<http://contextus.net/r+j/char/5>, <http://contextus.net/r+j/char/5>, Wherefore Art Thou <http://contextus.net/r+j/char/5>? – Crowdsourcing Linked Data From Shakespeare To Dr Who.

Dr K Faith Lawrence
Contextus.net
kf03r@ecs.soton.ac.uk

ABSTRACT
In this paper we present ongoing research into the development of a linked data repository of cultural narratives. Drawing from both modern and historical sources we expand on earlier work automatically generating RDF from TEI performance texts. Building on this foundation we employ a crowdsourcing methodology to correct errors in the descriptions and enhance the data with information which cannot be drawn from the text. We present this approach as a way to facilitate correction through the 'many hands' approach while also enabling us to analyse, explore and highlight the diversity of interpretation that exists for a text. In doing so we present an open resource aimed at supporting and inspiring new humanities, social science and computer science scholarship.

Categories and Subject Descriptors
H.3.1 [**Content Analysis And Indexing**]: Abstracting Methods

General Terms
Design, Experimentation, Human Factors, Theory.

Keywords
Linked Data, Digital Narrative, Crowdsourcing, Semantic Web, Ontology, TEI, RDF.

1. INTRODUCTION
The OntoMedia ontology was developed to describe the narrative components of media objects and the relationships between those components (Lawrence, 2007, Part IV). In this paper I lay out the way in which the ontology is being applied to the description of literary and modern sources with the intention of creating a repository of linked cultural narratives.

2. AUTOMATED DATA CREATION WITH TEI
Jewell (2010) developed a system to take TEI-encoded performance texts and generate RDF descriptions of the character interactions and the locations of those interactions. Although the experiment was a success, limitations were noted due to the minimal nature of the information stored in the encoding being used. Expanding on this initial experiment, we selected a cross-section of TEI-encoded Shakespeare plays from the Perseus Digital Library[1] and developed the RDF-generation code to recognise additional TEI attributes, most noticeably stage instructions related to the entrance and exit of characters (Lawrence, 2010). An event browser was developed to allow exploration of the events described within the RDF data by users without expert technical knowledge (See Figure 1).

Due to the multiple ways in which valid TEI documents can be created further adaptions will need to be implemented as more texts are prepared for processing. Currently under investigation are a selection of TEI-encoded plays by Oscar Wilde stored in the CELT[2] archive and, expanding into other markup languages, a selection of HTML transcripts from Doctor Who episodes (classic[3] and new[4]). It is hoped that this approach will prove to be applicable to many types of digital narrative, real and fictional, including hyper-textual, intertextual and re-mixed sources as well as media types beyond our initial selection which was limited to text.

3. CROWD-CORRECTION AND EXPANSION OF DATA
The production of metadata is potentially a time and resource intensive part of any linked data project. Automating this process can speed up creation. However, as described in the previous section, there are limitations as to what can be extracted from the encoded text. While sophisticated textual processing such as NLP may offer enhanced results via improved named entity recognition and anaphora resolution this reduces rather than solves the problem of accuracy. Especially in the case of fiction, with its many complexities and occasionally deliberate misinformation and ambiguity, some form of error checking is vital.

Beyond the issue of accuracy of the RDF generation process, the textual transcription of a performance, be it a play or a multimedia recording, may not contain all the desired information. Ad-libs, actor descriptions, visual effects or simple inaccuracies in the transcript mean that the resulting RDF description must be regarded as useful but fundamentally suspect.

WebSci '11, June 14-17, 2011, Koblenz, Germany.

[1] http://www.perseus.tufts.edu/hopper/

[2] http://www.ucc.ie/celt/

[3] http://homepages.bw.edu/~jcurtis/Scripts/scripts_project.htm#THE%20SCRIPTS

[4] http://who-transcripts.atspace.com/

The traditional literary approach would be to enlist experts to check the results or to create a canonical interpretation against which the results could be compared. In recent years there has been a rise in the use of crowdsourcing to distribute the workload and draw on the 'wisdom of the crowd'. To make this possible it was vital to create a system to allow non-technical users to be able to interact with and edit the data held within the triplestore. This interface (see Figure 2) is nearing completion and we hope to start collecting data shortly. In this we are also taking inspiration from the significant amount of effort and collective expertise that goes into online resources by media and literary fans.

4. QUESTIONS AND INTERPRETATION – MY VERSION VS YOUR VERSION

One criticism that can be leveled at crowdsourced data is that it represents opinion rather than an authoritative deliberation. Our contention is that in the case of narrative this represents an advantage because by gathering multiple interpretations we are can not only find the consensus values for the corrected data but also analyse the audience reaction to a given story and identify the points of contention and agreement. By offering contributing users the opportunity to create their own profile we open the way to begin investigating the demographics of groups that share similar interpretations of events or have shared interests based on editing patterns.

In addition to providing an interface for users to browse and query the repository, open access to the data will be made available via a SPARQL endpoint. Queries can be directed at a single graph (representing a single interpretation of a single text), across multiple graphs or across graphs generated to represent specific aggregated views of the text. In this way we intend to present a resource that offers a searchable reference for media and literary researchers.

5. CONCLUSION

In this paper we present our narrative repository and related methodology for seeding the repository with automatically generated RDF before opening the data up for correction, exploration and analysis. By taking this step we argue that we are not only creating the opportunity to harness the knowledge that exists outside the traditional accredited domain but also to engage with users of the repository and expose the multiple ways in which texts can be understood. Through this deliberate building of multiversal rather than universal interpretation we open the way for new investigation of the conceptual elements and relationships within and between narratives both in themselves and in the context of audience reception.

6. REFERENCES

[1] Jewell, M. O. 'Semantic Screenplays: Preparing TEI for Linked Data'. In Digital Humanities 2010, June 2010. Paper presented as part of panel: Scanning Between the Lines: The Search for the Semantic Story.

[2] Lawrence, K. F. 'The Web of Community Trust - Amateur Fiction Online: A Case Study in Community Focused Design for the Semantic Web'. PhD thesis, Electronics and Computer Science, University of Southampton, 2007.

Figure 1: Narrative Browser

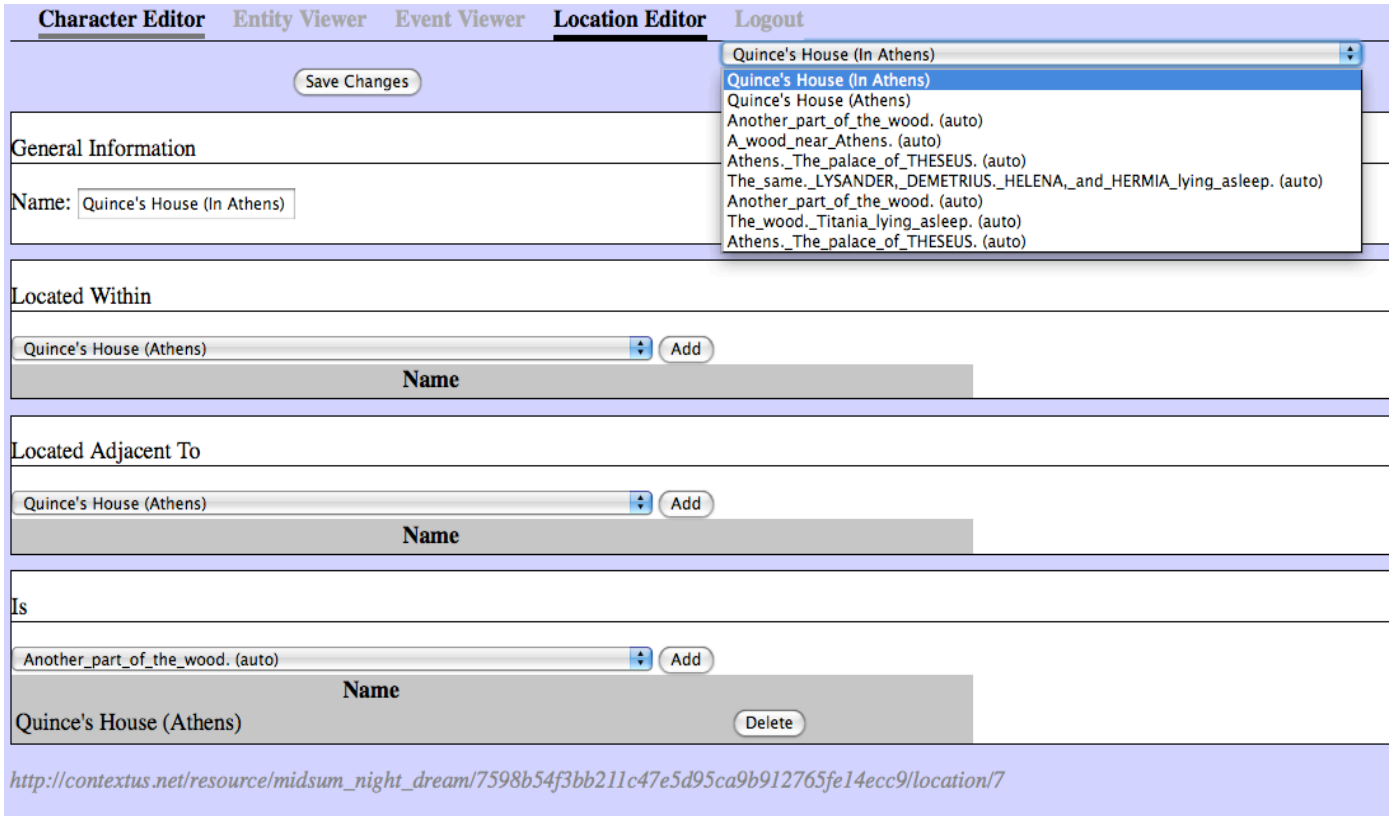

Figure 2: Editing and Expanding Location Data