

## COMPLEXITY AND SCALE IN AUDIO ARCHIVES

Jerry Goldman and Andrew Gruen

"I've been rich and I've been poor. Believe me, honey, rich is better."  
--Sophie Tucker

National archives and public broadcast archives in Europe and the United States hold millions of hours of spoken-word materials, the bulk of it in analog form. In 2005, a EU-US working group estimated that world holdings of audio materials in analog formats are on the order of 100 million hours. They will perish within a few decades unless we take steps to preserve them. Millions more hours come into existence in digital formats each year. Accelerating growth in spoken word documents will generate demand for efficient archiving and retrieval strategies. But these resources will prove stillborn if we do not identify ways to reveal their contents.

Our aim is to offer an approach to scale and complexity issues as we understand them in spoken word collections by relying on our experience with Supreme Court audio, which is metadata-rich. We will begin with the work we know well, suggest a step or two into the future, and then attempt to grasp the much larger issue of metadata-poor audio.

When the US Supreme Court installed a recording system in the courtroom in October of 1955, the resultant recordings were primarily used to assist justices and their law clerks. However, the court also archived the recordings at the National Archives and Records Administration for research and academic study. Today, the Archives hold about 9,000 hours of Supreme Court audio. Although researchers could physically visit these reel-to-reel recordings, the rise of on-demand audio streaming over the Internet made this system antiquated.

The court's written transcripts capture some of oral argument's content; hearing Rehnquist's coughs and voice deteriorate over his last year on the bench provides an additional layer of information for researchers. Although the Oyez Project <[www.oyez.org](http://www.oyez.org)> began as a way to deliver this audio to interested citizens in digital form, it is now much more. The project is now an attempt to prepare future audio archives of all types to store and deliver their holdings in a networked world.

At its inception the Oyez dataset was small. The project digitized recordings of the oral arguments from key cases in US constitutional law, wrote abstracts of each case and put them in a HyperCard stack. However as the project grew it began to directly confront issues of both scale and complexity. In addition to a streaming recordings and placing abstracts on the Web, Oyez began to collect other pieces of text and audio, along with metadata for both. The project also started to collect audio, text, photos, and videos that are related to the operation of the court but not to any individual case. The generation of multiple versions of audio items further complicated the task of curating the growing collection of Oyez materials.

The latest iteration of Oyez, version 5, contains three main data types: text, audio, and various forms of audio metadata including transcripts, time-synced transcripts, speaker information, speaker biographies, speaker photos, annotations and commentary.

As data collection drastically increased Oyez started looking for computational curation aids (instead of hiring more undergraduates). The project is turning to RDF in an attempt to improve both internal organization and to "future-proof" the dataset.

Oyez built an RDF schema that describes the structure and concepts within the data. Unlike building traditional taxonomies, creating the RDF schema was like building a waterfall from the bottom up; it started with the smallest constituent parts and then interrelated them into larger categories. As Oyez's focus is, primarily, to archive the US Supreme Court, and the work of the Court is broken into cases, the Oyez schema uses the individual case as its starting point. Cases are made up of events, people, and the roles those people play in each event. Each constituent part of all the case categories are described – there are many kinds of events, roles specific requirements, and people have names – but more importantly all the types are interrelated – events are tied to specific cases, some events require particular roles to be present, and people fill roles for any given case.

The benefits of RDF for the project are twofold. First, it should make Oyez data machine "understandable," that is it should add semantics to indicate, for example, that a string of text like 10-22-1956 is a date. Second, the schema will be publicly accessible and, thus, freely extensible. By marking up Oyez material with semantic metadata, researchers can begin to ask questions of the data where the answers are held implicitly. For example, although it has never been the project's aim to discuss the effects of aging on appellate

justices – and that data is not stored in the Oyez database explicitly – by using RDF, scholars can ask to look at trends in decision making of every justice when they were between the ages of 56 and 61. Because birthdays and the dates of events are known as dates, software can automatically locate records within the age range and produce meaningful results with minimal input from the researcher.

Because RDF schemas are public and extensible, students of related fields can both peer into how Oyez chose to organize its data and build upon the project's pre-existing structure. In the future, a congressional researcher could extend the above aging study to include both the Oyez data and other datasets, again without the need to have planned her data sets with such a study in mind. At the initiation of a new study, a scholar can compare two schemas, quickly identify points of comparison at the organizational level rather than a record level, and let a piece of software create a new dataset. As schemas begin to reference across domains, all other semantic metadata becomes more valuable because a piece of software can make inferences that were previously only found by extremely rare (and extraordinarily talented) interdisciplinary researchers.

Unfortunately a critical mass of RDF-organized information has not yet been reached. However, large academic research projects like Oyez can make a significant impact in defining the field. Academic archives can build the core of a Semantic Web by wrapping their content in semantic metadata and working together to develop interdisciplinary schemas.

Oyez is a metadata-rich prototype that may pave the way for others. But we also need to identify strategies that can tease semantic meaning from metadata-poor audio (the more typical archival object). Consider this example provided by Ant Miller at BBC-Archives:

Let us imagine an archive that has a set of audio assets about which we know very little. Perhaps the holdings are old, inherited or just 'found'. Chances are that before long the institution will have to digitize these holdings, and digitization means that the metadata will be essential-metadata the institution does not have. Once digitized as files, these assets will be good as lost without metadata.

One could, given sufficient resources, do some human detective work. Listen to the content, annotate it, transcribe it, try to find provenance, cross reference, index etc. In short, catalog it. From hard experience, most archive institutions know that done well this is expensive time-consuming work, and frankly, there may not be the time for these orphaned holdings

Or perhaps there is a way to use a combination of automated technologies and the vast accessible information resources of the internet to do some of what cataloguers do? We think in some domains and subject areas, this is possible, and here's how one can try.

Step one -- play your content. Digitize it now, and store it locally. Give it a global unique ID for your system. We think that it's a good idea to start building a metadata set right now- include all you can about the original carrier, the process of digitization (there's often a surprising amount of data available), and the organizational information associated with the asset- always a good idea to start a metadata set young we think!

Step two -- Speech to text. This can be a challenging area, and if you're looking for usable readable transcripts, often your results will be disappointing, but for the purposes of this process, 70 to 80% is fine. And better than that (which is possible) is an improvement that might be worthwhile. That transcript will be the key.

Step three -- Cross-reference. Take your transcript and cross-reference it to textual sources that are roughly contemporaneous with the assets. News library holdings, other transcripts, editorials, even literature referring to the period- all can be useful, because a proximal set of matches is what you're looking for.

Step four -- structure your matches. These proximal matches can be used within a taxonomic structure, and that's what real world cataloguers using something like UGC. But if you're going to be really smart, use an ontology. By using a semantic structure with meaningful relationships between the terms you can begin to extrapolate a set of useful reference terms beyond your original matching set. Berlin, Blockade, Tempelhof. Three useful terms. But what if you know, in your Ontology, that Tempelhof is an <airport> in the <city> of Berlin, <capital<devided>> of the <country> of <germany> that was used in the <airlift> to relieve the <blockade>? Now you can, with a good degree of confidence, build a much stronger set of key terms associated with your asset- you can place it within a richly structured information space.

Step five -- use your new index - here there are many options. They revolve around how rich and smart you want your catalog to be (and how impenetrable for untrained librarians), and how much you want your vast and now well indexed holdings to be accessible to the Google generation. Take these finely crafted ontological structures and flatten them to tags on MPEGs? It's an option- and delivers a powerful tool for now. We suggest you keep your ontology too, build it, grow it, try OWL and RDF implementations, and enjoy.

Either way now your orphaned audio can be referenced via words that were never recognized, maybe never even said, in the original.

The challenges are enormous but that's what makes some of us climb mountains or dive wrecks or make millions of hours of spoken-word collections accessible and useful in a data-rich but metadata-poor world.