BETH Bulletin 2026

No. 04

Hannie Riley, Esther Jeon,
and Jussi Hyvärinen

From Paper to FAIR: Using AI to Digitise and Curate Sensitive Research Data

Henrik Holtvedt Andersen

Research Librarian, MF Norwegian School of Theology, Religion and Society, Oslo, Norway

1. Introduction

The digital transformation of research has accelerated in recent years. Governments, research institutions, and funders are increasingly demanding greater transparency and the open sharing of research data. In Norway, the National Strategy on Access to and Sharing of Research Data articulates a clear goal: all publicly funded research data should, whenever possible, be made available for reuse under FAIR principles (Norwegian Ministry of Education and Research 2018). FAIR stands for Findable, Accessible, Interoperable, and Reusable. However, the reality in many research environments remains complex. Transforming data into FAIR data requires both technical skill and ethical awareness, and can be time-consuming.

Librarians are increasingly involved in research data management (RDM), yet many report inadequate data competencies (Antell et al. 2014). Researchers report a lack of time, skills, and knowledge regarding sensitive data management as barriers to sharing research data (Lillo 2022). One motivation for using AI in RDM is therefore to determine whether AI can be used to tackle these challenges and make RDM more effective, and to assess whether AI can assist librarians and researchers in developing data competence.

This paper builds on my experience as a research librarian supporting a project to digitise and manage the sensitive, paper-based datasets of theologian Erling Birkedal. The goal is to preserve the data for long-term archival storage and to share it openly in a research data repository upon the finalisation of the research project. FAIR data management is required to achieve this, and I have incorporated AI into the workflow. This project seeks to answer two main questions:

1. How can AI chatbots make the process of RDM more efficient, especially when dealing with analogue data?
2. How can AI chatbots support librarians and researchers in developing data competence for RDM?

The project also seeks to answer and raise awareness about these questions:

3. How can AI be used responsibly in projects involving personal and sensitive data?
4. What implications do AI-assisted RDM have for sustainability?

2. Research Data Management and the FAIR Principles

Research Data Management refers to the systematic handling of data throughout its lifecycle, from collection and processing, to storage, preservation, and dissemination. RDM ensures transparency, reproducibility, and compliance with legal and ethical standards (Science Europe n.d.). The FAIR data principles aim to make data more discoverable and reusable by humans and machines alike. FAIR data must be Findable, Accessible, Interoperable, Reusable (FAIR Principles n.d.).

3. The Birkedal Project: Religiosity from Childhood to Adulthood

Erling Birkedal's research project Religiosity from Childhood to Adulthood followed a group of Norwegian participants for over three decades. The project explored how individual beliefs evolve from adolescence into adulthood, and has produced several key works (Birkedal 2001; 2008; 2020). The project remains active, with plans to undertake one or more data collections in the future.

The dataset consists primarily of four rounds of questionnaires (1994, 1996, 2007, 2016) and complementary interview data. The interview data is not used in this project; only the questionnaires are used, as the interview data contains too much direct and indirect personal information to be shared openly. Anonymising this data would be time-consuming and require the deletion of important research data. The questionnaires contain reflections on faith, moral values, and personal experiences of religion, and these topics are categorised as sensitive personal data under GDPR. Since the respondents were minors during the first rounds of data collection, this data is also automatically considered sensitive personal data.

4 Methodology

4.1 Scanning and Secure Storage

All questionnaires were digitised using high-resolution scans in JPG format, as this worked better with ChatGPT than the complex PDF format. Scanned files were stored on the institutional OneDrive system, with sensitive material additionally encrypted with Cryptomator. This ensures secure access management and compliance with institutional data policies.

4.2 Anonymisation Procedures

Anonymisation was carried out manually. First, all direct and indirect identifiers were removed before any AI interaction took place. After AI interaction, some indirect data (e.g., names, dates, schools, locations) were reintroduced with generalised placeholders. For instance, a handwritten date like 26.03.2006 has become past six months, retaining research value while removing identifiability. In this way, all direct and indirect identifiers were anonymised. We had two linking keys, one on paper and another stored in an encrypted folder.

4.3 Prompt Engineering

AI transcription relied on ‘prompt engineering’, or the crafting of precise instructions

for ChatGPT to transform image-based responses into CSV tables. The prompts evolved over time from simpler, open prompts to more complex, instructive prompts to secure consistent, standardised output. They included:

- A clear description of the task and data context
- Column definitions and naming conventions
- Example outputs for reference (few-shot prompting)

Following these guidelines, ChatGPT produced fairly consistent, desirable output. Here is an example of a prompt used in the project, together with image uploads of a questionnaire:

> *use these columns:*
> *year,belief_in_god,belief_importance,religiosity,christianity,changed_opinion [...]*
>
> *This is a sample response from another questionnaire:*
> *2007,1,3,5,3, No, Have become more religious, Have faith in God [...]*
>
> *Follow this example and create a CSV file of the uploaded images.*
> *Do not invent values, leave cells blank if the questions from the form are unanswered-*
>
> *This is a questionnaire where informants have responded to the same questions (more or less) three times, once in 1996 and later in 2007 and 2016. Transform this into a dataset with rows and columns in CSV format. Think about FAIR data quality in the dataset. This will be shared as open research data and the data can be used by others. Do not invent values if questions are blank. Interpret the handwriting to the best of your abilities.*

4.4 Reviewing, Cleaning, and Curating

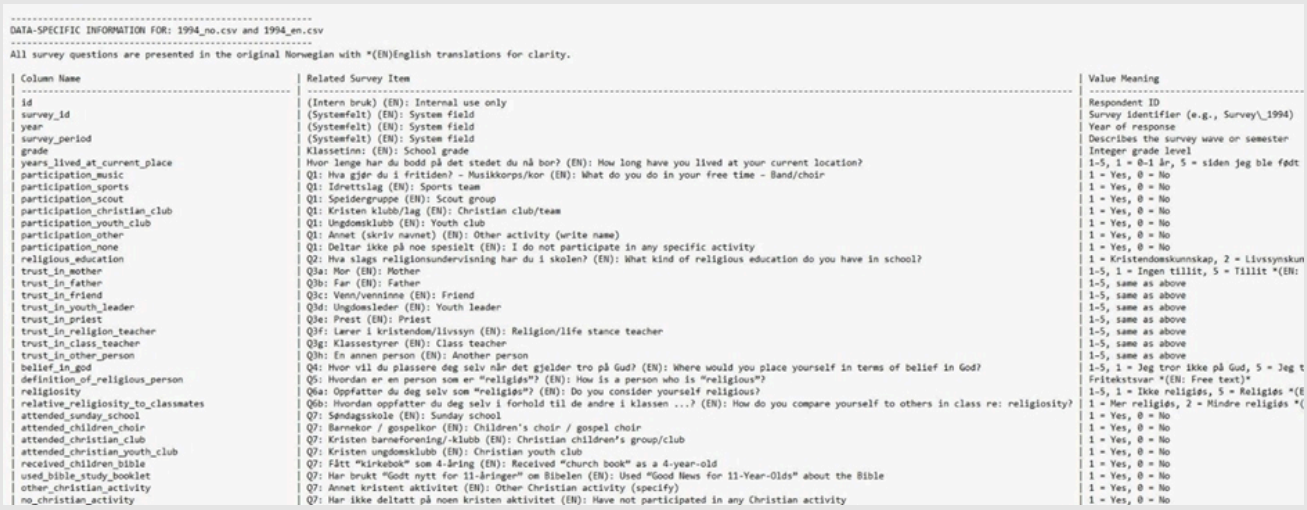
The raw transcribed data from ChatGPT was about 80 to 90% accurate. Whilst the use of AI significantly reduced the manual workload, human review was still essential. Errors, particularly misread handwriting or hallucinated values, were corrected manually and edited in OpenRefine. This review made the dataset achieve

a high standard of accuracy. Controlled cleaning and normalisation included normalising column names and values, capitalisation, trimming consecutive whitespace, and more. Some cleaning and normalisation were done using GREL commands in OpenRefine, generated by ChatGPT. Most cleaning tasks were performed effectively in OpenRefine because it provides reliable, deterministic cleaning operations. It is possible to get ChatGPT to perform many of these cleaning operations as well, and indeed, some instructions were given to this effect within the transcription prompt, with some degree of success. However, the main drawback is that ChatGPT's output must always be reviewed carefully because it is prone to error. This review is time-consuming, so the cost outweighs the gain for many of these cleaning operations using ChatGPT.

To facilitate longitudinal analysis, ChatGPT was used to merge all surveys into a single comprehensive dataset, referred to as the 'superfile'. All unique columns were preserved, and duplicates were avoided. The superfile contained many columns and empty cells because the questionnaire was not entirely consistent over the years. This complexity hindered human interpretability, so we produced a table with core questions throughout the years. Then another cleaning task became relevant: inputting NA (not applicable) in empty cells where questions were never asked, rather than in cells left intentionally blank. This input was done using ChatGPT because it could quickly and automatically detect the difference between those two types of empty cells by comparing the questionnaires and the CSV file. A review was also needed for these tasks, but in these cases, my experience was that the gain from using AI outweighed the cost.

4.5 README-file and Data Dictionary

To ensure adherence to the FAIR standard regarding Reusability, the dataset was accompanied by a README file and a data dictionary generated by ChatGPT.
A data dictionary was created by uploading all the image scans from each survey together with the CSV file (the tabular data) from the corresponding survey, with a prompt asking ChatGPT to map the questions to the column names and provide useful information about the values. The result was produced in .txt format and required some visual adjustments before it showed a neat table with a data dictionary.



DATA-SPECIFIC INFORMATION FOR: 1994_no.csv and 1994_en.csv

All survey questions are presented in the original Norwegian with *(EN)English translations for clarity.

Column Name	Related Survey Item	Value Meaning
id	(Intern bruk) (EN): Internal use only	Respondent ID
survey_id	(Systemfelt) (EN): System field	Survey identifier (e.g., Survey_1994)
year	(Systemfelt) (EN): System field	Year of response
survey_period	(Systemfelt) (EN): System field	Describes the survey wave or semester
grade	Klassetrinn: (EN): School grade	Integer grade level
years_lived_at_current_place	Hvor lenge har du bodd på det stedet du nå bor? (EN): How long have you lived at your current location?	1-5, 1 = 0-1 år, 5 = siden jeg ble født
participation_music	Q1: Hva gjør du i fritiden? – Musikkorps/kor (EN): What do you do in your free time – Band/choir	1 = Yes, 0 = No
participation_sports	Q1: Idrettslag (EN): Sports team	1 = Yes, 0 = No
participation_scout	Q1: Speidergruppe (EN): Scout group	1 = Yes, 0 = No
participation_christian_club	Q1: Kristen klubb/lag (EN): Christian club/team	1 = Yes, 0 = No
participation_youth_club	Q1: Ungdomsklubb (EN): Youth club	1 = Yes, 0 = No
participation_other	Q1: Annet (skriv navnet) (EN): Other activity (write name)	1 = Yes, 0 = No
participation_none	Q1: Deltar ikke på noe spesielt (EN): I do not participate in any specific activity	1 = Yes, 0 = No
religious_education	Q2: Hva slags religionsundervisning har du i skolen? (EN): What kind of religious education do you have in school?	1 = Kristendomskunnskap, 2 = Livssynskun
trust_in_mother	Q3a: Mor (EN): Mother	1-5, 1 = Ingen tillit, 5 = Tillit *(EN:
trust_in_father	Q3b: Far (EN): Father	1-5, same as above
trust_in_friend	Q3c: Venn/venninne (EN): Friend	1-5, same as above
trust_in_youth_leader	Q3d: Ungdomsleder (EN): Youth leader	1-5, same as above
trust_in_priest	Q3e: Prest (EN): Priest	1-5, same as above
trust_in_religion_teacher	Q3f: Lærer i kristendom/livssyn (EN): Religion/life stance teacher	1-5, same as above
trust_in_class_teacher	Q3g: Klassestyrer (EN): Class teacher	1-5, same as above
trust_in_other_person	Q3h: En annen person (EN): Another person	1-5, same as above
belief_in_god	Q4: Hvor vil du plassere deg selv når det gjelder tro på Gud? (EN): Where would you place yourself in terms of belief in God?	1-5, 1 = Jeg tror ikke på Gud, 5 = Jeg t
definition_of_religious_person	Q5: Hvordan er en person som er "religiøs"? (EN): How is a person who is "religious"?	Fritekstsvar *(EN: Free text)*
religiosity	Q6a: Oppfatter du deg selv som "religiøs"? (EN): Do you consider yourself religious?	1-5, 1 = Ikke religiøs, 5 = Religiøs *(E
relative_religiosity_to_classmates	Q6b: Hvordan oppfatter du deg selv i forhold til de andre i klassen ...? (EN): How do you compare yourself to others in class re: religiosity?	1 = Mer religiøs, 2 = Mindre religiøs *(
attended_sunday_school	Q7: Søndagsskole (EN): Sunday school	1 = Yes, 0 = No
attended_children_choir	Q7: Barnekor / gospelkor (EN): Children's choir / gospel choir	1 = Yes, 0 = No
attended_christian_club	Q7: Kristen barneforening/-klubb (EN): Christian children's group/club	1 = Yes, 0 = No
attended_christian_youth_club	Q7: Kristen ungdomsklubb (EN): Christian youth club	1 = Yes, 0 = No
received_children_bible	Q7: Fått "kirkebok" som 4-åring (EN): Received "church book" as a 4-year-old	1 = Yes, 0 = No
used_bible_study_booklet	Q7: Har brukt "Godt nytt for 11-åringer" om Bibelen (EN): Used "Good News for 11-Year-Olds" about the Bible	1 = Yes, 0 = No
other_christian_activity	Q7: Annet kristent aktivitet (EN): Other Christian activity (specify)	1 = Yes, 0 = No
no_christian_activity	Q7: Har ikke deltatt på noen kristen aktivitet (EN): Have not participated in any Christian activity	1 = Yes, 0 = No

Figure 1. Data dictionary from README-file

5. Legal, Ethical and Institutional Frameworks

5.1 GDPR and AI

The processing of sensitive research data via AI chatbots presents a complex legal challenge under the General Data Protection Regulation (GDPR) (European Union 2016). While pseudonymized data (data with a linking key) is still considered personal data under GDPR, recent legal developments have clarified how this applies to third-party AI processing.

The workflow for the Birkedal project relied on the principle of 'effective anonymisation' for third parties. Before any data was uploaded to ChatGPT, all direct and indirect identifiers were redacted. Whilst the researcher maintained a linking key to allow for future longitudinal study, this key remained strictly offline and stored on secure local infrastructure.

5.2 The SRB Ruling and Sikt Consultation

This approach aligns with the Court of Justice of the European Union (CJEU) ruling in European Data Protection Supervisor v. Single Resolution Board (2025). The court clarified that data held by a recipient (in this case, ChatGPT) is not necessarily 'personal data' if that recipient has no legal or practical means to access the additional information required to re-identify the individuals.

The anonymisation strategy was reviewed in consultation with the Norwegian Agency for Shared Services in Education and Research (Sikt n.d). Sikt's guidance confirmed that, due to the fact that the linking keys remained securely offline and out of reach for the AI provider, the data had effectively been anonymised for AI processing. Consequently, even if the researcher views the data as pseudonymised, the AI sees it as anonymous.

5.3. Institutional and National Variations

Although the project used ChatGPT Team, a 'closed' service that does not use data to train the underlying model, this does not automatically grant GDPR compliance in all research contexts. Under the EU Artificial Intelligence Act (AI) (EU Artificial Intelligence Act (AI) n.d.) and Norwegian national provisions, researchers must remain vigilant about where data is stored and processed. Although the principles of the SRB ruling provide a stronger legal basis for AI-assisted RDM, individual countries may have specific national requirements for handling sensitive research data.

6. Sustainability and the Energy Footprint of AI

One study estimates that generating a single large language model response can require up to 30 times more energy than a traditional web search (De Vries 2023). Training large models such as GPT-4 is associated with substantial CO_2 emissions. Google's reported emissions have increased by 51 percent since 2019 (Horton 2025), while Microsoft's emissions rose by 29 per cent between 2020 and 2023 (Kerr 2024). Google and Microsoft attribute much of their emissions growth to increased energy consumption in data centres (Horton 2025; Kerr 2024). AI use and infrastructure continue to expand; data centres are estimated to account for up to 4.5 percent of global electricity consumption by 2030, compared to approximately 1 per cent today (Chen 2025).

At the same time, promising developments are emerging. Newer models such as DeepSeek have reportedly reduced training-related energy use by approximately tenfold compared to ChatGPT (Williment 2025). The inference cost for systems

performing at the level of smaller capable models, such as GPT-3.5, declined more than 280-fold between November 2022 and October 2024 (The 2025 AI Index Report n.d.). Advances in data-centre efficiency are also underway, including the reuse of waste heat to supply nearby buildings (Sustainability Heats Up in Norwegian Data Centre Industry n.d.). In addition, energy efficiency at the hardware level has improved by approximately 40 percent annually (The 2025 AI Index Report, n.d.). Regulatory developments further signal a shift toward more sustainable AI practices, as the EU Artificial Intelligence Act introduces new reporting requirements related to energy use and sustainability (European Union n.d.). Finally, major technology companies have articulated long-term climate commitments, with Google aiming for carbon neutrality and Microsoft for carbon negativity by 2030 (Kerr 2024).

6.1. Practical Recommendations

Practical recommendations include limiting AI use to high-value, irreplaceable tasks. The transcription of the survey images in this project is arguably an example of this. For data cleaning and transformation, it is advisable to use non-AI tools such as OpenRefine or Python scripts where possible. In particular, OpenRefine is an easy, reliable, and 'low energy' tool. Not all AI tasks are equally energy-intensive; there is a big difference between having ChatGPT summarise a page and having it create a picture, as the latter is a lot more energy-intensive (Chen 2025). It is recommended to be aware of this difference and not make unnecessary energy-intensive prompts.

7. Discussion and conclusion

This project reveals that AI can be a useful tool in RDM. For non-coding professionals, ChatGPT provides access to data-processing methods without requiring deep programming knowledge. Automatic transcription reduced the time required to digitise Birkedal's surveys. Since ChatGPT can hallucinate and misinterpret, which is a concern for data quality and integrity, the subsequent need for review is a drawback as it reduces efficiency and sometimes causes the cost of using AI Chatbots in RDM to outweigh the gain. However, if the reliability of AI Chatbot output continues to improve, thus diminishing the need for additional human review, it would serve as an even more effective and useful tool in RDM.

It is highly recommended to develop competence in RDM before using AI in RDM. However, one can arguably develop or refresh RDM skills with the help of AI chatbots. When ChatGPT was prompted to give curating advice, it generally offered sensible, useful, and reliable answers. ChatGPT can also be used as a search engine that produces links to research articles and web pages with reliable information, which can then be used to fact-check the content in the output window. However, these links are not always reliable or relevant, so independent, external fact-checking is also necessary. Whilst AI can help with complex data programmatic tasks, it is also true that the output from AI chatbots often becomes less useful and accurate the more complex and niche your question is. This is due to its probabilistic mechanics and limited training data, which make most AI chatbots prone to overgeneralization. Some AI sceptics warn against overreliance on AI and argue that it can create intellectual laziness. Although these are serious and valid concerns, it is still possible to use AI in a critical and intellectually stimulating manner. In this particular project, the transcribing of questionnaires into a digital tabular form using AI is a more intellectually stimulating and creative process than manually creating the tables and plotting all the values by hand.

Great care should be taken when dealing with sensitive data and AI tools. For this project, there was a legal obligation to process the data using ChatGPT only after the data had been anonymised so that ChatGPT could not, under any reasonable circumstances, re-identify the respondents. We kept a paper linking key, meaning that the data remained pseudonymised for us, but for ChatGPT, it was effectively anonymised. However, researchers are advised to consult their institution’s data protection officer or national authority (e.g., Sikt or Datatilsynet in Norway) before applying these practices in their own contexts.

For the method outlined in this project, it would be beneficial to have secure, top-quality AI chatbots (e.g., media uploads) that allow for the processing of personal data, and perhaps even non-anonymised sensitive data. Not having to anonymise the material before AI processing would significantly increase the method's efficiency. To my knowledge, there was no such AI chatbot available in Norway as of 2025, though GPT UiO («GPT UiO n.d.) and the chatbot by Sikt («Sikt KI n.d.) partially met some of these criteria.

AI-assisted RDM, as performed in this project, is likely to be significantly more energy-intensive than RDM performed without the use of AI. This is concerning and provides a valid environmental argument against using AI in RDM. However, we are witnessing the emergence of AI models and infrastructure that are more energy-efficient than before, as well as green goals and commitments from AI companies and policymakers related to AI technology. This offers some hope for a greener AI future, though the increase in AI use is concerning and pulling in the opposite direction.

Declaration of AI Use:

Research project:
ChatGPT Team (OpenAI) was used during 2025 to support transcription and research data management, mostly relying on the platform's default model configuration. As OpenAI updates its models continuously, the underlying model version may have varied over time.

Article Writing:
The author used ChatGPT Team (OpenAI) with the default model configuration available between October 2025 and January 2026. The specific underlying model version may have changed during this period. The first draft originated from a PowerPoint presentation from the author at the 2025 BETH Conference in Stavanger, with the same content as this article. This was processed through ChatGPT to produce an approximately ten-page draft article. This AI-generated text was then substantially revised and expanded by the author. The abstract and parts of the introduction and theory section were lightly edited, while the rest were significantly rewritten and supplemented. Finally, Gemini Plus 3 Flash was utilised for linguistic refinement.

Bibliography

Antell, Karen, Jody Bales Foote, Jaymie Turner, and Brian Shults. 2014. "Dealing with Data: Science Librarians' Participation in Data Management at Association of Research Libraries Institutions." *College & Research Libraries* 75 (4): 557–74. https://doi.org/10.5860/crl.75.4.557.

Birkedal, Erling. 2001. *"Noen ganger tror jeg på Gud, men ... "?: en undersøkelse av gudstro og erfaring med religiøs praksis i tidlig ungdomsalder*. Tapir.

Birkedal, Erling. 2008. *«Kanskje jeg tror på en gud, men-»: en langtidsstudie av gudstro hos unge fra 13 til 25 år*. IKO-Forlaget.

Birkedal, Erling. 2020. *Religiøsitet fra barndom til voksen : tretten livshistorier om tro og livstolkning*. IKO-Forlaget.

Chen, Sophia. 2025. "How Much Energy Will AI Really Consume? The Good, the Bad and the Unknown." Nature 639 (8053): 22–24. https://doi.org/10.1038/d41586-025-00616-z.

De Vries, Alex. 2023. "The Growing Energy Footprint of Artificial Intelligence." *Joule* 7 (10): 2191–94. https://doi.org/10.1016/j.joule.2023.09.004.

EU Artificial Intelligence (AI) Act. n.d. Accessed October 7, 2025. https://enterprise.gov.ie/en/what-we-do/innovation-research-development/artificial-intelligence/eu-ai-act/eu-ai-act.html.

European Data Protection Supervisor v Single Resolution Board, Case C-413/23 P (ECJ 4. September 2025). Accessed Octboer 7, 2025. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62023CJ0413.

European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance). 119 OJ L. http://data.europa.eu/eli/reg/2016/679/oj.

Go FAIR. n.d. "FAIR Principles." Accessed October 22, 2025. https://www.go-fair.org/fair-principles/.

University of Oslow. n.d. "GPT UiO – UiOs Privacy Friendly GPT Chat." Accessed Oktober 20, 2025. https://www.uio.no/english/services/it/ai/gpt-uio/index.html.

Horton, Helena. 2025. "Google's Emissions up 51% as AI Electricity Demand Derails Efforts to Go Green." *The Guardian*, June 27. Accessed October 20, 2025. https://www.theguardian.com/technology/2025/jun/27/google-emissions-ai-electricity-demand-derail-efforts-green.

Kerr, Dara. 2024. "AI Brings Soaring Emissions for Google and Microsoft, a Major Contributor to Climate Change." *NPR*, July 12. Accessed October 20, 2025. https://www.npr.org/2024/07/12/g-s1-9545/ai-brings-soaring-emissions-for-google-and-microsoft-a-major-contributor-to-climate-change.

Lillo, Karen. 2022. "Deling av forskningsdata - forskerens perspektiv: En intervjustudie med forskere ved norske forskningsinstitusjoner." *NIBR-notat*. Oslo:OsloMet-Storbyuniversitetet. Accessed October 20, 2025. https://oda.oslomet.no/oda-xmlui/handle/11250/3017186.

Lo, Leo S. 2023. "The CLEAR path: A framework for enhancing information literacy through prompt engineering." *The Journal of Academic Librarianship* 49 (4): 102720. https://doi.org/10.1016/j.acalib.2023.102720.

Norwegian Ministry of Education and Research. 2018. "National Strategy on Access to and Sharing of Research Data." *Government.no*, February 13. Accessed October 20, 2025. https://www.regjeringen.no/en/documents/national-strategy-on-access-to-and-sharing-of-research-data/id2582412/.

Science Europe. n.d. "Research Data Management." *Science Europe*. Accessed October 22, 2025. https://scienceeurope.org/our-priorities/open-science/research-data-management/.

Sikt - Norwegian Agency for Shared Services in Education and Research. n.d. Accessed January 27, 2026. https://sikt.no/en/home.

Sikt KI. n.d. Accessed January 27, 2026. https://sikt.no/tjenester/sikt-ki.

"Sustainability Heats up in Norwegian Data Centre Industry." n.d. Accessed October 20, 2025. https://businessnorway.com/articles/sustainability-heats-up-in-norwegian-data-centre-industry.

The 2025 AI Index Report. n.d. Stanford Institute for Human-Centered Artifical Intelligence (HAI). Accessed October 27, 2025. https://hai.stanford.edu/ai-index/2025-ai-index-report.

Williment, Chloe. 2025. "AI's Energy Toll: Greenly Compares ChatGPT-4 and DeepSeek." *Energy Digital*, May 15, Accessed October 22 2025. https://energydigital.com/articles/ais-energy-toll-greenly-compares-chatgpt-4-and-deepseek.