

# Xaira: an XML-aware indexing and retrieval architecture

## 1 Introduction

This prospectus briefly summarizes the functionality of XAIRA, a general purpose XML text retrieval system, developed at Oxford University Computing Services for deployment in the next release of the British National Corpus. In its most recent form, we believe this software to be unique in its support for the vast majority of user requirements characterizing many areas of Humanities text-based research, not solely corpus linguistics. We would like to explore the general applicability of the system while addressing two important shortcomings of the current version: specifically, the restriction of the current version of the software to the Microsoft Windows platform, and its lack of thorough documentation.

We describe below the work that we would like to undertake this year in order to address these shortcomings. The bulk of this work comprises software development, to be carried out by the current Xaira developer in Oxford; development of both technical and user-level documentation, which will also largely be carried out in Oxford; and a period of user-testing and trialling, in which we plan to work closely with members of the Humanities Computing community world-wide. The development project would be managed within the Research Technologies Service ([www.oucs.ox.ac.uk/rts](http://www.oucs.ox.ac.uk/rts)) of Oxford University Computing Services (OUCS), which has been at the forefront of humanities research applications for information technologies for many years, and which would administer any grant awarded by the Foundation.

We plan to make the resulting software available for evaluation by as many relevant Mellon-funded text-creation projects as possible. The WordHoard project at NorthWestern, for example, seems to be in need of a tool precisely of this nature: we would like to work closely with developers and users of that project as typical members of the community of humanities users. Such users have repeatedly identified a need for sophisticated text searching tools, usable by non-experts, to complement the increasing sophistication of the digital textual resources now at their disposal. In Xaira we believe we have a unique tool which meets that need.

To date, development of Xaira has been funded from sales of the BNC, but this has now effectively ceased. To continue development of the system according to the plans described below will cost in the order of \$75,000, and could be accomplished within six to eight months. A more detailed cost breakdown is given in the work plan below. The work plan proposed could commence immediately on confirmation of the availability of funding: if for example this were available by 1 June 2004, we would aim to have a fully functional version of the system described below in place by 1 February 2005, with prototype versions available for testing in November 2004.

## 2 Xaira: what it does

Xaira is a suite of software tools developed at Oxford University Computing Services as part of the UK's largest publicly-available corpus building project the British National Corpus (BNC: <http://www.natcorp.ox.ac.uk>) and first released in 1994. Since that date, Oxford has sustained a long standing commitment to the maintenance of the BNC and SARA, its associated software, which is distributed with the corpus and is in use worldwide. Over the last two years, the SARA system has been completely rewritten as a general purpose tool for searching large XML corpora, with a particular focus on the needs of corpus linguists, with close attention to new XML-based encoding standards, and with the benefit of hindsight derived from a decade of feedback from hundreds of users world wide. It is intended to distribute this new version of the software with a new, XML-encoded, version of the BNC.

At present, the Xaira suite has the following components: (1) an indexer, which creates efficient multi-keyed indexes to a large collection of discrete XML documents; (2) a server, which handles interactions between client programs and the data files using a specialised XML query language; (3) a Windows client, which handles interaction with the user, providing a rich range of text query and manipulation facilities. This modular architecture permits the development of different specialized client programs for different applications or styles of usage.

The process of creating a new Xaira application is much simplified by a special purpose index building utility, called *xaira-tools*. This allows a user interactively to supply metadata about the corpus to be indexed, additional to that present in any TEI header, which is then used to drive the indexing process. An extended version of the TEI Header is used to store this corpus description for re-use by other components of the system.

Xaira is designed to work with any XML-encoded corpus, small or large. The more detail present in the markup, the more facilities are available to the client but the minimal requirement is only that the text be well-formed XML. If the corpus to be processed specifies a document type definition (DTD) or schema, then *xaira* will validate it at indexing time, and will not proceed if any validity errors are discovered. Unlike earlier versions of the program, any DTD may be used for this purpose, though Xaira was originally designed specifically to support the TEI family of DTDs.

Xaira can thus do something useful with the full range of digital material one might wish to build into a corpus. At one extreme, it can be used to provide basic searching facilities for a collection of Project Gutenberg style texts, innocent of any explicit descriptive markup at all; at the other, it can take full advantage of the rich annotation present in a multilingual corpus containing detailed feature structure analysis, POS-tagging, and explicit lemmatization. Because the semantics of the available markup are defined during indexing, the system builder can control how much, or how little, of the markup is to be made available to client applications, and how it is to be presented.

With XML support, comes Unicode support. Xaira uses Unicode internally to represent all character data: it can thus handle text in any language, and any combination of languages. We have so far tested it with Chinese, Eastern European, Medieval, and Ancient Greek scripts; the system should also support bidirectional and non-alphabetic character sets conformant to Unicode standards. In common with other XML systems, the system will correctly process character entity references found in the data, such references being retained untranslated in the underlying corpus index, but rendered as the appropriate Unicode code point when displayed in non-XML modes. The Xaira windows client includes a configurable input system, allowing selection of specific characters from a Unicode table by point and click, temporary mapping of keyboard keys, or complete redefinition of a new keyboard map, which can be loaded as needed.

Any XML or SGML aware search engine has the ability to locate specific tagged components and to carry out searches within the context of such components, for example to look for the word "crime" only where it appears in a headline. Most also have the ability to reorganize and display search results in a variety of forms. Xaira extends these facilities with some additional, more lexically-motivated, functions. These include: implicit or explicit tokenization and lemmatization of element content; definition of additional keys for index searching ; expandable automatic collocation search; a rich range of query types; definition of subcorpora and partitions.

Internally, Xaira uses a simple ECMAScript-like scripting language called *Sarascript*, which makes direct calls on an API. A key component of this is an XML-based query language, the *Corpus Query Language*, which is used to pass messages between the client and the server.

The interface supports searches for substrings or regular expression-style patterns, words, phrases, or the tags which delimit XML elements and their descriptive attributes. Tables of all attribute values used are constructed during indexing, and may be used in a number of ways by the client, for example to colour code parts of the displayed results, or to optimize certain kinds of searches. Tokens found in content are

used in many different ways: a search may simply check presence or absence in the lexical index maintained for the corpus (and inspect their frequency or collocational patterns), or they can be used as access points, to retrieve and display parts of the corpus.

Search targets can also use additional keys such as part of speech, semantic classification, or root form (lemma) of a token, as specified in the tagging of the texts. Queries may be scoped, searching for combinations of words etc. in particular contexts, which may be defined as XML elements, or as combinations of other identifiable entities, or as stretches of text. Such searches may be order-sensitive or insensitive. As with other query languages, different kinds of search can be combined to form complex queries of various kinds. The present client includes a range of query-definition facilities, ranging from the familiar Windows interface widgets, to a special “Query Builder” graphical interface, in addition to the scripting language mentioned above.

The results of a search can be displayed either one at a time or within a traditional KWIC style window which can be sorted, thinned, expanded etc. The context of hits located in this way can be explored by expanding it, up to the full text level if necessary. A range of formatting options can be specified, using a subset of the the W3C standard Cascading Stylesheets (CSS). Results can also be exported in a simple XML format for reprocessing, either by Xaira, or by any other XML-competent application, such as a word processor or XSLT engine.

Corpora can be reorganized or partitioned in a user-defined way, using the results of any query, the values of specified element/attribute combinations, or a manual classification. Searches carried out across partitioned corpora can be analysed by partition: so the client can display the relative frequencies of a given lexical phenomenon in texts of different categories identified in a corpus, for example to compare and contrast patterns of lexis in male and female speakers, or in texts of different genres, or from different periods.

### **3Xaira: The Next Phase**

In the next phase of development, our major objectives will be to remove the current system's dependence on a single operating system platform, to enhance its documentation and description, and to validate its usability in as wide a range of research applications as possible.

The original Xaira system architecture derived from a world in which relatively low-powered client machines accessed relatively high-powered server machines across a network, using a simple message-passing protocol. It made sense to assume that the client would be largely devoted to user-interface issues, and that the server would do all the serious computing. However, as client computers became more powerful, the demand for a standalone system led us to embed more of the system's basic functionality within the client. In today's computing environment, the pendulum is swinging back towards the notion of distributed processing, but with the important change that server and client machines are more intimately connected, by sharing a common object model. This enables us to support not only clients on different platforms, but clients with very different kinds of functionality.

With XML-based standards such as XML-RPC and WSDL it becomes possible for us to define the system in a way that is independent of any specific server implementation, and to expose its components over a network for use by any kind of client process. This means that searches or analyses of a Xaira system can be embedded in WSRP-compliant portal frameworks such as uPortal, or in any other web service aware applications, rather than requiring construction of complete client interfaces. Our vision is one in which diversity of application complexity and scope is as important as diversity of platform.

To that end, the first step in our work plan is to re-define the server as a set of objects and methods, which together we refer to as the Xaira Object Model. This model can then be implemented easily on different platforms, while subsequent development stages (detailed below) will enable us to move smoothly to a modern distributed architecture.

On completion of the workplan, our intention is to release the source code for the new server and indexer under an appropriate Open Source licence, together with detailed technical documentation; we would be happy to discuss the details of the eventual licencing and IPR arrangements with the Foundation. Our expectation is that the license should require that improvements to and derivative software based on the code developed would also be made available on an open source basis. We hope that this will facilitate further development of text analysis tools appropriate to the needs of different user communities, building upon the basic notions of the Xaira architecture

The technical documentation will be complemented by a set of user-focussed documentation for particular interfaces. The existing *BNC Handbook* (Aston and Burnard, Edinburgh University Press, 1997) contains a large number of practical exercises which introduce almost all the features of the SARA program in the context of language learning and teaching. Our intention is to produce comparable material for the new system, and to bring the existing Help files up to date. The new help files will be written as TEI conformant XML files from which online help in multiple formats and in different languages can be generated: it is hoped that this internationalization will be carried out as a collaborative project with other interested partners in Europe.

Finally, we aim to form a network of early adopters and experimenters who will assess the usability of the system for the needs of their particular projects. Such users will need some initial training in the customization facilities of the new system, and we envisage this being produced for a workshop to be organized towards the end of the project. We would aim to organize such a workshop in conjunction with some other relevant event, such as the annual TEI Members' Meeting (to be held this autumn in Baltimore), or the Digital Resources for the Humanities conference (to be held in September 2004, in the UK).

## 4The Context

Oxford University Computing Services has been at the forefront of work in the application of digital technology to Humanities Research for over thirty years. Its pioneering work in establishing the core of what has since become known as “humanities computing” needs no introduction. During the 1990s, the Humanities Computing Unit, a part of OUCS, developed such internationally-known resources as the Oxford Text Archive, the Text Encoding Initiative, and the Humbul Humanities Hub, which now form key components of the Research Technologies Service (RTS) at OUCS. The RTS embodies the synergistic vision characterizing the best of humanities computing, by bringing together insights and opportunities from different academic disciplines. In addition to the projects already mentioned, the RTS currently hosts nationally-funded research support projects such as OSSwatch, a national advisory service on the role of Open Source Software in academia; the Oxford eScience Support Centre, which co-ordinates Oxford's participation in the UK's eScience programme; and several JISC-funded projects and initiatives related to R&D of portal and web services.

OUCS is administratively part of the same division as the Oxford library services, and there is a good track record of collaboration between library and computing services in areas of common interest, most notably the development of digital library facilities and digitization projects; indeed, several key figures in Oxford's Digital Library are former OUCS staff. This collaboration is likely to expand further with the development of Oxford's Academic Services Portal project.

The British National Corpus project has been hosted at OUCS since its inception and OUCS continues to distribute and licence the corpus on behalf of the academic/industrial consortium responsible for it. The BNC and Sara are used in 30 countries world wide, notably Japan and continental Europe. As many other corpus building projects have tended to follow its design and encoding format to a large extent, it seems clear that there will be a large demand for software of this kind. In any event, OUCS is committed to continued support of the BNC and its associated software for the foreseeable future.