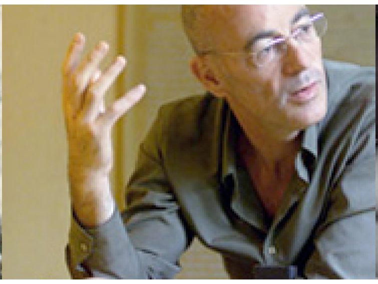
Culture-mining

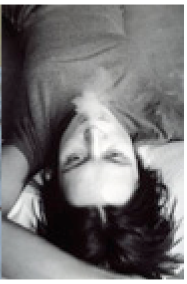
Time-based cultural documents and online meaning-based reSearch Tools

Grayson Perry

Julia Kristeva



Jaques Herzog



Sarah Lucas

Introduction

Artists, museums and the heritage sector are creating ever increasing amounts of audio-visual and digital content. One of the biggest consequent issues is how to manage and open this content to the public and to the education sector in order to provide optimum opportunities for retrieval and interpretation. Without effective methods for *reSearching* – retrieving, discovering, organising, contextualising and interpreting – the growing mass of cultural data risks becoming increasingly inaccessible.

Tate is working in collaboration with the Department of Computing of Goldsmiths College, University of London, to produce an open source application for *reSearching* audio-visual content online. This is framed within the project *Culture Mining*: "Creation of a flexibly searchable streaming media archive of contemporary and modern art theory and practice", currently funded by the AHRC under the ICT fund for cultural resource enhancement.

The work is driven by the Tate Online Events Archive – a collection of digital video recordings, gathered over a period of 6 years from webcast curated public events. The archive currently holds around 500 hours of content, including artist talks, cultural theory lectures, symposia, music and performance events, and continues to produce over 100 hours every year. The archive's material is organized as long play (contiguous) streaming media objects, of lengths generally between 30 minutes and 2 hours. The only means to research the archive is to play selected recordings and manually fast forward over irrelevant material.

The *Culture Mining* project aims to develop a user-centered tool that will support and encourage a deeper exploration of the interdisciplinary material. The tool will enhance the users' access to content by *guiding* their retrieval processes towards only those *fragments* (of the long recordings) that are relevant to their interests. Further, it will also provoke and inspire users to draw *new interpretations*, by presenting novel juxtapositions of different fragments of the material.

Aims

- The creation of a flexible, searchable streaming media archive of contemporary and modern, art, theory and practice
- The creation of an intelligent system for archiving and retrieving video and audio material
- Increased use and viewing of archived events
- Encourage a deeper exploration of the interdisciplinary material and cross references contained in the archive
- Consider maintenance, management and workflow solutions

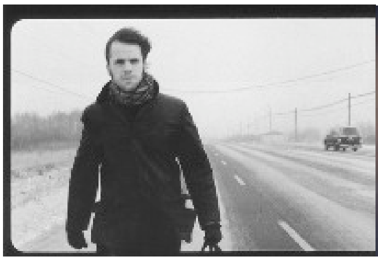
Outputs

- An intelligent system that allows easy access to fragments from Tate's digital video recordings of cultural theory lectures, artist talks, symposia, sound and performance
- A system that can semi-automatically fragment and tag audio-rich streaming media
- An ontology (a specialist vocabulary and logic) for representing, in a computer understandable format, the content of recorded talks, symposia and performances, mainly focused on modern and contemporary art;

- A search engine, that can reason with ontological descriptions in order to find relevant fragments;
- A user interface for the system, deployable in web browsers;
- A methodology for tagging the rest of the Tate material. This methodology will be applicable for any audio-rich streaming media

Project Team (alphabetical order):

- Kelli Dipple – Co-investigator (Tate Webcasting, Tate)
- Adrian Passow – Doctoral Researcher (Department of Computing, Goldsmiths University of London)
- Dr Tina Sherwell – Research Assistant, Art History (Department of Computing, Goldsmiths University of London)
- Dr Dan Tidhar – Research Fellow, Computing (Department of Computing, Goldsmiths University of London)
- Dr Marian Ursu – Co-investigator and Project Manager (Department of Computing, Goldsmiths University of London)
- Prof. Robert Zimmer (Principal Investigator, Department of Computing, Goldsmiths University of London)


Robert Frank Symposia

Yinka Shonibare, Turner Prize 04

Olafur Elliason, Unilever

Sean Scully, Tate Collection

Context

Contextual Factors that have defined and informed the project include: the nature and characteristics of the content, existing and potential audiences for the archive, available descriptions of event programmes, wider research into ontological structures, digital archiving methods and interfaces, increased user demand and expectation as well as Tate-wide digital asset management strategies.

Content Description

The Tate Online Events Archive (http://www.tate.org.uk/onlineevents/archive/), is a growing resource, focussed on modern and contemporary; art and culture, largely consisting of a selection of Tate's Adult Education Programmes that have been webcast live, and are available afterwards online, as long-play files. The audio/video archive is hosted in context of an ongoing live webcast programme and onsite performance and public events at Tate Modern and Tate Britain.

Tate has been webcasting a curated selection of public events for the last 6 years and has built up a substantial archive of recordings as streaming media objects. In the year 2004 - 2005, The Tate Online Events Archive and associated content received over 21 000 unique visitors, collectively watching up to 280 000 archives.

Largely the archive content consists of talks and presentations recorded in a video format. Ultimately the material defines itself on the basis of speech, text and ideas. Visual elements presented within the talks (slides, moving image, laptop presentations) are captured and edited together live; along with a 3 camera mix and event titles. Talks of 1 – 2 hours are available as long play files. Conferences and symposia are cut up into individual speaker's presentations and plenary discussions, 15 – 45 minutes each.

The content covers a wide range of overlapping topics including; for example: fine art history (modern and contemporary), cultural theory, visual culture, social science, media theory, design; as well as fine art, new media, performance, music and curatorial practice. Resources include, for example, talks by the cultural theorist Manuel DeLanda; artist Bill Viola; the filmmaker Anthony Manghella; performance-activists The Guerrilla Girls. Conferences and symposia include, for example, Open Systems: Rethinking Art c. 1970 and Thinking the City: Multidisciplinary Views on Urban Life and Culture.

Further Materials

Further audio/video resources are being produced in conjunction with Exhibition programmes and interpretation activities around Tate's Collection, including for example, Artist Interviews and Performance Documentation. Around 50% of this additional material also already exists in a digital format. This material is not collated in a central location and is often lost after the PR around a particular exhibition subsides. Whilst defining the scope of the project it was determined infeasible to include this additional content in the prototype model, however it is assumed that the future application of these tools would further aid the accessibility and visibility of online audio/video content that sits outside the Online Events programme. In the long run the inclusion of the widest range of audio/video content available, should be considered, in order to improve visibility and cross references, relevant to user demand.

Uniqueness of the Archive

The Tate Online Events Archive has existed as a digitized resource distributed via the internet since Tate started webcasting public events in the year 2001. Over-viewing similar collections has pointed this out as a unique attribute. At the time this research was undertaken many other collections either had a large amount of un-digitized material, or they had a small volume of digitized material. The scale of Tate’s existing digital archive and its ongoing growth, presented particular issues around the management and maintenance of material. However as an attribute, it also provided the project with an expansive and comprehensive set of ready-made digital resources, with which to test technical and ontological developments.

The archive is hugely diverse and interdisciplinary, which demands a suitable ontology to cover a wide cross section of domains and demonstrate a complex set of relationships. Much of the ontological research referenced, has focussed on specialist vocabularies and taxonomies for example, in context of New Media or Artists’ Film & Video.

The first stage of this project aims to cover only a selected portion of concepts within the archive. Wider implementation will require extensive ontological development. Therefore the structure of the prototype will be expandable and extendable with a view to longer-term maintenance.

The majority of technical research in the field of video-search tends to focus on visual referencing and frame difference. By understanding the Tate Online Event Archive’s uniqueness, by way of an emphasis on meaning through speech, discussion and natural language; this research has aligned more specifically with other research into text, speech and audio analysis.

Text-based descriptions of events currently available, often fail to describe accurately or in full, the content of the actual presentation. Descriptive texts are generally written previous to the event for marketing purposes. It is therefore an aim of the project, to implement a system that enables users to search natural vocabularies as used during the event presentations.

Given the shortcomings of available event descriptions, the project explored other means of description. Laboriously adding manual Meta-data descriptions after the event, was deemed infeasible to maintain, due to a lack of human resources. Other projects referenced, that had relational or inter-textual ontologies also largely had cumbersome manual tagging systems. This is a particular problem in the instance of this archive, due to the volumes of content produced. It has therefore been an aim of the project to develop a system with a layer of semi-automated features in order to alleviate manual updating tasks.

The archive consists of live, onsite events with live audiences. In addition to the representation of artists and other speakers, the archive offers rare documentation of the general public’s interaction with those artists and thinkers as they present their material. Here questions and responses to the speakers are very important raw materials, that may be of interest to the work of curators, educationalists and other researchers working in the areas of anthropology, sociology, cultural studies, art history etc. Such documentation would enable the study of how art and aesthetics were discussed and perceived by the public via an analysis of the question and answer sessions. These changes in public perception could be charted over a period of time and be coupled with the analysis of usage statistics, enabling an assessment of how public perception changed and engaged with curated programmes in context of wider trends in thinking.

Audience Analysis

Surveys were undertaken of different user groups' perceptions of the archive. Audience trends and user profiles were developed; based on server statistics, an online audience survey, direct dialogue with those engaging with the programme through website feedback, as well as in depth surveys on the use of the archive conducted with curators and post-graduate students.

The Surveys determined how the archive would be used and what language would be used for searches. It estimated the time that different user groups would spend using the tool, as well as suggested additional features desired by users from the interface.

Note that these summaries do not attempt to be comprehensive. The main purpose behind these outputs has been to refine requirements for functionality and interface, with regard to the use of the archive content by existing and potential audiences.

Audience Statistics and Trends

- In the year 2004 to 2005 use of the archive increased by over 300%
- During April 2005 a total of 9250 people visited the archive home page, out of which only 1710 actually played archived video clips. Those 1710 people collectively played a total of 23402 video clips across the month.
- The Tate Online Events Archive demonstrates an immense propensity for repeat visitors.
- The average viewing time for event archives is 30 minutes.
- Archives are accessed 24 hours a day, 7 days a week, 365 days of the year
- Geographical statistics suggest that audiences are diversely located, with the largest audiences, based in the UK, USA, Canada and Australia. Western Europe also appears heavily populated, closely followed by Northern and Eastern Europe. Asian audiences spread right across the continent; however they are often listed as individual users in specific locations, and not clustered like audiences in the UK, USA and Western Europe. This is also true of audiences from Africa, South America and the Middle East, who represent only a small portion of the overall audience.

The project investigated existing and potential audiences, in order to provide preliminary user profiles and a search case study, which in turn served to inform functionality requirements.

User Profile Summaries

Academics and Students are dedicated repeat visitors. They are sometimes live event attendees or in one way or another involved in the events themselves. In some cases students may review archives for events they could not otherwise attend. They are also good advocates for the archive, quoting references in papers and lectures and redistributing URLs to their students and peers. A unique attribute of this audience is that they are geographically clustered around institutions. Students are most likely to spend long hours searching and viewing content and may watch the same event more than once in order to quote effectively from an event.
Artists and Practitioners are sometimes involved in the events themselves. They may be positioning themselves into a peer group or listening to their mentors. Popular content includes other artists and curators talking first hand about their work and practice. Currently the group comes largely from a fine art background. This audience could expand to embrace more performance, live art, film, video and music practitioners, if the archive and its resources were communicated further and with more consistency into these sectors. Additional content such as Performance Documentation and Artist Interviews produced by Tate would be attractive additional content for this audience.

Professional Peers watch the programme closely but often do not have the time or inclination to view entire event archives, unless it directly relates to their current activities. They are on the mailing list and are keen to keep abreast of what Tate are doing, and how it compares to what they are doing. They travel regularly and are increasingly accessing email and the internet wirelessly or on mobile devices. They also often attend or participate in the actual live events. They may want to re-purpose archive resources.

A survey undertaken with Curators indicated that they, would use the tool mainly to search for artists or potential speakers, and would use speakers' names. Others would also use a concept based search or a specific artwork (title). Most suggest they would use specific words to search, not phrases. Most would spend between 5 - 15 minutes searching and up to 1 hour watching material. All would like to see biographies, bibliographies and additional references. The ability to save results was a popular feature. The main search site currently used by all is Google.

The *General Public* audience for the programme could be assisted by an interface that enabled users to explore and overview broad subject areas. This could be achieved by providing a lucky dip feature, a keyword cloud or a list of artists and presenters names. Content tasters should be visible upfront in the interface. The nature and depth of much of the content, appeals more often, to audiences with some existing knowledge of the topics. However the growth of this audience could be developed through an effective browse interface. Artist Talks and Interviews would be of interest to this audience. Curated selections from the archive may also be of assistance.

Search Case Study

Investigations into the nature and uniqueness of the archive's content, in conjunction with summaries of existing and potential audiences, has lead to a preliminary *Search Case Study*. This was necessary to support technical development. Further case studies will be produced in conjunction with future prototype versions.

This first case study is an attempt to capture expected results from a search query. The query used was < incarceration + new media + practice > as an example of a search, based on a broad concept in conjunction with an artistic form – from the perspective of an *Artist Practitioner Profile.*

This is just one example amongst many possible queries. It is expected that alongside technical development, these case studies will become increasingly refined, producing more accurate results with which to compare the automated results produced by the prototype system.

Benchmarking

Initial investigations involved benchmarking a range of cultural interfaces that engage the use of search functionality, or online multi-media and audio/video archives. The aim of the benchmarking was to compare and assess ways in which archives on these websites were presented, searched and navigated; through an assessment of their interfaces, search engines, structure and design.

Selection Criteria for websites in benchmarking document

1. Archival websites, containing audio-visual or multi-media materials on modern and contemporary art practice

2. Websites containing databases of information pertaining to audio-visual material or modern and contemporary art practice

3. Websites that had different methods of accessing the archives and databases, for example; semantic maps and category or keyword listings

4. Websites that were created by major institutions working in the field of modern and contemporary art or culture, such as Museums, Heritage Archives or New Media and Artists' Film & Video Collections

5. Many of the websites were part of museums or foundations, which had major art collections

Benchmarking Conclusions

Benchmarking of similar online archives, has brought about the following understandings and conclusions with regard to desirable features for functionality and interface:

- A scalable browser-like navigation system
- A customizable what's new / what's popular / or your folder 'welcome' interface
- A simple browser interface for non-registered users
- A playback interface
- A personalisation interface that includes technical, geographical, accessibility, general interest and keyword details associated with individual, registered users
- A shopping basket / personal folder interface, where search results can be saved, organized and returned to by individual, registered users
- A search interface with high level categorisations listed and a keyword field for domain researchers
- A search results listing, with expandable detail
- A personal assistant, adjunct to the search results interface, which if selected, provides users with further detail of related or associated events and subjects
- A printable summery of results selected and saved by registered users, with clear copyright information, in order to assist domain researches reference material effectively
- A help / FAQ interface
- Think maps or keyword clouds can aid both the user navigation as well as contribute to more streamlined interfaces for tagging material at the point of production.

Management and Workflow Context

Infrastructures and sustainable workflow methodologies that assist with the hosting**,** management**,** preservation**,** interfacing and distribution of these culture and heritage documents is a relevant concern for the museum sector. The ongoing management and maintenance of the archive is key to the feasibility of any future implementation. Therefore the project aims to integrate with wider infrastructure developments at Tate, in order to resolve effective preservation and distribution solutions that consider both internal and external use of the system.

The maintenance of the archive is problematic. To maintain by hand an index of the archive as it is currently being developed is not feasible. Therefore the usefulness of the archive is undermined by the difficulty of management on one end and the difficulty of navigability on the other. One of the primary goals is to develop systems that will aid in the management and maintenance of such archives. The maintenance will be enhanced by making very laborious tasks automatic

We aim to tackle both user navigation (search and interface) as well as internal management difficulties, making a section of the Tate Online Events Archive into a powerful resource for researchers. This resource will serve as a paradigm for later development in which the rest of the Tate Online Event Archive and associated audio/video content can be turned into a searchable well-indexed online media library.

The ambitious long-term goal of this research programme is to evolve a system that will semi-automatically fragment and tag multi-media material both at the time of production and at the time of delivery. In the scope of this prototype phase project, we will not be able to automate everything but we will create the infrastructure that will eventually lead to the automated creation of flexible multi-media archives.

Licensing and Copyright Considerations

Materials contained in the Online Events Archive are made available free from the Tate website for educational and non-commercial use only. The information, text, audio, video, performance, symposia and images (known collectively as the 'Content') are protected by copyright laws under the Copyright, Designs and Patents Act 1988, as amended 2003.

The majority of material is delivered in a streaming media format, which does not download onto an end user's hard disk. All rights to the content remain with the original authors, as licensed to Tate in a Contributors Agreement, whilst Tate hold copyright for the physical and digital recordings of that original content, as well as the collective works i.e. the Online Events Archive.

See here for a full copyright statement regarding Tate's Online Event Archive can be found at:
http://www.tate.org.uk/onlineevents/help.htm#copyright

The interface in development for this project, will consider the availability of clear copyright information, at various stages of the data mining process, in order to assist researches using the resource, handle and reference content appropriately.

In dialogue with Arts Council England and The University of Oxford, Tate prototyped models for the inclusion of downloadable content in the Online Events Programme, using a range of UK jurisdiction Creative Commons Licenses. In consultation with Creative Commons UK (http://creativecommons.org/) and with artists contributing to the programme, Tate developed a contract model that combines a tailored 'contributors agreement', similar to those used in open source software development, that works in conjunction with Creative Commons Licenses. The application of these licenses largely concerns the development of new content for the archive, rather than attempting to amend rights arrangements retrospectively.

Ontology

Key to the delivery of effective search results is the development of a vocabulary and ontology with which to accurately describe the content of the Tate Online Events Archive.

Existing ontologies were examined as part the benchmarking research. It is not the case that an ontology from one collection or archive can be transferred to another directly. The wide range of subjects and themes evident in the Tate Online Events Archive, as well as its continual transformation, as new events are added, requires that a suitable ontology encompass a wide cross-section of specialist and general vocabularies.

Investigation of ontologies used in art and cultural heritage websites, largely revealed the use of structures based on the categorization of an 'item' or 'object'. *The Getty* vocabularies, *The Dublin Core* and *CIMI* systems all encompass this type of structured classification. However, in comparison to the content of many other archives, the Tate Online Events Archive comprises time-based, audio and moving-images, the interpretation of which is different from artifacts.

Therefore the development of the ontology for the first prototype has drawn specifically from samples of the archive. The prototype ontology is concept lead, based on reasoning in relation to natural language and the interpretation of meaning from events such as artist talks or symposium discussions.

Ontology Development

Available descriptions of events contained in the archive were limited. Largely these descriptions consisted of marketing copy, written previous to the event. In order to make the content of talks and presentations available directly through the search engine, the project has aimed at producing structures and methods that help to define and describe the vocabularies of actual presentations.

In order to assist the system reason with the content of talks and presentations, the project took the direction of fragmenting individual events. Under the assumption that the system could then apply the knowledge and descriptions contained within the ontology to these *Event Fragments* and display the most relevant results.

Fragmentation

Event Fragments will ultimately serve as a navigation system for users to search through hundreds of hours of video content. The fragments will enable users to swiftly search and review select portions of events, which the system deems relevant to their query. This process allows users to preview a range of events quickly in order to make informed decisions about which 'whole events' are most relevant to their interests, before taking the time to watch the full proceedings. Not to underestimate the user's ability to make the most accurate decisions about what is relevant, this intelligent search tool aims to combine the knowledge of the system with the knowledge of the user, in order to produce accurate and sophisticated results.

At the same time, it is also important to recognize that we have users at two ends of the spectrum. Profiles consist of specialists with a high degree of existing domain knowledge and/or audiences who may have attended the actual event; who are now searching under the very specific theoretical language used in the talk itself, by referencing their own notes taken on the day.

However, profiles also consist of users with limited prior knowledge of topics. They may also be unfamiliar with Tate's Public Event programmes or have English as a second language. These users are more likely to search under broad subject areas using general terms. The *Ontology* and *Fragmentation* processes will assist the system deliver results for both the advanced search and browse features of the interface.

The Process of Fragmentation

There are several ways in which fragmentation can be approached. A template for fragmenting items in the archive has been developed in consultation with curators and in view of a detailed survey of sampled material. Enhancements will be implemented including the possibility, that the system could allow a user to move backwards and forwards from an automatically selected fragment, in order to assess the wider context in which the fragment is relevant. It should then be easy to move directly from that fragment to the next fragment, or to the entire event archive. It is important that all fragments are presented in context of information pertaining to the full event, such that the original and intended meaning is maintained and any references to the fragment may be framed accurately and appropriately.

Current Development

Stage 1: Themed grouping of events in archive - In order to select events and items for fragmentation and to assist the reasoning process, the different events were grouped into major themes. These were not intended to be all inclusive but to determine a primary level of commonality between events available in the Online Events Archive

Stage 2: Proposal of Fragmentation criteria and methodology in relation to duration and content in order to determine different formats, and different methods of fragmentation

Stage 3: Selection of events under the themes of Photography and Gender - The two themes one practice based the other concept based were selected in order to refine the selection of events and fragments for the first prototype and to develop reasoning methodology that works with the natural language used in the selected events.

Stage 4.: Selection of five hours of video material, made up of twenty minute sections, for the first prototype

Stage 5: Division of sections into 30 second (fixed -duration) fragments, for the first prototype. These are to be replaced with fragmentation based on sound-track pauses in the first place, and dynamic fragmentation (based on audio features and topic detection) thereafter.

Stage 6: Developing of reasoning methodology - via the development of the ontology, conclusions are drawn about the relation between concepts and the way they are expressed in natural language. These conclusions will form the basis for the creation of the tool, in the first prototype.

The project aims to evolve an ontological structure and vocabulary, which allows for an optimum level of lateral associations to be demonstrated. It is an ambition that the complex array of ideas and concepts illustrated in the archive content be represented in the resulting tool/system, in order to increase the use and viewing of archived events and encourage a deeper exploration of the interdisciplinary material.

Technical Development

The use of the archive as a research resource is limited by its poor indexing and the fact that the recordings are only available as long play files (0.5 – 2 hours in duration), their descriptions giving no indication of where, on a stream, to find a particular subject.

The main, long-term aim of the project is to develop a research tool/system that would provide a personalized environment whereby domain researchers can *find*, *discover, organize* and *interpret* relevant material from the Tate Online Events Archive. The novelty of this research consists in the employment of artificial intelligence techniques whereby the system becomes an *assistant* to the domain researcher, rather than a simple searching tool.

The system will be able to extract (from long recordings) only those sections that are relevant to a user's current interest. Furthermore, the system will be able to *personally assist* users, *directing* and *guiding* the reSearch activity with 'intelligent' guesses and suggestions.

Research and technical areas that have defined and informed the project include; multi-media archives (in particular from the cultural sector), indexing audio-visual material, intelligent search, ontological engineering, semantic web and knowledge-based systems.

Functionality and Intelligent Behaviour Modes

The (re)Search Tool will provide the user with the following basic functionality:

- searching, for users who have well defined retrieval intentions
- browsing, for users who want to explore the content of the archive without prior knowledge
- retrieval of relevant results and playback
- selecting results and organising them for subsequent use

This will be realised via a set of "intelligent tools" that will support:

- in-depth searches with multiple query modes
    - intent capture via talk-back mechanisms
    - long search processes (e.g. days)
    - semantic maps that further enable thematic searches

- tailoring results to the (assumed) users' interests
    - trim irrelevant items that result from a query string
    - augment with items that do not result directly from the query string

- explicit provision of guidance (mainly in searching, browsing, collecting and playback)
    - e.g. in the form of suggestions of further relevant material or queries

- provision of samples for motivating/inviting users to browse events
- provision of context dependent snapshots with different levels of detail
- organization of results and/or query strings in personal repositories

The "intelligent behaviour" will be implemented via a *domain reasoner*, that will hide from the user, most of the internal representations evident in the system, providing in the interface the appearance of *a personal assistant* to the archive.

Initial prototype

The initial implementation focuses on employing *ontology-based reasoning* for enhancing the quality and efficiency of the information retrieval component.

In contrast to many existing multimedia, information retrieval systems, our research focuses on searching the actual *content*, rather than searching conventional metadata (fields such as "artist", "date", etc). We go beyond retrieval of entire documents (i.e. filmed events), by extracting only those *fragments* that are most relevant to a user's interests. This is carried out via three processes (not necessarily independent):

- fragment definition and description;
- query refinement; and
- matching (via a relevance function/metric)

The domain ontology plays a major role in these processes.

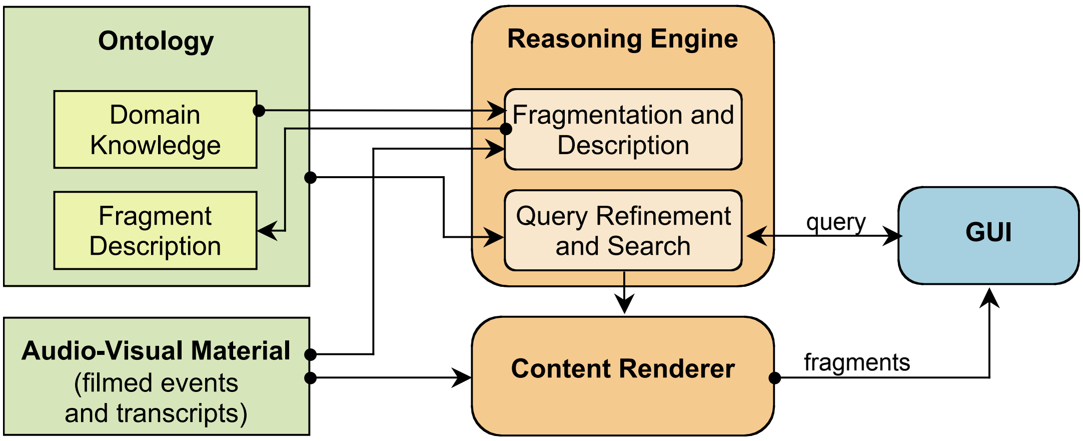
For the majority of the Tate Online Events Archive's material, spoken text is of primary importance. Spoken text compared to written text:

- is less structured and often involves spontaneous speech, thus becoming more difficult to fragment;
- includes cues such as accents, intonations, pauses, which are useful for fragmentation

In the current prototype retrieval is performed on transcribed text, extracted from the videos manually, but automatically in future versions. Since the text originates in spoken language, we argue that our retrieval scenario is significantly different from the mere textual one. Dealing with unstructured text, both on the thematic/argumentative level (e.g. speakers may jump between different subjects) and on the lexical level (the choice of words often differs from well-planned wordings), requires extensive application of world knowledge for meaning capture. This role is fulfilled by our approach to the ontology and our ontology-based reasoning mechanisms.

In the initial prototype, fragments are of fixed length. Indexing is done on the basis of the (key)words present in the transcript. We aim to improve *recall* by augmenting fragment indexing/description, so as to include related terms as well as the original ones. *Precision* is improved through query refinement, ambiguity resolution, and a relevance metric.

System architecture of the current prototype


Ontology
Domain Knowledge
Fragment Description
Reasoning Engine
Fragmentation and Description
Query Refinement and Search
query
GUI
Audio-Visual Material (filmed events and transcripts)
Content Renderer
fragments

Fragmentation and search are not concurrent processes. Fragmentation is an "off-line" process, carried out independently from the queries posed by the users. It uses the transcripts of the archive material and the domain ontology. It is completely transparent to users.

The queries posed by users via the GUI (in either searching or browsing modes) may be refined on the basis of the ontology. The results are compiled from the ontology (note that the fragments' description is stored as part of the ontology) and then passed onto the renderer, which will extract the actual content and send it to the GUI.

Simplifications considered for the initial prototype

- Manual transcripts (initially 5 hours of material)
- Fixed-duration fragments (30 seconds long)
- All material presented locally on disc

Examples for recall and precision enhancement

- Ontology-based indexing *enhancement* increases *recall*

e.g. "**gender**" appears in query --> fragments including "**male**" or "**female**" will be *added* to results

- Ontology based query *disambiguation* increases *precision*

e.g. "**race**"+ **"Yinka Shonibare**" in query --> "**athletic race**" will be *excluded* from results

Further steps within the second development phase

- Replace fixed-length fragmenting with variable-length, intelligent fragmenting. Fragmentation based on audio features (such as pauses) will be used in the first place, and ontology-related topic identification will be included afterwards.
- Further elaborate ontology-based reasoning
- Incorporate speech-recognition technology to replace manual transcriptions
- Move to client-server architecture and allow seamless integration with the existing Tate interface

Replacing the fixed-length fragmenting by an intelligent variable-length fragmenting method is deemed central to the second phase of prototype development. The task is not a trivial one, because fragment definition and description have to happen concurrently (what a fragment is about cannot be inferred until the fragment has been defined; which are the relevant fragments cannot be identified until a description is given). Our ontology based solution resolves this issue. The ontology serves as a common denominator between querying and fragmentation. The ontology defines topics. These are used in identifying atomic fragments. The topics are also used in query refinement. The matching algorithm aggregates atomic fragments as results to queries, calculating a relevance factor at the same time.

Multi-media Standards and Formats

The majority of content currently exists in Real Media format. Fragmentation and playback functionality development for this proprietary format presented some immediate problems. This issue is further emphasized by the wider range of content appearing in a number of different formats. A comprehensive format solution which avails itself to parallel open source developments and additional concerns with regard to preservation and multi platform delivery will be investigated further.

Dissemination

Secondary goals of the project include laying out the process for generalizing Tate's intelligent archive prototype into an open source, off-the-shelf software system and to disseminate the archive and methods into wider communities, including collection agencies, archives and distributors who hold material that is similar either in content or medium. And further, the dissemination of both source code and method, to the wider computer science and software developer communities.

References

Project Stakeholders

http://www.tate.org.uk Tate
http://www.goldsmiths.ac.uk/departments/computing/creative.html Goldsmiths College Department of Computing
http://www.ahrc.ac.uk/apply/research/sfi/ahrcsi/ict_in_arts_humanities_research.asp Arts and Humanities Research Council

Cultural Ontologies and Metadata

The Role of the Museum in the Digital Age & Collecting and Valuing Digital Content V2, 2003
"With the increasing availability of data and new forms of structuring digital archives, meta-data have become social, political and economical important instruments in an information sphere long considered value-free."
http://archive.v2.nl/v2_archive/projects/capturing/1_3_metadata.pdf
http://www.v2.nl/Projects/capturing/
http://framework.v2.nl/archive/archive/node/work/default.py/nodenr-130659
http://framework.v2.nl/archive/archive/node/work/default.xslt/nodenr-124777 Secure multi-media retrieval

Ontology Versioning on the Semantic Web
"Ontologies are often seen as basic building blocks for the Semantic Web as they provide a reusable piece of knowledge about a specific domain. However, those pieces of knowledge are not static, but evolve overtime…"
Michel Klein and Dieter Fensel, Vrije Universiteit, Amsterdam

http://www.knowledgesearch.org *Semantic Indexing - Creating tools to identify the latent knowledge found in text* and http://software.newsforge.com/article.pl?sid=06/09/19/1531258&from=rss Aaron Coburn, lead developer of the *Semantic Indexing Project* at Middlebury College

http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html *Ontology Development 101: A Guide to Creating Your First Ontology* by Natalya F. Noy and Deborah L. McGuinness, Stanford University

http://www.getty.edu/research/conducting_research/vocabularies/download.html Getty *Art and Architecture* Vocabularies
http://www.getty.edu/research/conducting_research/standards/ Getty Data Standards and Guidelines
http://www.cultos.org/ *Cultos*, Multi-media knowledge management for culture and arts
http://www.leeds.ac.uk/cedars/guideto/metadata/ *The Cedars Guide to Preservation Metadata*, March 2002,
http://web.nwe.ufl.edu/~mnorcia/articles/ArchiveFeverInstitution.html Norcia, Meg, *Institutionalising the Archive*, 2001
MPEG7 to describe multimedia in Museums and http://archive.dstc.edu.au/RDU/staff/jane-hunter/ECDL2/final.html *The application of Meta Data standards for video indexing* by Jane Hunter.
http://www.thinkmap.com/ *Think Map* – data visualization

Digital Culture

http://www.digicult.info/pages/Themiss.php *Technology Challenges for Digital Culture*
http://www.dlib.org/dlib/june03/miller/06miller.html *Understanding the international audiences for digital cultural content*
http://www.culturalcontentforum.org/intro.html Digital Cultural Content Forum
http://www.fondation-langlois.org/e/activites/zidarich/zidarich.pdf *Virtual Worlds as Architectural Space*

Preservation

Asset Management Integration of Cultural heritage In The Interexchange between Archives
"Digital Preservation: (fragile media) Once digitized content is often thought to be immortal. This is a popular fallacy: Bits are ageing too. Due to its medium, digital contents are exposed to degradation too, because the physical media e.g. disks or tapes are degrading. In contrast to analogue recordings where the degradation may be visible or audible digital recordings are degrading stealthy. Suddenly bits flip from a "1" to "0" or vice versa or they drop and are unreadable then. If such a bit error hits vulnerable areas of digital recordings such as file allocation tables a whole bunch of assets may get lost…"
http://www.amicitia-project.de

Permanence Through Change: The Variable Media Approach
"Entropy requires no maintenance. Entropy has its own poetry: it's all about delamination, disintegration, deteriation, degradation, decomposition and doddering decline." Depocas, Alain, Ippolito, Jon and Jones, Caitlin (editors) http://variablemedia.net/pdf/Permanence.pdf, 2003
http://variablemedia.net/e/welcome.html The Variable Media Network

http://www.tate.org.uk/research/tateresearch/majorprojects/mediamatters/ *Media Matters*, Tate / MOMA
http://rhizome.org/artbase/report.htm Rinehart, Richard, *Preserving the Rhizome ArtBase*, September 2002

Multi-media Standards and Formats

http://www.dstc.edu.au/Research/maenad-ov.html *Multimedia access across enterprises, networks and domains*
http://xml.coverpages.org/mpeg7.html mpeg 7 standards
http://www.dlib.org/dlib/september99/hunter/09hunter.html mpeg 7 overview
http://archive.dstc.edu.au/RDU/staff/jane-hunter/MW2002/paper.html *Combining CIDOC CRM*
https://www.helixcommunity.org/ Helix *Community*, Real Media developer site
http://www.jpeg.org/jpeg2000/ and http://www.jpeg.org/apps/culture.html *Jpeg 2000* for still and moving image preservation and distribution

Open Source Archives

http://www.dspace.org/ *dspace,* Digital Repository System Open Source Community, MIT
http://www.openarchives.org/ *Open Archives* Initiative
http://www.research.ibm.com/dx/ *Open DX* – open visualization data explorer

Community Archives and Networks

http://www.flickr.com/ *Flickr*
http://www.youtube.com/ *Youtube*
http://www.archive.org/index.php *Archive.org*
http://www.myspace.com/ *MySpace*

http://marcel.wimbledon.ac.uk/intro.htm *The Marcel Network*, Multimedia Art Research Centre
http://www.netzwerk-mediatheken.de/index_en.html Network of Multi-media Resource Centre

Culture and Heritage Interfaces: Video, Archive and Search functionality

Montevideo – Netherlands Media Art Institute Artists' Film & Video Online Search
http://netzspannung.org *Netzspannung*, developed by The MARS Exploratory Media Lab
http://www.imk.fraunhofer.de/mars
http://www.unesco.org/webworld/portal_archives/pages/Archives/Audiovisual_Archives/ - *Unesco* Archive Porthole
http://archive.v2.nl *V2 Institute for Unstable Media* – project archive
http://www.artandarchitecture.org.uk/ *Art and Architecture* Online Archive, Courtald Institute
http://rhizome.org/artbase/ Rhizome Artbase
http://artport.whitney.org/commissions/idealine.shtml Whitney Idealine
http://www.fondation-langlois.org/flash/e/index.php?Url=CRD/search.xml Daniel Langlois Foundation
http://www.aec.at/en/archives/navigator.asp ARS Electronica
http://www.britishpathe.com/index.html British Pathe
http://channel.walkerart.org/archive.wac Walker Art Gallery webcast archive
http://www.debalie.nl/dossierpagina.jsp?dossierid=10123# DeBalie, Media Archeology
http://on1.zkm.de/zkm/e/institute/mediathek/ ZKM Media Library
http://creativearchive.bbc.co.uk/ BBC Creative Archive
http://www.artsconnected.org/library_archives/
http://socks.ntu.ac.uk/dpa_search Digital Performance Archive
http://channel.walkerart.org/ Museum Education online distribution
http://www.fabchannel.com Paradiso music programme, The Netherlands

Audio Archives: Cultural Contexts, Audience Interfaces and playback solutions

http://www.ubu.com/ *ubuweb*, archive of sound poetry
http://www.mediascot.org/drift/html/index.php *Drift*, New Media Scotland: mp3 Archive - broadcasts and commissions
http://www.warprecords.com/?mart=WAP175 Warp Records
http://www.thewire.co.uk/ *The Wire*, video interviews, audio recordings, transcripts of talks and mp3 downloads
http://www.resonancefm.com/archive.htm *Resonance FM* streaming media and mp3 archive
http://on1.zkm.de/zkm/institute/mediathek/ideama/memorandum The International Digital Electroacoustic Archive
http://www.wps1.org/ WPS1 PS1 MOMA live radio webcasts
http://www.ps1.org/cut/tours.html *Bed of sound* playlist / playback pop up
http://www.frequencyclock.net Self Curating Net Radio

Related Online Events

Preliminary research undertaken in view of this project and wider cultural debates, began raising questions about how we select, interpret and present cultural artifacts and associated documents. Challenging traditional assumptions about the role of the curator and audiences, in an information and online content ecology.

Curating, Immateriality, Systems (http://www.tate.org.uk/onlineevents/archive/CuratingImmaterialitySystems/)
Tate Modern June 4 2005, A conference/webcast on curating digital media and immaterial culture.

c0dE 0f practice (http://www.tate.org.uk/onlineevents/archive/code_of_practice/) Tate Online June 13 - July 18 2005, An online panel discussion: How do we identify, sort, search and locate ourselves amidst the dynamic instability of immaterial culture and its artifacts?

Open Congress (http://www.tate.org.uk/onlineevents/archive/open_congress/) Tate Britain October 7 - 8 2005, An Open Congress that seeks to understand how methodologies derived from Free/Libre and Open Source Software [FLOSS] production can be deployed by those working in the area of art, and visual culture.

d_culture (http://www.tate.org.uk/onlineevents/archive/d_culture/) Tate Modern / Tate Online January 28 – March 31 2005, An online season of downloadable sound and debate, which prototyped new content and contract models, incorporating creative commons licensing for the redistribution and use of audio tools and resources. Audience data analysis in conjunction with this project, further assessed the tendencies of peer to peer exchange and the effect of viral distribution.

Print Publications

Benjamin, Walter, ‘The Work of Art in the Age of Mechanical Reproduction’ in Illuminations, London, Jonathan Cape Ltd, 1970

Caygill, Howard, ‘Meno and the Internet: between memory and the archive’, History of the Human Sciences, Vol. 12, no.2, 1999

Derrida, Jacques, Archive Fever, University of Chicago Press, London, 1996

Doane, Mary Ann, The Emergence of Cinematic Time – Modernity, Contingency, The Archive, Harvard College, USA, 2002

Foster, Hal, ‘Archives of Modern Art’, Design and Crime (and other diatribes), Verso, London, 2002

Foucault, Michel, ‘The Statement and the Archive’ in The Archaeology of Knowledge, Routledge, London, 2003

Heidegger, Martin, ‘The Question Concerning Technology’, from The Question Concerning Technology, New York, Harper & Row, 1977

Hooper-Greenhill, Eilean, Museums and the Interpretation of Visual Culture, London, 2002

Kenney, Anne, Rieger, Oya, Moving Theory into Practice: Digital Imaging for Libraries and Archives, Cornell University Library, California, 2000

Lynch, Michael, ‘Archives in formation: privileged spaces, popular archives and paper trails’, History of the Human Sciences, Vol. 12, no.2, 1999

Manovich, Lev, The Language of New Media, MIT, Cambridge, 2001

Osborne, Thomas, ‘The ordinariness of the archive’, History of the Human Sciences, Vol 12, no.2, 1999, pp51-64

Rose, Jacqueline, ‘The Archive’ in The Haunting of Sylvia Plath, Virago Press, London, 1991

Townsend, Sean , Chappell, Cressida, Struijvé, Oscar, Digitising History – A Guide to Creating Digital Resources from Historical Documents, Oxbow Books, Oxford, 1999