Tools and Principles for Creation in Interactive Storytelling: The Issue of Evaluation

Ulrike Spierling

Hochschule RheinMain, Unter den Eichen 5, 65195 Wiesbaden, Germany
ulrike.spierling@hs-rm.de

Abstract. For academic research and development of new authoring tools and concepts, there is the necessity to evaluate the quality of results and to generalize the findings. This position paper collects reasons why this is a challenge in the realm of evaluating authoring systems for the creative invention of novel kinds of interactive stories.

Keywords: interactive storytelling, evaluation, authoring tools, creation principles

1 Introduction

At recent issues of the ICIDS conference series and at related workshop events, a number of discussions on Authoring and Creation for Interactive Storytelling or Interactive Narrative Systems were raised. It seems that for over more than a decade, the topic of authoring stays alive as an unsolved problem in this research community. There have been different focuses in the discussion, for example: the so-called authoring bottleneck as a challenge to be addressed by automation, the denial of the importance of human authors in future storytelling systems, a large number of specific or generic authoring tools that are in need of a classification, the gap between disciplines of traditional storytellers and the technically challenging AI or interaction concepts, and more.

About a decade ago, also the need for evaluation had been expressed, such as by Jhala et al. [3] and within the IRIS NoE [2]. Jhala: *"Besides the actual design and implementation, we see more and more that authoring applications are not well evaluated or included in experiments. The authoring part is usually seen as an 'add-on'."* [3] By analysis of recent publications since then, we find a number of authoring concepts and tools published – however, with little focus on evaluation of the authoring process as a user experience, or rigorous evaluation of the tools. I argue that the academic creation of authoring tools has merit, in the light that new concepts for interactive storytelling are still to be invented, either depending on novel multimodal interaction technologies, or on novel smart (AI) technologies to accomplish dynamic storylines or story games reacting to users. These new (and also mostly yet ill-defined) goals for the final storytelling experience render it unlikely that traditional authoring systems that are currently more common in hypertext-style narratives (meaning, with a defined point-and-click interaction modality and structure) are suitable for the task [13].

PhD students who undertake research in (inventing) novel authoring methods for interactive storytelling need a way to prove their findings, by presenting an evaluation. The novelty of tool solutions at this research level may rarely lie directly in presenting yet another clickable editor for an otherwise well-known task. Instead, when new paradigms of media interaction are involved (for example, augmented reality, location-dependence, or AI-based chat/behavior and more) there is more to analyze than the 'clickability' of graphical editors. Test authors need to understand novel media concepts that in general are not yet well explored, and the success of the whole approach including conception and implementation is of interest. In the following, I present a number of challenges that are involved with this kind of academic evaluation.

2 Evaluation Challenges

The first thing to consider when conceiving an evaluation is a goal or a set of goals against which the concept or research software (as an artefact) is to be evaluated. Most likely, for authoring tools of interactive stories it involves users (subjects) in the role of authors. Depending on the goals of the evaluation, these authors need to fulfill tasks with the authoring tools, and depending on the complexity of these tasks, this may require a long-term endeavor, exceeding usual user test sessions or inquiries. In the following, a possible variety is sketched.

2.1 Subjects

When authoring tasks are to be evaluated in academic projects and case studies, there are different possibilities of acquiring subjects:

a) Tool developer. In many projects, it makes sense that the developer uses his/her own tools for testing, at least as its first user before asking anybody else. There are different constellations how this has been performed. The only possible evaluation goal to be achieved is the tool's principle effectiveness – that is, in the case of a successful achievement of a playable interactive story, the proof that it is indeed possible to accomplish a story result [12]. This kind of evaluation does not say anything reliable on the learnability or even efficiency of the concepts or the artefact concerning its editors.

b) Research team member hired for authoring (creative professional / internal team member or external with contract). There have been reported studies and experiences in which freelance storytellers or hired team members from creative disciplines take the role of a test user [11]. If this role is taken by a long-term team member, it is likely that he/she gets involved in a participatory design cycle of user-centred development of the authoring tools [5, 10]. This is an ideal case for the development process as such. In terms of evaluation, it has a similar reliability as in the first case. Although this team member can make a personal statement on learnability, while being a member of development, the judgement concerning generalizable usability features is biased and therefore limited. Moreover, the experience exceeds most probably the mere authoring tasks, as the tool and concept design has been performed at the same time. As such, it resembles traditional crafts and toolmaking from even long before the digital age, when the invention of tools was in the hands of the craftsmen and artists. Another limitation is the process's dependency on the talent, education and even personality of that team member.

c) Students of the development team. This is a typical case that happens as a result of the need for academic evaluation. To acquire a larger number of users who indeed have at least once used the tools in question, an authoring task gets assigned within a media design or interactive storytelling class over one teaching semester. The easiest way to do this is to supervise the assignment within the same faculty, which often means that the subjects are students of computer science or related studies, such as media informatics [7]. The value of the results is probably biased as these students cannot judge the approach from the point of view of a non-programming storyteller, and they rarely feel as frustrated as a storyteller from humanistic or creative disciplines, if they cannot fully express themselves due to constraints in the system. It is not surprising that reported evaluation results are often positive.

d) Students of different user communities. An alternative to the above case of student evaluation within the technical faculty, this involves more effort and interdisciplinary cooperation between colleagues. It will probably bring rich and unstructured feedback of a great variety, which needs sorting into categories. It also involves a higher success challenge with the tools that are mostly prototypes. As these prototypes are likely to have usability inefficiencies or bugs, it is hard to abstract from these to get real feedback on the conceptual accessibility. It is worthwhile to think of placeholder conceptual strategies before the digital tools are introduced (for example, card games and paper prototyping) [1]. Otherwise it is likely that a high proportion of the student feedback concerns the click usability (such as, position and color of buttons), which are mostly less interesting for the evaluation of the accessibility of the general concept of interactive storytelling.

e) Invited workshop participants. As part of a funded research project, it is possible to conduct workshops or summer schools of up to 7 days. Within a week, it is possible to get a group of people author interactive stories, including an introduction to the concepts and tools to be used and their evaluation. In workshops of only one day, this is hard to do, as most of the time is needed for introduction. It also has to be considered that goals against which to evaluate are mostly accessibility and learnability. It is unlikely that the authoring process gets into the phase in which efficiency plays an important role, and the results may consist of toy content that is less important for the authors. On the positive side, it is possible to acquire a wider range of demographics within the group of subjects, if the candidates of interest (e.g. outside academics) follow the invitation [2]. Feedback may again be unstructured and requires a thorough qualitative research strategy (for example, ethnographic observations, interviews and recordings, and structured content analysis).

f) Invited subjects for interview. The least time may be spent if interviewees are invited to get a demonstration of tools and concepts and their feedback is to be recorded and analyzed, or taken by a questionnaire. For evaluating concepts and tools in interactive storytelling this way, it is essential that the interviewees have expertise in this area, which is still not very common. In general, it is possible to get general feedback on ideas if the subjects are experienced in one way or another in authoring. The ideas should be presented at an abstract level and not involve the necessity for the subjects of clicking within a tool. However, if clicking is involved, it is most likely that the feedback concerns mostly the click usability.

g) A greater target group by online tools. Online tools would probably enable forms of quantitative evaluations. In the academic environment, this is often hard to accomplish, but not impossible, if the research is embedded in a long-term endeavor and possibly funded projects. It requires overhead for the researcher in terms of making the tools accessible, maintain a platform, provide tutorials, answer user questions, and still hope for structured feedback. To my knowledge, there are no serious evaluations reported that way in the ICIDS community. Even for tools such as Twine, which are widely known and accessible, it is hard to find concept evaluations.

2.2 Evaluation Goals

As mentioned above, there may be different goals for the evaluation. It is essential that these are known before the setup of an evaluation experiment. However, as also mentioned in the introduction, research in interactive storytelling is often embedded in an experimental environment in which the creative goals are unclear and messy, and cannot be defined before the research process is underway. There may be technical and structural ideas for which it is unclear how the intended end-user experience shall benefit or be changed, for example, by planning algorithms. Also, there are ideas for end-user experiences that raise yet unsolved technical challenges. Examples are free dialogues or meetings in an open space with virtual characters. Authoring concepts for these unsolved challenges often require either work-arounds or adaptation of the content ideas to a system's technical incapability.

In the following, some evaluation goals are sketched and discussed.

a) Click usability, efficiency. In many evaluations, the tool's usability is concerned, for which standards and heuristics exist. Many so-called authoring tools in the realm of interactive storytelling are in fact graphical editors that can be given to test users. Following basic HCI knowledge, these editors can also be analyzed by applying point-and-click heuristics, such as those of Nielsen [8].Click usability can be tested without a long-running authoring project, by giving users concrete single tasks to accomplish. I argue that while these evaluations are not at all irrelevant, they are not the most interesting in the area of novel inventions for interactive storytelling. The gap between hardly accessible technical concepts and classic storytelling knowledge is not addressed by improving the click usability of a tool that is otherwise hard to understand. It is just a different topic. However, if bad usability at this level hinders the usage of a tool and therefore the acquisition of new concepts, this is of course an issue.

b) Learnability, concept understanding. Besides click usability, most of the evaluations in authoring so far concern the aspect of learnability (if not only click usability). This is due to the fact that evaluation tasks often have to be limited in time, and test users are not getting to a level of professionalism that would enable them to judge anything else as their learning curve.

c) Quality of authoring results. When finally authored results of a certain quality exist, this can be assessed with end-users [6]. The take-up by an audience of created work with a tool is certainly an indicator that the tool is in a way successful. However, as in all other aspects mentioned above, this result depends on a number of factors that then have to be identified. These factors are the talent, creativity and knowledge/education of the authors, the time spent on finding problems and work arounds, the complexity of the result [4] in terms of possible end-user interaction and non-linearity, and more. The best accessible 'evaluation' would be kind of a 'post-mortem' or 'making-of' report of the whole process. For academic evaluation, there is a requirement here to profoundly investigate what findings may be generalizable (for example, by cross-reference to other published investigations), and to describe the limitations of this kind of research.

Conclusion

This position paper lists a number of challenges in evaluating authoring solutions that are part of academic research projects, considering that PhD students need evaluation results to prove their findings. It is meant as an input to an academic community discussion that may collect more opinions and define joint strategies how to properly assess research in this area. Authoring in interactive storytelling is part of a messy, creative field that not only depends on single tools, but also on environmental factors that are out of the control of the researchers. For example, this could be the question how the whole field is advancing as an industry calling for educated people, or the question whether novel courses of study take up these topics in their curricula, and lastly by novel advances in AI and interaction technology development.

References

1. Hoffmann, S., Spierling, U., Struck, G. (2011). A Practical Approach To Introduce Story Designers to Planning. In: Proceedings of GET 2011, IADIS International Conference Game and Entertainment Technologies, 22-24 July 2011, Rome, Italy
2. IRIS Repository: Authoring Tools and Creation Methods (2011). http://iris.interactive-storytelling.de/
3. Jhala, A., van Velsen, M. (2009). Challenges in Development and Design of Interactive Narrative Authoring Systems, a Panel. AAAI spring symposium, http://www.aaai.org/Papers/Symposia/Spring/2009/SS-09-06/SS09-06-011.pdf
4. Kapadia, M., Zünd, F., Falk, J., Marti, M., Sumner, R.W., Gross, M.: Evaluating the authoring complexity of interactive narratives with interactive behaviour trees. In: Foundations of Digital Games, FDG 2015 (2015)
5. Mateas, M. and Stern, A. (2005). Build it to Understand It: Ludology Meets Narratology in Game Design Space. In: Selected papers from the DiGRA 2005 International Conference: Changing Views, Worlds in Play. Vancouver, Canada.
6. Mateas, M., Stern, A. (2005). Procedural Authorship: A Case-Study Of the Interactive Drama Façade. In: Proceedings of Digital Arts and Culture (DAC), Copenhagen

7. Mehm, F., Göbel, S., & Steinmetz, R. (2012). Authoring of serious adventure games in StoryTec. In D. Hutchison (Ed.), Lecture Notes in Computer Science: Vol. 7516. E-learning and games for training, education, health and sports (pp. 144-154). doi:10.1007/978-3-642-33466-5_16
8. Nielsen, J. (1993). Usability engineering. New York, NY: Elsevier.
9. Slootmaker, A., Hummel, H., & Koper, R. (2017). Evaluating the usability of authoring environments for serious games. Simulation & Gaming, 48(4), 553-578. doi:10.1177/1046878117705249
10. Spierling, U. (2007). Adding Aspects of "Implicit Creation" to the Authoring Process in Interactive Storytelling. In: M. Cavazza, S. Donikian (eds.): Virtual Storytelling. Proceedings of the 4th International Conference ICVS 2007, Saint-Malo. LNCS vol. 4871, Springer-Verlag Berlin-Heidelberg 2007, pp. 13-25.
11. Spierling, U., Szilas, N. (2009). Authoring Issues Beyond Tools. In: Zagalo, N., Iurgel, I. Petta, P. (Eds.): Interactive Storytelling, Proceedings of ICIDS 2009, LNCS, vol. 5915, Springer Verlag Berlin-Heidelberg, pp. 50–61.
12. Szilas, N., Marty O., Réty, J.-H. (2003). Authoring Highly Generative Interactive Drama. In: Balet et al. (Eds.): Virtual Storytelling, Proceedings of ICVS 2003, LNCS, Vol. 2897, Springer-Verlag Heidelberg, pp. 37-46.
13. Twine. http://twinery.org/