PRESERVAÇÃO DE *CORPORA* DE FALA DO PORTUGUÊS BRASILEIRO: DIGITALIZAÇÃO, ACESSO E SEGURANÇA

BRUNO SILVÉRIO DE FREITAS
FLÁVIA CARNEIRO LEÃO

INTRODUÇÃO

O Centro de Documentação Cultural "Alexandre Eulalio" (CEDAE) é um órgão vinculado ao Instituto de Estudos da Linguagem (IEL), da Universidade Estadual de Campinas (UNICAMP), e foi criado há 30 anos como um "*um espaço que apresentasse condições apropriadas para a organização e conservação de materiais produzidos em pesquisas e projetos realizados pelos docentes do Instituto de Estudos da Linguagem*".

Assim, ao longo de seus anos de existência, o Centro incorporou ao seu acervo outros fundos documentais de grande importância para o estudo da literatura e da linguística brasileiras, merecendo destaque as coleções resultantes das pesquisas dedicadas às diversas manifestações das línguas faladas no Brasil.

Neste sentido, podemos afirmar que o acervo do CEDAE conta com alguns dos mais importantes levantamentos fonográficos constituídos para fins de pesquisa no Brasil, merecendo destaque: os levantamentos efetuados no contexto do projeto de "Estudo da Norma Urbana Culta do Português Brasileiro", as gravações produzidas pelo projeto de "Aquisição da Linguagem Oral" e as entrevistas geradas pelo projeto "Os Negros do Cafundó: Linguagem e Comunidade", que registram elementos remanescentes da língua banto presentes na fala de membros da comunidade quilombola do Cafundó, além de uma série de materiais que documentam línguas indígenas brasileiras extintas ou em vias de extinção.

Além disso, estão depositadas no CEDAE coleções de interesse preponderante para o conhecimento da cultura brasileira e da história recente do país, como os documentos do "Programa Certas Palavras" que registram entrevistas de personalidades expressivas da vida intelectual do Brasil contemporâneo.

Assim, longe da ocorrência esporádica em algum fundo ou coleção do acervo, os documentos sonoros constituem parte expressiva da documentação preservada pelo Centro e, deste modo, ações voltadas à sua preservação estão na origem do presente trabalho.

O CEDAE E A PESQUISA LINGUÍSTICA

Sendo a linguagem um objeto empírico escondido na mente humana, é impossível o desenvolvimento de pesquisas de descrição e história linguística sem que se recorra aos Centros de Documentação que, para os pesquisadores da área de Humanas, são como laboratórios.

Deste modo, como órgão vinculado ao Instituto de Estudos da Linguagem da UNICAMP e criado com o objetivo de preservar os materiais produzidos pelos docentes do IEL, o Centro hoje abriga um conjunto significativo de documentos sonoros que

registram a língua falada (português-brasileiro) e que são de grande relevância aos estudos linguísticos.

Registrar a língua em documentos escritos é um velho ofício, qualquer que seja a cultura a que nos referirmos e deste modo, a língua escrita, expressa em documentos textuais, vem sendo usada não só na composição de gramáticas, como também como fundamento de estudos diacrônicos.

Quanto à língua falada, somente a partir dos anos 70, com a invenção de gravadores portáteis, foi possível registrá-la, o que fez surgir por toda parte projetos de documentação e análise dessa modalidade.

E agora, com a *Internet*, parece que se esboça a experiência de mesclar a língua escrita à língua falada, produzindo-se documentos aparentemente de uma nova natureza, mas esta é matéria ainda para ser amplamente examinada pelos linguistas.

A CARACTERIZAÇÃO DA DOCUMENTAÇÃO SONORA DO ACERVO

Buscando atender à demanda por acesso aos documentos que compõem os *corpora* sonoros do acervo, procuramos conhecer os principais aspectos relacionados ao seu tratamento. Reunimos conhecimentos sobre o assunto, circunscrevemos as demandas, identificamos limitações, para então definirmos estratégias de ação, estabelecermos procedimentos, dimensionarmos recursos e equipamentos necessários.

Dessa maneira, começamos por mapear os documentos sonoros do acervo, seu volume, suas características e estado de conservação e, deste modo, constatamos a existência de uma grande variedade de formatos de fitas magnéticas analógicas.

Tais fitas apresentavam configurações distintas no que se refere aos formatos, dimensão dos carretéis (com diâmetro entre 12,7 e 18 cm), constituição física, tecnologia utilizada para o registro do som (captação/audição), velocidade empregada (captação/audição) etc.

Percebemos, também, que a alta frequência de audição das fitas, gerada pela elevada demanda de acesso, era um dos fatores a gerar danos ao material, seja pela abrasão gerada pelo contato com o mecanismo dos equipamentos, como pelo manuseio.

Assim, no que se refere à conservação dos áudios, além dos fatores acima apontados, identificamos outros fatores de degradação relacionados aos materiais constitutivos das fitas, como o acetato –utilizado na produção de fitas de rolo magnéticas entre 1948-1965–, por exemplo.

Além disso, diante da obsolescência das tecnologias empregadas para o registro de áudio, a dificuldade de contar com equipamentos adequados à sua audição foi mais um dos aspectos avaliados e que limitava o uso dessa documentação.

Some-se ao exposto, a demanda por acesso simultâneo a um mesmo documento, restrição que gerou ao longo dos anos um grande volume de reproduções analógicas (em fitas cassete), com vistas à ampliação da possibilidade de acesso. Longe de resolver o problema, esta iniciativa ampliou o volume do acervo e instituiu a necessidade de agendamento para a consulta, pois mesmo dispondo de cópias para acesso, a falta de disponibilidade de equipamentos suficientes para audição ainda era um fator limitador.

A partir dessas constatações passamos a considerar a digitalização como uma estratégia para o acesso seguro aos documentos, para a ampliação desse acesso e para a sua conservação.

Assim, decidimos estabelecer critérios para o desenvolvimento desse trabalho, priorizando os documentos que:

• apresentavam risco ou que
• estavam vinculados a algum tipo de sistema/tecnologia sem suporte comercial/obsoleta, e/ou que
• tivessem uma demanda de consulta elevada.

De acordo com esses critérios, iniciamos o trabalho pelo material mais antigo, ou seja, pelas fitas de rolo analógicas que, como se supunha, apresentavam a maior variedade, seja de formato, de constituição, como de velocidade e sentido de gravação e, desta maneira, as cassetes, os mini cassetes, etc. foram deixados para a segunda etapa do projeto.

Os equipamentos para a captação do som das fitas foram igualmente analisados, tendo-se como referência a dimensão dos cabeçotes de leitura, a capacidade de operar em sentidos invertidos e em velocidades específicas, sendo que a conclusão desta análise apontou para a inadequação dos equipamentos disponíveis, que apresentavam desgaste de componentes, falta de manutenção etc. Além disso, o reparo ou a aquisição de "novos" equipamentos somou-se às demais dificuldades.

A CONSERVAÇÃO DAS FITAS DE SOM MAGNÉTICAS

Desde o seu aparecimento os métodos de gravação analógica têm gerado uma série de materiais e suportes que, em maior ou menor grau, estão sujeitos à degradação e consequentemente ao risco de perda de informações.

Considerando que o principal paradigma da conservação de documentos sonoros é a "preservação do original" e que os suportes sonoros têm uma expectativa de vida curta, a conservação passiva, que protege os documentos, sobretudo dos agentes ambientais, é insuficiente e, deste modo, a sobrevivência do documento só pode ser "garantida" se renunciarmos à sua materialidade, através de um processo contínuo de transferência de informação para novas mídias. (Sterne, 2003, p. 323-325).

Se a estabilização do processo de degradação é tida como a principal meta na conservação de documentos sonoros, a operação de transferência do som do meio analógico para o digital é uma ação conservativa que promove uma mutação midiática do documento e, neste sentido, digitalização e preservação, por vezes, se sobrepõem, uma vez que a digitalização é parte integrante do processo de conservação. (Schüller, 2003, p. 259-260).

Em contraposição, existem razões para não digitalizar esses materiais, como o rápido desenvolvimento da tecnologia, que inevitavelmente leva à obsolescência do hardware[1], dos formatos digitais e das mídias/suportes.

Isso nos leva a pensar que a digitalização poderia ser uma solução definitiva para os problemas de conservação; no entanto, é de se esperar que novos formatos apareçam na medida em que o desenvolvimento científico e tecnológico avança e que novos materiais, assim como novas tecnologias, nos tragam novos problemas.

Por outro lado, embora existam recomendações nacionais e internacionais voltadas à digitalização de documentos históricos, não encontramos um padrão universalmente aceito para os valores básicos a serem utilizados na digitalização de documentos sonoros, como o formato de gravação, a profundidade de *bits* e a taxa de amostragem. E, finalmente, é sabido que as mídias digitais, que derivam do material digitalizado da mídia analógica, têm uma expectativa de vida muito baixa, quando pensamos em longa permanência.

[1] *Hardware* é um conjunto de unidades físicas, componentes, circuitos integrados, discos e mecanismos que compõem um computador ou seus periféricos.

Além disso, se a transição do analógico para o digital pressupõe a perda de dados, com a digitalização, no entanto, a produção de cópias do digital para o digital reduziria este problema, pois uma vez digitalizados, os originais ficariam então protegidos.

Cientes de que a digitalização é sobretudo um modo de dar acesso a materiais raros, em risco de desaparecimento e distantes geograficamente, e não apenas uma solução permanente para a preservação, decidimos pela digitalização das fitas de rolo analógicas e buscamos definir parâmetros com base em recomendações internacionais e, embora tenhamos verificado a inexistência de um consenso a este respeito, observamos também que a divergência entre elas é pequena.

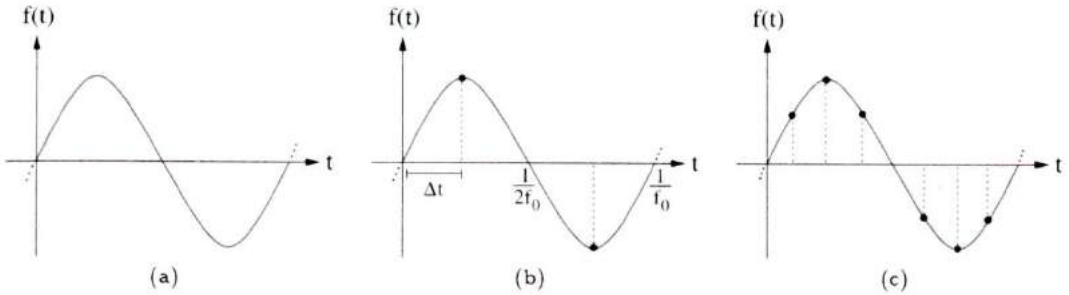
Assim, para a definição adequada do parâmetro de frequência do sinal sonoro, utilizamos o *teorema de amostragem de Whittaker-Shannon* e o *limite de Nyquist* discutidos por Pedrini e Schwartz (2008, p.16). No teorema, a taxa de amostragem adequada pode ser obtida a partir de um conjunto de amostras que satisfaçam completamente o intervalo e o limite da frequência pela equação:

$$\Delta t \leq \frac{1}{2f_o}$$

onde Δt representa o intervalo de amostragem com banda limitada no domínio $[-f_o, f_o]$ do espaço de frequências.

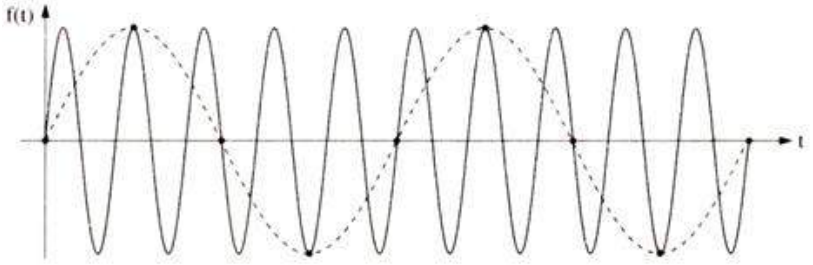
O resultado da equação deve possibilitar a obtenção de pelo menos duas amostra no intervalo da frequência de sinal completo.

Utilizando os modelos de representação de Pedrini e Schwartz (2008), podemos observar que a figura 1(a) contém o sinal original em um determinado espaço, tempo e amplitude. Na figura 1(b) tem-se o sinal com a indicação dos pontos de amostragem mínimos requeridos para a reconstrução adequada do sinal, segundo o *limite de Nyquist*. A figura 1(c) demonstra uma taxa de amostragem quatro vezes maior do que o limite mínimo necessário para a reconstrução do sinal.



Taxa de amostragem de um sinal periódico (Fonte: PEDRINI; SCHWARTZ, 2008, p. 17)

Pedrini e Schwartz (2008, p. 17) demonstram que, caso o intervalo de amostragem seja menor que o *limite de Nyquist*, o resultado será um sinal distinto do original, conhecido como fenômeno de *aliasing*. Assim, na figura 2, podemos observar que a linha tracejada, representando o sinal reconstruído a partir dos pontos de amostragem definidos, não corresponde com a frequência original do sinal.



Fenômeno de aliasing (Fonte: PEDRINI; SCHWARTZ, 2008, p. 17)

Dessa maneira, a taxa de amostragem de meio período é a mais adequada, considerando a relação entre o volume de amostras geradas e a quantidade mínima necessária para a reconstrução equivalente ao sinal original.

Além disso, a capacidade auditiva humana percebe frequências que se encontram numa faixa entre 20 e 20.000 Hz (Amaral, 2009, p. 56), entretanto, a configuração do som apresentada pelas fitas do acervo não ultrapassam 25.000 Hz.

Logo, para termos a melhor qualidade sonora, o valor da taxa de amostragem deveria ter o dobro do valor da frequência que o ouvido humano é capaz de escutar e, deste modo, optamos por adotar uma taxa de amostragem de 48.000 Hz, seguindo o que sugere a *International Association of Sound and Audiovisual Archives* (IASA). (International Association of Sound and Audiovisual Archives, 2009, não paginado).

Além da taxa de amostragem, a prescrição para a profundidade dos *bits* foi analisada, pois é ela que quantifica e fornece a fidelidade da amostra, indicando os *bits* que serão utilizados para representar cada ponto do sinal sonoro a ser digitalizado em cada instante da amostragem (de 8 a 64 *bits*).

Para tal definição, consultamos pesquisadores da Universidade –linguistas e engenheiros de som, que atuam na área de fonética, de fonologia e de engenharia–, que analisam tais gravações em *softwares*[2] especializados, para esclarecermos a demanda de taxa de amostragem adequada às suas pesquisas. A resposta foi que para a análise de falas 16 *bits* eram suficientes, do mesmo modo que a recomendação da IASA. (International Association of Sound and Audiovisual Archives, 2009, não paginado).

Por fim, vale mencionar que essas reflexões se pautaram também pela relação custo x benefício, pois dado que o volume de arquivos gerados é grande e os custos de armazenamento altos, não poderíamos promover a digitalização dos documentos com base em parâmetros acima do necessário para a garantia da qualidade digital do som dos documentos.

No que se refere ao formato do arquivo digital a ser gerado, o formato **.wav** é hoje uma prática dominante, tendo se tornado um padrão de fato. Este formato é oficialmente recomendado pelo Comitê Técnico da IASA e, deste modo, foi adotado pelo Centro para a produção de suas matrizes digitais. (International Association of Sound and Audiovisual Archives, 2009, não paginado).

Quanto aos parâmetros para a geração de cópias derivadas para acesso, optamos pelo formato **.mp3**, por ser compacto e de fácil transmissão e "leitura", o que favorece e agiliza o seu acesso através da *Web*.

[2] *Software* é um conjunto de instruções lógicas interpretadas por um computador com o objetivo de executar tarefas específicas.

A PRESERVAÇÃO DE DIREITOS EM DOCUMENTOS SONOROS

Outro aspecto a ser considerado diz respeito à legislação que envolve os documentos de arquivo. Referimo-nos em especial aos direitos autorais e ao direito à privacidade.

Por se tratarem de *corpora* de falas, gravadas em situação de informalidade: em casa, no trabalho, na escola, na rua etc., o contexto dos registros muitas vezes expõe situações privadas. . Neste sentido, consultamos o Comitê de Ética da Unicamp, que nos orientou na preparação de um termo de autorização escrito e assinado em papel e, sendo assim, adotamos o procedimento de autorização com relação a cada um dos falantes registrados. Quanto ao cumprimento da lei que protege os direitos do autor, os pesquisadores responsáveis pelos registros autorizaram igualmente a publicação das gravações através do site do CEDAE.

No entanto, apesar das autorizações obtidas, julgamos ainda necessária a proteção dos documentos contra a cópia direta e, desta maneira, buscamos desenvolver uma solução que franqueasse o acesso aos documentos sonoros sem que fosse necessário o seu *download*, ou seja, a cópia dos documentos.

O ACESSO E A DIFUSÃO

Uma vez definidos os parâmetros para a digitalização, passamos a nos dedicar ao desenvolvimento de uma solução eletrônica que possibilitasse não só o acesso às derivadas digitais, mas também às transcrições dos áudios, de modo a promover a melhor difusão das coleções sonoras do CEDAE.

O Acesso às derivadas digitais

Com a geração das derivadas digitais[3], novas necessidades e demandas surgiram em relação ao gerenciamento e à disponibilização desses representantes. O tratamento digital adequado às matrizes e derivadas requereu uma infraestrutura de TI segura, estável e com o máximo de disponibilidade para esse enorme volume de dados. Além disso, as derivadas foram concebidas para permitir o acesso rápido, fácil e múltiplo aos documentos através de ambiente eletrônico.

Nesse caso, diversas ações foram realizadas para preparar e garantir um ambiente de *hardware* e *software* adequado e alinhado às normas das áreas de Tecnologia da Informação e de Arquivos.

O primeiro passo das ações foi definir a infraestrutura base de *hardware* para o correto armazenamento dos dados digitais. Sendo assim, o projeto utilizou equipamento com redundância de componentes, como fontes de alimentação de energia, disco rígido (HD) e sistema de resfriamento (*cooler*), compostos por tecnologia *hot-swap*[4] que permitem, em caso de falha, serem substituídos sem necessidade de parada do equipamento. Além desses recursos, o projeto contou com equipamentos *nobreak*[5], com o objetivo de minimizar a ocorrência de falhas em sistemas e componentes pelo desligamento abrupto em decorrência de oscilação ou queda de energia elétrica.

[3] O processo de captura digital, a partir dos documentos originais, gera representantes digitais de alta e de baixa resolução, denominados respectivamente, Matrizes e Derivadas.
[4] *Hot-swap* é uma tecnologia que permite a troca de componentes defeituosos sem que seja necessário desligar o computador.
[5] Equipamento capaz de fornecer energia elétrica por meio de baterias a um sistema por um determinado período de tempo, no caso de interrupção do fornecimento de energia da rede pública.

O armazenamento dos dados foi definido em discos com tecnologia de redundância RAID[6] nível 5, com implementação via *hardware*, o que possibilita eficiência de escrita e de leitura e um comprometimento de espaço menor.

Para a implantação da base de *software*, desde o sistema operacional até o sistema *Web*, o projeto também levou em consideração requisitos de segurança e independência de soluções proprietárias, para que não demandassem grande volume de recursos financeiros para a aquisição, manutenção, treinamento etc. Além disso, a dependência por tecnologia proprietária, que segue as tendências de mercado e empresas, pode ser classificada como um fator de risco e vulnerabilidade para um projeto voltado a documentos históricos.

Na concepção do projeto, a Resolução GR-052/2012, que estabelece as Normas e Procedimentos para o Uso dos Recursos de Tecnologia da Informação e Comunicação na Universidade Estadual de Campinas, serviu como referência, em especial o artigo 80 do Capítulo X, que define:

> **I** – os mecanismos de acesso a sistemas e serviços eletrônicos institucionais devem evitar impor uma plataforma (*hardware* e *software*) particular aos usuários finais;
> **II** – caso o acesso se dê através da *Web*, então ele deve ser viável a partir de pelo menos dois dentre os navegadores mais usados na *Internet*;
> **III** – se houver necessidade de *software* cliente nos equipamentos dos usuários, sua instalação e uso não devem onerar os Órgãos/Unidades responsáveis por tais equipamentos; (UNIVERSIDADE ESTADUAL DE CAMPINAS, 2012, p. 72-74)

Sendo assim, os sistemas selecionados para compor a chamada base principal do ambiente virtual de acesso, como: o sistema operacional, o *backup*, o controle de acesso a arquivos, o controle de tráfego de rede etc são do livres e de código aberto, ou seja, utilizam formatos e padrões que, em sua maioria, são usados por uma gama elevada de instituições públicas e privadas.

Dentre esses sistemas, definimos o uso do Ubuntu Server/Linux[7], para o sistema operacional, o Samba[8], para gerenciar e auditar os arquivos, o Apache[9], para gerenciador *Web*, o Mysql[10], como base de dados e o Bacula[11] para a realização de *backup*.

Consequentemente, com a escolha de padrões e formatos abertos, estáveis, consolidados e não proprietários, podemos afirmar que o percentual de compatibilidade com *hardwares* e outros sistemas, até mesmo os proprietários, tende a ser maior.

Para a delimitação do escopo do projeto de *software*, baseado na "Engenharia de *Software*" de Pressman, foram definidos 8 macro requisitos funcionais e 5 requisitos não funcionais, conforme abaixo:

Requisitos Funcionais: Usuário não precisa realizar cadastro para acessar os dados, Lista de documentos sonoros por Fundo/Coleção, Lista de documentos sonoros Total, Busca por documento sonoro, Busca por Fundo/Coleção, Ordenação de Listas por Fundo/Coleção, Unidade de Descrição, Título, Código e Duração, Visualização/audição de arquivos MP3, Visualização simultânea de transcrições de áudio.

Requisitos não funcionais: Segurança de dados (acesso, *download* etc.), Interface simplificada, Compatibilidade com os dois principais navegadores *Web,* Independência de

[6] RAID é a sigla de *redundant array of independent disks* (Conjunto Redundante de Discos Independentes), sendo um mecanismo criado com o objetivo de melhorar o desempenho e segurança dos discos rígidos em computadores, através do uso de discos rígidos extras.
[7] http://www.ubuntu.com/
[8] https://www.samba.org/
[9] http://www.apache.org/
[10] https://www.mysql.com/
[11] http://blog.bacula.org/

hardware específico do usuário, Independência de *softwares* específico do usuário. (Pressman, 2006, p. 116 – 129).

A solução para acesso Web

A escolha do ambiente virtual, que chamaremos de *framework*[12], baseou-se em pesquisas de soluções implementadas em instituições congêneres, que pudessem atender as características definidas anteriormente. Porém, mesmo após a ampliação das pesquisas à instituições nacionais e internacionais, verificamos a inexistência de soluções para o cenário proposto.

Diante da constatação da inexistência de soluções adequadas, procuramos identificar soluções com características e/ou funcionalidades correspondentes às especificações definidas, ou seja, flexíveis para a modificação, de baixo custo e com comunidade de mantenedores sólida. Dentre os diversos sistemas pesquisados, observamos que o Ampache[13] possuía atributos demandados pelo projeto para *streaming*[14] de áudio.

Nesse sentido, a característica do Ampache como aplicação para o uso de *streaming* era uma funcionalidade que atendia à necessidade de segurança, por impedir o *download* direto aos documentos.

Entretanto, para atender a todos os requisitos, a aplicação demandou ajustes, aperfeiçoamentos e complementos para receber os documentos sonoros da forma mais adequada ao gerenciamento técnico dos arquivos e aos acessos realizados pelos pesquisadores internos e externos à instituição.

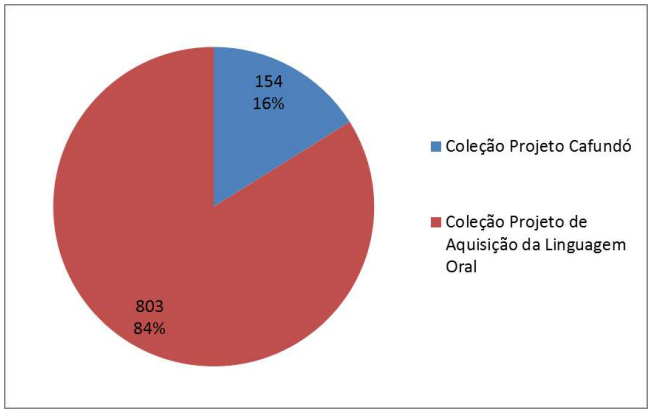
Nesse sentido, foram utilizados conceitos da engenharia reversa, (Pressman, 2006, p. 688 - 690) com o objetivo de mapear as funcionalidades e os fluxos de atividade internos à aplicação.

A aplicação também demandou ajustes na tradução de conteúdo, correção de distorções na identificação de campos, ajustes de funcionalidades, aperfeiçoamento da interface do usuário, inclusão de recursos extras, como por exemplo, visualização simultânea das transcrições de áudio e controle de acesso externo à Universidade.

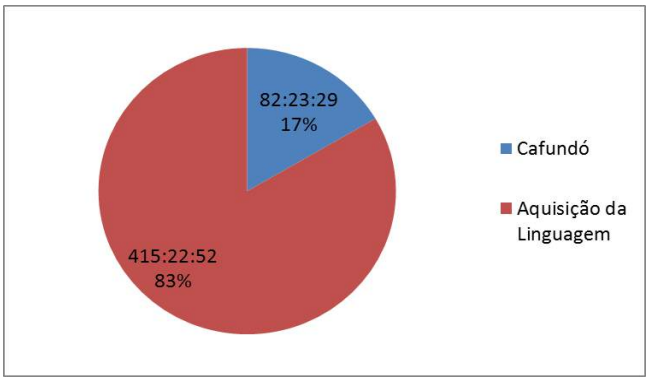
Esse conjunto de intervenções demandaram aproximadamente 4 meses de trabalho até a fase interna de teste da aplicação. Ao final do trabalho, a aplicação recebeu o nome de Plataforma de Documentos Sonoros ou PDS, tendo sido disponibilizada para acesso na *Internet* em agosto de 2013.

Como citado anteriormente, as coleções "Cafundó" e "Aquisição da Linguagem Oral" foram selecionadas e digitalizadas para serem inseridas na base de dados e disponibilizadas para consulta ao público através da *Internet*. Juntas, as duas somam um total de 957 (Vide Gráfico 1) documentos sonoros, que correspondem a mais de 497 horas de gravação (Vide Gráfico 2). Além disso, aproximadamente 42% dos documentos sonoros presentes na PDS possuem transcrições disponíveis para visualização, num total de 401 documentos e 15.000 páginas. Sendo assim, atualmente, o volume total dos dados armazenados em disco é equivalente a ½ *Terabyte*.

[12] *Framework* é uma abstração que une códigos comuns entre vários projetos de *software* provendo uma funcionalidade genérica.
[13] http://ampache.org/
[14] *Streaming* é uma forma de distribuição de dados em uma rede através de pacotes, que permite que o usuário reproduza conteúdos protegidos por direitos de autor na *Internet*, sem a violação desses direitos, pois não é possível a realização de *download*. Portanto, em *streaming*, as informações não são armazenadas pelo usuário em seu próprio computador.



Volume de documentos sonoros por coleção.
(Fonte: Banco de dados da PDS – CEDAE, 2014)



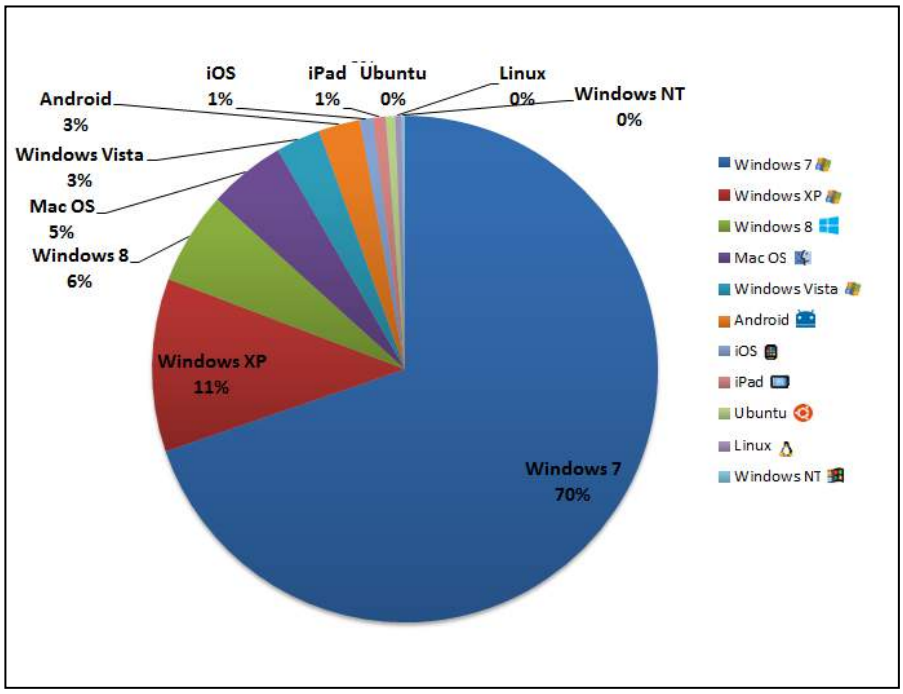
Total de horas de gravação por coleção.
(Fonte: Banco de dados da PDS – CEDAE, 2014)

Desde o lançamento ao público, a PDS tem passado por diversas análises de tráfego e de compatibilidade de acesso, com o objetivo de validar se a aplicação vem atendendo aos requisitos especificados no projeto inicial.

Para a obtenção dos dados de navegação e medições *web* utilizamos a ferramenta Piwik[15], instalada junto com a PDS, para aplicarmos o processo de *web analytics*[16].

Através do Piwik, nos primeiros seis meses de funcionamento, pudemos realizar análises de compatibilidade de *software* do usuário que confirmaram a hipótese de que a grande maioria dos acessos era realizada por indivíduos utilizando o sistema operacional *Windows*, conforme demonstrado no Gráfico 3.

[15] Piwik é uma aplicação de código aberto para coleta, medição e geração de relatórios de dados *Web* que pode ser encontrada em: http://piwik.org/.
[16] *Web analytics* refere-se à medição, coleta, análise e a geração de relatórios da internet com o objetivo de compreender e aperfeiçoar o uso de usuários às páginas *Web*.



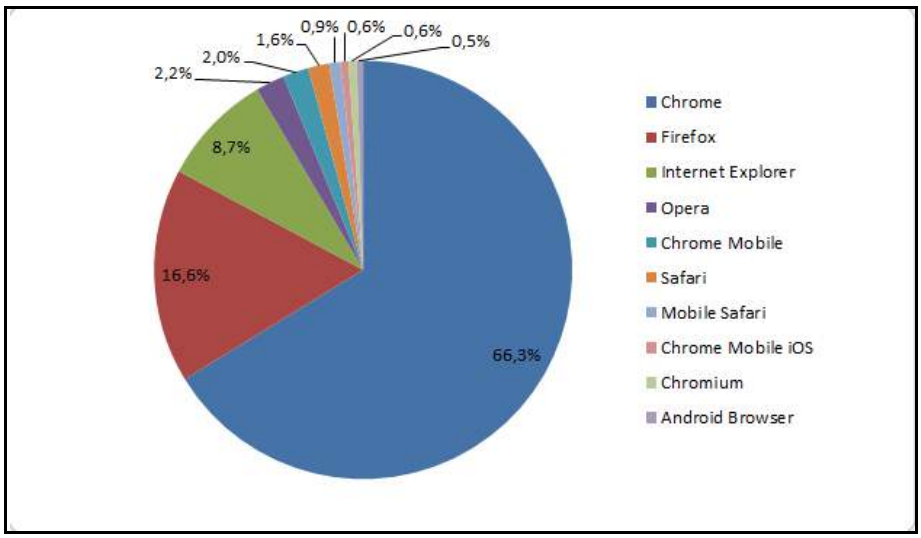
-Percentual de acesso por sistema operacional do usuário (Fonte: Piwik CEDAE, 2014)

A diversidade de sistemas operacionais utilizados pelos usuários, observada nos percentuais, valida o requisito de compatibilidade da aplicação com qualquer sistema operacional, independentemente do ambiente de hospedagem. Outro detalhe importante a destacar, refere-se ao fato de que um número considerável de usuários ainda utilizava *Windows XP*[17] para acessar a PDS. Nesse caso, apesar de fatores mercadológicos, mantivemos a possibilidade de acesso ao maior número de usuários e/ou versões de sistemas operacionais.

Por ser uma aplicação *Web*, a PDS depende de um *hardware* mínimo do usuário, que possa executar algum sistema operacional com navegador *Web* instalado. Sabemos que os sistemas operacionais possuem pelo menos um navegador para acesso à *Internet*. Por isso, o projeto foi codificado em HTML (*HyperText Markup Language*) e PHP (*Hypertext Preprocessor*), utilizando recurso de formatação por CSS (*Cascading Style Sheets*), e JAVASCRIPT com o objetivo de aperfeiçoar as interações com o usuário.

O gráfico 4 demonstra que o acesso à aplicação ocorre por diversos navegadores, atendendo ao requisito de viabilidade de acesso por pelo menos dois dos principais navegadores da *Internet*.

[17] Em 2014 a Microsoft anunciou o fim do sistema operacional Windows XP. Ver: http://windows.microsoft.com/pt-br/windows/end-support-help.



Número de acessos por navegador (Fonte: Piwik CEDAE, 2014)

A movimentação de grandes volumes de dados é um risco para a estabilidade de aplicações *Web*, pois a demanda por processamento e tráfego de rede é intensa. Sem o devido monitoramento, esses fatores podem comprometer o funcionamento da aplicação. Assim, medições realizadas frequentemente mostram que o tempo médio de geração de páginas para acessos nacionais é de aproximadamente 0,20 milissegundos, enquanto que acessos internacionais levam em média 0,49 milissegundos. Os números mostram que, apesar da diferença apresentada entre usuários nacionais e internacionais, os valores estão em um intervalo satisfatório para a especificidade do material acessado. A tabela 1 mostra o tempo médio para acesso às principais páginas da aplicação realizada por acesso internacional.

Páginas *WEB*	**Tempo médio de exibição (milissegundos)**
Principal	0,53
Lista de Álbuns	0,45
Coleção Cafundó	0,35
Lista de Áudios	0,59
Coleção Aquisição da Linguagem	0,68
Busca Avançada Álbum	0,40
Busca Avançada Áudio	0,41

Tempo médio de exibição de páginas por usuário estrangeiro. (Fonte: Piwik CEDAE, 2014)

CONCLUSÃO

Podemos dizer que, do ponto de vista da preservação, os suportes de áudio estão mais ameaçados do que os documentos de texto convencionais, devido a sua instabilidade química, a vulnerabilidade aos perigos externos ou a mera repetição do acesso.

Assim, medidas para o acesso e a preservação desses suportes têm, portanto, de ser tomadas. Isto vai além das medidas básicas de prevenção, tal como estabelecido para a maioria dos documentos convencionais e, neste sentido, a adesão à todas as normas e práticas recomendadas acima não é garantia contra a perda acidental de documentos.

Equipamentos para acesso, mesmo bem conservados, de forma inesperada podem falhar e destruir uma fita.

Quanto aos documentos digitais, sabemos que eles podem desaparecer sem qualquer pré-aviso e a qualquer momento. Por conseguinte, é preciso contar com, pelo menos, duas cópias de cada documento e, assim, fitas magnéticas originais, vulneráveis e instáveis, devem ser copiadas para formatos de arquivo robustos e confiáveis.

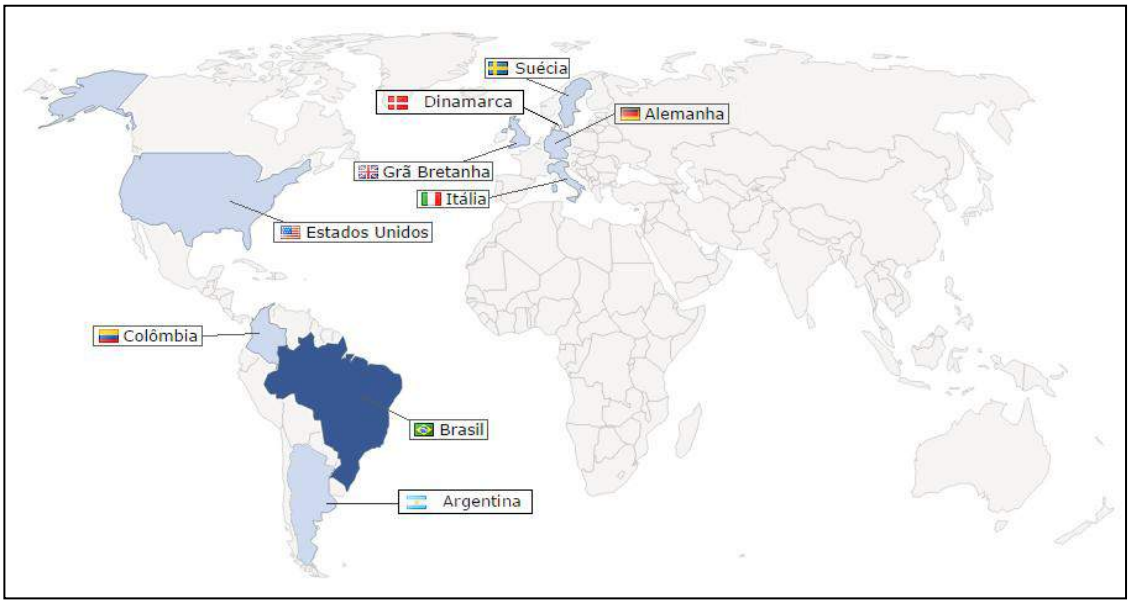
Portanto, pelo menos duas cópias devem ser feitas e armazenadas em locais diferentes, para proteção adicional da informação em caso de desastre.

Para o acesso, cópias de trabalho apropriadas ou derivadas de acesso também devem ser feitas. Ao contrário de documentos textuais, matrizes digitais de áudio não devem nunca ser manipuladas por pesquisadores, mas apenas por pessoal treinado.

A escolha de acesso em ambiente *web* de *streaming* de áudio *open source* deve ser igualmente considerada, pois as modelações criadas podem ser alteradas e aperfeiçoadas no caso de ser necessária a implantação de novas funcionalidades.

Além disso, as soluções para acesso *Web* a documentos sonoros devem atender não só à preservação dos documentos, mas também dos direitos previstos na legislação brasileira e, neste sentido, a garantia de segurança contra cópia direta dos conteúdos disponibilizados é fundamental.

Concluímos que os resultados obtidos têm colaborado de forma marcante para a intensificação da pesquisa com fontes primárias, em especial com *corpus* de registros sonoros, pois favorece a preservação, o acesso e auxilia os arquivos a cumprirem suas funções sem as restrições de espaço físico. Pelo contrário, avaliamos que o espaço virtual potencializa e multiplica a capacidade de atendimento ao usuário, que já não coincide, necessariamente, com a capacidade do local de guarda e preservação dos documentos (vide Infográfico1).



Acesso por Países (Fonte: Piwik CEDAE, 2014)

Por fim, observamos que preservar a informação na era digital requer novas formas de trabalho, tais como a atualização de competências e conhecimentos do pessoal que trabalha nos arquivos.

REFERÊNCIAS

AMARAL, Mauro Sérgio da Rosa. *Migração de suporte de fitas magnéticas de áudio cassete:* um estudo preliminar do Tribunal Regional da 4ª região – TRF4. Porto Alegre, 2009. 89 f. Monografia (Bacharelado em Arquivologia) – Faculdade de Biblioteconomia e Comunicação Social da Universidade Federal do Rio Grande do Sul.

ARELLANO, Miguel Angel. Preservação de documentos digitais. *Ci. Inf.*, Brasília , v. 33, n. 2, p. 15-27, ago. 2004. Disponível em: <http://www.scielo.br/scielo.php?pid=S0100-19652004000200002&script=sci_arttext>. Acesso em: 20 abr. 2015.

BOSTON, George (Ed.).*Safeguarding the documentary heritage:* a guide to standards, recommended practices and reference literature related to the preservation of documents of all kinds. Paris: United Nations Educational, Scientific and Cultural Organization, 1998. Disponível em: <http://www.unesco.org/webworld/mdm/administ/en/guide/guidetoc.htm>. Acesso em: 22 abr. 2015.

BUARQUE, Marco Dreer. Estratégias de preservação de longo prazo em acervos sonoros e audiovisuais. In : ENCONTRO NACIONAL DE HISTÓRIA ORAL (9:2008; São Leopoldo, RS). *Anais*. Rio de Janeiro: Associação Brasileira de História Oral ; São Leopoldo, RS : UNISINOS, 2008. 9f. Disponível em: <http://cpdoc.fgv.br/producao_intelectual/arq/1718.pdf>. Acesso em: 10 maio 2015.

CONSELHO NACIONAL DE ARQUIVOS (Brasil). *Resolução nº 6, de 15 de maio de 1997*. Dispõe sobre diretrizes quanto à terceirização de serviços arquivísticos públicos. Disponível em: <http://www.conarq.arquivonacional.gov.br/cgi/cgilua.exe/sys/start.htm>. Acesso em: 24 abr. 2015.

_____. *Resolução nº 25, 27 de abril de 2007*. Dispõe sobre a adoção do Modelo de Requisitos para Sistemas Informatizados de Gestão Arquivística de Documentos - e-ARQ Brasil pelos órgãos e entidades integrantes do Sistema de Arquivos-SINAR. Disponível em: <http://www.conarq.arquivonacional.gov.br/Media/publicacoes/earqbrasilv1.pdf>. Acesso em: 24 abr. 2015.

_____. *Recomendações para digitalização de documentos arquivísticos permanentes*. 2010. Disponível em: <http://www.conarq.arquivonacional.gov.br/media/publicacoes/recomenda/recomendaes_para_digitalizao.pdf>. Acesso em: 24 abr. 2015.

INTERNATIONAL ASSOCIATION OF SOUND AND AUDIOVISUAL ARCHIVES. *TC04 Guidelines on the Production and Preservation of Digital Audio Objects:* standards, recommended practices, and strategies. 2 ed. 2009. Disponível em: <http://www.iasa-*Web*.org/tc04/audio-preservation>. Acesso em: 07 maio 2015.

MATZ, Judith (Ed.). *Sound Savings:* preserving audio collections. Association of Research Libraries, Austin, 2004. Disponível em: <http://www.arl.org/storage/documents/publications/sound-savings.pdf>. Acesso em: 10 maio 2015.

PEDRINI, Hélio; SCHWARTZ, Willian Robson. *Análise de imagens digitais*: princípios, algoritmos e aplicações. São Paulo: Thomson Learning, 2008.

PRESSMAN, Roger S. *Engenharia de Software*. 6 ed. São Paulo: McGraw-Hill, 2006.

SCHÜLLER, Dietrich: The role of audiovisual documents for safeguarding cultural and linguistic diversity. In: INTERNATIONAL SYMPOSIUM ON PRESERVATION OF CHINESE ETHNIC GROUPS, 2003, Beijing. *Proceedings of the International Symposium on Preservation of Chinese Ethnic Groups*. Beijing: Chinese Academy of Arts, dez. 2003. p. 259-260.

_____. Preservation of audio and video materials in tropical countries. *IASA Journal*, n. 7, p. 35-45, 1996. Disponível em: <http://www.unesco.org/webworld/audiovis/reader/7_5.htm>. Acesso em: 07 maio 2015.

STERNE, Jonathan. *The Audible Past:* cultural origins of sound reproduction. Durham: Duke University Press, 2003.

THE NATIONAL RECORDING PRESERVATION BOARD. *Capturing Analog Sound for Digital Preservation:* report of a roundtable discussion of best practices for transferring analog discs and tapes. Washington, D.C., 2006. Disponível em: <http://www.clir.org/pubs/reports/reports/pub137/pub137.pdf>. Acesso em: 10 maio 2015.

UNIVERSIDADE ESTADUAL DE CAMPINAS. *Resolução GR nº 052, de 21 de dezembro de 2012*. Estabelece as Normas e Procedimentos para o Uso dos Recursos de Tecnologia da Informação e Comunicação na Universidade Estadual de Campinas. *Diário Oficial do Estado de São Paulo*, São Paulo, 27 dez. 2012. Seção 1, p. 72-74.

WEBB, Colin; CANBERRA NATIONAL LIBRARY OF AUSTRALIA.*Guidelines for the preservation of digital heritage*. Information Society Division, United Nations Educational, Scientific and Cultural Organization,

2003. Disponível em: <http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>. Acesso em: 22 abr. 2015.

RESUMO: *Este trabalho descreve e analisa as atividades desenvolvidas pelo Centro de Documentação Cultural "Alexandre Eulalio" (CEDAE), para a estruturação de procedimentos padronizados para a preservação de documentos sonoros. Trata da conservação de fitas magnéticas de rolo por meio da digitalização, bem como as questões relativas ao uso da tecnologia para o acesso e a disseminação de tais documentos na Internet.* **PALAVRAS-CHAVE:** *Documentos sonoros, Digitalização, Conservação, Fitas magnéticas de rolo, Migração de suporte.*

RESUMEN: *Este trabajo describe y analiza las actividades desarrolladas por el Centro de Documentación Cultural "Alexandre Eulalio" (CEDAE), para la estructuración de procedimientos estandarizados de preservación de documentos sonoros. El proyecto trata de la conservación de las cintas magnéticas de carrete abierto a través de su digitalización, y también plantea cuestiones relacionadas con el uso de la tecnología para el acceso y la difusión de estos documentos en Internet*. **PALABRAS-CLAVE:** *Documentos sonoros, Digitalización. Conservación. Cintas magnéticas de carrete abierto, Migración de soporte.*