underrepresented in most design decisions of modern online editions which usually see the publication as the natural goal of their project, a publication which usually only allows for reading and searching as the typical forms of text usage.

Our paper will describe the baseline encoding format for some of the text types supported by Textgrid at the moment including the metadata format and discuss in what ways the three requirements are met by them.

One of the aims of our paper is to put our arguments and design decisions up for discussion in order to test their validity. Another aim is to reflect on the consequences of this approach for others like the TEI, especially the idea to define important text types for the humanities and provide specific markup for them.

Breaking down barriers: the integration of research data, notes and referencing in a Web 2.0 academic framework

Ian R. Johnson
johnson@oci.arts.usyd.edu.au
University of Sydney, Australia

In this paper I argue that the arbitrary distinction between bibliographic data and research data – which we see in the existence of specialised library catalogues and bibliographic systems on the one hand, and a multitude of ad hoc notes, digitized sources, research databases and repositories on the other – is a hangover from a simpler past, in which publication and bibliographic referencing was a well-defined and separate part of the research cycle.

Today published material takes many different forms, from books to multimedia, digital artworks and performances. Research data results from collection and encoding of information in museums, archives, libraries and fieldwork, as well as output from analysis and interpretation. As new forms of digital publication appear, the boundary between published material and research data blurs. Given the right enabling structures (eg. peer review) and tools (eg. collaborative editing), a simple digitized dataset can become as valuable as any formal publication through the accretion of scholarship. When someone publishes their academic musings in a personal research blog, is it analogous to written notes on their desk (or desktop) or is it grey literature or vanity publishing?

The drawing of a distinction between bibliographic references and other forms of research data, and the storing of these data in distinct systems, hinders construction of the linkages between information which lie at the core of Humanities research. Why on earth would we want to keep our bibliographic references separate from our notes and developing ideas, or from data we might collect from published or unpublished sources? Yet standalone desktop silos (such as EndNote for references, Word for notes and MSAccess for data) actively discourage the linking of these forms of information.

Bespoke or ad hoc databases admirably (or less than admirably) fulfill the particular needs of researchers, but fail to connect with the wider world. These databases are often desktop-based and inaccessible to anyone but the user of their host computer, other than through sharing of copies (with all the attendant problems of redundancy, maintenance of currency and merging of changes). When accessible they often lack multi-user capabilities and/or are locked down to modification by a small group of users because of the difficulties of monitoring and rolling back erroneous or hostile changes. Even when
accessible to the public, they are generally accessible through a web interface which allows human access but not machine access, and cannot therefore be linked programmatically with other data to create an integrated system for analyzing larger problems.

For eResearch in the Humanities to advance, all the digital information we use—bibliographic references, personal notes, digitized sources, databases of research objects etc.—need to exist in a single, integrated environment rather than in separate incompatible systems. This does not of course mean that the system need be monolithic—mashups, portals and Virtual Research Environments all offer distributed alternatives, dependant on exposure of resources through feed and web services. The ‘silo’ approach to data is also breaking down with the stunning success of web-based social software such as the Wikipedia encyclopaedia or Del.icio.us social bookmarking systems. These systems demonstrate that—with the right level of control and peer review—it is possible to build substantial and highly usable databases without the costs normally associated with such resources, by harnessing the collaborative enthusiasm of large numbers of people for data collection and through data mining of collective behaviour.

To illustrate the potential of an integrated Web 2.0 approach to heterogeneous information, I will discuss Heurist (HeuristScholar.org)—an academic social bookmarking application which we have developed, which provides rich information handling in a single integrated web application—and demonstrate the way in which it has provided a new approach to building significant repositories of historical data.

Heurist handles more than 60 types of digital entity (easily extensible), ranging from bibliographic references and internet bookmarks, through encyclopaedia entries, seminars and grant programs, to C14 dates, archaeological sites and spatial databases. It allows users to attach multimedia resources and annotations to each entity in the database, using private, public, and group-restricted wiki entries. Some entries can be locked off as authoritative content, others can be left open to all comers.

Effective geographic and temporal contextualisation and linking between entities provides new opportunities for Humanities research, particularly in History and Archaeology. Heurist allows the user to digitize and attach geographic data to any entity type, to attach photographs and other media to entities, and to store annotated, date-stamped relationships between entities. These are the key to linking bibliographic entries to other types of entity and building, browsing and visualizing networks of related entities.

Heurist represents a first step towards building a single point of entry Virtual Research Environment for the Humanities. It already provides ‘instant’ web services, such as mapping, timelines, styled output through XSLT and various XML feeds (XML, KML, RSS) allowing it to serve as one component in a decentralized system. The next version will operate in a peer-to-peer network of instances which can share data with one another and with other applications.

The service at HeuristScholar.org is freely available for academic use and has been used to construct projects as varied as the University of Sydney Archaeology department website, content management for the Dictionary of Sydney project (a major project to develop an online historical account of the history of Sydney) and an historical event browser for the Rethinking Timelines project.