Conversational Interaction with Historical Figures: What's it good for?

Xin Qian, Douglas W. Oard, and Joel Chan

University of Maryland, College Park, 20740, USA
{xinq, oard, joelchan}@umd.edu

Abstract. Historical records can tell us about the knowledge, opinions, and actions of individuals, but direct use of historical records is a complex process that is better attuned to the needs of scholars than the interests of the general public. This paper explores the potential for conversational interaction with representations of historical figures that are constructed from primary and secondary source materials. An interview study, supported by a computational prototype as a provocation, is used to elicit expert assessments of the potential uses of such systems in historical museums. Qualitative analysis of study results reveals three broad classes of design implications: situating design in archival assurance, creating immersive user experiences, and supporting active inquiries.

Keywords: Information retrieval · Question answering · Digital archives · Cultural institutions · Natural language dialogue

1 Introduction

Over the last two decades of the 20th century, the General Electric corporation advertised using the slogan "We bring good things to life!" By this, they meant both that they gave life to good things (precisely, machines), and that those good things could enrich people's lives. Now, two decades after GE retired the slogan, we stand at the verge of a third possible meaning as we begin to draw on the vast digital and digitizable traces that people have left behind to reanimate some aspects of their lives. In particular, we are interested in providing the possibility of conducting meaningful and informative conversations with historical figures who are no longer alive, and thus no longer able to speak for themselves.

Since the inception of writing, people have left written traces of some of their activities. Over time, those traces have been enriched in many ways, both in content and form, to the point where today we have not just written records but also spoken records and vast stores of digital activity records. Not all of this makes it into the future in a usable form, but today's profusion of digital content, coupled with the rapidly declining costs of storage, make it likely that very substantial quantities will. The usual way of using this information involves three distinct steps: first finding that which might be expected to be useful, then assimilating what actually is useful from among what's found, and finally using what has been learned in some way [26]. Iteratively applying such

a process effectively is an acquired skill, and thus it is no surprise that doctoral students are typically better at it than high school students [14,19]. There are, however, many more high school students than there are doctoral students, suggesting that there might be a market for approaches that allow people to interact somewhat more naturally with the historical record. The search-assimilate-apply process dates back a few millennia to the invention of archives [18], and later of libraries [40]. However, humans have been optimized by evolution to acquire and assimilate information through conversation since perhaps hundreds of millennia before that. Our goal, therefore, is to support conversational interaction with representations of historical figures from our past.

When a conversational agent emulates the interaction style of someone who is no longer alive, this has been called "virtual immortality" [5]. Early systems that demonstrated the potential for conversations with historical figures such as Charles Darwin [50] or Richard Nixon [12] were hand-engineered, able to support a conversation with a specific figure on a limited range of topics. However, two technologies are now converging to permit broader support for these kinds of conversational interaction. One thread is a type of knowledge-grounded conversational agent that has led to systems that are today popularly referred to as "retrieval-based chatbots" [55]. The key idea in a retrieval-based chatbot is that when the system reaches a point in a conversation where it needs some information, it turns to a search engine to find what it needs [15]. For a simple example, ask Amazon's Alexa where the third-largest pyramid is; Alexa will tell you that it found the answer on the Web. The second thread is a line of work on automated style rewriting in which the goal is to transform written or spoken content into forms of expression that more closely resemble those of some specific person. A third piece of the puzzle that makes it possible to connect these two lines of work is the increasing availability of content in both digital and digitized primary and secondary sources. We can use all of that content as a basis for retrieval, and we can use the first-person writing and speaking we find there as a basis for rewriting retrieved content into a form that the person we wish to represent might actually have produced.

The fact that we could now build such systems leads to the question of whether we should. That question, in turn, raises at least two concerns. First, what would we do with such a system—what purposes could it serve? Our principal goal in this paper is to explore that question in the context of one type of cultural heritage institution: a history museum. A second equally important question is how best to navigate the critical ethical questions that arise. We would, after all, essentially be putting words in the mouth of the historical figure, and we would be doing so with imperfect technology that might sometimes hallucinate things that they would not actually have said. We largely leave questions about how best to address such concerns to future work, noting here only that Amazon's Alexa points to one possible approach—be clear with those who use your system, what you are doing, and how you are doing it.

With that as background, the remainder of this paper is organized as follows. The next section reviews related work on virtual immortality, retrieval-based

chatbots, and style rewriting. We then describe a text collection, and a system to retrieve content from that collection that we created to illustrate what might be possible. We used the system as a boundary object [51] to facilitate an interview study in which four academic experts in libraries, archives, and museums reacted to our vision, and to our initial prototype's embodiment of some aspects of that vision, offering their ideas on how such systems might actually be used. We conclude with a qualitative analysis of the results of that study.

2 Related Work

In this section, we review related work on virtual immortality, retrieval-based chatbots, and style rewriting.

2.1 Virtual Immortality Demonstration Systems

Virtual immortality is a rather grand name for systems that have sought to demonstrate the potential for conversing with representations of specific people, specifically historical figures or fictitious characters [5,38,47]. To date, all such systems have been hand-engineered, with curated content that is designed to answer questions that the designers anticipate might be asked. Prominent examples include a systems that can respond with recordings of actors playing Charles Darwin's [50] or Albert Einstein's ghost [33], a virtual President Nixon [12], interactive storytelling by Holocaust survivors [3,52,2], a representation of the playwright August Strindberg serving as a tour guide for Stockholm [24,6], and a fictional character (Sergeant Blackwell) who answers questions after a training session [31]. Figure 1 summarizes some the characteristics of these systems.

2.2 Retrieval-based Chatbots

Retrieval-based chatbots are a flexible approach to building automated systems for conversational text interactions that can converse knowledgeably with people on a range of topics. The responses generated by a retrieval-based chatbot are typically based on existing material found in some text collection(s), perhaps augmented by content from a database or knowledge graph [54]. There has been evolutionary development from single-turn question answering [15,41,54] to multi-turn conversations [55]. One approach to developing such systems has therefore been to first build systems that do well at retrieving the right response for a single turn, and then to extend that system to interpret subsequent user requests in the context of prior user requests and system responses. The key to that first step, successfully retrieving a suitable response, lies in matching user requests with candidate responses from the collection. The matching problem is core to research in information retrieval (IR), and in particular the sub-field of open-domain question answering (QA) [53]. State-of-the-art QA systems build on neural "deep learning" BERT (Bidirectional Encoder Representations from Transformers) models that use self-attention to model both the linguistic context

	Charles Darwin's Ghost (Same applies to Albert Einstein's)	Ask the President	New Dimension in Testimony (NDT)	August System	Sergeant Blackwell
Figure	Charles Darwin	President Nixon	Holocaust survivors	August Strindberg	Sergeant Blackwell
Real records?	Curated: writers create the narratives	Video sequences	Yes, elicited w/ protocol in film studios	Curated	Fictional
Scope of answerable questions	Any of 199 questions, from philosophical to personal	Any of 280 questions	Questions that are answerable by recorded dialogs	Restaurants and facilities in Stockholm, information about KTH, the research, the system, and August Strindberg	57 pre-defined subject lines including the identity of the character, its origin, and questions about the training

Fig. 1. Virtual immortality demonstration systems.

of each question and document term and relationships between terms that appear in questions and terms that appear in documents [16]. These deep learning models are trained on massive quantities of language use, including many examples of questions and answers, and they have demonstrated substantially better accuracy [32] than traditional retrieval techniques based on question rewriting and term matching [46]. However, the massively interconnected self-attention networks in BERT result in slower system response, so the usual approach is to do an initial rapid search and then further refine that smaller set of search results using BERT. This is the approach taken by monoBERT [36], which we use in Section 3.

2.3 Style Rewriting

The general goal of style rewriting is to change the form of a passage of text or speech to match specific style requirements of an application [20]. Early work on this task used hand-crafted rules to perform, for example, paraphrase and simplification [11]. More recent models have used statistical approaches, for example, guided by a word co-occurrence graph learned from a large collection of text in the desired style [4,28]. As with QA, the most recent work uses Transformer-based models for sequence-to-sequence transfer, treating style rewriting in a manner akin to the way that modern systems like Google Translate convert text from one language to another. Such techniques have been used to transfer text into styles used by characters from "Star Trek" [28] or those of more abstract personas [49]. Similar techniques have also been used to rewrite text with specific degrees of formality [42] or to obfuscate gender [43]. We have not implemented

style rewriting for the prototype described in Section 3, but such a capability will ultimately be necessary for deployed systems.

3 A Chatbot Prototype

To illustrate some of the capabilities we ultimately envision providing to participants in our interview study, we started by assembling a text collection for one historical figure whom we expected our participants would know sufficiently well. We first describe the materials that we collected, and then how we used those materials to build a single-turn retrieval-based chatbot.

3.1 President Reagan Collection

In the United States, presidential libraries typically co-locate a historical museum built around the legacy of a presidential administration with the National Archives and Records Administration (NARA) staff that manages the records of that administration. Materials from these presidential libraries are naturally of interest to scholars (e.g., [45,34]), but for our present purpose it is the museum function of a presidential library that most interests us. The libraries of more recent presidents hold larger quantities of digital and digitized records, but it is the records of administrations further in the past that are able to provide access to the largest fractions of their holdings. There is thus currently a sweet spot that starts roughly with the Carter administration and continues through roughly the Clinton administration from which substantial digital or digitized materials are now available. Among that window, we chose to work with records from the Reagan administration, the first of the two-term (i.e., 8-year) administrations.

The Ronald Reagan Presidential Library and Museum website[1] contains records of the Reagan Administration from 1981 to 1989, such as speeches and reports, records of the President's daily activities, and donated personal collections. Additionally, there are a number of secondary sources on President Reagan that draw on these materials, including a biography [7] and an annotated collection of Reagan's letters [48] that can be used for research purposes (without public redistribution) under the fair use provisions of U.S. copyright law. We assembled from these Web and published materials a diverse text collection of

[1] `https://www.reaganlibrary.gov/`

Table 1. Statistics of the President Reagan collection.

Document type	Count
Public Papers Documents	8,148
Personal Diary Entries	2,902
Interview Transcripts	248

Table 2. Some questions that could be asked of the President Reagan collection.

Question type	Example
Domestic policy	*What do you think employment levels will be next year?*
International policy	*How will the United States deal with Cuba?*
News event	*What do you think of Doonesbury's critiques of you?*
Personal affairs	*How would you like to be remembered?*

substantial size that can provide different types of insights into President Reagan's thinking and his statements. Table 1 gives statistics of the collection that we assembled.

3.2 A Prototype Single-turn Chatbot

The goal of the chatbot that we created for this study was to retrieve content relevant to user questions for a single conversation turn. Given a user question, the system uses two sequential components, the first of which seeks to find passages (i.e., parts of documents) that contain an answer, and the second of which seeks to extract the best answer from a retrieved passage. Figure 2 illustrates this framework. A user question can be on any aspect of Reagan's experiences or attitudes that are documented in the collection (see Table 2 for examples).

Retrieval. The retrieval component first uses non-neural relevance matching to get an initial set of highly ranked passages, and then it reranks those passages with a neural text matching model [37]. To accommodate the length limitations of the computationally complex reranker, all documents in the collection are first segmented into 300-word passages, which are indexed for efficient term-matching retrieval using ElasticSearch v.6.5.4. Reranking is performed using monoBERT [36], a neural, Transformer-based BERT text matching model [16]. monoBERT is fine-tuned on the MS MARCO passage retrieval task, and therefore re-purposed for pointwise classification of passage relevance. We

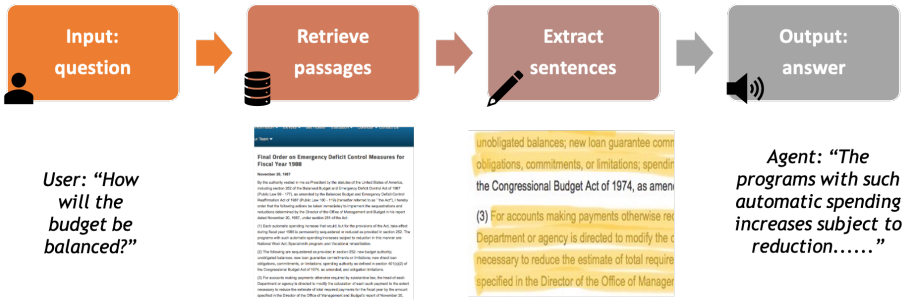

Fig. 2. The chatbot prototype uses two sequential components, retrieve and extract.

use a BERT-base version of monoBERT as the reranker, made available by the original authors.[2] The inputs to the reranker are a concatenation of the user's question with each passage. The reranker cross-encodes the inputs. The output is a matching score for that question-passage pair. Passages that had been ranked highly by ElasticSearch get reranked by the reranker. The scores by the reranker are sorted such that passages most likely to contain an answer are at the top of the resulting ranked list.

Extraction. The second component extracts relevant sentences from retrieved passages. We model our approach on techniques used in the Machine Reading Comprehension (MRC) task in open-domain question answering [13], which seeks to extract relevant spans from a specified text as answers to an input question. While the more general MRC setting allows spans of arbitrary length (but usually just several words) as answers, our current implementation extracts spans at the sentence level, as an answer sentence selection task [56]. Since the participants in our study could examine sentences that precede and follow the selected sentence, we felt that a sentence selection task would suffice to support our study. For an answer sentence selection task, WikiQA [56] provides a suitable data collection to train a model for this component. It includes 3,047 questions from Wikipedia-related user questions sampled from Bing query logs, paired with a total of 29,258 sentences from summary paragraphs in Wikipedia pages. The extraction model to train on this data collection is again a neural, Transformer-based BERT text matching model. We use the Huggingface's transformers library [3] v2.4.1 to fine-tune it as a classification task ourselves, with the default training procedure and hyperparameters. Same as how the retrieval component's neural text matching model gets fine-tuned, the extraction model cross-encodes the user's question with each sentence from a passage, and then predicts a label. At inference time, the model works with the user's question and each sentence from retrieved passages of the Presidential Reagan collection. We use spaCy [27] v2.2.4, specifically its Sentencizer module, to identify the sentences in each passage.

Web Interface. The final component of our prototype is a Web interface that incrementally shows the results of the initial ElasticSearch term-matching passage retrieval, the neural passage reranking, and the answer sentence selection stages. Table 3 summarizes the typical time between issuing the question and the availability of each element of the displayed response. We focus here on effectiveness rather than efficiency; in a deployed system, these components could be optimized for efficient response, with sub-second latency.

[2] `https://github.com/nyu-dl/dl4marco-bert`
[3] `https://huggingface.co/transformers/`

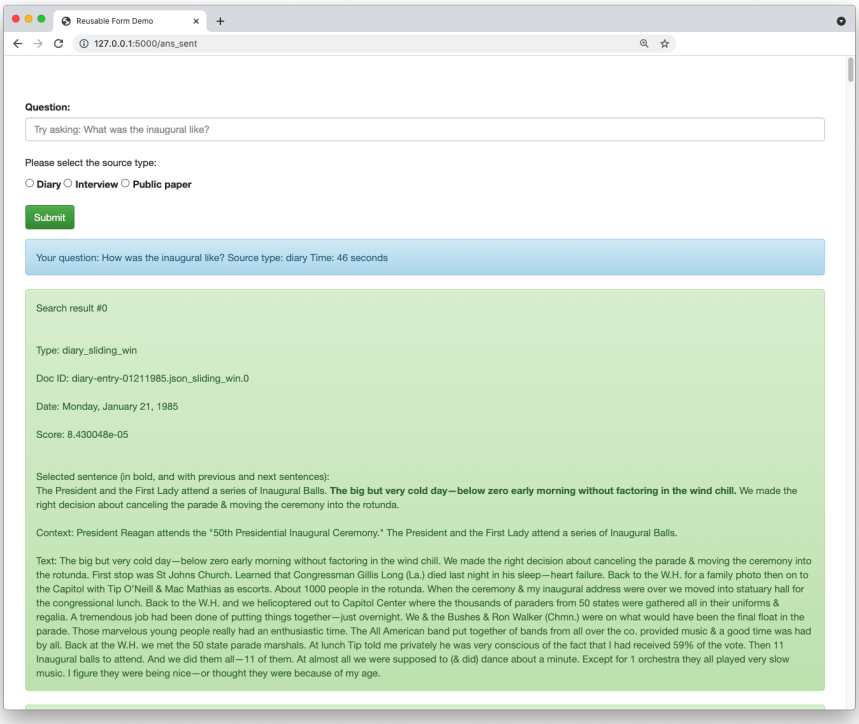

Fig. 3. The prototype Web interface. This screenshot shows the final stage, answer sentence selection, on diaries.

4 Expert Interviews

Historical museum exhibits impart a presentation of history to visitors to fulfill a role of public education, to which selected artifacts from the collection of historical figures contribute. We chose a presidential library as a putative setting for our envisioned system. This setting occupies a middle ground in a continuum

Table 3. Empirical differences between three elements of the prototype.

	Passage retrieval		**Answer sentence selection**
	Non-neural term-matching	**Neural reranking**	
Format	Passage of up to 300 words		Single sentence
Response time	∼0 seconds	10–15 seconds	∼60 seconds

Table 4. Biographical sketch and research focus (rephrased for anonymity) for each participant.

ID	Biographical sketch	Research focus
P1	Female, PhD., English literature	Designing digital systems for trans-media storytelling in the virtual world
P2	Male, J.D.	Public access to archival materials, including presidential library materials
P3	Male, Ph.D., Information Studies; Graduate certificate in Museum Studies	Understanding, access and use of cultural heritage collections
P4	Female, PhD., Museum Anthropology	How technologies could enable heritage institutions to share knowledge with communities

of scenarios for accessing the legacy of historical figures, ranging from formal learning, to everyday entertainment. How can systems of the type we envision system support information access in this setting? The system also sits at an intersection of computer science and museum studies. This position requires a substantial translation between participants from both fields, only after which a shared vision can emerge [51]. To facilitate this translation, we used our prototype as a boundary object to support in-depth, semi-structured, contextual interviews [30,17] with academic experts in libraries, archives, and museums (LAM). Our research questions were:

- **RQ1:** Could the envisioned system, as demonstrated by our prototype, return results about historical figures that LAM experts would perceive as potentially useful?
- **RQ2:** What are the opportunities and challenges for embedding our envisioned system into real information access scenarios, exemplified by that of historical museums?

Institutional Review Board approval was obtained in advance of our study.

4.1 Participants

Recruitment of interview participants was based on convenience sampling from public research universities in the US. We chose four academic experts in libraries, archives, and museums. Participants' ages ranged from 18 to 66. Two are female, while two are male. Three identified as Caucasian, one as Asian. Table 4 describes their backgrounds. Despite the sample's small size, the sample has diversity regarding individual expertise.

4.2 Interview Protocol

Each participant scheduled a one-to-one interview with the first author of this paper through email, and the interview was conducted remotely using Zoom. Interviews were audio- and screen-recorded. Each interview lasted approximately 45 minutes, with no monetary compensation.

The interview began with a tutorial video about the prototype system. The participant then tried the live system, but verbally speaking questions that the interviewer typed on their behalf. The participant and the interviewer then together inspected the results from each stage as they appeared. Participants were asked to think aloud to give their reactions to the results.

After using the prototype, the interviewer and participant began a semi-structured interview conversation. The conversation began with a general question about their impression towards the system: *"What kind of challenges and/or value do you see in a system like this?"* To foster a natural conversational flow, the interviewer often prompted participants to elaborate on their answers from perspectives related to their own expertise. In addition to questions tailored to participants' experience, the interview also included a small set of more general questions that could be asked wherever appropriate, often as follow-ups to capture participants' passing thoughts in earlier responses. For example, *"How do museums engage visitors, of whom a large body only made a single visit, in a longer connection to the museum's subject?"* could be asked if the participant mentioned museum visitors; *"Text answers can be in the form of long-form narratives. What strategies do museums use to efficiently and effectively present them,"* could be asked if the conversation reached a visual aspect of museum exhibitions; or *"Museum visitors receive information from curated contents on display, how do museums decide those contents?"* could be asked if topics related to curation arose. These procedures were documented in an interview guide.

4.3 Analysis Approach

Overall, the analysis approach was based on iterative open-coding. The data for this was four interviews of a total duration of 3 hours and 7 minutes, auto-transcribed. An initial pass of memoing happened during and right after each interview, where the interviewer took margin notes and then created a bullet-point summary on the printed interview guide. The notes include questions answered versus skipped, terms mentioned in participants' responses to describe or pinpoint references, and general observations. Answered interview questions from early interviews partially informed the types of questions answerable and interesting for later interviews, especially for experts with similar research directions (e.g., P3 to P4).

The four interview transcripts were imported to NVIVO v.1.4.1, then each was open-coded in separate sessions and in reverse order of interviewing (i.e., P4 to P1), to identify themes. In these open-coding sessions, the first author of this paper tagged related text spans with initial codes. Operations included creating new codes or grouping spans into existing codes. Some initial codes adopted the exact wording as noted in memos. Grouping into existing codes often involved cross-checks with the code's existing text spans, and slight updates to the code. The author decided span boundaries usually as turn-taking points, or at transition phrases, such as *"and I think [PAUSE]"* or *"if that makes sense"*. Some initial codes from these sessions included *"controversy in archives: ownership"* and *"archives need language (later revised as, cross-lingual) support"*. Each

open-coding session was considered completed when the author perceived both a reasonable coverage of coded spans and an exhaustion of discovering existing or new codes throughout an entire transcript.

A more focused coding session was then conducted on the tagged initial codes. This was an iterative process that aimed towards highest-level implications, engaging with the raw text spans as "data", in the context of each participants' background and interests. It yielded a nested structure from the initial codes, where top-level codes are close to the implications we summarize below in Section 4.6. A more refined sort-through adjusted that structure into a temporal or spatial logic sensible for presentation [8]. For example, P2 described a process for the inquiring users, while P3 and P4 thought through the question of who are the users. The author subsequently made the former code subsidiary to the later code.

The qualitative interviewing methodology requires describing the author's position as the research investigator, i.e., reflecting the researcher as the "instrument" for producing knowledge from personal narratives [8]. While the author doing the coding has expertise in system design and development and has been influenced by views of technologies as socially constructed [39], they lack systematic training in museum studies. A syllabus in museum studies[4] also informed the analysis.

4.4 Replicability

We acknowledge an inherent complexity in our interview study protocol, where the study both uses a computational prototype (an instrument novel in design), and recruits academic experts with specific expertise. Given these considerations, conceptual replication [1] would be appropriate. To support conceptual replication, we describe the prototype configuration (Section 3) and summarize domain experts' backgrounds and demographics (Section 4.1), while keeping confidential actual participant identities pursuant to our IRB-approved study protocol.

4.5 Results from the Prototype Exploration

On average, participants spent 11.3 minutes interacting with the system, and issued 1.5 questions. Participants asked questions impromptu (i.e., from the top of their minds) based on their past knowledge. Participants rarely commented directly about result quality being good, bad, relevant, or irrelevant. Instead, they often dug directly into result contents and spoke about insights. For example, upon seeing results that matched their expectation, P3 showed a contented smile and added that, *"Yeah, so say no to drugs was a popular slogan during the Reagan administration. So it was the Reagan project."* Subsequent conversations were positive towards the robustness of results. Half of the participants proactively connected or compared-and-contrasted the system with mainstream search engines (P1: *"If I were querying Google, I would expect to get many news*

[4] HIST 691: Museum Studies by Dr. Spencer R. Crew from George Mason University.

reports. What I might be able to extract from this system that I could not from Google is searching certain genres or types of information. And the addition of diaries could be very valuable..."; P2: *"Google and others have done smaller-scale projects... It seems that your project as the digitization initiatives scale up, would become much more interesting for researchers, but you have enough."*)

Despite the promising results, one noteworthy instance of poor results came in response to one question by P2. The question involved a named entity with a common English first name. Results mentioned a different person from Reagan's cabinet. P2 attached IR relevance as the meaning of result quality in the protocol, and pinpointed those results as *"false positives"*. The error suggests that the system should adequately handle named entities due to the specificity of social circles in the materials of historical figures.

Overall, participants expressed a generally positive impression towards the system. P1 affirmed the value of a system that *"speaks from primary source information."* P4 associated the system's functionality with a recent, massive full-text search tool—a platform that digitized books.[5] All participants articulated potential use cases for the system, including the system being a display for users to wrap up with summative questions at the end of an exhibition (P1), being an automated desk assistant in a museum library for locating materials in binder books (P2 and P3) and being an interface that indigenous leaders could use to distill cultural values from indigenous archive collections (P4). We unify these cases into themes to be taken as implications, and characterize the target user population throughout Section 4.6.

4.6 Design Implications

Analysis of the interview study identified three design implications.

Implication 1: Creating immersive experiences. Selecting museums as the venue underscores the importance of user experience design. P3 reflected that any possible means of visual design accompanying the textual content would enhance the system's value as a museum exhibit. P3 further suggested incorporating *"pictures, videos, maps, rather than just text,"* implying the multi-sensory requirement for effective interaction.

Some recent literature supplements P3's suggestion and highlights the importance of museums exhibiting continuity of time and place [10,44] to contextualize the history, which we summarize as "in-the-moment." That phrasing differentiates this idea from well-implemented "of-the-moment" displays, such as the Colonial Williamsburg Visitor Center for the eve of the American Revolution. Featuring an "in-the-moment" experience might be appropriately sensible for a museum subject of historical figures. For example, visitors to the US Holocaust Memorial Museum in Washington, DC receive "identification cards" of Holocaust experiencers, to align their visit with experiencers' life narratives in

[5] `https://opentexts.world/about`

sequence. If museum text materials have metadata annotations describing time and place, that could help instantiate several instances of our envisioned system for distribution across the physical museum layout. P1, an expert in digital humanities, directly pictured an "in-the-moment" system competency as if the retrieval system had provided time-travel immersion. Reading from retrieved texts, they reported that the system restored intensive memory, as if bringing them *"right back in a very emotional way"* to their earlier memories of the Reagan administration's silence on the AIDS epidemic.

Implication 2: Considering archival assurance. *Assurance* refers to steps that prove item value or physical existence, and records item information [21]. Reacting to the results of a query on Reagan's Berlin Wall address, P3 critiqued the contents for being potentially unverifiable by the user of such a system since *"no visitors in the museum will be actually in that Berlin Wall address!"*

The diversity of issues that might arise precludes any universal solution. For example, arranging manuscripts into a public-facing collection sometimes involves more sensitive factors than the obvious format conversion operations [22]. Developing a conversational system on top of those public-facing collections without awareness of these factors could be problematic. P4 provided an example of this, noting that indigenous materials could have been illegally confiscated from tribal groups and removed from their original cultural contexts [25], and that *"the records are usually described by predominantly white institutions and white systems and often colonial systems, i.e., systems of power that do not tend to highlight those stories or those historical figures"*.

On a related point, gaps in the records that are available for digital processing could also be a concern. Digitization remains a continuing strategic challenge for archives [35]. Nowadays, public access to archives in a National Archives and Records Administration (NARA) research room still employs a system of notebooks and finding aids, making only limited use of computer databases (P2). Consistently, P4 noted that working with digitized text would be the first challenge, that currently *"very few archival collections are digitized ... at all, let alone transcribed or ... text searchable in any way. "* On the other hand, there has been a significant focus on the processing of personal papers and manuscripts [22], where the duty is not exclusive to archivists. For example, between 1987 to 2018, the Einstein Papers Project[6] released multiple volumes of Einstein's writing and correspondence.

Implication 3: Engaging inquiring museum visitors into community-based scholar roles. Many museums include a library room or reading room that provides access to related materials [23,9]). Visitors who read in these libraries are a specific population, active community members willing to spend additional time for more in-depth inquiry. For convenience, we describe them as "library room" scholars.

[6] `https://www.einstein.caltech.edu/index.html`

Our envisioned system provides a progression point that could entice single-visit visitors to become library room scholars. P1 also imagined visitors to *"come across the system at the end of the exhibition... primed to ask, well-formulated queries based on the knowledge acquired."* P1 further suggested that museum curators might well want to collaborate closely with the system designers to foster just this sort of synergy. As much as the system could provide a successful initial interaction of single-visit visitors with the archival content, a growing sense of curiosity or resonance could move some of the people towards long-term, scholarly work. For example, P2 pointed out that *"having a narrative"* would be beneficial to the younger population who learn history through mass media instead of books with details, who *"frankly are not well versed in history, of what happened in the Cuban Missile Crisis or the assassination of Martin Luther King."*

Connecting this with Implication 2, we could imagine community-based library room scholars reciprocating in archival assurance. P4 acknowledged the system being intriguing for *"genealogists as a huge portion of users, trying to reconstruct their family histories, e.g., Black or Latin community members."* For example, one recent change at the Smithsonian's National Anthropological Archives has been welcoming indigenous researchers to correct metadata and description in archives when they came across errors in the reading room [9].

5 Conclusion and Future Work

In the movie *Field of Dreams*, a voice invites us to imagine that "if you build it, they will come." In our study, we have sought to replace, or at least to augment, that voice with the voice of experience. We have drawn on the insights of experts in libraries, archives, and museums. The insights help us imagine how systems that would allow museum visitors to converse with a representation of an actual historical figure might actually be used, and what issues might arise in the context of that use. To do this, we found it helpful to actually build it, at least in prototype form. We see this intersectionality between the technical and the applied as both useful inspirations for our next steps, and as emblematic of the value of bringing different types of expertise together to capitalize in reflective ways on emerging opportunities.

Records of prominent historical figures provide enduring traces of their life. Seeing a convergence of conversational technologies, we explored the possibility of conversational interactions with historical figures, as represented in their records. We assembled a text collection for Ronald Reagan, and developed a prototype retrieval-based chatbot as a boundary object to support interviews with academic experts. Our study yielded implications addressing immersive experiences, assurance, and supporting a continuum of use that could help some visitors engage more deeply with the content over time,

While this interview study offers preliminary insight into how our envisioned system might be used, there remains a gap between our ultimate vision and our current prototype. Thus, we plan next to focus our technical work on style

rewriting, and then on multi-stage dialog management. We also plan to focus on the ethical concerns that arise from systems posthumously representing historical figures; for this we hope to conduct a similarly structured interview study with ethicists. Looking further ahead, we might also explore other applications of this technology, such as interactive textbooks [29] that might embed our envisioned system in a sidebar allows spontaneous questions scaffolded from course materials. And perhaps one day, we might be able to bring historical figures to a Reddit "Ask Me Anything" session. If we are able to build *that*, then perhaps they will come.

Acknowledgements

The authors are grateful to the anonymous participants in our research study, who contributed their time and expertise to this study. This research has been supported in part by the National Science Foundation under grant number 1618695. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Aguinis, H., Solarino, A.M.: Transparency and replicability in qualitative research: The case of interviews with elite informants. Strategic Management Journal **40**(8), 1291–1315 (2019)
2. Artstein, R., Gainer, A., Georgila, K., Leuski, A., Shapiro, A., Traum, D.: New Dimensions in Testimony Demonstration. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. pp. 32–36 (2016)
3. Artstein, R., Traum, D., Alexander, O., Leuski, A., Jones, A., Georgila, K., Debevec, P., Swartout, W., Maio, H., Smith, S.: Time-offset Interaction with a Holocaust Survivor. In: Proceedings of the 19th International Conference on Intelligent User Interfaces. pp. 163–168 (2014)
4. Banerjee, S., Biyani, P., Tsioutsiouliklis, K.: Transforming Chatbot Responses to Mimic Domain-specific Linguistic Styles. In: Second Workshop on Chatbots and Conversational Agent Technologies (2016)
5. Bell, G., Gray, J.: Digital Immortality. Commun. ACM **44**(3), 28–31 (Mar 2001)
6. Bell, L., Gustafson, J.: Utterance Types in the August Dialogues. In: ESCA Tutorial and Research Workshop (ETRW) on Interactive Dialogue in Multi-Modal Systems (1999)
7. Brands, H.W.: Reagan: The Life. Anchor (2016)
8. Brinkmann, S.: Qualitative Interviewing. Oxford University Press (2013)
9. Buchanan, R., George, K., Gibson, T., Hung, E., Labinsky, D., Marsh, D.E., Menyuk, R., Norton-Wisla, L., Ortego-Chiolero, S., Sowry, N., et al.: Toward Inclusive Reading Rooms: Recommendations for decolonizing practices and welcoming indigenous researchers (2021)
10. Cafaro, F., Ress, S.A.: Time Travelers: Mapping museum visitors across time and space. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16): Adjunct. p. 1492–1497 (2016)

11. Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J.: Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In: Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology. pp. 7–10 (1998)
12. Chabot, L.: Nixon Library Technology Lets Visitors 'Interview' Him (1990), https://www.latimes.com/archives/la-xpm-1990-07-21-mn-346-story.html, [Online; accessed 12-April-2021]
13. Chen, D., Fisch, A., Weston, J., Bordes, A.: Reading Wikipedia to Answer Open-domain Questions. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1870–1879 (2017)
14. Cole, C.: Inducing Expertise in History Doctoral Students via Information Retrieval Design. The Library Quarterly: Information, Community, Policy **70**(1), 86–109 (2000)
15. Dalton, J., Xiong, C., Callan, J.: TREC CAsT 2019: The conversational assistance track overview (2020)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
17. Dixon, E., Lazar, A.: Approach Matters: Linking practitioner approaches to technology design for people with dementia. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2020)
18. Duff, W.M., Johnson, C.A.: Accidentally Found on Purpose: Information-seeking behavior of historians in archives. The Library Quarterly: Information, Community, Policy **72**(4), 472–496 (2002)
19. Fidel, R., Davies, R.K., Douglass, M.H., Holder, J.K., Hopkins, C.J., Kushner, E.J., Miyagishima, B.K., Toney, C.D.: A Visit to the Information Mall: Web searching behavior of high school students. Journal of the American society for Information Science **50**(1), 24–37 (1999)
20. Fu, Z., Tan, X., Peng, N., Zhao, D., Yan, R.: Style Transfer in Text: Exploration and evaluation. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (Apr 2018)
21. Glazer, A.S.: Auditing Museum Collections (1991), http://archives.cpajournal.com/old/11287210.htm, [Online; accessed 12-April-2021]
22. Greene, M., Meissner, D.: More Product, Less Process: Revamping traditional archival processing. The American Archivist **68**(2), 208–263 (2005)
23. Griffin, C.: The Museum Library. University Museum Bulletin **19**(2), 23–27 (1955)
24. Gustafson, J., Lindberg, N., Lundeberg, M.: The August Spoken Dialogue System. In: Sixth European Conference on Speech Communication and Technology (1999)
25. Harth, M.L.: Learning from Museums with Indigenous Collections: Beyond repatriation. Curator: The museum journal **42**(4), 274–284 (1999)
26. Hjørland, B.: Epistemology and the socio-cognitive Perspective in Information Science. Journal of the American Society for Information Science and Technology **53**(4), 257–270 (2002)
27. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020), https://doi.org/10.5281/zenodo.1212303
28. Jena, G., Vashisht, M., Basu, A., Ungar, L., Sedoc, J.: Enterprise to Computer: Star Trek chatbot. arXiv preprint arXiv:1708.00818 (2017)
29. Koike, H., Sato, Y., Kobayashi, Y., Tobita, H., Kobayashi, M.: Interactive Textbook and Interactive Venn Diagram: Natural and intuitive interfaces on augmented

desk system. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '00). p. 121–128 (2000)
30. Kvale, S.: Doing Interviews. Sage (2008)
31. Leuski, A., Pair, J., Traum, D., McNerney, P.J., Georgiou, P., Patel, R.: How to Talk to a Hologram. In: Proceedings of the 11th International Conference on Intelligent User Interfaces. pp. 360–362 (2006)
32. Lin, J., Nogueira, R., Yates, A.: Pretrained Transformers for Text Ranking: BERT and beyond (2020)
33. Marinelli, D., Stevens, S.: Synthetic Interviews: The art of creating a 'dyad' between humans and machine-based characters. In: Proceedings 1998 IEEE 4th Workshop Interactive Voice Technology for Telecommunications Applications (IVTTA'98). pp. 43–48 (1998)
34. McGranahan, C.: A Presidential Archive of Lies: Racism, twitter, and a history of the present. International Journal of communication **13**, 19 (2019)
35. National Archives: Strategy for digitizing archival materials (2014), `https://www.archives.gov/digitization/strategy.html`, [Online; accessed 12-April-2021]
36. Nogueira, R., Cho, K.: Passage Re-ranking with BERT. arXiv preprint arXiv:1901.04085 (2019)
37. Nogueira, R., Yang, W., Cho, K., Lin, J.: Multi-stage Document Ranking with BERT. arXiv preprint arXiv:1910.14424 (2019)
38. Parkin, S.: Back-up Brains: The era of digital immortality (2015), `https://www.bbc.com/future/article/20150122-the-secret-to-immortality`, [Online; accessed 12-April-2021]
39. Pinch, T.J., Bijker, W.E.: The Social Construction of Facts and Artefacts: Or how the sociology of science and the sociology of technology might benefit each other. Social Studies of Science **14**(3), 399–441 (1984)
40. Prabha, C., Connaway, L.S., Olszewski, L., Jenkins, L.R.: What is Enough? Satisficing information needs. Journal of Documentation (2007)
41. Radlinski, F., Craswell, N.: A Theoretical Framework for Conversational Search. In: Proceedings of the 2017 conference on Computer Human Information Interaction and Retrieval. pp. 117–126 (2017)
42. Rao, S., Tetreault, J.: Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, benchmarks and metrics for formality style transfer. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 129–140. New Orleans, Louisiana (Jun 2018)
43. Reddy, S., Knight, K.: Obfuscating Gender in Social Media Writing. In: Proceedings of the First Workshop on NLP and Computational Social Science. pp. 17–26. Association for Computational Linguistics, Austin, Texas (Nov 2016)
44. Ress, S., Cafaro, F., Bora, D., Prasad, D., Soundarajan, D.: Mapping History: Orienting museum visitors across time and space. Journal on Computing and Cultural Herit. **11**(3) (Aug 2018)
45. Riley, R.L.: Presidential Oral History: The Clinton presidential history project. The Oral History Review **34**(2), 81–106 (2007)
46. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and beyond. Foundations and Trends in Information Retrieval **3**(4), 333–389 (Apr 2009)
47. Savin-Baden, M., Burden, D.: Digital Immortality and Virtual Humans. Postdigital Science and Education **1**(1), 87–103 (2019)
48. Skinner, K.K., Anderson, A., Anderson, M.: Reagan: A life in letters. Simon and Schuster (2004)

49. Song, H., Wang, Y., Zhang, W.N., Liu, X., Liu, T.: Generate, Delete and Rewrite: A three-stage framework for improving persona consistency of dialogue generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5821–5831. Online (Jul 2020)
50. Spice, B., Sloss, E.: Carnegie Mellon's Synthetic Interview Technology Enables Virtual Chats with Darwin's Ghost (2009), `https://www.cmu.edu/news/archive/2009/February/feb5_darwininterview.shtml`, [Online; accessed 23-February-2021]
51. Star, S.L., Griesemer, J.R.: Institutional Ecology, Translations' and Boundary Objects: Amateurs and professionals in berkeley's museum of vertebrate zoology, 1907-39. Social Studies of Science **19**(3), 387–420 (1989)
52. Traum, D., Jones, A., Hays, K., Maio, H., Alexander, O., Artstein, R., Debevec, P., Gainer, A., Georgila, K., Haase, K., Jungblut, K., Leuski, A., Smith, S., Swartout, W.: New Dimensions in Testimony: Digitally preserving a holocaust survivor's interactive storytelling. In: Schoenau-Fog, H., Bruni, L.E., Louchart, S., Baceviciute, S. (eds.) Interactive Storytelling. pp. 269–281. Springer International Publishing, Cham (2015)
53. Voorhees, E.M., et al.: The TREC-8 Question Answering Track Report. In: TREC. vol. 99, pp. 77–82 (1999)
54. Wang, H., Lu, Z., Li, H., Chen, E.: A Dataset for Research on Short-text Conversations. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 935–945. Association for Computational Linguistics, Seattle, Washington, USA (Oct 2013)
55. Wu, Y., Wu, W., Xing, C., Zhou, M., Li, Z.: Sequential Matching Network: A new architecture for multi-turn response selection in retrieval-based chatbots. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 496–505. Vancouver, Canada (Jul 2017)
56. Yang, Y., Yih, W.t., Meek, C.: WikiQA: A challenge dataset for open-domain question answering. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2013–2018 (2015)