The Authoring Tool Evaluation Problem

Charlie Hargood and Daniel Green

Bournemouth University, UK
{chargood, dgreen}@bournemouth.ac.uk

Abstract. Authoring tools, the software used to create, edit, and develop Interactive Digital Narrative (IDN), are a critical part of both IDN authorship and research. These tools, their features, interface paradigms, visualisations, and user experience (UX) can impact the authoring process and the resulting works, and consequently must inform our wider understanding of IDN context. While IDN research has widely explored data models for authoring tools, feature sets, and demonstrated a variety of developed tools for a range of IDN forms, it has done comparatively very little to evaluate and study the UX of these tools and their impact on authors and their works. In this chapter we survey the existing work on authoring tools and explore the scale of this problem, the reasons for it, how the community has documented this issue, and how we might begin to tackle it. We conclude that existing methods for the study of UX are poorly suited for the study of authoring tools, and that as well as making the study of tool UX a priority we must also develop new methods of evaluation.

"We shape our tools and thereafter our tools shape us" - John Culkin

Interactive Digital Narrative (IDN) is crafted by writers and narrative designers (hereafter simply referred to collectively as "authors") using a myriad of tools and technologies. These technologies are writing tools, domain specific languages, story logic compilers and they vary from roughly hewn in house tools to polished commercial products, and from research prototypes to proprietary studio software. Collectively we can call this collection of technologies "Authoring Tools". There is some debate as to what exactly fits within this definition [61,22] - but for the purpose of this chapter we will be adopting Green's definition [22] of tools created specifically for the purpose of creating or designing the story component of an IDN rather than those that merely could be used.

These tools are critical to IDN - they are the interface through which the ideas and designs of the author are filtered into the work. The vast majority of works of Hypertext fiction or game narrative or electronic literature was, at one point or another, pushed through the aperture of an authoring tool and it cannot be overstated how the design of that tool may have shaped the resulting work. The tool itself may contain functions or interfaces that lend themselves to one style of storytelling or form of interaction over another, but also the User eXperience (UX) of that technology may have an effect on the author such that it changes their workflow or creative thinking which in turn changes their work.

Furthermore the usability of these tools might even impact whether works are created at all, a new author tempted by the medium might be dissuaded by a difficult tool or persuaded by an accessible tool - their works may or may not come to pass because of the design of the tools of our domain. It is for all these reasons that the design of authoring tools is critical to the field of IDN and remains a frequent grand challenge, and topic of debate, within the space of relevant conferences such as ICIDS[1] or workshops such as NHT[2] and AIS[3]. Over a decade ago Spierling and Szilas explained "Authoring is still considered a bottleneck in successful Interactive Storytelling" [67], and further back still Adams [1] was highlighting the challanges in multimedia authoring tool design - these bottlenecks and challenges remain.

UX research is a substantial part of HCI within computer science, concerning itself with understanding how users interact with technology and the impact of design decisions on their use, usability, and user satisfaction. This is critical for both understanding the value of these tools and the impact of their application. As a substantial field UX research is well documented in books such as those by Goodman et al [19], and conferences such as ACM CHI[4], but also in domain specific publications such as the work of Drachen et al [11] and ACM CHIPlay[5] for games UX. However, while a substantial part of the wider technology world recognises the importance and value in UX research, in our own field of IDN its application is limited to understanding "the reader experience" through the evaluation of works and experiences (such as those called for by O'Flynn [53] or reviewed by Revi [57]) and less "the author experience" - where authoring tools often go unevaluated. There are some discussions of the general challenges in IDN authorship [67,37], and Emily Short is a significant voice in the community reviewing these tools [62], however these do not amount to a formal study of the UX of authoring tools. This is a claim that suggests we do not understand the tools on which the work in our field relies upon, but it is also a claim that demands explanation.

1 Problem

As discussed above there are a myriad of IDN authoring tools [61,22]. Some of these tools are proprietary, or remain sealed in their studios beyond scrutiny, meaning it is impossible to tell if UX studies of these tools have been conducted or what the results of those studies might be. However, a great many are open source, freely available research prototypes, or otherwise accessible and yet we do not see a wealth of UX understanding of these tools. A survey of authoring tools from wider IDN community as listed in table 1 reveals a significant majority have not been studied with regards to their user experience, and even for those

[1] https://ardin.online/conferences/icids-interactive-storytelling/ as of 24/01/22
[2] http://nht.ecs.soton.ac.uk/ as of 24/01/22
[3] http://narrativeandplay.org/ais/ as of 24/01/22
[4] https://sigchi.org/conferences/conference-history/chi/ as of 25/01/2022
[5] https://sigchi.org/conferences/conference-history/chiplay/ as of 25/01/2022

that have, the level of evaluation is somewhat modest. In should be noted that our survey excludes tools that might be used for a small part of the authoring process, but not for the creation of the story proper, such as Story Validator [75] (which is used in IDN analysis) or procedural generators such as PaSSAGE [72] or SPHINX [50]. While tools for procedural systems or emergent narrative are not included here that is not to say their UX is any less important or that IDN research should not aspire to explore this space (emergent narrative is addressed in another chapter in this section) - these are however fundamentally very different forms with a different concept of authorship that we do not complicate our initial survey with. We categorise the current state of evaluation of these tools into three groups:

- **Example**: Presentation of tool with examples of use and function.
- **Partial**: Some form of evaluation but does not fully consider the UX.
- **Evaluated**: Those who undergo an evaluation that does explore the UX.

Example		**Partial**	**Evaluated**
StoryPlaces [47,26]	StorySpace [5,6]	Mímisbrunnur [68,69]	SVC [79]
Villanelle [42]	ASAPs [39,40]	CANVAS [35,33,34]	SWB [56,55]
DraMachina [10]	EmoEmma/DSL4MAS [8]	StoryTec [17,16]	GHOST [25]
FAtiMA [41]	GAIA [36]	IDtension [70]	Deig [13,12]
HyperDyn [49]	Scenejo [77,15]		Inform 7 [51]
ABL [43]	ICT Story Manager [20]		Quest [76]
VHE [18]	NSL [74,73]		Articy:Draft [2]
Twine [38]	Scribe [44]		
Art-E-Fact [65,30]	Cyranus [31,32]		
Generator [54]	Creactor [32]		
INSCAPE [3]	TADS [58]		
ADRIFT [78]	Inklewriter [29]		
Ren'Py [59]	Timeline [63]		

Table 1. Authoring tools in 3 evaluation groups: example, partial, evaluated

It is important to note that the genealogy of these technologies is such that one technology is often based on the advances developed in another and where this happens we consider evaluations to be transitive forward, but not backwards. For example, if system B was based on an earlier system A, then we would consider an evaluation of A relevant in part to B, but not vica versa, and we are careful to explain below where we feel an evaluation is transitive or not.

Furthermore it is also important at this stage to stress that it is not our intention to shame any of the scholars, developers, or creators behind these tools (indeed, one of this chapters authors own tools is top of this list). Many of these works have contributed immeasurably to IDN research and practice, and there are often good reasons for the absence of evaluation which we will explore later in this chapter. However, first we need to continue to explore the current state of understanding and evaluation in these three groups of tools.

1.1 Example Group

This group includes the majority of authoring tools in our survey, including some of the arguably most significant tools (in that they are commonly discussed), such as Twine [38] and Inklewriter [29]. Tools in this group are presented without any form of evaluation beyond examples of their use and functionality (thus the label 'example'). They are commonly presented as a discussion of the tool presenting its functionality and design, such as in work by Bernstein [5,6] or Mitchell [49], and in some cases an example of a story created in the tool is presented as a case study as seen in Martens' work [42], or work by Kim [36]. In some cases the tools are presented separately from case studies demonstrating their use, such as Weiss's work presenting Scenejo [77] and Glock's examples [15]. Even a tool used as widely as Twine [38] does not appear to have had its UX studied beyond some anecdotal personal experiences of working with the tool as seen in the works of Miles [46] and Schlauch [60], or use in non-UX evaluations in other domains such as Sørensen et al's work [64] on Inklewriter [29] in education.

This approach is less an attempt of evaluation and more one of demonstration - the authors do not claim to evaluate the tools, but do wish to clarify their functionality and potential through an example. While we learn much about the potential of the technology from this approach, and in some cases exciting new developments in authoring paradigms, we do not learn anything about the author experience of using such tools, their usability, or how the design of these tools and their innovations might influence authorship or resulting works.

In some cases these works were formative and built upon in later systems that were partially evaluated, such as in how INSCAPE [3] and VHE [18] led to StoryTec [17] (which is discussed in the partial group below). However, there is enough difference between these systems that we cannot see such an evaluation as transitive. Furthermore, in other cases it is worth noting that the reader experience of the resulting stories has been later evaluated, such as for StoryPlaces[6] [48] or Art-E-Fact [27], but again this is the UX of the story not the tool, and this speaks to a fundamental problem in the field of the scholarly attention prioritising reader experience over author experience.

1.2 Partial Group

Some tools do go beyond just exampling functionality and features, and evaluate their tools. However, in the case of the tools in this group they stop short of might what be called a full UX study of the author experience, often just measuring one part of it or exploring a single limited aspect.

Stefnisson in their work on Mimisbrunner [68] conducts an evaluation of the tool [69] using the Creativity Support Index (CSI) [9] in order to show their tool is more supportive than directly programming an experience. While a quantitative measure of support for creativity such as CSI is undeniably valuable, it does not

[6] It is to be noted that, at time of writing, a project with a UX evaluation of StoryPlaces' authoring tool has begun but is not yet complete

give us a full picture of the user experience, how the design impacted the authors, or how it affected their workflow. Consequently while quantitative measures provide valuable indicators of potential issues, affordances, or phenomena it is difficult to learn from such an evaluation the reason for, or explanation of, any impact. Indeed, in this case it limits our conclusions merely to the fact that the tool supports creativity more then pure programming.

StoryTec [17], like the group above, demonstrates its functionality through an example in the 80 days work [16] but its original publication does also include a usability evaluation. This evaluation draws an impressive number of participants (n=86) but does not go beyond a simple self-report quantitative measure of usefulness and a couple of user quotes. While we can draw some conclusions from this as to the usability of this particular system, we again are unable to learn anything of substance about the details of the author experience or the impact of the tool.

We see a similar approach in Kapadia's work on CANVAS and their IBT projects. The original IBT work is evaluated in terms of the users "difficulty" in using the tool via a self-report 1-5 difficult rating and the clicks/time to create a story [35] but again while we might learn a limited impression of the usability of this tool we are no wiser as to why or the specifics of the author experience of using IBTs here. Kapadia's further work using this in CANVAS [33] and elsewhere [34] relies upon this earlier evaluation of IBTs being transitive and does not further evaluate the user experience of the authoring tool. However, CANVAS later formed part of SWB [56] which was more extensively evaluated and is discussed below [55].

Finally, we have Szilas' seminal work on IDtension [70,71], a complicated technology framework that is part procedural generator but also includes authorship components. For the most part IDtension fits within the prior group in that the work on it communicates functionality and examples, however Szilas' work with Marty does go beyond this into collaboration with an author that explores the author experience. This evaluation is somewhat informal, includes a single author, and falls short of a rigorous UX study, but there is some consideration for the experience here that can help us learn about the impact of the tool's design, such as the resulting systematic and fractal method of writing.

Consequently, in this group we can see some studies adopting an approach to tool evaluation - demonstrating relative ease of use in a limited quantitative fashion. But, from these studies we do not learn about the specifics of the author experience or how the tools design might impact their use or results in a way that could inform the creation of future tools or our understanding of their use.

1.3 Evaluated Group

In some cases authoring tools have been more substantially evaluated in a way that explores the author experience, as seen with the tools in this group. The majority of this work comes from 3 teams of researchers: the Zurich team of Zund and Poulakos et al, the Skovde team of Engstrom et al, and the authors of this chapter (Hargood and Green of Bournemouth).

Beginning with our own work this is atypical in that it is not evaluating our own tools but the tools of others, specifically Inform 7 [51], Quest [76], and Articy:Draft [2]. This study [23] was motivated in part by the very problem that we are presenting in this chapter, and all of these tools which would have otherwise fallen into the example group without evaluations beyond examples of their use. This makes this study somewhat less typical (all others being evaluations of the developers' own tools) and we will discuss it (and our observational and interview approach) later in this chapter.

The Zurich team, alongside Kapadia (whose work on CANVAS we cover in the previous section), have completed evaluations of their Story Version Control (SVC) [79] and Story World Builder (SWB) [56,55] systems. Their evaluations rely on the System Usability Scale (SUS) [7] - a well-established, and long used, survey from the UX research world. While SUS only provides a quantitative measure of usability (similar to the quantitative approaches in the partial group above), Zund and Poulakos do go beyond this to explore the author experience through interviews and discussion, drawing conclusions on how authors' practice was impacted (such as reusing content) and the usability of paradigms (such as graphs). However, even here the evaluation and analysis of the data is still somewhat brief, and the evidence provided amounts to no more than a couple of pages of discussion rather then a full qualitative dissection of the experience and its explanation. Consequently, while we absolutely can learn about the author experience from these studies, our understanding remains incomplete.

Engstrom's work on Deig [13] and the Deig Writing Companion [12] is significantly more extensive. Here we see a longitudinal study where a substantial group of writers (n=19) spend an extended period writing using the tool (5 days) in the case of Deig [13] and partial use through a full game development project and 3 writers in the case of Deig Writing Companion [12]. The result is a rich collection of evidence on the author experience, usability, intuitiveness, and structure for the Deig tools that represents the gold standard in author experience studies. While the results here are exceptional, so too are the costs - UX research is built upon a foundation of pragmatism and the necessity of keeping methods achievable [19,11], and not all research projects have the resources or the opportunities for longitudinal studies - such that insisting on such an approach would do nothing to address the problems discussed in this chapter. This does not diminish the impressive results of Engstrom's work, but it does mean mean that his approach is not necessarily a solution for the problem.

Finally, outside of the publications of these 3 teams there is one other tool that fits this group: GHOST and the work by Guarneri et al. [25]. Their survey and interview method goes beyond the quantitatively focused work in the partial group exploring author experience issues such as complexity and speed. However, it is still an extremely brief discussion appended to the presentation of the tool, and participants were only reporting on a 15 minute experience with the technology making the conclusions based only on the very start of a project. This makes this study informative but still of limited value in terms of fully understanding the author experience.

2 Explanation

So we have a problem in IDN research. Authoring tools are an essential part of our practice, UX is an essential part of understanding technology and its impact, but only a tiny handful of tools have been through any form of UX evaluation - and for most of those it is brief and/or does not explore the author experience. We are wielding hammers that may be shaping us and our work in ways we don't understand. Why?

While the authors of this chapter do not pretend to have interviewed every IDN academic on the reasons for this issue, from the work discussed above we can infer four potential reasons for this issue:

2.1 Collaboration between research and authors is hard

UX evaluation demands users, and for authoring tools that means authors. This is a skilled and limited participant set. Participant recruitment always raises challenges for any user study, even when the participants may belong to broad demographics, but when we constrain potential participants to a limited group of skilled individuals who already have access to similar tools to those that you are offering access to finding suitable participants becomes a genuine problem. This was something identified as a key challenge for IDN research in Spierling and Szilas' seminal work on IDN authorship [67] where they also point out that authors that are attracted to the project are often direct collaborators, co-designers, or even developers whose proximity makes them unsuitable as a UX evaluation participant. Consequently, faced with the challenge of finding authors there is a temptation to focus on the reader instead where participants are much less limited. This may explain why we see many more reader experience evaluations [57] than author experience evaluations, ,bringing us neatly to our next explanation.

2.2 Reader focus

A lot of IDN work focuses on the reader rather than the author. This is not without reason; the reader experience is to some extent the ultimate outcome of a IDN project, and what impacts their understanding and experience is undeniably important. IDN is also a research field that explores a broad range of mediums from Hypertext to parser fiction, to 3D worlds, to VR, AR, and locative narrative - and the field's understanding of this range of storytelling mediums and their impact on the reader is far from complete and undeniably important. Consequently, a significant proportion of IDN projects may consider authors, and authoring tools, as merely a means to an end - a step along the road to a particular piece of work or story deployment, and that any evaluation will seek to evaluate those stories rather than the tools used along the way. This may also be why, in our survey above, the most common form of authoring tool evaluation is "example" - stories created in the tools as proof of its functionality, but lacking scrutiny of their UX or authorial impact.

2.3 UX is not a priority

IDN research covers a range of questions and areas of study far beyond UX, and in exploring these mediums, poetics, and technologies , as discussed above, it is possible authoring tools are often created as a means to an end towards answering these other questions. Given the limited resources of a research project, and the cost of running user studies, we might postulate that a motivation for not exploring the UX of tools is their cost, given that they don't represent the priority for the project. There are two issues at play here - the first is the author experience as a priority (something this chapter is trying to address) but also the cost of UX. As discussed earlier, UX research often highlights the importance of pragmatism in methods [19,11], and UX researchers in other parts of the digital creative industries such as Medlock [45] and Huguenin [28] often stress the need for pragmatic approaches that recognise limited resources and "quick and dirty" UX methods in a world that potentially recognises the value of UX but does not necessarily prioritise it. Consequently, difficult and/or expensive methods such as the longitudinal studies presented by Engstrom [13,12] cannot be the sole answer to this problem at scale. While these types of study provide invaluable insight, we also need the "quick and dirty" methods seen elsewhere in UX research to make studies as accessible and pragmatic as possible. This brings us to our final explanation.

2.4 UX methods are poorly suited to the study of authoring tools

Modern UX and HCI research has been criticised for being overly dogmatic. Greenberg and Buxton highlighted this in their seminal work [24] and also called for UX research to develop new methods custom - suited to the focus of their study rather than always adopting established protocol. Established UX study best practice often focuses on task-based usability tests [19] - have a user use your product as it is supposed to be used in a number of set tasks and record their performance, response, and experience. This may work find if you are developing a car or a shopping website, where the common usage is clear and achievable in a short period of time. But what are the common tasks for IDN authorship? And how long do they last? Were we to use this standard approach with an authoring tool we might face the problem of telling an author to sit down and create an IDN, which might take days or even years. It is for this reason we see some works like Engstrom's [13,12] using longitudinal methods, and also why studies such as Guarneri et al's [25] are problematic as a typical short task method means your participants only barely begin to create their stories when that early set up is not typical of a substantial part of the writing process. Breaking authorship down into sub tasks comes with the challenge of identifying representative tasks for something as wide reaching and varied as a creative work. Even were we to achieve that we are still faced with the problem that all of those tasks put together is still a study that lasts the length of the creative process and beyond the resources of most researchers. Such long studies also hinder the application of many of a UX researchers most useful tools - methods such as the verbal

protocol, described as "the single most valuable usability engineering method" by the UX pioneer Jakob Nielsen [52], become impossible to apply over such exercises robbing us of the valuable experience data they might provide. There are other solutions to longitudinal studies like this in UX, such as the diary study method [19], but these are very expensive as they effectively demand commissioning a full writing project - limiting the study to very a small n and failing to provide the pragmatism demanded above.

This absence of pragmatic user experience evaluation methods combined with the availability of short quantitative approaches such as SUS [7] and CSI [9] potentially leads researchers into relying on these brief measures of usability instead, such as we saw above in many of the works of our survey's "partial" group. It is to be noted that survey based quantitative measures play a valuable role in UX evaluation, helping to identify phenomena for study and form broad initial observations. However, used alone they leave us with a relative number confirming relative usability but not an in-depth understanding of the authoring experience or the impact of the tool. Consequently, established UX methods have not only failed to serve author experience evaluation, they have arguably laid a quantitative baited trap potentially tempting researchers into avoiding its qualitative challenges.

3 Solutions

The nature of this chapter, indeed this book, is to highlight unsolved problems and unanswered questions. Consequently the authors of this chapter do not claim to have a solution. However, that doesn't mean we cannot discuss potential ways forward to begin to address this issue.

3.1 We need new methods

As described above, there are a number of problems with the application of established UX methods on authoring tools:

- It is challenging to break down a complex creative process such as IDN authorship for task-based usability tests
- Authorship is a lengthy process
- Shortening IDN authorship risks only evaluating story set up, which isn't typical of the full authoring process
- Longitudinal studies lose access to some evidence and are not always pragmatic
- Quantitative measures are useful but insufficient by themselves to fully understand the author experience

Consequently, in the spirit of Greenberg and Buxton [24], we need to develop new bespoke methods for authoring tool evaluation. These methods should seek to provide insight into traditional UX concepts such as usability, accessibility,

and performance - but also the impact of the author experience on author workflow and practice, and the impact on the work itself. Furthermore, these methods should tackle not just the impact of the interface paradigms, but also authoring tool features and functionality, and the impact of underlying models and structures. All of these variables are important to understanding the impact of our tools, and that impact can come from all of these design sources.

While we don't have a solution to all of this yet, the authors of this chapter have been working towards some new methods in this space to address this problem, and we call on the community to assist in these efforts. We have developed [23], and continue to refine [21], a new method for understanding the author experience of IDN authoring tools. This method aims to both be pragmatic in being deliverable in a 1-2 hour study, but still target a representative sample of the authoring process, and gather quality data on the author experience.

The underlying principle of our method is completing an incomplete story. The participant will be given an authoring tool with an IDN story that is part written but missing a significant part of the story. They will then be asked to complete this missing part with guidance notes on what it should include, but creative freedom to interpret those as they see fit. In summary, the protocol is as follows:

1. Participants complete a brief pre-study demographic survey
2. Participants are given a short video and notes training them in basic use of the authoring tool.
3. Participants are given the authoring tool with a recognisable story that is partially complete but missing a section in the middle
4. Participants are given guidance on what this section should include but given creative freedom to complete the work.
5. While completing the work the participant thinks aloud (verbal protocol), their screen is recorded to log interactions, and the researcher documents their behaviour, response, and attitude.
6. Following the exercise, the participant is interviewed on their experience targeting the impact of different parts of the tool design and their process.
7. Study closes. Data gathered includes the demographic survey, resulting story, screen recording of the exercise, audio recording of think aloud and interview, and researcher notes on behaviour during the study.

The specifics of the story selected (for our studies we deployed the often used "Little Red Riding Hood" and selected scenes from "Mass Effect" for different studies) and specific interview questions can vary depending on study - but this is the outline of the method we have been developing. There is a benefit to using a story with which participants are familiar in order to avoid a further training burden of familiarizing them with the story, but there is also a benefit in controlling the content of the story to be representative of the form of IDN being explored. "Little Red Riding Hood" has previously been used as a staple test story IDN by the research community [66], and the interactive fiction community has similarly made use of Cloak of Darkness [14] in a similar way. Benchmarking,

and the role of stories such as Little Red or Cloak of Darkness as standard stories to test in evaluations (and their suitability), is naturally a research topic in its own right, and while expanding on that here is outside the scope of this chapter that does not make it any less an important, and further work there would be valuable.

We have explored this method, and variations of it, on a set of authoring tools which (prior to our studies) were both prominent and unevaluated beyond examples of use: Inform 7 [51], Quest [76], and Articy:Draft [2] have all been evaluated using the method [23]. We have also applied iterations of the method to prototype interfaces we have been working on for new tools [21]. These studies have shown our method to be effective, returning a wealth of useful authoring experience data, but also reveal weaknesses within the method as well.

In terms of weaknesses there are two key issues here. The method still requires tool training, and this is both time consuming and a problem in terms of what is sufficient training for a genuine test. This problem could be avoided by recruiting prior users - but this would not help for new tools, and we were keen to target the author recruitment problem discussed in the previous section by targeting as broad a section of participants as possible. As such, we recruited people with a professional interest in IDN (and as such would at least be comfortable with the concepts involved) but still trained them in individual tools. Only about 14% of our participants felt the training was insufficient for the exercise [21] - but this remains a constraint of the method that might influence results. The second weakness is the tension between structure and creativity. In attempting to retain the consistency of task-based usability methods we wanted to try to have authors complete a repeatable and representative part of the story. To do this, while avoiding the problem of participants only setting up their story, we developed an approach that had participants complete a middle section of a partially complete story that guided them in terms of its content to ensure representative IDN content was explored, such as introducing characters, dialogue, exploring a space, and other common patterns. However, early feedback on the methodology criticised the artificial nature of the exercise in being too constrained given the inherent free form creativity involved in writing, so consequently we adjusted our content instructions to be mere suggestions, and while we kept these suggestions, participants were given more creative freedom. This is a tension in the methodology that is as yet unresolved - guidance ensures evaluation consistency and repeatable exercises but constrains the fundamental creativity in genuine authorship. Similarly, the more we give participants creative freedom the more genuine the exercise but the more exercises diverge, and the less we can be sure of representative content.

In terms of strengths, we are pleased that the "completing a partially completed story" approach keeps the study length modest, and that by training participants we can recruit broadly. Both of these ensure a pragmatic approach to the method often called for in UX evaluation. At the same time, our studies have shown this method can return rich author experience data from which we were able to detect impacts on the authors' workflow, impacts on their resulting

stories, and impact on their attitudes and experience [23,21]. Furthermore, the story-finishing approach helps to ensure the exercises are more representative of the writing process by avoiding an exercise that is story set up only, and guiding the content to representative patterns (although as discussed above there is a tension in this part of the method design). Finally, the rich array of qualitative data gathered - from recordings of use, to the verbal protocol, to the interview - meant we were able to explore a range of areas of impact such as interface paradigms, functionality, and underlying models. As we continued to use this method we tweaked the story content being completed, or the interview questions, to focus on the parts of the tool most in need of scrutiny - for example in a later study we focused part of our protocol on story testing functions in order to better impact the author experience impact there [21]. Not only was our work enlightening in terms of the author experience of the tools we explored, but we were able to use it to draw together a set of principles for future authoring tool design [21].

Ultimately, we do not pretend this is "one method to rule them all" - the approach shows promise, but has weaknesses, and needs further refinement. Furthermore, to attempt to establish a new author UX evaluation orthodoxy would be contrary to the very call to action that inspired us to develop new methods [24]. Indeed, while pragamtic qualitative methods, such as those we propose, are part of the way forward there are other paths that also demand attention - such as quantitative methods, and model evaluations. Furthermore, pragmatic lighter methods such as ours do not replace the qualitative value of larger scale free writing longitudinal studies, such as those demonstrated by Engstrom [12,13] - which are more genuine, and less artificial, than what we propose here (if significantly more expensive). However, our method does represent an example of the beginnings to a potential approach to address part of the problem this chapter explores. There will be a need for other bespoke methods that address this problem, and different mediums will demand their own bespoke methods. As previously mentioned evaluating a procedural tool is very different to the more conventional authoring tools explored in this chapter - and bespoke methods for different forms is at the heart of both Greenburg's motivating work [24] and our intent.

As noted, the approach we describe here is principally qualitative, however as stated above quantitative measures also have an important role to play. While long established methods in this area such as SUS [7] already provide valuable instruments, here we may further develop these into bespoke quantitative measures for authorial experience - indeed, the previously mentioned CSI [9] is such a development, and might be explored further.

Finally it is also important to confront the assumption of what is being evaluated - a tool, such as interface paradigms and functionality, or underlying models, such as how that tool understands the components and connections of an IDN. One model (such as caligraphic hypertext [4]) could be used by many tools (such as Twine [38], Storyspace [6], and others). No authoring tool evaluation is completely divorced from either the tool or the model - as the author must

experience both in order to create. However, another gap in our field lies in the development of new methods to explore specifically the impact of one or the other on the author experience through direct comparison.

3.2 The author experience needs to become a priority

It is possible that, despite recognizing the value of understanding the author experience, the absence of tool evaluations of depth is not just due to the challenges of those studies and their methods, but also due to the priority of challenges and questions to the research community. Consequently, while we feel methodological challenges represent a big part of this problem (as discussed above), all the methods in the world won't help if answering the question of the author experience is not a priority for the community. Consequently, if this chapter does anything beyond highlighting the problems of authoring tool evaluation, it is to call the community to action to address this problem. As we have laid out here authoring tools are a critical part of IDN and our understanding of them, and how they impact authors and their works, is lacking. Consequently, beyond our call for new methods above, we conclude this chapter with three grand challenges to the community:

1. **Study new tools**: As the community continues to develop new tools we need to make understanding the author experience of these technologies a priority and a part of research plans that include tool development. We should be developing new methods to do this, and iterating on prior methods, to improve our understanding of the consequences of new technology.
2. **Study old tools**: Understanding the author experience needs to not only be part of projects that include tool development, but also something that concerns itself with the unevaluated tools of the past, particularly those widely used and adopted by the community. Furthermore, as these studies begin to appear we should, as good scientists, be repeating those studies to confirm, refute, and modify our findings and understanding. Understanding the author experience cannot just be a part of research that happens to include a tool for another focus - it needs to become a priority in its own right.
3. **Study the impact on both author and work**: We need to broaden our approach to the author experience beyond mere usability. This does not mean to disregard usability, it remains a critical issue, but to go beyond it to understand how the workflow and practice of authors is affected in their UX more broadly. While some studies we have discussed here have begun to do this we also need to go beyond even this to explore the impact on resulting works and explore how stories are changed by the tools that develop them.

This is not to suggest that this should be the only priority for IDN research, but rather that it should be a priority, and current work in the field (as laid out in this chapter) does not support that it is. This chapter also lays out why we feel it should be a priority, in that the author experience of tools has direct impact on the works created and the accessibility of the field. Consequently author experience not only influences reader experience but the existence of works, structural designs, and styles of writing in our medium as a whole.

Authorship, and authoring tools, remain a critical part of IDN research [67], and UX remains a critical part of the study of technology [19], creating a combined challenge in our field. Our understanding of our tools needs to go beyond examples of their use, and our approach to UX needs to mature beyond a simple metric of usability. We need a richer understanding of the author experience, how we shape our tools, and how they shape us.

References

1. Adams, B., Venkatesh, S.: Authoring multimedia authoring tools. IEEE multimedia **11**(3), 1–6 (2004)
2. Articy-Software: Articy: Draft, https://www.articy.com/en/, accessed: 25/01/2022
3. Balet, O.: Inscape an authoring platform for interactive storytelling. In: International Conference on Virtual Storytelling. pp. 176–177. Springer (2007)
4. Bernstein, M.: Card shark and thespis: exotic tools for hypertext narrative. In: Proceedings of the 12th ACM conference on Hypertext and Hypermedia. pp. 41–50 (2001)
5. Bernstein, M.: Storyspace 1. In: Proceedings of the thirteenth ACM conference on Hypertext and hypermedia. pp. 172–181 (2002)
6. Bernstein, M.: Storyspace 3. In: Proceedings of the 27th ACM Conference on Hypertext and Social Media. p. 201–206. HT '16, Association for Computing Machinery, New York, NY, USA (2016)
7. Brooke, J., et al.: Sus-a quick and dirty usability scale. Usability evaluation in industry **189**(194), 4–7 (1996)
8. Charles, F., Pizzi, D., Cavazza, M., Vogt, T., André, E.: Emoemma: Emotional speech input for interactive storytelling. In: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2. pp. 1381–1382 (2009)
9. Cherry, E., Latulipe, C.: Quantifying the creativity support of digital tools through the creativity support index. ACM Transactions on Computer-Human Interaction (TOCHI) **21**(4), 1–25 (2014)
10. Donikian, S., Portugal, J.N.: Writing interactive fiction scenarii with dramachina. In: International conference on technologies for interactive digital storytelling and entertainment. pp. 101–112. Springer (2004)
11. Drachen, A., Mirza-Babaei, P., Nacke, L.E.: Games user research. Oxford University Press (2018)
12. Engström, H.: 'i have a different kind of brain'—a script-centric approach to interactive narratives in games. Digital Creativity **30**(1), 1–22 (2019)
13. Engström, H., Brusk, J., Erlandsson, P.: Prototyping tools for game writers. The Computer Games Journal **7**(3), 153–172 (2018)

14. Firth, R.: Cloak of darkness. IF Wiki, https://www.ifwiki.org/Cloak_of_Darkness, accessed: 25/01/2022
15. Glock, F., Junker, A., Kraus, M., Lehrian, C., Schäfer, A., Hoffmann, S., Spierling, U.: "office brawl" a conversational storytelling game and its creation process. In: Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology. pp. 1–2 (2011)
16. Göbel, S., Mehm, F., Radke, S., Steinmetz, R.: 80days: Adaptive digital storytelling for digital educational games. In: Proceedings of the 2nd international workshop on Story-Telling and Educational Games (STEG'09). vol. 498 (2009)
17. Göbel, S., Salvatore, L., Konrad, R.: Storytec: A digital storytelling platform for the authoring and experiencing of interactive and non-linear stories. In: 2008 International Conference on Automated Solutions for Cross Media Content and Multi-Channel Distribution. pp. 103–110. Ieee (2008)
18. Göbel, S., Schneider, O., Iurgel, I., Feix, A., Knöpfle, C., Rettig, A.: Virtual human: Storytelling and computer graphics for a virtual human platform. In: International Conference on Technologies for Interactive Digital Storytelling and Entertainment. pp. 79–88. Springer (2004)
19. Goodman, E., Kuniavsky, M.: Observing the user experience: A practitioner's guide to user research. Elsevier (2012)
20. Gordon, A., van Lent, M., Van Velsen, M., Carpenter, P., Jhala, A.: Branching storylines in virtual reality environments for leadership development. In: Proceedings of the national conference on Artificial Intelligence. pp. 844–851. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999 (2004)
21. Green, D.: Don't Forget to Save! The Impact of User Experience Design on Effectiveness of Authoring Video Game Narratives. Ph.D. thesis (2021)
22. Green, D., Hargood, C., Charles, F.: Define "authoring tool": a survey of interactive narrative authoring tools. In: Authoring for Interactive Storytelling Workshop 2018 (2018)
23. Green, D., Hargood, C., Charles, F.: Use of tools: Ux principles for interactive narrative authoring tools. Journal on Computing and Cultural Heritage (JOCCH) **14**(3), 1–25 (2021)
24. Greenberg, S., Buxton, B.: Usability evaluation considered harmful (some of the time). In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 111–120 (2008)
25. Guarneri, A., Ripamonti, L.A., Tissoni, F., Trubian, M., Maggiorini, D., Gadia, D.: Ghost: a ghost story-writer. In: Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter. pp. 1–9 (2017)
26. Hargood, C., Weal, M.J., Millard, D.E.: The storyplaces platform: Building a web-based locative hypertext system. In: Proceedings of the 29th on Hypertext and Social Media, pp. 128–135 (2018)
27. Hassenzahl, M., Ullrich, D.: To do or not to do: Differences in user experience and retrospective judgments depending on the presence or absence of instrumental goals. Interacting with computers **19**(4), 429–437 (2007)
28. Huguenin, J.: Running user tests with limited resources and experience. In: Games User Research. Oxford University Press (2018)
29. Ingold, J.: Inklewriter, https://www.inklestudios.com/inklewriter/, accessed: 07/04/2022
30. Iurgel, I.: From another point of view: art-e-fact. In: International Conference on Technologies for Interactive Digital Storytelling and Entertainment. pp. 26–35. Springer (2004)

31. Iurgel, I.A.: Cyranus–an authoring tool for interactive edutainment applications. In: International Conference on Technologies for E-Learning and Digital Entertainment. pp. 577–580. Springer (2006)
32. Iurgel, I.A.: An authoring framework for interactive narrative with virtual characters. Ph.D. thesis, Technische Universität (2008)
33. Kapadia, M., Frey, S., Shoulson, A., Sumner, R.W., Gross, M.H.: Canvas: computer-assisted narrative animation synthesis. In: Symposium on Computer Animation. pp. 199–209 (2016)
34. Kapadia, M., Shoulson, A., Steimer, C., Oberholzer, S., Sumner, R.W., Gross, M.: An event-centric approach to authoring stories in crowds. In: Proceedings of the 9th International Conference on Motion in Games. pp. 15–24 (2016)
35. Kapadia, M., Zünd, F., Falk, J., Marti, M., Sumner, R.W., Gross, M.: Evaluating the authoring complexity of interactive narratives with interactive behaviour trees. Foundations of Digital Games (2015)
36. Kim, S., Moon, S., Han, S., Chan, J.: Programming the story: Interactive storytelling system. Informatica **35**(2) (2011)
37. Kitromili, S., Jordan, J., Millard, D.E.: What authors think about hypertext authoring. In: Proceedings of the 31st ACM Conference on Hypertext and Social Media. pp. 9–16 (2020)
38. Klimas, C.: Twine, https://twinery.org/, accessed: 25/01/2022
39. Koenitz, H.: Extensible tools for practical experiments in idn: the advanced stories authoring and presentation system. In: International Conference on Interactive Digital Storytelling. pp. 79–84. Springer (2011)
40. Koenitz, H., Chen, K.J.: Genres, structures and strategies in interactive digital narratives–analyzing a body of works created in asaps. In: International Conference on Interactive Digital Storytelling. pp. 84–95. Springer (2012)
41. Kriegel, M., Aylett, R., Dias, J., Paiva, A.: An authoring tool for an emergent narrative storytelling system. In: AAAI Fall Symposium: Intelligent Narrative Technologies. pp. 55–62 (2007)
42. Martens, C., Iqbal, O.: Villanelle: an authoring tool for autonomous characters in interactive fiction. In: International Conference on Interactive Digital Storytelling. pp. 290–303. Springer (2019)
43. Mateas, M., Stern, A.: A behavior language for story-based believable agents. IEEE Intelligent Systems **17**(4), 39–47 (2002)
44. Medler, B., Magerko, B.: Scribe: A tool for authoring event driven interactive drama. In: International Conference on Technologies for Interactive Digital Storytelling and Entertainment. pp. 139–150. Springer (2006)
45. Medlock, M.C., Wixon, D., Terrano, M., Romero, R., Fulton, B.: Using the rite method to improve products: A definition and a case study. Usability Professionals Association **51**, 1963813932–1562338474 (2002)
46. Miles, A.P., Jenkins, K.: (re) born digital–trans-affirming research, curriculum, and pedagogy: An interactive multimodal story using twine. Visual Arts Research **43**(1), 43–49 (2017)
47. Millard, D., Hargood, C., Howard, Y., Packer, H.: The storyplaces authoring tool: pattern centric authoring. In: Authoring for Interactive Storytelling Workshop 2017 (2017)
48. Millard, D.E., Packer, H., Howard, Y., Hargood, C.: The balance of attention: The challenges of creating locative cultural storytelling experiences. J. Comput. Cult. Herit. **13**(4) (dec 2020)
49. Mitchell, A., McGee, K.: The hypedyn hypertext fiction authoring tool. In: Proceedings of the 2nd workshop on Narrative and hypertext. pp. 19–22 (2012)

50. Morgan, L., Haahr, M.: Honey, i'm home: An adventure game with procedurally generated narrative puzzles. In: International Conference on Interactive Digital Storytelling. pp. 335–338. Springer (2020)
51. Nelson, G.: Natural language, semantic analysis and interactive fiction. IF Theory Reader **141**, 99–104 (2006)
52. Nielsen, J.: Usability engineering. Morgan Kaufmann (1994)
53. O'Flynn, S.: Media fluid and media fluent, e-literature in the era of experience design. Hyperrhiz (20) (2019)
54. Pope, J.: Generator, https://genarrator.org/, accessed: 25/01/2022
55. Poulakos, S., Kapadia, M., Maiga, G.M., Zünd, F., Gross, M., Sumner, R.W.: Evaluating accessible graphical interfaces for building story worlds. In: International Conference on Interactive Digital Storytelling. pp. 184–196. Springer (2016)
56. Poulakos, S., Kapadia, M., Schüpfer, A., Zünd, F., Sumner, R.W., Gross, M.: Towards an accessible interface for story world building. In: Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference (2015)
57. Revi, A.T., Millard, D.E., Middleton, S.E.: A systematic analysis of user experience dimensions for interactive digital narratives. In: International Conference on Interactive Digital Storytelling. pp. 58–74. Springer (2020)
58. Roberts, M.: Tads, https://www.tads.org/, accessed: 25/01/2022
59. Rothamel, T.: Ren'py, https://www.renpy.org/, accessed: 31/05/2022
60. Schlauch, M.: The unlucky hans. the difficulties of adapting fairy tales as text-based games for young readers. gamevironments (15) (2021)
61. Shibolet, Y., Knoller, N., Koenitz, H.: A framework for classifying and describing authoring tools for interactive digital narrative. In: International Conference on Interactive Digital Storytelling. pp. 523–533. Springer (2018)
62. Short, E.: Writing in collaboration with the system (Oct 2014), https://emshort.blog/2014/10/29/writing-in-collaboration-with-the-system/
63. Silva, P., Gao, S., Nayak, S., Ramirez, M., Stricklin, C., Murray, J.: Timeline: An authoring platform for parameterized stories. In: ACM International Conference on Interactive Media Experiences. p. 280–283. IMX '21, Association for Computing Machinery, New York, NY, USA (2021)
64. Sørensen, B.H., Levinsen, K.T.: Evaluation as a powerful practices in digital learning processes. Electronic Journal of E-learning **13**(4), pp290–300 (2015)
65. Spierling, U., Iurgel, I.: " just talking about art"–creating virtual storytelling experiences in mixed reality. In: International Conference on Virtual Storytelling. pp. 179–188. Springer (2003)
66. Spierling, U., Iurgel, I., Richle, U., Szilas, N.: Workshop on authoring methods and conception in interactive storytelling. In: Joint International Conference on Interactive Digital Storytelling. pp. 356–357. Springer (2009)
67. Spierling, U., Szilas, N.: Authoring issues beyond tools. In: Joint International Conference on Interactive Digital Storytelling. pp. 50–61. Springer (2009)
68. Stefnisson, I., Thue, D.: Mimisbrunnur: Ai-assisted authoring for interactive storytelling. In: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. vol. 14 (2018)
69. Stefnisson, I.S.: Mímisbrunnur: A Mixed-Initiative Authoring Tool for Interactive Storytelling. Ph.D. thesis (2018)
70. Szilas, N.: Idtension: a narrative engine for interactive drama. In: Proceedings of the technologies for interactive digital storytelling and entertainment (TIDSE) conference. vol. 3, pp. 1–11 (2003)
71. Szilas, N., Marty, O., Réty, J.H.: Authoring highly generative interactive drama. In: International Conference on Virtual Storytelling. pp. 37–46. Springer (2003)

72. Thue, D., Bulitko, V., Spetch, M., Wasylishen, E.: Interactive storytelling: A player modelling approach. In: AIIDE. pp. 43–48 (2007)
73. Ursu, M.F., Zsombori, V., Wyver, J., Conrad, L., Kegel, I., Williams, D.: Interactive documentaries: A golden age. Comput. Entertain. **7**(3) (sep 2009)
74. Ursu, M.F., Cook, J.J., Zsombori, V., Kegel, I.: A genre-independent approach to producing interactive screen media narratives. In: AAAI Fall Symposium: Intelligent Narrative Technologies. p. 174 (2007)
75. Veloso, C., Prada, R.: Validating the plot of interactive narrative games. In: 2021 IEEE Conference on Games (CoG). pp. 01–08 (2021)
76. Warren, A.: Quest, http://textadventures.co.uk/quest, accessed: 25/01/2022
77. Weiss, S., Müller, W., Spierling, U., Steimle, F.: Scenejo–an interactive storytelling platform. In: International conference on virtual storytelling. pp. 77–80. Springer (2005)
78. Wild, C.: Adrift, https://www.adrift.co/, accessed: 25/01/2022
79. Zünd, F., Poulakos, S., Kapadia, M., Sumner, R.W.: Story version control and graphical visualization for collaborative story authoring. In: Proceedings of the 14th European Conference on Visual Media Production (CVMP 2017). pp. 1–10 (2017)