First Class Objects and Indexes for Chant Manuscripts

Louis W. G. Barton
University of Oxford
St Anne’s College
Oxford OX2 6HS; U.K.
louis.barton.je.77@aya.yale.edu

Peter G. Jeavons
Oxford Univ. Computing Laboratory
Wolfson Building, Parks Road
Oxford OX1 3QD; U.K.
peter.jeavons@comlab.ox.ac.uk

John A. Caldwell
Emeritus Fellow of Jesus College
University of Oxford
Koon Shan Barry Ng
The NEUMES Project

ABSTRACT
We discuss a crucial part of infrastructure for the Web-delivery of medieval chant resources. Although widely accepted by software professionals, the distributed-content model is sharply opposed by some chant scholars. We advocate for a paradigm of the Web as a massive database where each “first class object” acts like a record; metadata about, and links to such objects are compiled in virtual libraries. Scholarly-edited indexes determine which objects are in libraries, and unreliable content is excluded. Special metadata ontologies can be defined without modifying the primary content.

Categories and Subject Descriptors
H.3.7 [Information Storage and Retrieval]: Digital Libraries – Standards; J.5 [Applications]: Arts and Humanities – Performing arts (music)

General Terms Design, Standardization.

Keywords Distributed content, scholarly-edited index, search, medieval chant manuscripts, neumes, images, transcription.

1. CONTEXT
In [2] and [1] we proposed that digital transcriptions of medieval chant manuscripts should be hosted on the Web as distributed content. Influential voices in the chant research profession oppose our work, for the stated reason that centralised-database storage is necessary to authoritative content. We reply that they use database systems inappropriately and the term ‘database’ indiscriminately.

Although a digital library of chant manuscript transcriptions does not yet exist, many digital images of chant manuscripts are online. These images are being created by libraries independently and are hosted on the Web in a distributed manner. This has given us an opportunity to demonstrate in a concrete way several points about distributed content, quality control by scholarly-edited indexes, and specialised metadata ontologies, that we argued just in theory.

Our “Distributed Digital Library of Chant Manuscript Images” [4] is intended to demonstrate these software-infrastructure ideas, methods of searching for content, and software implementations. It does not pretend to be a definitive resource for scholarly research. By “first-class object” we mean a stand-alone, integral file that is visible to indexing robots and that can be downloaded.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.


2. THE DEBATE
Below are objections that have been raised against our proposal for chant transcriptions as first class objects, and short responses.

• O: We are ignoring “a major international, distributed project which has been in existence since the early 1950s,” i.e. RISM (Répertoire International des Sources Musicales).
• R: RISM uses the centralised-content model, although several country locations have their own RISM databases. There is a misunderstanding over what is meant by “distributed content.” These databases are ‘distributed’, but the content is centralised.

• O: We are “criticizing an existing working system, which contains over 300,000 records,” i.e. the CANTUS database [3].
• R: We do not criticise CANTUS regarding its integrity or the extent of its content. Our point is that the content of CANTUS is hidden on the Web as records inside a relational database. This content could be exported to XML as first-class objects.

• O: “The centralized model would be simpler because only one system of indexing and access needs to be implemented.”
• R: Often, the simplest and most obvious technical solution does not produce the best results for the end-users of software. For example, a GUI (graphical user interface) allows users to click icons or hyperlinks, drag-and-drop, copy-and-paste, etc., but the programming for it is highly complex. The command line interface is much easier to program, but it is difficult for users.

• O: If chant databases were OAI-compliant, then they could be ‘harvested’ by OAI harvesters. This would enable users to search for content across multiple OAI repositories.
• R: OAI (the Open Archives Initiative Protocol for Metadata Harvesting) uses fixed ontologies of metadata. An assumption (from Leibniz’s law, standardised in Codd’s relational model) is that an entity can be described by a finite set of attributes. This is valid for formal entities (e.g. an employee record) but it is not valid for actual, real-world entities (e.g. the person who is an employee). Semantics of actual objects may be indeterminate or involve subjective judgements (e.g. “find examples of beautiful neumation”), when the attribute in question is not contained in, or derivable from the object itself. Different ontologies may be needed for different scholarly purposes; paleographers, for instance, index objects by highly specialised criteria that are not appropriate to a ‘standard’ ontology. An ontology might have conflicting values: two scholars could assign different dates to a chant manuscript or identify its neume notation differently. It is better for resources to be first class objects, bundled with basic metadata, and visible for cataloguing by specialised indexes.

This research project is supported by Eduserv.
• O: There are important differences between chant transcriptions and photographs of chant manuscripts: a distributed library of images is not a valid paradigm for scholarly transcriptions.

R: We agree that reliable transcription of manuscripts involves more scholarly judgements than does photography. Attribution, however, may require expertise in both cases. Already, several websites are trying to be a one-stop, comprehensive site for chant manuscript data; central control cannot stop this. Quality control in such an environment is best effected via scholarly indexing, not by keeping content hidden in a controlled database.

• O: If everyone is allowed to put chant content on the Web, the absence of any editorial review will cause problems.

R: Unreliable content is a fact of life on the Web. End-users need to judge the credibility of content providers, or search for peer-reviewed content only.

3. CONTENT SEARCH

When data are intended for the Web but are inaccessible as first class objects, often they are called 'dark matter' (a metaphor of the dark matter theorised to exist in outer space). Usually these are data stored in database management systems (e.g. Microsoft Access), where a Web portal enables users to query the database.

For example, consider the behaviour of the CANTUS database [3] in a search. CANTUS is a Web-based research tool that is highly esteemed by chant musicologists. Most of its information is stored in an Access database. Figure 1 shows three of the eight results from querying this database on 'cao8414d' (which is a standard chant ID). Clicking an 'Incipit' displays a database record.

We remark that a database query involves a dialogue between user and database. The user composes or selects criteria for a search, motivated by the intention to find some particular information. A precondition is that the end-user must possess some knowledge of the information s/he is seeking (such as a chant ID or an incipit).

Figure 2 shows the result of a Google search on 'cao8414d'. The CANTUS database has eight records matching this criterion, but they are not indexed by Google. Google does not possess the background information needed for a dialogue with the database.

If we change the search to 'cao 8414d' (space after 'cao') the result is different. Google finds matching content just on the NEUMES Project website (Figure 3). Result #2 (check marked) is a transcription in NEUMES encoding, wrapped in a NeumesXML descriptor file. It is a first class object and it contains the chant ID.

3. CONTENT SEARCH

When data are intended for the Web but are inaccessible as first class objects, often they are called 'dark matter' (a metaphor of the dark matter theorised to exist in outer space). Usually these are data stored in database management systems (e.g. Microsoft Access), where a Web portal enables users to query the database.

For example, consider the behaviour of the CANTUS database [3] in a search. CANTUS is a Web-based research tool that is highly esteemed by chant musicologists. Most of its information is stored in an Access database. Figure 1 shows three of the eight results from querying this database on 'cao8414d' (which is a standard chant ID). Clicking an 'Incipit' displays a database record.

We remark that a database query involves a dialogue between user and database. The user composes or selects criteria for a search, motivated by the intention to find some particular information. A precondition is that the end-user must possess some knowledge of the information s/he is seeking (such as a chant ID or an incipit).

Figure 2 shows the result of a Google search on 'cao8414d'. The CANTUS database has eight records matching this criterion, but they are not indexed by Google. Google does not possess the background information needed for a dialogue with the database.

If we change the search to 'cao 8414d' (space after 'cao') the result is different. Google finds matching content just on the NEUMES Project website (Figure 3). Result #2 (check marked) is a transcription in NEUMES encoding, wrapped in a NeumesXML descriptor file. It is a first class object and it contains the chant ID.

4. REFERENCES


