3D virtual worlds as search, discovery and retrieval engines

Rob Warren
Big Data Institute, Dalhousie University
Halifax, Canada
rhwarren@dal.ca

David Evans
Department of Computing and Mathematics
University of Derby, UK
d.f.evans@derby.ac.uk

ABSTRACT

The search for "relevant" information has long been driven by keyword searches and some basic visualizations. The increase in both the amounts of data available and in the breath of data that is not a discrete text document is creating new opportunities for non-traditional means of information retrieval (IR). In this paper, we present a prototype system where a 3D virtual world is used to access and discover information in semantic web databases. A prototype that focuses on the period of the Great War is discussed.

1. INTRODUCTION

Information retrieval research has traditionally focused on keyword searches of text documents represented using the "bags of words" model. Other models have evolved, including searching by image similarity, audio search, recommender systems and most recently faceted search. All are vibrant areas of research that try to enhance people's ability to find the information they are looking for while lowering the cost of doing so. The mass digitization projects of historical archives, born-digital information, low cost storage and computing power are challenging these methods to do better with a problem space that is increasing in volume and complexity.

A number of current successes in IR are made possible because the methods depend on implicit structures: a) Their underlying mechanics were tied to a specific use case or content type, b) their operation was tied to a specific collection or c) they were tied to a specific culture and/or language. Taking the specific case of a library, what happens when every digital "book" within the library becomes linked into every critical edition every written and the word-net ontology? Where the book authorship and editorship is just a subsection of a loosely distributed social networking platform? How do we deal with data that does not fit the normal definition of a document? For example, statistical data sets have traditionally been treated like a black box as only an entry in the catalogue. How can we reference and make available the contents of the data set as yet another set of documents?

We present here a 3D virtual simulation that is an early prototype in filling this need by representing the information known about a particular point in space and time as a virtual world. We assume the use of semantic web (Linked Open Data, [2]) technologies to access and interact with the data. While the use of this technology that unifies both data and meta-data representation makes the proposed approach possible, we fully realize that its widespread adoption may never occur. The methods proposed here are generic and can be reimplemented with other data management frameworks if needed.

To the best of our knowledge the use of Linked Open Data to automate the online creation of data-driven procedurally generated 3D simulations of historical and current locations is a novel contribution. Previous contributions were tools supporting the designer in creating environments that would then be statically used within a game. Here the virtual world is regenerated each time the game resets using the information as available at start-up time.

2. PREVIOUS WORK

A full literature review would be too lengthy but we briefly touch upon a few areas of interest. This work builds upon the early work done on the Semantic Virtual Environment by Karsten Otto [11], Grimaldo et al. [7] and Kallman et al. [9] on encoding the relationships between objects and the environment using semantic web technologies. Early text-based games encouraged players to explore their virtual world, and tools such as the Z-machine [3] pushed game designers into creating procedurally-generated content (though typically in response to player actions only). This was continued in early online games such as Multi-User Dungeons (MUDs), with the incorporation of simple external data such as the time of day [4]. More recently, Togelius et al. [1, 13] and Hartsook [8] both worked on procedural game generation on a more complex scale. Our work merges these areas into a unified methodology that procedurally generates a 3D simulated world on an online (runtime generation) basis based on information gathered within libraries and archives.

3. PROBLEM DESCRIPTION

In our work our objective was to study how effectively a simulation, displayed as an automatically generated 3D virtual world, could be used to represent the data within digital libraries and archives. We thus built a 3D simulated world prototype with which to conduct initial experiments.

Search is Over '14 London, UK
.

Traditional electronic information seeking and discovery is carried out over a catalogue that points to data objects such as books. Gross relevancy is insured through the selection of a relevant collection, fond, library or archive. Following interaction with this catalogue, a separate process deals with the data objects themselves.

The current wave of content stores being merged with catalogue information is generating a new opportunity in locating micro-documents, such as a specific line item in an attendance ledger of a 1904 temperance hall association, the bill of materials to build a locomotive with, a candlestick used as a murder weapon in a "whodunnit" or even a large collection of picture annotations in an online museum. The use and retrieval of these micro-documents is difficult without a customized application owing to the extremely pointed context of its use. Furthermore, the ad-hoc nature of context at this level of granularity makes it unlikely that a single catalogue holds all the necessary information. Thus, a user-centered application is needed that can interact with multiple data sources to answer the user's need.

We propose making use of a virtual 3D world simulations as a means of searching and retrieving information, basic relevance being conveyed naturally through physical and spatial proximity. The underlying rationale is that while a certain amount of training is required to use any non-trivial retrieval engine, a 3D simulation is intuitive to interact with and the majority of people have previously played a computer game.

The 3D virtual world becomes a mediator or agent to help the user interpret the data. This interpretive function has previously been taken up by authors, editors and designers of games whom would have been sensitive to the needs of the intended audience. An architect might focus on stylistic elements, an art historian on the mouldings of a building while an engineer might be interested in its structure. The general public might wish for an aesthetic and entertaining simulation that would clash with the evaluation criteria of an expert. We note that some documentary makers for example "intend that the audience take the relevant film scenes as reliable representations of some element of the actual world from which true beliefs can be formed about the film's subject" [12]. It is in this same spirit that we suggest using 3D computer generated representations of the data.

There has been some resistance in using such simulations as information retrieval in that some professionals wish to see the raw data for themselves along with a natural reluctance to entrust a machine with these decisions. This desire, while commendable, is becoming infeasible to satisfy. We are at the point where the rate at which new information is being generated in some scholarly fields is greater that the domain expert's capacity to consume it. At present, web browsers and search engines silently negotiate access to information, making "raw data" increasingly inaccessible. The simulations we advocate are merely the next step in this, though this time at the semantic level.

In implementing our prototype we opted to define realism not through the visual accuracy of the rendering but in terms of the veracity of the data presented. In other words, the events, places and things that are present within the virtual world must be based on traceable facts. No metaphor is used for data items; the world represents things using generic 3D assets for objects.

Each trench, tree and discrete element in the simulation can be clicked on to obtain the source document that defined the existence of that object. Even when some elements such as foliage and fauna are estimated through models, as done by by Dussel et al. [5], the model parameters can be properly documented. We therefore incorporate data provenance into the virtual world, making verification and traceability of information a natural part of exploration and discovery.

4. PROTOTYPE

Our prototype is implemented in the popular Unity game engine and represents a section of the Vimy Ridge trench system during the Great War. The Library of Congress subject headings list over 100,000 books on the great war and a tremendous amount of archival material is being made available on an unprecedented scale. Combined with Linked Open Data sets on the Great War such as the Muninn Project[1], we used this as our test case.

The simulation is driven by a location and a date. We now review three elements of the simulation: its use as an information retrieval method, its use as an information discovery mechanism and the opportunities that Linked Open Data presents in documenting information used within projects.

4.1 3D Virtual World as Information Retrieval Engines

Time and space are two of the key components of how people think about events and topics. For our prototype, we selected the site of Vimy Ridge in France in early 1914. This data was retrieved from multiple Linked Open Data projects including Muninn WWI, DBPedia and the French and German National Libraries.

Figure 1: The main simulation window with time and location (A), the scanned period map (B) is being used both as a virtual "prop" map (and trench geometries) and (C) a live mini-map tracks the users movement in part of the virtual world.

Figure 1 is a screen shot of the prototype where the user is walking through the German trenches of the Vimy area during the Great War. The simulation allows the user to navigate the environment of Trench warfare on the Western Front: the protective trenches are higher that the person can see, the environment claustrophobic and navigation is difficult. The sun provides only a rudimentary sense of direction and time as it moves in the small area of the sky visible to the user.

[1]http://www.muninn-project.org/

The trench geometries are extracted from a trench map depicting this location and this specific time while the image of the trench map itself becomes a map prop within the simulation.

What makes this simulation interesting for information retrieval purposes is that it allows answering poorly defined questions that would otherwise be hard to answer, such as "Who has the advantage of terrain at this stage of the war?". We also note that while we focus here on the Great War, this same method can be used for any past or present location with the necessary data.

4.2 3D Virtual Worlds as Discovery Engines

Information discovery can be hampered in its search for relevant documents by a lack of a discovery mechanisms or finding aids within the collection or document. A book on botany may be relevant to camouflage techniques during the Great War but without the logical linkages the machine will never know about this relevancy. Similarly, the push to digitize archives on a massive scale means that a large number of documents that have never been individually accessed are now available online. Given that the descriptive data needed to access these document is in many case missing, there is a large chance that important documents are hiding in plain sight.

Figure 2: The use of litter within virtual environments is an opportunity to reuse random archival material. Here, an image of the French newspaper Figaro, dated two weeks earlier than the environment chronology, is used.

A technique that is used often within virtual environments to add an aesthetic element of visual realism is the dispersal of litter such as rocks, notices posted on buildings or discarded newspapers or trash. Usually, these elements are added as details by designers manually using stock materials and objects. Because of the availability of online documents, we can now automate the retrieval of documents that can serve as litter or props. We can ensure contextual relevancy through the selection of the right collection without concerning ourselves with query relevance. This is acceptable because the user will cognitively ignore it as trash to begin with while allowing for the possibility that it might be of interest to the user.

Figure 2 is a screenshot of our prototype where we generate discarded newspapers within the Entente trenches. Through basic Linked Open Data interfaces, we can ensure that we retrieve an outdated newspaper from the date of the user query which ensures that it represents old news. While the discarded paper adds aesthetic value to the scene, it also serves a secondary role of intentionally crowd sourcing the review of the entire collection. Should any document be of obvious importance, it would be located over time.

A study at Cornell University reported that 61% of books circulated at least once from 1990 to 2010 [6] which means that in 10 years about 39% of the collection was not reviewed. Since this number is likely to be higher in archival resources where the entire collection is not fully indexed, we propose that this use is a good means of ensuring that all data within the collection is "seen" at least once for items that may be of high value.

For cost reasons, archives and libraries tend to digitize content at a large scale without reviewing the contents of the data. A rare and valuable manuscript may have gone unnoticed or a lost letter may be identified. Over time, it can be ensured that the content is reviewed by many different persons with different backgrounds for something of value.

4.3 Giving Credit for Data Use

The access of data-sets through a 3D environment is a process that can appear to be opaque even with clickable provenance links: there exists so much information that is being queried at different levels of granularity that some of it will be unreachable by the user.

We have experimented with the automated creation of "credits" at the end of the simulation that scroll through contributions in a manner not unlike movie credits. We can also easily create a bibliography from the queried Linked Open Data that can be used to suggest further reading, but the rolling credits give us an additional mechanism to report on the relative importance of the information used in the simulation.

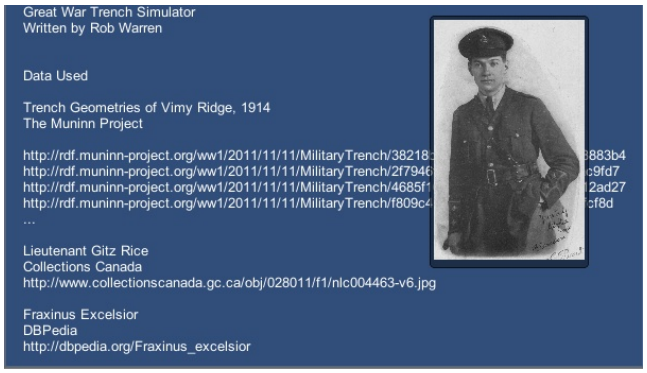

Figure 3: Credits roll at the end of the 3D simulation.

Publication mastheads, television and movie credits will report contributors in an order that is intended to communicate the importance of the contribution or a contractual obligation. Due to the dynamic nature of an online 3D simulation, the credits will change every time the simulation is run because the content will have changed. We therefore can use a mixture of text grouping, ordering and size to communicate the data sources used to create the simulation.

For aesthetic purposes, we place the downloaded images within the credits at random locations and play whichever musical score was retrieved within the simulation in the background. The screenshot in Figure 3 is a sample of how credits are currently rendered. We opted on a simple method where the data sources are ordered according to the amount

of information (triple count) used from the data source.

5. DATA CONSUMPTION ISSUES

The creation of this initial prototype has highlighted some issues with current practices in describing data within online collections. We have discovered four issues that become acute as a consequence (direct or indirect) of the collection's identity, and that of its catalogue, no longer being used a priori to decide relevance.

The first is that descriptive vocabularies are still being used with the expectation that they will be processed by human beings instead of machines. This is partly a result of "mapping" approaches to converting data where text is being inserted into new descriptive vocabularies without necessarily using the correct identifiers. A typical example is the use of Dublin Core[2] vocabularies with text strings, which are not appropriate when dealing with large multi-national datasets. Effectively, a search query must account for every language (<dc:type>image</dc:type> ∪ <dc:type>bild</dc:type> ∪ ... <dc:type>изображение</dc:type>) in order to have a appropriate coverage which is not practical or efficient.

The second issue is that there is still a reliance on generalizable specifications when a specialized, specific vocabulary is necessary. The case of media recording is a poignant example of under-specification. Consider the festive holiday song "O holy night". While the song itself is culturally French, there are translations of the lyrics in at least 5 languages. It is sometimes impossible for software to determine whether a particular library is listing the song as being culturally French, whether the song is being sung in French, whether the entry refers to the recording or media (CD) published in France or whether the interpreting artist is of French nationality.

Third, the distribution of library and archives information is often done through a data dump or an API unique to the library. Given that at any given time a small percentage of the catalogue is of interest to the client, data dumps are not ideal as they must be transferred in order to keep results up to date. The Extraction, Transformation, Load (ETL) costs soon become unmanageable when multiple clients are attempting to update data-sets in the multi-terabyte range. Similarly, the creation of ad-hoc APIs is counter-productive since the client must implement a new interface component for each new data-source. In our implementation we made use of SPARQL endpoints to query the data at the source and hence lower data transfer amounts.

6. CONCLUSION

In this paper we have presented a prototype 3D virtual world that functions as an information retrieval tool on Linked Open Data sets. While the prototype is basic, we have demonstrated how a simple document can function as both a data source in itself as well as a visual asset within the environment. We also demonstrated that the use of Linked Open Data makes it possible to document and present the source data in a detailed fashion, including data provenance, even when it is not a traditional publication.

Evaluation remains a problem: the performance of an information retrieval systems is typically measured over a limited collection (corpus) of documents in order to normalize results using a predetermined performance metric. Creating an evaluation framework for these types of tools will be an important step in their widespread adoption.

In future work we will focus on generating scenarios or narratives about a certain topic or event from data contained within online archive and libraries. Questions of historical accuracy still need to be resolved, but we also look forward to studying how lessons can be created automatically in terms of setting the learning objectives of a game or simulation.

[2]http://purl.org/dc/elements/1.1/

7. REFERENCES

[1] E. Anderson, L. McLoughlin, and et al. Developing serious games for cultural heritage: a state-of-the-art review. Virtual Reality, 14(4):255–275, 2010.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. International Journal on Semantic Web and Information Systems, 5(3):1–22, 2009.

[3] K. Bracey and J. C. Penney. The Z-Machine standards document version 1.1. http://inform-fiction.org/zmachine/standards/z1point1/index.html, retrieved 2014-07-21, February 2014.

[4] P. Curtis. Mudding: social phenomena in text-based virtual realities. In P. Ludlow, editor, High Noon on the Electronic Frontier: Conceptual Issues in Cyberspace, pages 347–373. MIT Press, Cambridge, MA, 1996.

[5] O. Deussen, P. Hanrahan, B. Lintermann, R. Měch, M. Pharr, and P. Prusinkiewicz. Realistic modeling and rendering of plant ecosystems. In 25th Annual Conference on Computer Graphics and Interactive Techniques, pages 275–286, 1998.

[6] R. Entlich, G. Green, P. Hirtle, S. Rockey, D. Schnedeker, P. Stevens, K. Tancheva, K. Walker, and J. Alberts. Report of the collection development executive committee task force on print collection usage cornell university library. Technical report, Cornell University Library, October 2010.

[7] F. Grimaldo, F. Barber, and et al. Semantic virtual environments for interactive planning agents. In International Digital Games Conference, volume 17, 2006.

[8] K. Hartsook and et al. Toward supporting stories with procedurally generated game worlds. In Computational Intelligence and Games, pages 297–304, 2011.

[9] M. Kallmann and D. Thalmann. Modeling objects for interaction tasks. In Proc. Eurographics Workshop on Animation and Simulation, pages 73–86, 1998.

[10] P. Lebling and M. Blank. Zork: A computerized fantasy simulation game. IEEE Computer, 12(4):52–59, April 1979.

[11] K. A. Otto. The semantics of multi-user virtual environments. In Workshop Towards Semantic Virtual Environments, pages 35–39, 2005.

[12] C. Plantinga. What a documentary is, after all. The Journal of Aesthetics and Art Criticism, 63(2):105–117, 2005.

[13] J. Togelius, M. Preuss, and et al. Towards multiobjective procedural map generation. In Workshop on Procedural Content Generation in Games, page 3, 2010.