

TAPoR Needs Assessment: Final Report

Prepared by Sue Fisher, Electronic Text Centre, University of New Brunswick
October-November, 2002

Introduction

In October-November 2002, Sue Fisher of the Electronic Text Centre at the University of New Brunswick conducted interviews with stakeholders in the TAPoR project in order to

- synthesize areas of commonality and discrepancy around a core set of discussion points,
- make recommendations to TAPoR principal investigators based on this synthesis; and
- make recommendations to the Electronic Text Centre with respect to the development of a metadata framework for the TAPoR Portal

The results of this process are included in this document.

Three separate documents provide the groundwork for this report and are attached as appendices:

- TAPoR Needs Assessment Discussion Document: This document formed the framework for the telephone/in-person interviews with TAPoR researchers.
- TAPoR Needs Assessment Summary: This document is a point form transcription of the recorded interviews with TAPoR researchers.
- TAPoR Needs Assessment Summary: External Researchers' Comments: this document is a rough transcription of interviews conducted with humanities computing researchers external to TAPoR

The interviews were conducted in confidence and with the understanding that only the synthesized report would be made public. Consequently, the above summary documents should remain internal to the Electronic Text Centre.

Discussion Participants

TAPoR Researchers:

University of Victoria

Michael Best

Scott Gerrity

University of Alberta

Terry Butler

Stéfan Sinclair

McMaster University

James Chartrand
Geoffrey Rockwell

University of Toronto

Ian Lancashire
Elaine Toms

University of New Brunswick

Alan Burk
Brad Nickerson

Note: Despite efforts to schedule an interview with the Université de Montréal's Daniel Poulin, the interview did not take place in time for the completion of this report.

External Researchers

Susan Hockey, University College, London
John Unsworth, University of Virginia

Graduate Student Representative

Natasha Nunn, University of Alberta

Governing comments:

I preface this report with 3 comments made by the external researchers I spoke with during this assessment process. Each highlights the dynamic nature of our discipline and stresses the need for text analysis researchers to be on the forefront of trends within humanities computing.

1. Susan Hockey: “where does TAPoR see itself in 10 years time? Once you start a project like this it will start to move in directions you had never anticipated. The best way to handle this is to be clear in your initial vision so that you recognize the nature and significance of the moves that come about.”
2. John Unsworth: “the current offerings in text analysis tools are dead. Their utility is limited and they haven't been updated in years. What I hope this project will do is build new kinds of tools—tools that we might describe as text analysis—but that let us do intelligent things with intelligent text. What these new tools spit out may be data but it may also be insights into electronic text publishing based on the use and frequency of markup, or perhaps provide us with means of visualizing the implications of markup.”
3. Susan Hockey: “The only humanities scholars that are interested in text analysis anymore are the corpus linguists. Most humanities scholars are interested in the electronic text only as a cultural object. Text analysis must adapt to creative new forms of publishing.”

Key Issues

In synthesizing my discussions with TAPoR members I endeavoured to keep in mind practical issues that confront the TAPoR project. Key among these is the nature of current funding structures. The Canada Foundation for Innovation has provided funds for the creation of the TAPoR infrastructure (Portal design and creation and the establishment of 6 independent Nodes) but it does not provide long-term funding for the continued management of this infrastructure. All recommendations provided in this report take this fiscal reality into account. Despite the fixed nature of this funding, however, the discipline of Humanities Computing is continuously evolving; as such the architectural specifications of TAPoR will need to be dynamic and scalable. The architecture upon which the portal and nodes are built should be comprehensive and forward thinking so as not to become obsolete.

Findings and Recommendations

The TAPoR infrastructure comprises two major components: Portal and Nodes. In my discussions with TAPoR researchers it became apparent that each of these components has its own set of users. The differing user needs will have an impact on the overall architectural design for each component. General consensus among TAPoR researchers is that Portal users might come from any constituency and range from being complete novices in field to expert users wishing to perform quick-and-dirty research through the established interface provided by the portal. Conversely, users at the Nodes are most likely to be experts in the field; for the most part they will be the principle investigators in TAPoR or affiliated researchers carrying out large-scale humanities computing projects. Due to the diverse nature of these two user groups, I have made two categories of recommendation.

TAPoR PORTAL

In order to assist novice users, the Portal will want to create an introductory/tutorial environment. This environment will closely resemble the original Portal design proposed by Geoffrey Rockwell and Stéfán Sinclair wherein users will be able to match representative texts with representative text analysis tools such that they can learn the function and relative merit of text analysis. Such an environment, however, should be only one option provided by the Portal given the full range of research that humanists will want to conduct using the TAPoR Portal. A group of representative texts, no matter how robust, will never satisfy the diverse needs of humanities researchers. Furthermore, a representative tool set will need to evolve as the discipline advances. Given this, good Portal design will have built-in mechanisms for the ingestion of texts and tools. I use the term "text" in this document in its broadest possible sense; the electronic texts that populate TAPoR will vary in media type, format, and encoding style.

Broadening the range of Texts and Tools

The consensus among TAPoR participants was that, although TAPoR is not a comprehensive digital library, its purpose and function does bear a distinct similarity to that of the digital library. Rather than being a depository for texts from a wide range of sources, TAPoR should be able to accommodate the temporary ingestion of text to suit

individual researchers' purposes. Such text might be uploaded by the individual researcher or it may come from any one of a number of electronic text archives on the World Wide Web (such as the Oxford Text Archive).

Texts and tools that do reside on the TAPoR infrastructure (those contributed by participating institutions) and that are made available to researchers beyond the individual node level should be accepted and made available according to a written and publicly available collections development policy.

It is not only in its texts that TAPoR bears resemblance to a digital library but in its offerings of tools relevant to the field of humanities computing. Though not an exhaustive holding pen for the full range of tools that might be of interest within the discipline, TAPoR's holdings may well range beyond traditional text analysis tools to include tools that support the creation and publication of electronic text and that support the management of large-scale Humanities Computing projects. Furthermore, TAPoR will need to accommodate tools that have not yet been imagined or developed but that will form the next generation of text analysis tools.

Recommendations

Based on the above observations, I make the following 4 recommendations:

1. TAPoR will need to standardize a series of processes for the collection of texts and tools that reside on TAPoR servers. Issues include
 - The creation of written collections development policies for both texts and tools. Factors to consider include desired range of media, format, subject coverage, markup standards, editorial authority, or any other factors deemed important to TAPoR principal investigators. Inevitably, TAPoR will receive request to deposit texts or tools on its servers (by researchers both internal and external to the participating Nodes). Such a policy will allow TAPoR to make decisions about what kinds of texts and tools can or cannot be accepted.
 - the ability to add metadata to the text or tool at the time of input either through web templates or automated uploading processes (for texts that already contain metadata)
 - the ability to add a rights management description to a resource or collection of resources
 - the creation of a human management infrastructure and approval process that will allow material to submitted to the Portal, reviewed, and made public
2. TAPoR will need to establish processes for querying key electronic text sites and the WWW in general from the Portal interface for the existence of external texts that could be used by researchers using TAPoR tools. A future research project beyond the CFI mandate might be to explore metadata harvesting for the TAPoR context. Such an undertaking would be a large scale initiative requiring original research and testing in addition to working in conjunction with international initiatives in the metadata field (e.g. METS <http://www.loc.gov/standards/mets/>, and OAI <http://www.openarchives.org/>)
3. Given the diverse range of media, format, and markup being created by TAPoR researchers, the Portal design will need to take into account that not all texts will

work with all tools. Researchers should be easily able to match a text with an appropriate tool and vice versa.

4. TAPoR should establish a standardized presentation for tools documentation and other user training and help material. At its simplest, a common template should be used for a FAQ file for each tool made available through TAPoR. Additional funding should be solicited to develop support materials for Portal users.

TAPoR NODES

The primary users at the TAPoR Nodes will be either experts in the field of humanities computing or humanities computing scholars in training (either students or working scholars guided by training staff at each Node). The diversity of such users is great:

- scholars creating electronic texts
- scholars creating open source tools
- scholars conducting text analysis research
- scholars creating digital publishing and delivery models
- scholars extending the humanities computing mandate to multi-media
- scholars dealing with notions of archival editions vs. working edition

The list could go on.

It is important to note that all these pockets of research are independent and must be carried out at Nodes that have autonomy of function. Having said this, however, there are common issues that affect the broad diversity of this research. Commonality of process and function across all 6 Nodes is worth pursuing for issues that have an impact on the offerings of the general Portal. In the opinion of several TAPoR researchers interviewed, the key strength of TAPoR is its ability to develop the Humanities Computing research community in Canada and beyond. Much of this community building can begin around common research issues that affect the interplay of the TAPoR Nodes with the Portal.

Recommendations

I recommend that informal working groups be established within the TAPoR community to address the following common research concerns. Because CFI does not provide funds for the ongoing management of TAPoR, these groups could serve as advisory bodies recommending where the need for further major collaborative funding is needed.

Group 1: Metadata

The immediate needs of TAPoR are for adequate collections-level metadata processes that can be integrated into the initial architecture for the Portal. Templates for the creation of item-level metadata that can be made accessible to researchers at the individual nodes are also desirable. Future funding possibilities in the area of metadata include metadata harvesting, and user documentation for the adequate creation of metadata records in the TAPoR environment.

Group 2: Rights Management

The first round of TAPoR will need to establish standard protocols for articulating the ownership of and access restrictions to any given collection housed by TAPoR and for

communicating this information to the end-user. Further research in this area may include the evaluation of rights management systems and the implications of commercial models for the distribution of copyrighted TAPoR resources.

Group 3: Open Source Tools Development

The long-term viability of TAPoR lies in its ability to foster the development of tools that will be useful and accessible to the widest possible range of Humanities Computing researchers. Several researchers within the TAPoR community (led by the University of Alberta) have shown a marked interest in developing TAPoR's potential as a host for the development of open source tools for the field. This working group could, among other things, establish policies for the creation of open source, generalizable, well-documented, peer-reviewed software. There is great potential for individual tools-development projects and funding applications to grow out of this working group.