From Web Data to DIY Ideas: Material-Based Project Generation Systems

Shengyan Xu *

Golisano College of Computing and Information Sciences, Rochester Institute of Technology, Rochester, United States

* Corresponding Author Email: shengyan.xu94@gmail.com

Abstract. This article investigates and explores the prototype of a material based project generation system, aiming to transform online DIY content into structured design knowledge for sustainable reuse. Based on an unfinished system developed by the author in his early years, this work reviews the methods of collecting unstructured data from the Internet, extracts key information through lightweight natural language processing, and normalizes it into a unified model suitable for automatic generation. Discussed baseline strategies for retrieving and editing project generation, as well as evaluation metrics that consider constraint satisfaction, feasibility, and user experience. The paper also investigates practical applications in personal creativity support, maker education, and community recycling, supported by initial human in the loop implementation and interactive segmentation code that links material understanding with basic robot operations. The challenges of multimodal alignment, richer constraint modeling, and robust dataset construction have been identified, highlighting opportunities for future integration with digital twins and automated assembly. Although the original system is still incomplete, the study concludes that a practical and impactful pipeline is achievable, demonstrating how artificial intelligence can help users recycle materials more effectively and contribute to sustainable, resource-saving designs.

Keywords: Upcycling; Natural Language Processing; Material-based Generation; Sustainability; DIY Systems; Web Data Extraction; Creative Design Automation; Human-in-the-loop Learning.

1. Introduction

In recent years, driven by the global trend of sustainable development and circular economy, interest in "upcycling" – creatively using waste or other used materials to create new, high-value products – has increased. [1] Upgrading cycles has become a way of life for many people now and also shows how communities and industries are rethinking resource use and waste reduction. Although there has been an increase in momentum recently, most of the upgrade and renovation projects are still carried out manually. People often do not know how to use the available materials well, and as a result, they are unable to make many new things.

At the same time, there is also a large and expanding library of DIY (do-it-yourself) and upgrade recycling projects on the Internet. Projects, as well as their materials, reasons and the final results, have been posted on the website and forum. However, these resources are scattered, disorderly, and not uniform in form. Raw HTML pages generally contain descriptive text, sequential instructions, images within text sections, and different terms for the same information. Due to the absence of a standard format, it is difficult for computers and people to search, organise and use this information.

This paper built an early system for crawling DIY websites, extracted essential information, and standardised material data. Read it as an all-encompassing review and framework recommendation for material-based project generation systems. The main parts of such a system are introduced in this paper: web data collection, information extraction, pattern design and baseline generation strategies. It will also investigate the possible applications and put forward some main research problems, such as multi-modal understanding, richer constraint modeling, and future integration with robots. With the progress of natural language processing, data extraction and generative design, this work intends to build the foundation for intelligent systems that can convert network data into practical ideas to help design things more sustainably and use resources more efficiently.

2. Web Data Collection and Preprocessing for DIY Projects

This website has a large number of DIY and upgrade recycling platforms, and documents creative reuse projects. The contents of these websites are diverse, including stories told in blogs, bullet-point summaries and step-by-step instructions. Maker forums, blogs on sustainable development and commercial bicycle websites are examples. A large number of sources can be used to gather data for all kinds of projects and materials. There are also problems with data quality, format uniformity and meaning ambiguity.

Recursive Domain-Oriented Web Crawlers can be employed to collect this diverse content. Crawlers can systematically follow the internal links of a well-planned DIY and upgrade website to obtain the original HTML pages for subsequent processing. The output of this stage is intentionally wide and unfiltered; it contains all kinds of text, embedded multimedia and repeated content. Different people have written these sources in different styles and with varying degrees of detail, so the raw data is often messy and inconsistent.

2.1. Information Extraction from Unstructured Content

Once collected, the original HTML page needs to be converted into structured information. [2] We can use simple natural language processing (NLP) tools to extract the most important parts from each project. [3] Four types of data are particularly important for downstream use:

Material: A list of items required to complete the project, preferably including quantity (e.g. "2 vegetable cans").

Explanation: A structured sequence of steps, usually in the form of numbers or bullet points, describing how to complete a project.

Final artifact: A brief description of the completed object or result, typically derived from the project title or conclusion.

Source and media: Original source URL, as well as image URLs related to specific steps (if available).

By tagging and segmenting text into these core categories, the system transforms heterogeneous and unstructured data into meaningful machine-readable components. This makes it easier for subsequent algorithms to analyze, compare, and mix project information.

2.2. Normalization and Data Cleaning

After extraction, standardization is crucial for improving data quality and ensuring consistency between projects. Common cleaning steps include changing material names to lowercase, converting plural forms to singular forms, and matching nicknames with standard names (such as "Duct Tape" to "duct Tape"). In addition, the quantity of expressions can be parsed as a structured attribute. These steps reduce repetition, fix language differences, and standardize data, making future matching and project generation more accurate.

2.3. Structuring for Downstream Use

Once a clean and structured DIY project dataset is created, the next step is to design a system that can generate new project ideas based on user provided materials. This section describes the key components of the system, including pattern design, baseline generation strategy, evaluation metrics, and human centered validation. These components together form the core framework of the material based project generation framework.

3. Core Components of Material-Based Generation Systems

3.1. Schema Design and Data Representation

A clear data schema is the key to converting messy DIY project data into a consistent computer-readable format. The typical patterns for material based project generation include the following fields:

Material: List of required items, including names, standard units, and quantities.

Step: Explain in order how to complete the project, remove template language and irrelevant text.

Final artifact: A concise description of the final item, such as a "desktop pen holder" or a "garden flowerpot"

Source and media: Original source URL and related images, preserving the source and allowing for richer multimodal analysis.

Optional comments: safety instructions, expected time, or required tools.

This structured format makes it easier for subsequent parts of the system to search, compare, and create project plans. It also ensures that future updates (such as adding images or user feedback) can be included without the need to reconstruct the entire data model.

3.2. Baseline Generation Strategies

Material-based project generation is different from traditional text generation in that it must satisfy certain constraints for the input; that is, all user-provided materials need to be included in the output, and no external components should be used to ensure logical consistency [4]. To address the above problems, some non-training baseline strategies can be employed before using more advanced models:

Template-based copying and editing: Find the most similar existing items based on material overlap and reuse their content as a template, then perform light modifications to match the new material list. When the examples in the dataset are too similar, this method is relatively simple and straightforward.

Retrieval-editing fusion: Instead of using a single template, obtain the top k relevant items and combine their steps to create a logical outline. System check to ensure all input materials are included and to remove any extra parts. This way will have more diverse results and work better without a fixed template.

Tip-based light-weight generation (optional): Prompt a small language model to use the retrieved examples as context for project idea generation. Then, the order can be reversed to promote the output with a large amount of material coverage and fewer external parts. A hybrid mode has achieved both the creative and controllable.

Although these ways are relatively simple, they can still be used to establish an all-encompassing workflow and performance goals for future optimization work. They offer an initial reference for the evaluation of other advanced learning methods.

3.3. Evaluation Metrics and Performance Assessment

To know whether the created project can meet people's needs reasonably, one can test it. Some typical kinds of quantitative and qualitative indexes are used.

Material Coverage: The proportion of materials provided by users in the generated project. A high coverage indicates good adherence to the constraints.

Foreign Material Rate: The proportion of materials in the output that are not provided as inputs. A smaller value is closer to being fully constraint-satisfying.

Vocabulary overlap: A simple text similarity score of the generated output and the reference item, indicating how similar the system's output is to the existing examples.

Clarity and Feasibility: Manually rate the comprehensibility and operability of the generated steps.

Together, the above indicators can offer us an all-around view of the system's operating condition and demonstrate whether it meets the demands and is practical in use..

3.4. Human-Centered Evaluation

Automation indices are practical and unbiased; otherwise, human assessment needs to be performed on the resulting projects. Small-scale user research can be carried out to determine whether the generated ideas are feasible, safe and practical. Qualitative feedback on creativity and practicality from the participants will also be collected for optimization in future development of the power generation system. The above studies help to improve the practical value of the algorithms and meet the demands of actual applications by users.

4. Applications and Case Studies

4.1. DIY Creativity Assistance for Individuals

One of the first is to help people find ideas for using old things in a new way. Many people have kept old houseware such as glass jars, cardboard boxes and plastic bottles, and have no idea how to use them anymore. A system based on generated materials can be used as a personal design assistant; the user inputs a list of available materials, and then the system returns several project suggestions along with a brief introduction and necessary steps. It will be simpler for the recycling group to organise and more people will join the sustainable living community. In addition, recommend projects with specific materials to promote resource-saving creativity and reduce waste at the individual level.

4.2. Educational and Maker-Space Applications

These systems are not only for personal use but also have a large-scale impact on education and maker spaces; [5] schools, libraries, community workshops, and so on are beginning to introduce design-based learning and creation activities. The project generator can be used as a learning aid in science, technology, engineering and mathematics (STEM) courses to learn about material properties, construction methods and environmental impacts.

As shown in Table 1, the overall statistical data of the education and makerspace initiatives, along with examples of how institutions such as MIT and other university makerspaces have incorporated sustainability and creative reuse into their facilities, are presented here.

Teachers can add the system to the long-term development course to teach students how to use daily objects as functional tools or works of art. Maker Spaces and Community Labs can use this system to accelerate the ideation process in workshops and offer participants structured prototypes and experimental starting points.

Maker/Educational Space	Location	Makerspace Area	Number of Design/Build Project Spaces	Key Sustainability Notes	Source
MIT Makerspaces	USA	~12,077 m²	> 40 design/build spaces	47 % of students spend >5 h/week making; 58 % wanted training.	Lee S.H., "Carbon-Neutrality in Makerspaces – Circular Makers" (2021)
University-based Makerspaces (Survey)	Various	—	—	Identified barriers to student participation (time, purpose, certification).	Andrews M.E. et al. (2024)

Table 1. Statistics of educational / maker-space initiatives.

4.3. Community Recycling Platforms

Materials-based power generation can help promote all-around development initiatives at the community level. Municipal recycling centres or local non-profit organisations can add such systems to digital platforms to record the recyclable materials residents wish to donate or exchange.

Table 2 shows some typical DIY and community recycling platforms, and their main users, re-use purposes, and data availability for material-based generation research.

Then, based on the materials from all users, the platform will suggest collaborative projects for the community and make full use of shared resources. Together, they will help build a circular economy by extending the life of the materials beyond a single home and integrating them into a community of creation.

Platform	Primary Audience	Key Functionality	Material Reuse Focus	Data Source
Instructables	Maker Community	DIY project tutorials, community sharing	Generic; allows upcycling categories	"10 Instructables Projects That Blew Our Minds"
Library Makerspaces	Public/Educational	Community making, reuse of materials	Emphasis on hands-on reuse and building	"Library makerspaces: revolution or evolution?"

Table 2. Comparison of DIY and community-recycling platforms.

4.4. Comparison with Existing Solutions

Although some online platforms, such as Unbreakable and Pinterest, have offered DIY creativity, they are mainly static collections. Users have to find and learn about projects themselves; there is no specific teaching on using these materials. At the same time, a dynamic and interactive material-based project generation system can be built to meet all the demands of different users and materials. The new kind of tool will be able to suggest content and services that users are interested in by analysing their behavior and preferences in a more advanced way.

5. Future Directions and Challenges

5.1. Multi-modal Understanding and Visual Analysis

Most of the existing methods are based on text and cannot collect a large amount of DIY project data. Images in online instructions usually contain essential context, such as the shape of the material, the order of assembly or space arrangement, etc., which are difficult to express through text alone. Add computer vision technology to recognise and interpret these pictures better in the system, thus learning about the project content and producing more precise and practical instructions [6]. The next generation of systems will link written instructions with pictures to help users learn how to assemble objects, and more detailed steps will be provided.

5.2. Richer Constraints and Context-Aware Generation

The main reason for the current generation strategy is a lack of materials; however, other factors in actual construction will also affect it. The following are reasons why a project may not be possible at present. Future work will explore how to integrate these richer constraints in the generation process and put forward recommendations that are both consistent with the material and contextually appropriate [7]; multi-objective optimisation methods or constrained perceptual decoding strategies can be used to balance creativity and practicality.

5.3. Human-in-the-Loop Design

Although there have been many achievements in the field of automatic power generation technology, human creativity and judgment are still needed. Machines can analyse data and produce a first draft; however, people add imagination, emotion and background that computers cannot fully grasp. In the near future, the design of systems should incorporate a human-in-the-loop approach to collaborate with users and algorithms rather than functioning independently [8]; that is to say, users can guide the process and modify materials, ideas or styles after the system generates new content.

5.4. Bench marking and Open Datasets

There are currently no open or standardised datasets available in this area. At present, most of the DIY or upgraded recycled content is spread across the Internet in many different forms. Some websites only provide short text descriptions, and others have images or videos without organised information. Many of these resources are also under copyright or licensed, so they are not suitable for research. Therefore, it is difficult for researchers and developers to compare their systems or reproduce each other's results, and thus the actual progress has been hindered.

Set up a standard benchmark and open dataset to address the problem. Shared datasets can have a large number of well-structured DIY projects, such as materials, tools, steps and results, and follow a unified format. Using the above resources, the different models can be tested and compared fairly under the same circumstances. It will be more open to the experiment, and we will find out which way works better [9].

5.5. Integration with Robotics and Automated Systems

Looking ahead, the link between material-based design systems and robots will be needed to a greater extent. After creating the project, the robots will be used to help with some parts of the construction work, such as cutting, assembly and 3D printing of components. A sensor- and-vision-based system will be employed to identify materials and accurately obtain information on the size and shape of objects for high-precision control of the robot.

Digital Twin Technology can also be used to drive this change. A digital twin is a virtual model of a real object or system. In the future, all materials and DIY projects will be able to keep their own digital records of history, attributes and previous usage. In conjunction with the above information, sensors will be used to gather data on material behaviour; then, a new assembly plan can be tested virtually, and the simulated results obtained before actual construction will help reduce waste and prevent design errors early on [10].

6. Conclusion

In this paper, the key components of the project generation system based on materials are investigated, including data collection and pre evaluation, pattern design, baseline generation strategy and evaluation method. The article also explores their applications in the real world, from personal creativity support to community recycling platforms, and outlines the main research challenges that still exist in the future, such as multimodal understanding, richer constraint modeling, and potential integration with robots. Finally, material-based generation systems can link artificial intelligence, creativity, and sustainability, transforming raw network data into practical design knowledge and helping to build a future that utilizes resources more wisely.

Reference

[1] Ferraro, A., Karatzas, K., & Jones, M. (2021). Text mining for sustainability: A systematic review. Journal of Cleaner Production, 288, 125583.

[2] Etzioni, O., Banko, M., Soderland, S., & Weld, D. S. (2008). Open information extraction from the web. Communications of the ACM, 51(12), 68–74.

[3] Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. Proceedings of EMNLP 2014, 740–750.

[4] Dong, L., Wang, S., & Zhang, J. (2021). Constraint-based text generation: A survey. ACM Computing Surveys, 54(5), 1–34.

[5] Sung, K., Cooper, T., & Kettley, S. (2014). Individual upcycling practice: Exploring the possible determinants of upcycling based on a literature review. Sustainable Development, 22(2), 123–135.

[6] Liu, Y., Wan, X., & He, J. (2022). Multimodal knowledge extraction and generation: Recent advances and challenges. IEEE Transactions on Neural Networks and Learning Systems, 33(10), 5034–5051.

[7] Gao, X, Y., (2022). A Hybrid Search Model for Constrained Optimization. Wiley Journal.

[8] Sun, Y., Li, Z., & Zhang, H. (2020). Human-in-the-loop machine learning: A survey. ACM Computing Surveys, 53(2), 1–31.

[9] Buechley, L., & Eisenberg, M. (2008). The LilyPad Arduino: Toward wearable engineering for everyone. IEEE Pervasive Computing, 7(2), 12–15.

[10] Sun, Y., Li, Z., & Zhang, H. (2020). Human-in-the-loop machine learning: A survey. ACM Computing Surveys, 53(2), 1–31. Turner, C. J., Emmanouilidis, C., & Tiwari, A. (2021). Digital twins for sustainable smart manufacturing: A review. Advanced Engineering Informatics, 50, 101431.