# How to Report the Contributions of a CC System?

**Anna Kantosalo, Simo Linkola, and Tomi Männistö**
Department of Computer Science
University of Helsinki, Finland
anna.kantosalo / simo.linkola / tomi.mannisto @helsinki.fi

**Abstract**

We argue that the lack of well established reporting practices for applied Computational Creativity systems is hindering progress in the field. We consider that the current lack of reporting details – and variation in form and content – makes it difficult for third parties to reliably evaluate and compare systems based on publicly available information. This hinders forming an understanding of the similarities, differences and relative qualities of these systems. We propose a set of building blocks for robustly reporting the contributions of computationally creative systems to promote visibility and clarity in the field.

## Introduction

The field of Computational Creativity (CC) is growing and reaching new levels of maturity. As the field attracts new audiences and new participants, it needs to make the research approachable and easy to understand through transparency. One of the key issues for transparency in applied creative systems is establishing field specific reporting practices. As the field matures, we have seen some gradual change towards better reporting practices. For example, Jordanous (2012b) has suggested practices for reporting CC evaluation. However, no comprehensive guide exists so far to support structuring CC reports for various audiences considering the basic elements of an applied system from a CC perspective. Therefore, we sketch out building blocks to support the transparent reporting of applied creative systems. These building blocks can be directly applied to support authoring specific sections of an applied CC paper.

Applied CC is set apart from theoretical CC, consisting of philosophy and methods of CC, by its focus on implementing systems that generate, evaluate or both generate and evaluate creative artefacts. The systems can be autonomous, interact with humans, or consist of several (autonomous) agents interacting with each other. They are often built to demonstrate a specific new CC method and increasingly deployed in real world contexts to aid real world creators. We outline reporting principles for such applied systems to improve communication within the field of CC and with the general public.

We argue that good reporting of applied creative systems should support *transparency* (see e.g. Fidler and Wilcox (2022) or Tearse, Wardrip-Fruin, and Mateas (2010)), which is a requirement for *reproducibility* (see e.g. Fidler and Wilcox (2022)) and allows for *system comparison*. As a stretch goal we consider that great reporting practices should also support communication to scientists, practitioners and the general public and relate new discoveries to previous progress in the field, following principles previously found useful in design science (Johannesson and Perjons 2014).

We propose three building blocks to support good reporting practices for applied creative systems. These building blocks can help authors to decide what to include in their applied CC research papers. Our aim is to supplement existing writing guides from related fields. Our building blocks are tailored to include aspects specific to creative systems, such as definitions for creativity. We next present our contributions and then discuss how they connect to general principles of good research and current practices in the field.

## Building Blocks for Describing Computationally Creative Systems

We consider that at the heart of a successful applied CC paper is a robust description of the CC system and its contributions. We argue that the *description* of the system and its *evaluation* should go hand in hand with a *definition* of creativity fit for the *context* the system operates in. It is the mapping between this definition and the system description that allows the reader of an applied CC paper to contextualise the system and its contributions in the larger framework of CC research. We discuss these three parts in detail below.

### Building Block 1: Definition of Creativity

A working definition of creativity allows the reader of the applied CC paper to situate the work within the larger scope of the CC field. A well chosen definition also allows readers from other, connected disciplines, as well as laypeople to understand how the applied research connects to our general understanding of creativity. In short, a well selected working definition for creativity manages the reader's expectations.

What sets a *working definition of creativity* apart from a *general definition of creativity* is that the definition does not need to be exhaustive: It can focus on a specific aspect of creativity, which is of interest to the researchers developing the applied system, or the domain the system operates in.

The best working definitions are short and refer to the larger body of literature on defining creativity (see e.g. Runco and Jaeger (2012)).

The authors of the applied CC paper should explicitly argue how the working definition of creativity connects to the *creativity goals* of the system. These are goals directly linked to the creativity of the system. If creativity is not the main goal, or the only goal of the system, authors should argue how the creativity goals significantly support the other goals of the system. For example the goal of a co-creative system may be to aid a user in a design task. This goal can be attained in many ways, but an important sub-goal, directly linked to creativity, could be the generation of valuable and novel design suggestions for the user. This way, in addition to setting expectations, motivating research, and connecting new research to existing research on CC, an explicit working definition of creativity promotes the transparent selection of suitable evaluation criteria for a CC system (Jordanous 2012b).

## Building Block 2: System Description

A successful description of a CC system consists of several parts. The importance of each part depends on the *the scope of the system*, the stage of its *life-cycle* and *system goals*. Defining these explicitly is important to direct the readers' attention, manage expectations, limit scope, and set the context of the work.

*The scope of the system* should clarify if it is a full system, a part of a larger system or possibly a system embedded in a larger context or ecosystem of other systems. *The system life-cycle stage* should describe if the system is new, or a more established one, setting the expectations for the description and evaluations of the system. Finally, the *system goals* should connect to the chosen *creativity goals* and the *definition of creativity*.

**The Generation-Evaluation Process.** Typically a computationally creative system includes a part that generates creative outputs. The description of the *generator* should be detailed enough to enable the reproduction of a similar generator. The authors should at least answer the following questions: What kind of artefacts does the generator produce? What are the properties, and desired properties of the artefacts? What methods does the generator use to produce the artefacts? What kind of an architecture does the generator have and how does it connect to the rest of the system? Which data sets (or inspiring sets) does the generator use? If the system relies on a generative model requiring training, how was the model trained and what kind of parameters were used? If pre-trained models were used, what were they trained with and why are these models suitable for the creative purpose?

Correspondingly, many creative systems contain an internal evaluation component, or a component evaluating the creative contributions of other members of creative collectives. The *evaluator* should be documented with similar scrutiny to the generator.

If the generator and/or the evaluator are key contributions of the paper, the description of them must help the readers to understand how they work exactly. This requires comparing the generator/evaluator to existing generators/evaluators, which either produce or evaluate artefacts of the same kind or use similar processes in different domains, and explicitly pointing out the differences. If the generator or the evaluator consists of multiple parts, ablation studies are a good way to show how each of the subcomponents of the generator/evaluator affect the produced artefacts. This may require building mock or dummy implementations of each of the subcomponents. While ablation studies may seem like extra work, they tremendously support the transparency and comparison of the systems, and should be seriously considered in any system where the generator and/or the evaluator is part of the contributions.

**Interfaces & Communication.** For systems that interact either with humans or other systems, documenting the interaction *interfaces* is equally important. A short use case and/or a diagram illustrating how a human (or a machine) would interact with the system can be used to describe many aspects of an interface in an easy to understand manner. For visual interfaces this can be augmented with images and samples of other types of interface modalities can be included in external materials, such as video or audio. Whether a system interacts with a human or another system, it is also important to consider the following questions: Why does the system communicate with others? How does it happen? With whom? What kind of information is sent, and received? And finally, what triggers communication?

**System History.** Depending on the life cycle stage of the system some amount of the *history* of the system may be required for understanding it. History is especially important for studies building on existing systems: What version of the system is used? How does it differ from previous versions of the same system? In most cases it is good to explicitly answer the question: "What is the new contribution this version of the system makes (also for creativity)?" For papers that primarily demonstrate improvements to existing systems, it is important to also document changes made to the algorithms and models used in detail. The reader should have a clear idea how the system components are changed compared to older versions and what the expected (or assessed) benefits of the changes are.

Ideally the history can also include core elements of the design process of the system: What important design decisions were made during the development of the system and how do they support the system goals and its creativity? A design decision can be for example what data set is used as an inspiring set for the system. It is important to document the expected benefits of the chosen approach with respect to the creativity goals of the system.

Finally an increasing number of systems learn and change during their life-cycle. These adaptive systems should describe also what changes during the run of the system, how the changes are triggered, and what contributes to them.

## Building Block 3: Evaluation & Contributions

Evaluation in CC can refer to several different concepts: Internal evaluations conducted by the computationally creative

system, or external evaluations aimed at summative or formative judgements of the quality and development areas of the system, possibly in a specific context. Similarly evaluation can be conducted by not only the system itself, but by system developers, or a third party, such as experts or laypeople. Full details on methods of evaluating applied CC systems is beyond the scope of this paper and there are several perspectives to evaluation that can be taken, including not only the evaluation of the creativity of the system, but also the fit of the system in the overall creative context it operates in. We refer the interested reader to Agres, Forth, and Wiggins (2016) or Jordanous (2012b) for more detail. Here we focus on evaluation as a relevant part of communicating the contributions of an applied CC system.

At minimum, documentation of an evaluation should explain what is evaluated, by whom, how, where and why. These questions help readers to assess if the evaluation of a system is robust, if it generalises to other audiences and contexts, possible sources of bias and if the evaluation is relevant. The documentation of the procedure also allows for reproducing the evaluation or conducting a similar evaluation on another system contributing to reproducibility and comparison of CC systems.

To be meaningful and relevant, the evaluation must be tied to the *creativity goals* of the system. As a core, extraordinary claims demand extraordinary evidence, therefore the chosen evaluation method and metrics should support the claims the authors of the system make about its creativity. This means the authors should document what metrics were used in the evaluation of the system and how do these link to the goals of the system and the chosen definition of creativity. So far there are very few established evaluation metrics presented in the field and some authors develop their own metrics or loan metrics from related fields. Echoing Jordanous, (2012b) it is important to establish why these metrics work in the chosen context so that the relevance of the evaluation can be assessed. Similarly, for an author to claim a system is creative, it is also important to document the self-evaluation metrics used by the system.

## Discussion

We start with a brief discussion of the scientific objectives of the building blocks. We then discuss how the building blocks fit to the larger context of academic writing advice and connect with reporting practices from related fields.

We consider applied CC research as a discipline under the umbrella of Design Science. Similar to applied CC research, Design Science is a research paradigm that seeks explanations, predictions and descriptions for the current world, by actively trying to improve it through the creation of new systems (Johannesson and Perjons 2014, p.1).

### Scientific Objectives for the Building Blocks

The purpose of our building blocks is to support three key ideas: *transparency, reproducibility, and comparing contributions within CC*. We consider that current weaknesses in reporting threaten these ideals and therefore hold back the progress of the field.

*Transparency* is a facet underlying the other two key ideas we wish to support. With transparency we refer to making information about the analysis and methods used accessible to the reader in a way that supports constructing an unbiased understanding of the applied CC system. The content of the blocks supports attaining this goal as the reader of a paper following the suggested block structure is more easily able to find the related information and make meaningful comparisons between systems.

Transparency is closely related to reproducibility in empirical science. Lack of transparency and completeness in method reporting (Fidler and Wilcox 2022) or datasets (Tearse, Wardrip-Fruin, and Mateas 2010) hinders reproducing previous experiments and the re-creation of systems. In addition, lack of transparency can render some CC evaluation methods useless, and impede with the independent evaluation of systems and research results.

For example Ritchie's (2007) criteria for evaluating creative outputs requires knowing the inspiring dataset used by the generator, as well as having access to a sufficient sample of results. If these are not stated, an independent evaluator cannot evaluate the applied CC system built by another, hindering for example, the use of the system as a baseline for future evaluations.

Similarly important for independent evaluation is to know the objectives of the research and the definition for the type of creativity the researchers are striving to implement with their system; In her seminal paper on standardised evaluation in CC Jordanous (2012b, p.1) argues for "stating what it means for a particular computational system to be creative, deriving and performing tests based on these statements". The lack of defining creativity makes it difficult especially for a layperson to evaluate creativity (Jordanous 2012b), which may limit the use of applied CC research results by general audiences. Therefore, announcing a working definition for creativity would improve both use and verification of results, but still many applied CC papers only make implicit assumptions about creativity.

Moreover, applied CC research seems to rarely record and publish negative results. Jordanous' evaluation of five CC presentations showed that developers typically focus on a few specific aspects of creativity, leaving multiple aspects impossible to review (Jordanous 2012a, pp.217-219). Only in one case of the five systems Jordanous' evaluated with her colleague was information sufficient to give a poor review of an aspect of creativity (Jordanous 2012a, pp.217). By pointing out their working definition on creativity, authors can communicate their focus to the external evaluator, as well as more reliably report also their negative findings on a specific aspect of creativity.

*Reproducibility* of experiments is a cornerstone of credible science. The so-called replication crisis brought the validity of results in medical, life and behavioral sciences into question in the 2010s (Fidler and Wilcox 2022). The definition of reproducibility varies between fields (Fidler and Wilcox 2022), here we refer to the ability to redo computations or whole experiments in principle and in practice, with the expectation of producing the same or sufficiently similar results. It can be further described as conceptual replications

focused on verifying underlying hypotheses and direct replications aimed at controlling for samples, artifacts, fraud or generalization (Fidler and Wilcox 2022).

Similar to design science, replication in applied CC research can foster the accumulation and development of design theories and to encourage the reuse of designed systems and existing theories (Brendel et al. 2021). Currently the failure to reuse systems and connect studies to existing knowledge is limiting the contributions and effect of design science research (Brendel et al. 2021). We find this to be true for applied CC research as well: Especially the lack of robust documentation hinders progress and replication in the field, valuable knowledge lost, when specific systems loose their financial support and the systems and the related infrastructure is abandoned. It is of immediate concern that many of these tools cannot be reproduced as sufficient documentation of their development is not provided.

Replication studies in applied CC research are very scarce, and difficult to conduct as well. One of the few studies that could be considered a replication study in applied CC, is the re-creation of the Minstrel system reported by Tearse, Wardrip-Fruin, and Mateas (2010). This attempt to reconstruct a seminal system in computational story generation struggled with lack of original documentation, as for example the dataset used by Turner in developing the original system was undocumented. We argue that there are several other system, the recreation of which would be impossible, as we lack not only the data used in their creation, but sometimes also sufficient detail of the system architecture and implementation.

Finally, the lack of robust documentation hinders *comparing contributions made within CC*. This can mean the comparison of computationally creative systems overall, comparison of systems within the same creative subdomain, or even the comparison of a system with its earlier installations. The practical development of systems is driven especially by formative feedback (Jordanous 2012b). More documentation is required for formative evaluation tools such as SPECS (Jordanous 2012b) to be applicable to systems by outside evaluators. Alternatively, evaluations conducted by researchers themselves should be reported more openly and thoroughly. Similarly, for the purposes of scientific integrity, different editions of the same system should clearly document differences among the different editions of the system so that specific data can be connected with a specific implementation of the system creating a more robust system history benefiting practitioners in developing similar systems in the future.

### The Building Blocks as Writing Advice

The building blocks suggested above could also be alternatively titled as a CC system documentation checklist, for they are largely based on the authors' experiences in participating in peer reviewing processes for papers describing CC systems. The critique presented most often seems to deal with establishing what it means for the system to be creative (Block 1), documenting the generation procedures in enough detail (Block 2), or showing a meaningful evaluation of the results (Block 3).

The blocks are also linked to the larger concept of academic writing advice. As we consider the field of applied CC inherently as a part of design science, the CC system including the generator naturally becomes one of the key items to document in research communications. Here we have only focused on aspects related to CC specifically. We therefore refer the interested reader to more specific advice on writing papers for design science (Johannesson and Perjons 2014, p.153-154).

We are also aware that as a multidisciplinary paradigm applied CC research has a lot to draw from related disciplines. We would for example argue that to a degree the adoption of neural nets in the generators offers great chances to draw from well established documentation practices in that specific area of artificial intelligence. Similarly in building interactive or co-creative computational systems, we have learned and adopted practices of evaluation with humans from interaction design. The purpose of this writing guide is therefore not to be definite, but we hope it works together with experiences from other disciplines to support a more robust reporting practice in applied CC research.

## Conclusions and Future Research

While we do not particularly focus on evaluation here, it is clear that the diverse reporting practices contribute to the 'methodological malaise' in CC evaluation identified by Jordanous (2012b) and others. The lack of sufficient, accurate and accessible reporting of CC systems is contributing to a situation where reproduction of systems, and transparent evaluation by third parties, or the comparison of different systems or the different editions of the same system cannot be conducted. This hinders progress as we cannot leverage the full potential of applied CC research and build on the findings and work of others, establishing a robust, continuous base of evidence for improving machine creativity.

We have suggested three building blocks: a definition of creativity, description of the CC system and its evaluation to support applied CC researchers in communicating the contributions of their systems to different audiences. To further support transparency in CC research we encourage developing more formal languages for the description of CC systems in ways that can also be archived for future research. This could include experimenting with existing descriptive languages like UML, or ontologies such as OWL. We would also like to encourage authors to share implementations of applied CC systems. Good implementations could be gathered and made accessible online for example similar to the deep learning Model Zoo[1] project. In the future, we intend to conduct a literature review to further examine the weak points in the reporting practices of applied CC systems.

## Author Contributions

The original idea for the paper came from the authors AK and SL. AK wrote the majority of the paper with SL. TM commented on the idea and the draft, with insights on documentation in software engineering and defining the scope of the work.

[1] https://modelzoo.co/ accessed 23rd of May 2022.

## Acknowledgments

Funded by Academy of Finland (Grant #328729). We would like to thank our colleagues Dr. Anna Jordanous and Prof. Hannu Toivonen and the anonymous reviewers for their comments on the draft.

## References

Agres, K.; Forth, J.; and Wiggins, G. A. 2016. Evaluation of musical creativity and musical metacreation systems. *Comput. Entertain.* 14(3).

Brendel, A. B.; Lembcke, T.-B.; Muntermann, J.; and Kolbe, L. M. 2021. Toward replication study types for design science research. *Journal of Information Technology* 36(3):198–215.

Fidler, F., and Wilcox, J. 2022. Reproducibility of Scientific Results. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2022 edition.

Johannesson, P., and Perjons, E. 2014. *An Introduction to Design Science*. Springer International Publishing, 1st ed. 2014. edition.

Jordanous, A. 2012a. *Evaluating computational creativity: a standardised procedure for evaluating creative systems and its application*. Ph.D. Dissertation, University of Sussex.

Jordanous, A. 2012b. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:76–99.

Runco, M. A., and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity Research Journal* 24(1):92–96.

Tearse, B.; Wardrip-Fruin, N.; and Mateas, M. 2010. Minstrel remixed: Procedurally generating stories. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 6(1):192–197.