, the highest similarity between RAG (GPT-4.0) and GPT-4.0 may be because the length of the text generated by GPT-4.0 is greater than that generated by GPT-3.5 (Figure 3), and the text segments originally generated by GPT-4.0 are more similar to each other than those generated from the retrieval results by RAG. In this way, comparing the text length and cosine similarity between cases where the RAG is used and those where it is not used might be helpful for evaluating the effectiveness of the RAG.

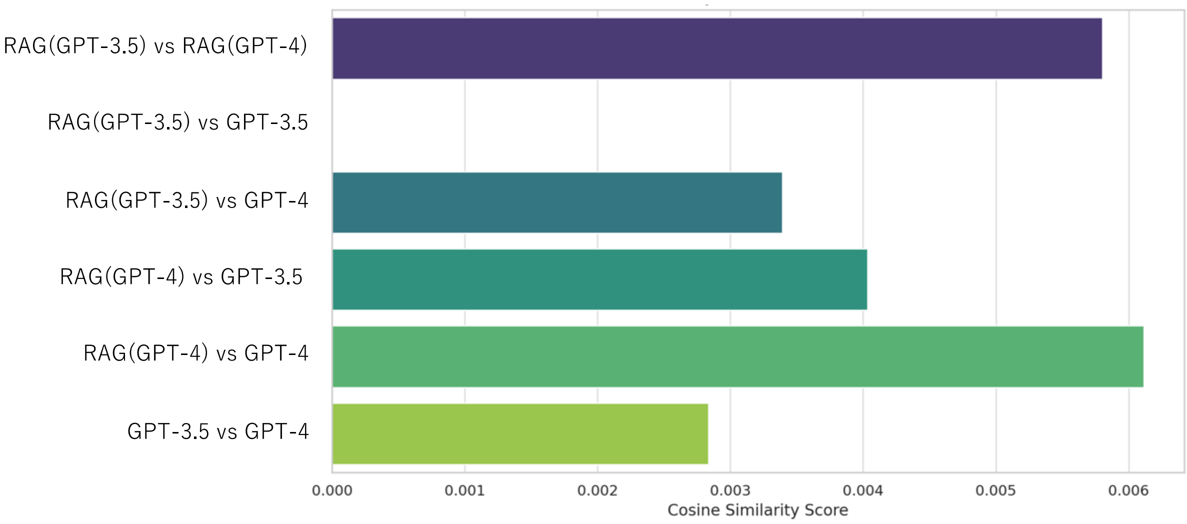


Figure 4. Cosine Similarity Score in each model.

5. Limitations

One limitation of this study is the rapid development of LLMs, which poses the risk that the results obtained from the models such as GPT-3.5 and 4.0, will become outdated. Therefore, it is necessary to validate these findings using the latest LLM models. To achieve this, it is crucial to develop evaluation datasets for LLMs in local languages such as Japanese. This includes automatic evaluation benchmarks like llm-jp-eval (https://github.com/llm-jp/llm-jp-eval) developed by LLM-jp (LLM-jp et al., 2024), which develops LLM with strong Japanese knowledge. The question list (Table 1) used in this study may be useful for creating such evaluation dataset.

Another limitation of this study is that the evaluation of RAG was based solely on simplistic metrics, such as text length and cosine similarity. To determine what constitutes a good response for policymakers, researchers, or school practitioners, it is necessary to implement a chat system that applies RAG and conducts user evaluations of the LLM's responses.

6. Conclusions

In this study, it was possible to confirm that LLMs can provide answers to local education-related knowledge, which has not been adequately addressed before, using RAG with open public education data. In the context of formulating educational policies, collecting evidence (data that can serve as the basis for policies) is an important step in the initial process of policy formulation (Bardach & Patashnik, 2023). Applying RAG to public education data, as conducted in this study, could not only contribute to the collection of such evidence but also brainstorming utilizing local knowledge that is not incorporated into the training of large language models (LLMs), providing concise responses based on local knowledge. This capability can be leveraged to assist in policymaking aligned with the specific characteristics of different countries and regions. However, since the accuracy of retrieval results and the extent to which the generated outputs are based on retrieval results and are useful for policymakers, researchers, or practitioners have not yet been verified, future development and verification of RAG chat systems could enable the effective use of LLMs with local educational data, potentially aiding in the formulation of educational policies.

Acknowledgements

This study was supported by JSPS JP23K17012 and 23H01001.

References

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, *32*(4), 399–418. https://doi.org/10.1016/j.giq.2015.07.006

Bardach, E., & Patashnik, E. M. (2023). *A Practical Guide for Policy Analysis: The Eightfold Path to More Effective Problem Solving*. CQ Press. https://play.google.com/store/books/details?id=WXG7EAAAQBAJ

CKAN. (n.d.). *The world's leading open source data management system*. 2006-2023. https://ckan.org/

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2312.10997

Grobit. (2008-2023). https://github.com/kermitt2/grobid

LangChain. (2023). https://github.com/langchain-ai/langchain

LLM-jp, Aizawa, A., Aramaki, E., Chen, B., Cheng, F., Deguchi, H., Enomoto, R., Fujii, K., Fukumoto, K., Fukushima, T., Han, N., Harada, Y., Hashimoto, C., Hiraoka, T., Hisada, S., Hosokawa, S., Jie, L., Kamata, K., Kanazawa, T., … Yoshino, K. (2024). LLM-jp: A cross-organizational project for the research and development of fully open Japanese LLMs. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2407.03963

National Institute for Educational Policy Research (NIER). (2023). *Education Data Platform*. https://edpportal.nier.go.jp/

OpenAI. (2022). *Introducing ChatGPT*. https://openai.com/blog/chatgpt/

Ragas. (2023). *Ragas*. https://docs.ragas.io/en/stable/

Takami, K. (2023). Exploring ChatGPT Performance on PISA Multiple Choice Sample Questions Comparing English and Japanese Expression. *31th International Conference on Computers in Education Conference Proceedings*, 27–32.

TEI Consortium. (2023). *Tei p5: Guidelines for electronic text encoding and inter change*. https://tei-c.org/guidelines/p5/