

# Towards Managing Design Science Knowledge with Large Language Models

*Emergent Research Forum (ERF) Paper*

**Rodrigo Folha**  
Center of Informatics  
Federal University of Pernambuco  
rbf2@cin.ufpe.br

**Arthur Carvalho**  
Farmer School of Business  
Miami University  
arthur.carvalho@miamioh.edu

## Abstract

This study explores the utilization of Large Language Models (LLMs) in managing Design Science Research (DSR) knowledge. It addresses the absence of a consolidated knowledge base for DSR artifacts, proposing an AI-driven approach for extracting and structuring information from scientific articles. Three experiments were conducted to evaluate the effectiveness of LLMs — specifically GPT 3.5 and GPT 4 — when the requested information is 1) spread across multiple articles, 2) contained within a single article, and 3) well-structured inside a single file. Our results indicate the potential of AI tools in efficient information extraction and real-time data processing. However, they also underscore the limitations of current AI models, such as underperforming when extracting DSR-related information spread across many articles. Overall, our research contributes to the understanding of AI capabilities in enhancing DSR knowledge management while identifying promising areas for further exploration.

## Keywords

Artificial intelligence, design knowledge, design science, large language models.

## Introduction

In Information Systems (IS), the design science research (DSR) methodology is employed to incrementally build a knowledge base and artifacts that can be shared with new researchers. Despite having a highly productive scientific community with well-established associations, conferences, and journals, there is currently no de facto shared knowledge base where artifacts and principles can be systematically indexed. A potential reason is that although editorial guidelines for journals and conference proceedings often establish a standard format for article structure, the same does not apply to the communication of research within the article itself. In the particular context of DSR articles, many authors do not follow standards when communicating their design knowledge, thus making the indexing activity extremely challenging. Given the above issues and research gap alongside the rise of generative artificial intelligence (AI), we report in this article a preliminary attempt to build a knowledge management tool based on state-of-the-art large language models. Thus, our overarching research question is, *can AI enhance the recognition of design knowledge descriptions and contribute to the construction of a DSR knowledge base?*

## Research Background

Design science research in information systems involves the creation and evaluation of design knowledge that leads to artifacts intended to solve problems in various fields ranging from healthcare (Anderson, Carvalho, and Merhout, 2022) to logistics (Folha et al., 2022). Ideally, these artifacts are represented in a structured form, which can vary from formal logic and rigorous mathematics to informal natural language descriptions. Some frameworks have been proposed to guide design science researchers in their endeavors. For example, Peffers et al. (2014) introduced a guideline for IS researchers to implement a DSR methodology through a series of well-defined activities, namely *problem identification and motivation*,

*definition of objectives for a solution, design and development, demonstration, evaluation, and communication.* Focusing on the first activity, the authors elaborated that there should be an effort on the part of researchers to transform identified problems into objectives that the system must address, commonly known as *design requirements* (DR). Another term frequently employed in DSR is *design principles* (DP), which are “*prescriptive statements that indicate how to do something to achieve a goal*” (Gregor, Kruse, and Seidel, 2020). Finally, *design features* (DF) elucidate concretely how to build the artifact. Collectively, DRs, DPs, and DFs are called *design knowledge descriptions* (DKDs).

A recurrent concern in DSR is that there is no structured approach for accumulating knowledge, specifically in refining or expanding upon prior design knowledge (Reining et al., 2022). Trying to fill that gap, Gau et al. (2023) created a mining mechanism to automatically extract some data from published DSR articles, such as title, year, journal or conference name, author names, keywords, design requirements, principles, and features. That information is stored in a database that supports searches based on specific keywords. However, mining unstructured data in academic articles is not a straightforward process, as authors do not usually follow a standard template for reporting their results. Even when narrowing down to design requirements, principles, and features, there is no standard writing pattern. The solution by Gau et al. (2023) was to search for specific acronyms, such as DP, DR, or DF, followed by a detailed description of the respective terms. Naturally, that approach fails to capture the several cases when acronyms are not used. We contribute to the above discussion and underlying literature by reporting on preliminary experiments about how modern AI tools based on large language models (LLMs) may help manage design knowledge.

## Experiments

To answer our research question, we performed three different experiments. The first experiment investigated whether LLMs can replace the entire process of storing scientific articles and processing these documents to answer corpus-related questions that are relevant to a design science researcher. The second experiment had the more specific aim of measuring how good an AI-based tool can be at extracting knowledge from single articles that may be in various formats and use different nomenclatures. The third experiment recreated the first one under more favorable conditions, namely by having the whole corpus of articles inside a single well-structured file.

In our experiments, we used OpenAI API Assistant alongside the base models GPT-3.5-turbo-1106 and GPT-4-1106-preview, which were the only state-of-the-art models available at the time of the experiments that allowed attaching documents as a source of external data. Regarding the data, we collected several design science articles from two different venues popular among design science researchers, namely the Business and Information Systems Engineering (BISE) journal and the International Conference on Design Science Research in Information Systems and Technology (DESRIT). When running the experiments, OpenAI API Assistant only allowed to store and query 20 files, with each file’s maximum size not exceeding 512 MB. That was enough to accommodate the 20 selected articles, as they had an average size of 1.28 MB. In terms of article selection, we focused on maximizing the diversity of application domains. In particular, we ignored articles whose topics were about improving or studying DSR as a methodology, and we focused instead on applied work that relied on DSR as the research methodology. We ended up selecting six relevant articles from BISE and 14 articles published in the DESRIT proceedings. We manually extracted the title, abstract, keywords, authors, year of publication, and design requirements, principles, and features from those articles. The articles did not follow a reporting standard, meaning that the retrieved information could be, for example, inside the main text, tables, images, or even a mixture of various elements. We looked at the accuracy of the studied LLMs by comparing their results against those based on the manually extracted information, a common practice in natural language studies (see, for example, Carvalho and Xu, 2021, and Carvalho et al., 2019). Finally, all of our prompts were zero-shot attempts, meaning we did not provide examples to the LLM on acceptable answers to the queries. We return to this point later in the article.

### Experiment #1

Our first experiment involves requesting complex information about the DSR knowledge contained in a collection of articles. The goal is to understand whether LLMs can be potentially used to store and extract relevant information from many DSR articles simultaneously. To that end, we asked the following questions (verbatim):

Question #1: List the titles of all articles that contain any design requirements about document and knowledge management.

Question #2: List the authors' names of all articles containing design principles related to healthcare.

The results obtained for the first question can be seen below. They show that modern LLMs still struggle with complex requests in a single prompt when relevant data is dispersed among many files.

Expected answer: "A Design Science Approach to Blood Donation Apps," "A Digitization Pipeline for Mixed-Typed Documents Using Machine Learning and Optical Character Recognition," "Guiding Refugees Through European Bureaucracy," "Designing a Trustworthy Mobile App for Document Management," and "Supporting Product Development by a Trend Analysis Tool Applying Aspect-Based Sentiment Detection."

Answer by GPT 3.5: "Just What the Doctor Ordered - Towards Design Principles for NLP-Based System in Healthcare."

Answer by GPT 4: "Can Artificial Intelligence Help Used-Car Dealers Survive in a Data-Driven Used-Car Market."

The request in the second task was less vague and, arguably, easier than the first one, but still, only 33% of the expected results were returned by the LLMs. We have some evidence that the models limited themselves to identifying authors from only a single article whose design principles are related to healthcare.

Expected answer: "Helena M. Muller, Melanie Reuter-Oppermann," "Florian Onur Kuhlmeier, Ulrich Gnewuch, Stefan Luttke, Eva-Lotta Brakemeier, Alexander Madche," and "Marvin Braun, Aycan Aslan, Till Ole Diesterhoft, Maike Greve, Alfred Benedikt Brendel, Lutz M. Kolbe."

Answer by GPT 3.5: "Marvin Braun, Aycan Aslan, Till Ole Diesterhoft, Maike Greve, Alfred Benedikt Brendel, Lutz M. Kolbe."

Answer by GPT 4: "Marvin Braun, Aycan Aslan, Till Ole Diesterhoft, Maike Greve, Alfred Benedikt Brendel, Lutz M. Kolbe."

## **Experiment #2**

Given the previous observation that LLMs might struggle to retrieve information spread across a corpus of different articles, our second experiment involved four questions related to bibliographic metadata and three related to DKDs, all of them concerning single DSR articles. Specifically, the questions we asked were:

- Can you extract the title from article X?
- Can you extract the names of the authors from article X?
- Can you extract the abstract from article X?
- Can you extract the year it was published from article X?
- Can you extract the design requirements from article X?
- Can you extract the design principles from article X?
- Can you extract the design features from article X?

Table 1 summarizes the obtained results in terms of overall accuracy, i.e., the percentage of correctly answered questions across all 20 articles. In particular, Table 1's first row includes all questions about metadata. Note that the accuracy rate for both models is relatively high, as title, author names, abstracts, and year of publication tend to be all well-formatted data. The second row in Table 1 shows the accuracy for design knowledge descriptors only. Naturally, the accuracy of both models drops significantly since, as we previously suggested, the reporting of DKDs does not always follow a well-structured format, especially when DKDs are embedded inside images.

	GPT 3.5	GPT 4
All metadata	77%	91%
DKD	18%	74%

**Table 1. Overall Accuracy of LLM models.****Experiment #3**

The results from the second experiment show that the structure of an article may help with information retrieval. That observation led us to the third experiment, which involves requesting information from a structured file. Specifically, that file follows a well-organized hierarchical format using nested key-value pairs in a JSON format to store the metadata and the DKD from all scientific articles in our DSR knowledge base. We then repeated the same process as in Experiment #1 to understand whether LLMs could now provide better results; that is, we started by asking the following questions:

Question #1: List the titles of all articles that contain any design requirements about document and knowledge management.

Question #2: List the authors' names of all articles containing design principles related to healthcare.

A priori, we expected that the AI-based tool would perform excellently in answering complex questions based on a well-structured file. However, the obtained results show otherwise. The answers to the first question can be seen below.

Expected answer: "A Design Science Approach to Blood Donation Apps," "A Digitization Pipeline for Mixed-Typed Documents Using Machine Learning and Optical Character Recognition," "Guiding Refugees Through European Bureaucracy," "Designing a Trustworthy Mobile App for Document Management," and "Supporting Product Development by a Trend Analysis Tool Applying Aspect-Based Sentiment Detection."

Answer by GPT 3.5: "System Design Principles for Intergenerational Knowledge Sharing."

Answer by GPT 4: " 'A Design Science Approach to Blood Donation Apps' by Helena M. Muller and Melanie Reuter-Oppermann. This article presents design requirements for effective smartphone applications for blood donation and a conceptual model in the form of mock-ups and 'Identifying Digital Transformation Paradoxes' by Lieselot Danneels and Stijn Viaene. While this article does not explicitly mention design requirements, it presents design knowledge to identify organizational paradoxes in a digital transformation context."

In the answers above, we see articles that were not included in the manual response by the human expert, but the GPT 4 model explains the reason for their inclusion. Nevertheless, it is noticeable that several articles were excluded from the model's response.

In the second question, both models arrived at the same result. Although the number of responses is limited to two, the provided answers are nonetheless correct. However, not all possible answers are presented, indicating that the tool might understand the question and can extract information through context, but it restricts content due to potential time/processing constraints.

Expected answer: "Helena M. Muller, Melanie Reuter-Oppermann," "Florian Onur Kuhlmeier, Ulrich Gnewuch, Stefan Luttke, Eva-Lotta Brakemeier, Alexander Madche," and "Marvin Braun, Aycan Aslan, Till Ole Diesterhoft, Maike Greve, Alfred Benedikt Brendel, Lutz M. Kolbe."

Answer by GPT 3.5 model: "'Design Science Approach to Blood Donation Apps by Helena M. Muller and Melanie Reuter-Oppermann' and 'Just What the Doctor Ordered -Towards Design Principles for NLP-Based System in Healthcare by Marvin Braun, Aycan Aslan, Till Ole Diesterhoft, and Maike Greve.'"

Answer by GPT 4 model: "'Just What the Doctor Ordered - Towards Design Principles for NLP-Based System in Healthcare: Marvin Braun, Aycan Aslan, Till Ole Diesterhoft, Maike Greve' and 'A Design Science Approach to Blood Donation Apps: Helena M. Muller, Melanie Reuter-Oppermann.'"

## Conclusion

Through three different experiments, this preliminary study demonstrates the potential and limitations of Large Language Models (LLMs) in managing Design Science Research (DSR) knowledge. The experiments revealed that while LLMs like GPT 3.5 and GPT 4 show promise in extracting and processing information from DSR articles, their accuracy varies significantly depending on the structure and complexity of the data. In particular, the models performed better with well-structured data but struggled with complex queries whose results are spread across multiple documents. In terms of future research, a natural extension of this work is to investigate the performance of LLMs under few-shot prompting, i.e., by first providing the model with a small set of examples to guide its behavior. Moreover, it is worthwhile exploring a more diverse set of LLMs. This latter point also relates to investigating the cost-benefit trade-offs of different LLMs in various scenarios. For example, although GPT 4 is considerably more accurate in our experiments, using its API is also substantially more expensive than using GPT 3.5. Other important facets concern the rise of powerful, free, and open-source LLMs, such as LLaMa 3, and the fact that DSR researchers might have different tolerance for accuracy and speed, as well as limited financial resources. Finally, our work also speaks about the necessity of standardizing the reporting of DSR knowledge. Although attempts at the standardization of reporting design principles (Gregor et al., 2020) and design theories (Gregor and Jones, 2007) have been proposed, other design knowledge components can still be reported and even interpreted drastically differently by different researchers. Overall, we believe developing methodological and reporting standards can greatly help tackle the issue of knowledge accumulation and retrieval in design science.

## REFERENCES

- Anderson, C., Carvalho, A., and Merhout, J. 2022. "Patient Consent for Health Information Exchange: Blockchain-driven Innovation," in *Proceedings of the 28th Americas Conference on Information Systems*.
- Carvalho, A., Levitt, A., Levitt, S., Khaddam, E., and Benamati, J. 2019. "Off-the-Shelf Artificial Intelligence Technologies for Sentiment and Emotion Analysis: A Tutorial on Using IBM Natural Language Processing," *Communications of the Association for Information Systems* (44), pp. 918-943.
- Carvalho, A., and Xu, J. 2021. "Studies on the Accuracy of Ensembles of Cloud-Based Technologies for Sentiment Analysis," in *Proceedings of the 27th Americas Conference on Information Systems*.
- Folha, R., Times, V. C., Carvalho, A., Araujo, A., Viana, F, and Couto, H. 2022. "Towards a Novel Business Process Model for Food Delivery Services Using Blockchain Technology," in *Proceedings of the 28th Americas Conference on Information Systems*.
- Gau, M., Maedche, A., and vom Brocke, J. 2023. "Accessing the Design Science Knowledge Base - A Search Engine for the Accumulation of Knowledge Across Decentrally Organized Publications," in *Design Science Research for a New Society: Society 5.0. DESRIST 2023*.
- Gregor, S., and Jones, D. 2007. "The Anatomy of a Design Theory," *Journal of the Association for Information Systems* (8:5), pp. 312-335.
- Gregor, S., Kruse, L. C., and Seidel, S. 2020. "Research Perspectives: The Anatomy of a Design Principle," *Journal of the Association for Information Systems* (21:6), pp. 1622-1652.
- Peffer, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2014. "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* (24:3), pp. 45-77.
- Reining, S., Ahlemann, F., Mueller, B., and Thakurta, R. 2022. "Knowledge Accumulation in Design Science Research: Ways to Foster Scientific Progress," *ACM SIGMIS Database: The DATABASE for Advances in Information Systems* (53:1), pp. 10-24.
- Schoormann, T., Möller, F., and Hansen, M. R. P. 2021. "How Do Researchers (Re-)Use Design Principles: An Inductive Analysis of Cumulative Research," in *The Next Wave of Sociotechnical Design: 16th International Conference on Design Science Research in Information Systems and Technology*. Springer, Cham, pp. 188-194.