Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciências
Escola Superior de Desenho Industrial

Pedro Herzog

Sistema para indexação e visualização de depoimentos de história oral: o caso do Museu da Pessoa

Rio de Janeiro
2014

Pedro Herzog

Sistema para indexação e visualização de depoimentos de história oral: o caso do Museu da Pessoa

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-graduação em Design, da Universidade do Estado do Rio de Janeiro.

Orientador: Prof. Dr. Marcos André Franco Martins

Coorientadora: Prof.ª Dra. Lívia Lazzaro Rezende

Rio de Janeiro

2014

CATALOGAÇÃO NA FONTE
UERJ/REDE SIRIUS/BIBLIOTCA CTC/G

H582Herzog, Pedro.

Sistema para indexação e visualização de depoimentos de história oral: o caso do Museu da Pessoa / Pedro Herzog. - 2014.
89f. : il.

Orientador: Marcos André Franco Martins.

Dissertação (Mestrado). Universidade do Estado do Rio de Janeiro, Escola Superior de Desenho Industrial.

1. História oral - Teses. 2. Indexação- Teses. 3. Visualização da informação - Teses. 4. I. Martins, Marcos.II. Universidade do Estado do Rio de Janeiro. Escola Superior de Desenho Industrial. III. Título.
CDU 930,2

_____ _____

Assinatura Data

Pedro Herzog

Sistema para indexação e visualização de depoimentos de história oral: o caso do Museu da Pessoa

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-graduação em Design, da Universidade do Estado do Rio de Janeiro.

Aprovada em 26 de fevereiro de 2014.

Banca Examinadora:

_____

Prof. Dr. Marcos André Franco Martins (Orientador)

Escola Superior de Desenho Industrial da UERJ

_____

Prof.ª Dra. Verena Alberti

Fundação Getúlio Vargas – FGV / CPDOC

_____

Prof. Dr. Washington Lessa

Escola Superior de Desenho Industrial da UERJ

Rio de Janeiro

2014

DEDICATÓRIA

Este trabalho é dedicado ao meu filho Miguel e ao meu irmão João.

AGRADECIMENTOS

À minha família, pelo apoio e compreensão.

A Marcos Martins, meu orientador, por ter me aceitado como orientando e confiado na proposta. Obrigado pelas críticas precisas e pelas palavras de incentivo, ambas fundamentais para conclusão dessa dissertação.

A Lívia Lazzaro, minha amiga e co-orientadora, pela seriedade e pelas observações sempre construtivas.

A Sergio Boiteux, meu amigo, pela generosidade nas longas conversas sobre possíveis abordagens para este projeto. Pela sua ajuda fundamental em várias etapas e pelas inestimáveis contribuições e referências.

A Durval Amorim e Thiago Silveira, webdesigner e desenvolvedor web respectivamente, pelo profissionalismo e dedicação.

A Karen Worcman, e toda a equipe do Museu da Pessoa, pela abertura do seu acervo e disponibilidade para o desenvolvimento desse projeto.

Aos professores Rodolfo Capeto, Noni Geiger e Eliane Jobim, pelos comentários valiosos. Aos professores Washington Lessa e Verena Alberti, que aceitaram compor a banca de defesa.

A todos os amigos que tiveram que me ouvir explicando a pesquisa ou mesmo reclamando da vida, obrigado pela atenção e paciência.

Acredito que a única esperança a longo prazo para a humanidade é construirmos um mundo em que se reconheça o quanto temos em comum nas nossas necessidades, medos e sonhos. Ouvir histórias de vida é um dos mais prazerosos meios de se aproximar dos outros.

Paul Thompson

RESUMO

HERZOG, Pedro. *Sistema para indexação e visualização de depoimentos de história oral*: o caso do Museu da Pessoa. 2014. 90f. Dissertação (Mestrado em Design) – Escola Superior de Desenho Industrial, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2014.

Esta dissertação apresenta a estruturação de um sistema para indexação e visualização de depoimentos de história oral em vídeo. A partir do levantamento de um referencial teórico referente à indexação, o sistema resultou em um protótipo funcional de alta fidelidade. O conteúdo para a realização deste foi obtido pela indexação de 12 depoimentos coletados pela equipe do Museu da Pessoa durante o projeto Memórias da Vila Madalena, em São Paulo (ago/2012).

Acervos de História Oral como o Museu da Pessoa, o Museu da Imagem e do Som ou o Centro de Pesquisa e Documentação de História Contemporânea do Brasil / CPDOC da Fundação Getúlio Vargas, reúnem milhares de horas de depoimentos em áudio e vídeo. De uma forma geral, esses depoimentos são longas entrevistas individuais, onde diversos assuntos são abordados; o que dificulta sua análise, síntese e consequentemente, sua recuperação.

A transcrição dos depoimentos permite a realização de buscas textuais para acessar assuntos específicos nas longas entrevistas. Por isso, podemos dizer que as transcrições são a principal fonte de consulta dos pesquisadores de história oral, deixando a fonte primária (o vídeo) para um eventual segundo momento da pesquisa.

A presente proposta visa ampliar a recuperação das fontes primárias a partir da indexação de segmentos de vídeo, criando pontos de acesso imediato para trechos relevantes das entrevistas. Nessa abordagem, os indexadores (termos, *tags* ou anotações) não são associados ao vídeo completo, mas a pontos de entrada e saída (*timecodes*) que definem trechos específicos no vídeo.

As *tags* combinadas com os *timecodes* criam novos desafios e possibilidades para indexação e navegação através de arquivos de vídeo. O sistema aqui estruturado integra conceitos e técnicas de áreas aparentemente desconectadas: metodologias de indexação, construção de taxonomias, folksonomias, visualização de dados e design de interação são integrados em um processo unificado que vai desde a coleta e indexação dos depoimentos até sua visualização e interação.

Palavras-chave: História Oral. Indexação. Visualização de Dados. Folksonomia.

ABSTRACT

HERZOG, Pedro. *System for indexing and visualizing oral history testimonials:* the Museu da Pessoa´s case. 2014. 90f. Dissertation (Master in Design) – Escola Superior de Desenho Industrial, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2014.

This work presents the construction of an interface for visualizing and navigating the many narratives of oral history testimonials. Collections such as those belonging to the CPDOC/FGV, the Museu da Imagem e do Som and the Museu da Pessoa, contain thousands of hours of audio and video interviews. Each one of them covers many subjects, which complicates its analysis, synthesis, indexing, and consequently its retrieval.

This proposal aims to facilitate the retrieval of primary sources (audio and video) by indexing specific excerpts of testimonies. To accomplish this, technologies and methodologies from areas such as: tagging, content analysis, text mining, thesauri construction and data visualization will be applied. Hence the need for an approach that consolidates these various project phases into one unified process in which the interdependencies of each step are clear and transparent.

As case study, we will use 12 testimonials collected in late 2012 by researchers from the Museu da Pessoa. By indexing these videos, we will create an interface for navigating the interview segments, now categorized by topics.

Keywords: Oral History. Indexing. Data Visualization. Folksonomy.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

ABCDM – Arquivologia, Biblioteconomia, Ciência da Informação, Documentação e Museologia

ABNT – Associação Brasileira de Normas Técnicas

CMS – *Content Management System*

CPDOC – Centro de Pesquisa e Documentação de História Contemporânea do Brasil

FGV – Fundação Getúlio Vargas

HTML – *Hyper Text Markup Language*

NOBRADE – Norma Brasileira de Descrição Arquivística

OHDA – *Oral History in the Digital Age*

SQL – *Structured Query Language*

UNISIST – *United Nations International Scientific Information System*

XML – *Extensible Markup Language*

SUMÁRIO

INTRODUÇÃO

Esta dissertação apresenta a estruturação de um sistema para indexação e visualização das diversas histórias presentes em depoimentos de história oral registrados em vídeo. Como resultado deste trabalho, é apresentado um protótipo funcional de interface para visualização e navegação entre entrevistas coletadas pela equipe do Museu da Pessoa para o projeto “Memórias da Vila Madalena”.

Acervos de história oral reúnem milhares de horas de entrevistas em áudio e vídeo. Cada uma dessas entrevistas contém diversas histórias com diferentes assuntos, o que dificulta sua análise, síntese, indexação e, consequentemente, sua recuperação.

A transcrição dos depoimentos oferece a base para realização de buscas textuais nos acervos de história oral, possibilitando um rápido acesso aos assuntos de interesse nas longas entrevistas. Sendo assim, a “visualização” da história oral sempre se deu prioritariamente através do registro escrito, deixando as fontes primárias (aúdio e vídeo) para um eventual segundo momento da pesquisa.

A internet ampliou o alcance dos projetos de história oral e ofereceu novas possibilidades para sua veiculação e prática. Com o aumento da capacidade de transferência de dados surgiu também um grande potencial para difusão de vídeos, introduzindo novas questões para a indexação desses depoimentos. Quais trechos das entrevistas devem ser selecionados para edição? Como indexar e relacionar esses trechos de vídeo?

Diante disso, diversas instituições de história oral em todo o mundo vem desenvolvendo metodologias próprias de acordo com os recursos e as necessidades de cada projeto. Em 2012, o *Oral History in the Digital Age (OHDA)*[1]*,* apresentou recomendações gerais para curadoria da história oral digital.

Para Mark Tebeau (2012), colaborador do *OHDA*, embora não se possa eleger uma metodologia ou "melhor prática” para o desenvolvimento de projetos de história oral, já podemos ter uma ideia de como devemos pensar a curadoria de

---

[1] Oral History in the Digital Age - OHDA é uma iniciativa do *Institute of Museum and Library Services – IMLS* com a *Oral History Association* e a *Library of Congress*.
Disponível em: http://ohda.matrix.msu.edu/best-practices/. Acessado em: 16/01/2014.

história oral digital, de forma a oferecer uma interação mais rica com as mídias primárias.

Entre os pontos apontados, destacamos:

1. A importância do reconhecimento de que história oral é fundamentalmente uma experiência aural. O áudio ou vídeo revelam aspectos que não podem ser traduzidos plenamente pela transcrição (texto).
2. Um depoimento de história oral pode ser analisado em segmentos e não apenas no nível da entrevista completa. Isto é, a indexação deve ser associada a um intervalo de tempo, criando marcadores que levam o usuário diretamente para um trecho específico do depoimento.
3. Esses segmentos podem ser relacionados entre entrevistas.

As novas possibilidades para difusão e acesso de arquivos de vídeo demandam novas ferramentas para indexação e visualização dos depoimentos de história oral.

Para atender essas demandas, esse trabalho apresenta a estruturação de um sistema que integra conceitos e técnicas de áreas aparentemente desconectadas: metodologias de indexação, construção de taxonomias, folksonomias, visualização de dados e design de interação são integrados em um processo unificado, que vai desde a coleta e indexação dos depoimentos até sua visualização e interação.

O objetivo final deste projeto consiste na criação de uma interface interativa que possibilite aos usuários uma navegação através dos segmentos de vídeo e, ainda, fornecer aos pesquisadores novas ferramentas para indexação. A forma final desta estruturação será o protótipo de uma interface interativa que visa possibilitar aos usuários uma navegação através dos segmentos de vídeo e, ainda, fornecer aos pesquisadores novas ferramentas para indexação.

Privilegou-se, aqui, o empenho no desenvolvimento funcional do protótipo como forma de investigação a respeito da aplicabilidade, em termos práticos, dos conceitos levantados. Considerando-se o tempo requerido por tal desenvolvimento, reservei, para um tempo futuro, os ensaios de validação e testes de usabilidade, inviáveis nos prazos disponíveis.

No capitulo 1, apresento um breve histórico da história oral com ênfase na sua relação com as tecnologias e ferramentas para registro e reprodução de

depoimentos. Também apresento o Museu da Pessoa, objeto de estudo deste trabalho. Ainda nesse capítulo são detalhados os novos desafios para a curadoria de história oral digital que nortearam o desenvolvimento deste projeto.

No capítulo 2, faço um levantamento das linguagens de indexação, descrevo o processo de indexação do Museu da Pessoa e, finalmente, apresento a abordagem integrada para Design de Informação Computacional proposta pelo americano Benjamin Fry, expert em visualização de dados, para lidar com a complexidade resultante do processo de indexação.

No capítulo 3 aplico a abordagem integrada descrita no capítulo anterior, para o desenvolvimento de um protótipo. Cada uma das etapas é ilustrada com capturas de tela e exemplos práticos que evidenciam a interdependência entre as etapas.

O capítulo 4, diz respeito aos resultados obtidos, apontando possíveis melhorias e desdobramentos para o projeto.

1 **HISTÓRIA ORAL E TECNOLOGIA**

1.1 **Introdução**

História oral é a coleta e o estudo de informações históricas sobre indivíduos, famílias, eventos importantes, ou mesmo da vida cotidiana usando registros de áudio, vídeo, ou transcrições de entrevistas. Resultado dos avanços da tecnologia, principalmente dos meios eletrônicos como gravador, o vídeo e o computador, a história oral se apresenta como forma de captação das experiências de pessoas que participaram de, ou testemunharam eventos e cujas memórias e percepções devem ser preservadas em um registro aural. De acordo com a Profa. Verena Alberti (2005), “História oral é um método de pesquisa (histórica, antropológica, sociológica, etc.) que privilegia a realização de entrevistas com pessoas que participaram de, ou testemunharam acontecimentos, conjunturas, visões de mundo, como forma de aproximar do objeto de estudo.”

A prática da história oral reúne procedimentos que se iniciam com a elaboração de um projeto e continuam com a definição de um grupo de pessoas a serem entrevistadas, com o planejamento e condução das gravações, com a transcrição, com a conferência da transcrição, com a autorização para o uso, arquivamento e, finalmente, publicação das entrevistas como fonte de consulta para outros estudos.

1.2 **Os primeiros registros de áudio**

Os primeiros registros de áudio foram realizados por antropólogos, folcloristas, etnomusicólogos e lingüistas, ainda no início do século XX com o fonógrafo criado por Thomas Edison em 1899.

Figura 1, O fonógrafo e o cilíndro Edison.

Fonte: Wikipedia. Disponível em: http://en.wikipedia.org/wiki/Phonograph_cylinder. Acessado em: 22/01/2014

Ainda hoje, existem cerca de 100 mil documentos de pesquisa nesse formato em todo mundo. Esses cilindros eram feitos de uma mistura de diferentes ceras e sua superfície relativamente mole acabava sendo alterada a cada reprodução. Apesar disso, graças a técnicas especiais, foi possível transferir esses registros para o formato digital, recuperando dados que de outro modo estariam irremediavelmente perdidos.

Figura 2, Béla Bartók realizando o registro de música folclórica nas pequenas cidades das regiões rurais da Hungria e Romênia. (1907).

Fonte: BÓNIS, Ferenc, 1972, p.80

Em suas pesquisas, o compositor Béla Bartók, no intuito de registrar e estudar as tradições da música folclórica romena e húngara, utilizou um fonógrafo para capturar o áudio executado pelos camponeses.

Figura 3, Béla Bartók realizando uma transcrição de áudio para notação musical (c.1940)

Fonte: BÓNIS, Ferenc, 1972, p.203

Bartók realizou a transcrição do áudio para a notação musical. A transcrição oferece a base para uma leitura analítica da música. Além disso, embora não seja possível capturar todas as nuances e sutilezas da música na partitura, cada vez que os cilindros eram reproduzidos, o registro ia se desgastando.

Então, neste caso, a transcrição tem dupla função: aprofundar o estudo daquela música folclórica e preservar os cilindros de cera. Graças aos esforços para transcrição e preservação dos cilindros, esses registros foram digitalizados e agora podemos escutá-los livremente.

Embora este exemplo não seja propriamente um caso de história oral, as questões de preservação e reprodução que preocupavam os etnomusicólogos, ilustram aspectos do registro de áudio que se aplicariam mais tarde às entrevistas de história oral. De forma análoga às preocupações de Bartók, as transcrições das entrevistas de história oral, além de permitirem um estudo detalhado dos depoimentos registrados, preservavam as mídias originais (fitas magnéticas) do desgaste pela reprodução.

1.3 **Breve histórico**

O desenvolvimento de projetos de história oral sempre esteve associado aos recursos tecnológicos disponíveis para sua realização. A chamada moderna história oral [2], coincide com a popularização do gravador portátil em 1947, que facilitou os

2 Termo cunhado por Allan Nevins, da Columbia University

registros de entrevistas em fitas magnéticas. Nessa época, o rádio já era um importante meio de difusão e as entrevistas tornaram-se populares.

Mais tarde, já na década de 80, as câmeras de vídeo passaram a fazer parte do processo de captação dos depoimentos, permitindo o registro não somente do áudio como também da imagem dos depoentes.

No início da década de 90, com a crescente presença dos computadores pessoais, as fitas magnéticas foram sendo substituídas por *compact disks* e *hard disks*, o que propiciou novas soluções para o armazenamento e reprodução dos arquivos de áudio e vídeo agora em formato digital.

O desenvolvimento da internet ampliou o alcance dos projetos e ofereceu novas possibilidades para o avanço da prática de história oral.

Diante disso, diversas instituições de história oral em todo o mundo desenvolvem metodologias próprias de acordo com os recursos e as necessidades de cada projeto.

1.4 **O Museu da Pessoa**

O Museu da Pessoa é um acervo de histórias de vida colaborativo, aberto à participação de toda pessoa que queira compartilhar sua história a fim de democratizar e ampliar a participação dos indivíduos na construção da memória social.

Fundado em São Paulo, em 1991, o Museu da Pessoa é hoje uma rede internacional, com iniciativas em Portugal, Estados Unidos e Canadá.

Desde sua criação, o Museu da Pessoa é uma organização da sociedade civil que atua para registrar, preservar e transformar em informação histórias de vida de toda e qualquer pessoa da sociedade. A partir de metodologias próprias, capta, organiza e edita conteúdos disseminados em publicações, programas de rádio e TV, exposições e no portal[3].

Com programas nas áreas de memória institucional, educação, comunicação e desenvolvimento comunitário, o Museu da Pessoa já realizou 220 projetos de memória que visam multiplicar e democratizar sua metodologia e seu acervo, que

---

[3] Portal do Museu da Pessoa. Disponível em: http://www.museudapessoa.net. Acessado em: 16/01/2014.

inclui aproximadamente 15 mil histórias de vida e 72 mil documentos e fotos digitalizados.

1.4.1 6 Bilhões de Outros

O projeto 6 Bilhões de Outros[4], de Yann Arthus-Bertrand e da Fundação GoodPlanet, realizou 5.600 entrevistas em 78 países do mundo. O Museu da Pessoa participou da produção de algumas entrevistas em São Paulo. Neste projeto, os trechos (depoimentos editados) foram agrupados em temas como: estar em casa, deixar o seu país, histórias de amor, fazer o amor durar, desafios da vida, perdoar, felicidade, sentido da vida, etc. A reunião dos trechos em cada um desses temas deu origem a pequenos filmes de aproximadamente 30 minutos, onde pessoas das mais variadas origens falam sobre determinado assunto. Como as entrevistas foram realizadas a partir de um mesmo roteiro de perguntas, a mesma pessoa acaba aparecendo em diversos filmes. A exposição realizada em São Paulo em 2011, foi a inspiração para o desenvolvimento deste trabalho. Os trechos de vídeo agrupados por temas e apresentados sucessivamente, revelaram o conteúdo rico e dinâmico adormecido nos acervos de história oral.

Figura 4, Imagem extraída da vinheta do projeto “6 Bilhões de Outros”.

[4] Disponível em: http://www.7billionothers.org/. Acessado em: 16/01/2014.

1.5 **Novas possibilidades e desafios**

A produção e o consumo de vídeo vem crescendo gradativamente com o surgimento de serviços como o YouTube. De acordo com um estudo realizado pela empresa Cisco, em 2015 dois terços de todo o tráfego na internet será tomado por vídeos.
Com o aumento da capacidade de armazenamento e transferência de dados, surge também um grande potencial para difusão de vídeos, introduzindo novas questões para acervos de história oral no mundo inteiro.

> A medida em que os arquivos de vídeo tornam-se imediatamente acessíveis, novas formas para navegação nas entrevistas e entre entrevistas são cada vez mais desejáveis. (LAMBERT, D.; FRISCH, M., 2012) [5]

As fontes primárias, agora em formato digital, podem ser livremente reproduzidas. Servidores de mídia (*streaming*) possibilitam o acesso direto a um ponto do vídeo sem precisar "baixar" o arquivo inteiro. Os vídeos das entrevistas podem ser acessados simultâneamente por pessoas em todo mundo. Entretanto, a acessibilidade oferecida pelo formato digital e pela internet encontra no tamanho dos acervos de história oral e na duração das entrevistas individuais em vídeo, um desafio. Nas basta disponibilizar o arquivo de vídeo sem oferecer um acesso direto aos trechos mais relevantes, e para seleção e anotação desses trechos, é necessário o trabalho humano.

Embora algumas tecnologias para *taggeamento* automático possam prometer a redução deste fator, elas ainda não se provaram úteis para esse tipo de anotação.

> O sistema de palavras-chave provavelmente também mais criará do que resolverá problemas, porque os conceitos que precisarão ser indexados podem ser expressos por muitas palavras diferentes ou, na verdade, por alusões indiretas, ou até mesmo por abstenção e omissão. Como os computadores pensam com uma coerência rigidamente tacanha, as palavras-chave têm que ser editadas no texto para que possam ser localizadas. Isso significa que, com a maioria das coleções de história oral, usando ou não um computador, a indexação será um processo mais próximo do índice de nomes e assuntos de um livro comum (THOMPSON, Paul. 1978, 1988. A voz do passado: história oral)[6]

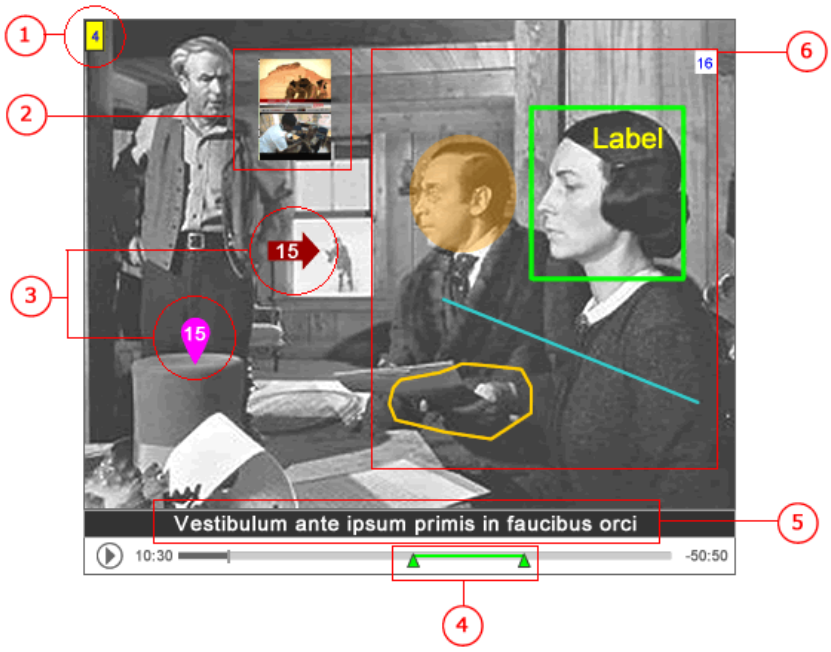
---

[5] Lambert, Doug. e Frisch, Michael. Meaningful Access to audio and vídeo passages: a two-tiered approach for annotation, navigation, and cross-referencing within and across oral history interviews. Em *Oral history in the digital age*, editado por Doug Boyd, Steve Cohen, Brad Rackert e Dean Rehberger. Washington, D.C.: Institute of Library and Museum Services, 2012
Disponível em: http://ohda.matrix.msu.edu/2012/06/meaningful-access-to-audio-and-video-passages-2/. Acessado em: 16/01/2014.

6 Thompson, Paul. 1978, 1988. A voz do passado

Além disso, as ferramentas para adicionar anotações ou comentários para segmentos específicos do vídeo ainda são limitadas. De acordo com o *Annotations at Harvard* [7], os requisitos para um sistema de anotações em vídeo devem ser:

Figura 5, Requisitos para um sistema de anotações em vídeos.



1. Anotações ou comentários sobre o vídeo completo
2. Camadas de imagens ou vídeos
3. Marcadores customizados pelo usuário
4. Intervalo de tempo (começo e fim) para cada anotação
5. Legendas automáticas podem ser usadas como anotações de base
6. Camadas de gráficos e textos gerados pelos usuários

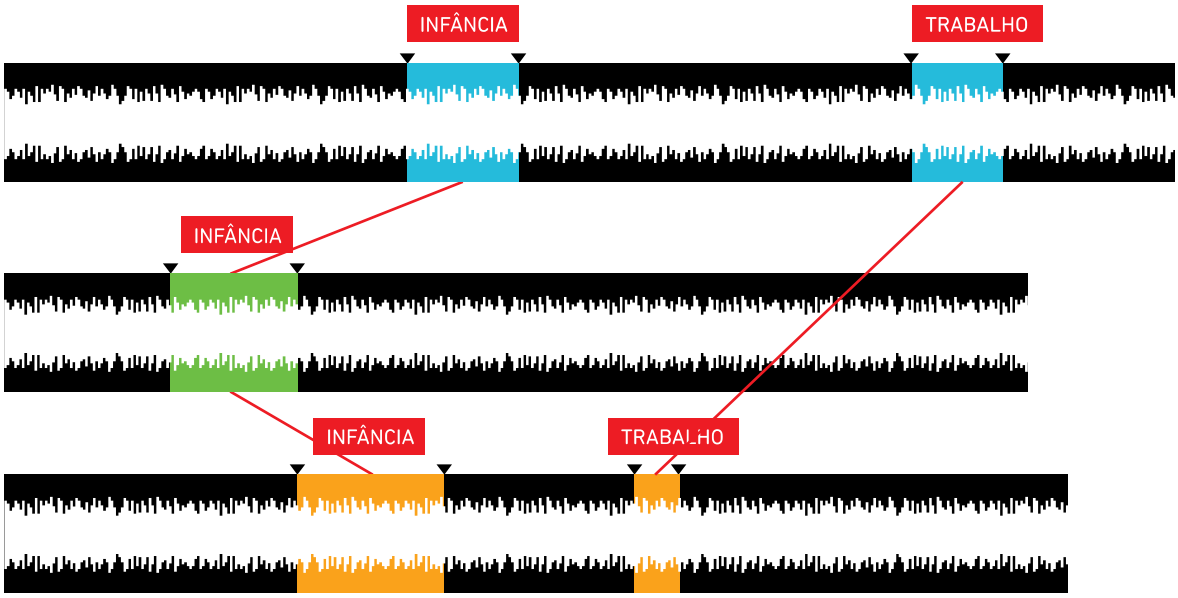
Com a adição dessas camadas de informação, os arquivos de vídeo digital oferecem novos pontos para busca e interação. Embora todas as possibilidades elencadas possam trazer benefícios para a prática de história oral, a presente dissertação concentra-se nas anotações de segmentos de vídeo (item 4 do gráfico acima), para anotação de trechos das entrevistas.

---

[7] *Annotations at Harvard é* uma iniciativa da Universidade de Harvard para criação de um sistema colaborativo de anotações entre professores e alunos. Disponível em: http://www.annotations.harvard.edu/icb/icb.do?keyword=k80243&pageid=icb.page466612. Acessado em: 16/01/2014.

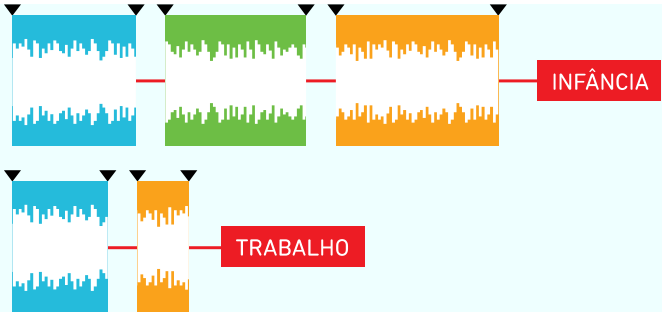
A figura abaixo ilustra a indexação de segmentos em 3 depoimentos de história oral. As anotações (ou *tags*) são utilizadas para o relacionamento entre os trechos selecionados nas diferentes entrevistas.

Figura 6, A indexação de trechos de vídeo e o relacionamento entre entrevistas.



A partir dessa indexação, pode-se agrupar os trechos que compartilham a mesma *tag*. Esses trechos ordenados em sequência (veja a figura abaixo), reproduzem o resultado obtido no projeto 6 Bilhões de Outros, em que os filmes apresentavam sucessivamente trechos de diversas entrevistas.

Figura 7, Os trechos que compartilham a mesma tag ordenados em sequência.



Dessa forma, indexando os segmentos das entrevistas, surgem novas possibilidades para interação com acervos de história oral, e amplia-se a disponibilização do registro aural, objetivo deste trabalho.

2 **METODOLOGIAS DE INDEXAÇÃO**

Neste capítulo faço um levantamento das linguagens de indexação, suas vantagens e limitações. Em seguida analiso a metodologia de indexação utilizada pelo Museu da Pessoa no projeto “Memórias do Comércio de São Paulo”. A partir desta análise, apresento uma proposta híbrida para indexação dos trechos de vídeo, combinando a linguagem livre (utilizando *tags*) com a linguagem controlada (utilizando taxonomias).

Ainda neste capítulo, apresento a abordagem integradora proposta por Ben Fry para o Design de Informação Computacional, como um roteiro para o desenvolvimento do protótipo no capítulo seguinte.

2.1 **Indexação**

2.1.1 Definições

Indexação, para biblioteconomia ou ciência da informação, é a ação de descrever e identificar um documento de acordo com o seu assunto. Segundo Vieira (1998), “Indexação é uma técnica de análise de conteúdo que condensa a informação significativa de um documento através da atribuição de termos, criando uma linguagem intermediária entre o usuário e o documento.”

Para o UNISIST / *United Nations International Scientific Information System* (1981) e a ABNT / Associação Brasileira de Normas Técnicas (1992) indexar é descrever e identificar o conteúdo de um documento a partir de termos representativos dos seus assuntos. Faz-se importante por permitir a recuperação do documento por meio de sua representação temática.

Embora não haja consenso sobre como se dá o processo de indexação, em geral, são descritas três operações básicas inerentes à atividade[8]:

1. Análise conceitual
2. Síntese (identificação dos conceitos)
3. Representação (tradução desses conceitos em termos de indexação)

2.1.1.1 Análise conceitual

A determinação do assunto do documento inicia-se por meio da análise conceitual, realizada pela leitura e segmentação documental (no caso deste projeto, a leitura e segmentação dos vídeos das entrevistas). A análise conceitual consiste na identificação dos assuntos do documento, ou seja, na compreensão do seu conteúdo temático. É considerada a etapa mais importante no trabalho do indexador, tratando-se de um processo subjetivo e intelectual. Lida com análise, interpretação e definição do que será indexado, isto é, com a tomada de decisão envolvendo inclusive o contexto para o qual o documento está sendo indexado.

2.1.1.2 Síntese

A partir da análise conceitual, são identificados os conceitos mais adequados para indexação dos documentos, visando sua recuperação por um público específico.
Segundo Lancaster (2004), a indexação de assuntos é normalmente feita visando a atender às necessidades de um grupo específico de pessoas, ou seja, é preciso que se tome uma decisão não somente quanto ao que é tratado no documento, mas por que ele se reveste de provável interesse para determinado grupo de usuários.

2.1.1.3 Representação

A representação, ou tradução, consiste na representação dos conceitos em termos de indexação. Quando esses termos estão presentes no próprio documento, diz-se que a indexação é feita por *extração*, usando a linguagem natural. Quando o

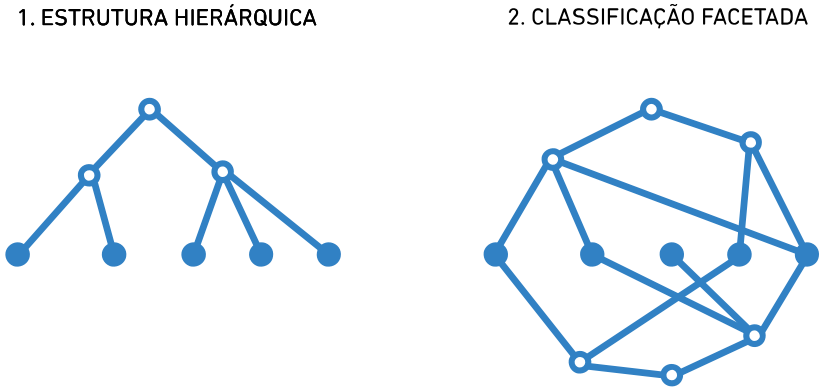
[8] Cf. RUBI, M. P. *Política de indexação para construção de catálogos coletivos em bibliotecas universitárias*. 2008. 166. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2008.

indexador escolhe palavras de fontes externas ao documento, a indexação é por *atribuição*, por meio da linguagem controlada ou livre (como veremos a seguir).

2.1.2 Estruturas de classificação

A biblioteconomia e a ciência da Informação têm uma longa tradição na organização de recursos em uma variedade de estruturas de classificação. Em princípio, de acordo com Quintarelli[9], existem duas estratégias complementares:

Figura 8, Estrutura hierárquica (*top-bottom*) x classificação facetada (*bottom-up*). Entidades (objetos indexados) são representadas pelos círculos cheios e os termos indexadores são representados pelos círculos vazios.



1. A abordagem **hierárquica** (*top-bottom*), arranja a totalidade das classes em classes e subclasses progressivamente mais específicas. Uma relação transitiva direta implica na herança das propriedades dos grandes grupos para os mais específicos. Um exemplo clássico é a classificação lineana dos organismos em espécies (figura 13, página 34).

2. A abordagem **facetada** (*bottom-up*), consiste em esquemas de propriedades geradas através de um processo de análise e síntese. A construção dessas estruturas começa no nível fundamental das entidades (objetos indexados) e suas propriedades. Esses conceitos são então organizados em grupos mutuamente excludentes, com base em similaridades conceituais, e esses grupos são associados a agrupamentos sucessivamente maiores para formar

[9] Disponível em: http://www-dimat.unipv.it/biblio/isko/doc/folksonomies.htm#overview
Acessado em: 20/01/2014

facetas ou aspectos. Por exemplo: um vocabulário facetado para classificação de carros contaria com facetas para “cor” (branco, preto, prata), “categoria” (sedan, conversível, *offroad*), e “câmbio” (automático, manual).

A princípio, ambos os tipos de classificação podem ser transformados um no outro. A diferença conceitual, entretanto, é que a abordagem hierárquica resulta em uma ordem específica de classes e entidades, onde propriedades gerais estão localizadas no topo da hierarquia e as propriedades mais específicas ficam na base. Estruturas facetadas são mais flexíveis, permitindo a associação de entidades através de uma rede de relacionamentos. Como veremos mais adiante, ambas as abordagens serão utilizadas no sistema de indexação aqui proposto.

2.2 **Linguagens de indexação**

2.2.1 Linguagem natural (*por extração*)

A linguagem natural, sinônimo para discurso comum, representa o vocabulário normalmente usado na fala e na escrita. Sistemas de informação que adotam linguagem natural possibilitam a realização de buscas textuais em grandes volumes de texto. Utilizando mecanismos de busca, os usuários fornecem ao computador uma seqûencia específica de caracteres para varredura dos textos. Como resposta (recuperação da informação), o computador apresenta uma lista de documentos que contém aquela mesma sequência de caracteres. Para Krippendorff, no entanto,

> O termo Recuperação da Informação não deveria ser aplicado. O termo Busca Textual descreve com mais precisão o que o computador faz, separando a questão da qualidade semântica dos resultados de busca. (...) Os caracteres de texto somente podem se transformar em informação quando lidos por uma pessoa. (KRIPPENDORFF, 2012)

No caso dos acervos de história oral, as transcrições dos depoimentos, permitem a realização de buscas textuais. Entretanto, buscas textuais são limitadas pelo fato de serem explícitas, isto é, os termos buscados precisam estar contidos literalmente no texto. Enquanto as buscas textuais são úteis quando buscamos por um termo específico – quando sabemos o que estamos procurando – elas oferecem pouca ajuda quando queremos apenas explorar o acervo sem uma questão prévia.

Além das buscas textuais, as transcrições também oferecem a base para extração automática de termos. Ferramentas para processamento de texto em linguagem natural como o *Alchemy API*[10], podem ajudar na indexação de grandes volumes de texto, identificando lugares, pessoas, organizações, atividades, etc. Mas embora poderosas, essas ferramentas ainda dependem de uma pessoa para monitorar o resultado obtido automaticamente.

2.2.2 Linguagem controlada ou documentária (*por atribuição*)

Linguagem controlada é um conjunto de termos, símbolos e regras pré-estabelecidos para indexação de assuntos. Também conhecida como vocabulário controlado, esta linguagem procura garantir uma padronização dos termos indexados e a fim de melhorar a recuperação da informação.

> A qualidade de um sistema de recuperação de informação é normalmente medida pela **revocação** [*recall*] (o número de documentos relevantes recuperados em uma busca, comparado com o número total de documentos relevantes presentes na coleção), e pela **precisão** (o número de documentos relevantes comparado ao número de irrelevantes). (STEFANER: 2007)

Diferentemente da linguagem natural, em que os termos aplicados são extraídos do texto, na linguagem controlada os termos atribuídos podem não estar literalmente presentes no texto. Isso é particularmente útil para indexação de entrevistas de história oral, onde muitas vezes os assuntos estão implícitos ou subentendidos nos depoimentos.

O principal objetivo de um vocabulário controlado é garantir que cada conceito seja descrito de uma única forma. Se existem múltiplas formas, elas devem ser controladas ou normatizadas para que a informação fornecida pelo usuário não seja dispersada no sistema, mas reunida em um só ponto de acesso.

A utilização de um vocabulário controlado permite o tratamento de propriedades típicas da linguagem que geram ambigüidades na terminologia. A polissemia (mesma grafia para diferentes significados) e a sinonímia (diferentes grafias para o mesmo significado), podem ser controladas com estruturas relativamente simples.

---

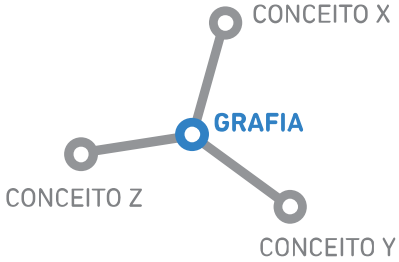
[10] Disponível em: http://www.alchemyapi.com/ Acessado em: 20/01/2014

Polissemia

Mesma grafia para diferentes significados.

A diferenciação de homógrafos (polissemia) pode ser feita com uma simples lista ou glossário com os termos especificados em linguagem natural.

Figura 9, Controle da polissemia.



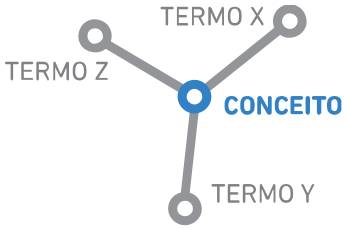
Exemplo: Manga (fruta) / Manga (de camisa) / Manga (Carlos Manga, pessoa).

Sinonímia

Diferentes grafias para o mesmo significado.

O controle de sinônimos é feito ao optar-se por um único termo padronizado, sendo as outras formas ligadas a ele através de remissivas.

Figura 10, Controle da sinonímia.

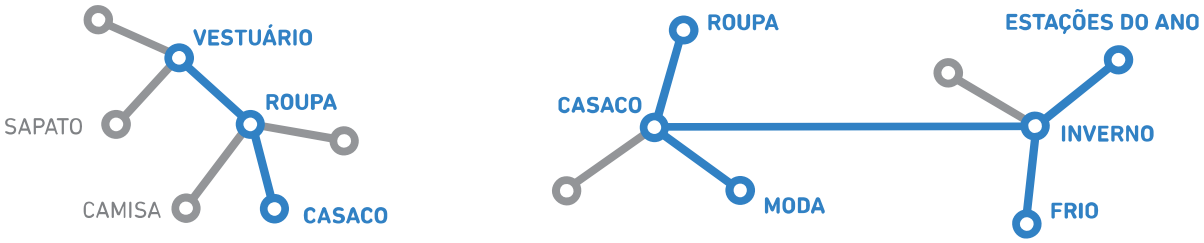


Exemplo: Automóvel / Carro / Veículo

Relacionamentos hierárquicos e associativos (ou facetados)

Além dos casos descritos acima, os vocabulários controlados podem ser utilizados para criar outros tipos de relacionamentos entre os termos. Utilizando-se estruturas mais complexas, podem ser criados relacionamentos hierárquicos e associativos (ou facetados).

Figura 11, Estruturas hierárquicas e associativas (ou facetadas).



Exemplo: “casaco” relaciona-se de forma hierárquica a “vestuário”, pois é uma especificidade deste, e ao mesmo tempo associa-se a “inverno” (um relacionamento não-hierárquico ou facetado).

De acordo com o *National Information Standards Organisation* / NISO (2005), vocabulários controlados são estruturados para oferecer diferentes tipos de relação entre os termos neles contidos. Existem 4 (quatro) diferentes tipos de vocabulário controlado:

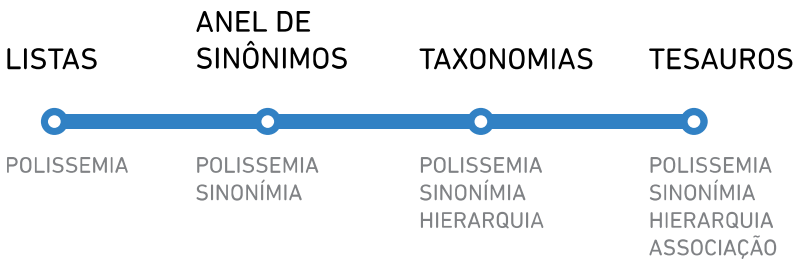
. Listas

. Anéis de sinônimos

. Taxonomias

. Tesauros

Figura 12, Estruturas para construção de vocabulários controlados por ordem de complexidade (da esquerda para direita).



A figura acima apresenta as estruturas para construção de vocabulários controlados por ordem de complexidade. Essa complexidade é ditada pelos tipos de relação que cada estrutura deve dar conta. A figura mostra também, que as estruturas mais complexas (taxonomias e tesauros) englobam as estruturas mais simples (listas e anéis de sinônimos). Por exemplo, um tesauro inclui dispositivos explícitos para controle de sinônimos, organização de hierarquias e criação de

relações associativas, enquanto uma lista é um simples conjunto de termos sem qualquer relação entre eles.

A arquivologia, a biblioteconomia e a ciência da informação tem longa tradição em organizar acervos utilizando essas estruturas. É importante fazer uma distinção clara entre esses conceitos.

2.2.2.1 Listas

Uma lista é um conjunto limitado de termos ordenados alfabeticamente ou de acordo com alguma outra lógica evidente. Listas são utilizadas para descrever aspectos das entidades com um número limitado de possibilidades.

Exemplo 1: Lista alfabética

Alabama
Arkansas
California
Connecticut
Delaware

Exemplo 2: Lista lógica

Mercurio
Venus
Terra
Marte
Jupiter
Saturno
Urano
Netuno
Plutão

2.2.2.2 Anéis de sinônimos

Embora um anel de sinônimos seja considerado um tipo de vocabulário controlado, ele desempenha um papel diferente das outras estruturas aqui descritas. Anéis de sinônimos não são utilizados durante o processo de indexação. Mas podem ser aplicados no momento da busca. Anéis de sinônimos garantem que um conceito que

pode ser descrito por múltiplos sinônimos ou termos equivalentes sejam recuperados caso qualquer um dos termos seja buscado.

Um anel de sinônimos é, portanto, um conjunto de termos que são considerados equivalentes para uma determinada busca.

2.2.2.3 Taxonomias

A taxonomia serve para classificar informação em um hierarquia (árvore) utilizando o relacionamento pai-filho (generalização ou “tipo-de”). Um exemplo clássico de taxonomia é a classificação de humanos segundo a taxonomia lineana, ilustrada abaixo:

Figura 13, Taxonomia lineana dos seres vivos. Classificação dos humanos.

REINO: *Animalia*
FILO: *Cordata*
SUBFILO: *Vertebrata*
CLASSE: *Mammalia*
SUBCLASSE: *Theria*
ORDEM: *Primata*
SUBORDEM: *Anthropoidea*
FAMÍLIA: *Hominidae*
GÊNERO: *Homo*
ESPÉCIE: *Sapiens*

Repare que todos os termos da taxonomia estão integrados através do relacionamento de generalização, ou seja, um mamífero é um tipo de vertebrado que, por sua vez, é um tipo de cordado, que, por sua vez, é um tipo de animal. De acordo com Ora Lassila e Deborah McGuiness[11], essa taxonomia é classificada como uma “hierarquia tipo-de formal”, isto é, os relacionamentos de generalização são respeitados integralmente.

Outro exemplo de taxonomia são as estruturas de diretórios (pastas) nos computadores. Como foi visto, em uma taxonomia os itens são organizados através de relacionamentos de generalização (pai-filho, classe-subclasse), no caso do

---

11 Cf. LASSILA, O. McGUINESS, D. The Role of Frame-Based Representation on the Semantic Web, 2001. Disponível em: http://www.ida.liu.se/ext/epa/ej/etai/2001/018/01018-etaibody.pdf Acessado em: 20/01/2014.

exemplo, diretório e subdiretório. Neste caso, a classificação fica a critério do usuário, que decide o nível de formalidade que vai impor aos relacionamentos de generalização. Conceitos relacionados podem ser agregados a uma categoria, mesmo que não respeitando integralmente o relacionamento de generalização. Um exemplo de organização de pastas seria: “2005 / fotos / privado”. Repare que o termo “fotos” não é um tipo de “2005”. Essa taxonomia seria, portanto, classificada como “hierarquia tipo-de informal” segundo a categorização proposta por Lassila e McGuiness.

A primeira coisa que as pessoas aprendem quando trabalhando com computadores é a metáfora visual arquivo-pasta, utilizada para organizar os arquivos, os favoritos ou e-mails. O problema dessas estruturas arquivo-pasta, é que enquanto ela facilita a entrada no mundo digital através de uma analogia com objetos do mundo real, a longo prazo, é frequentemente fonte de frustração.

Estruturas taxonômicas requerem não apenas muito esforço para serem construídas, mas especialmente para serem reorganizadas. Elas funcionam bem em casos específicos, com categorias pré-definidas, estáveis e itens restritos com claras diferenças entre as classes. Nessa situação, as taxonomias são ferramentas para busca e recuperação da informação. O problema é que a informação digital com que interagimos está sempre sujeita a mudanças e re-priorizações.

De fato, experiências mostram que os nomes de pastas denotam propriedades ou facetas (“fotos”, “2005”, “privado” etc.), mais do que conceitos ou classes. A metáfora do desktop nos permite utilizar as pastas dessa forma, uma vez que a relação de ordem direta denota “A contém B”, sem restringir a semântica dos conceitos envolvidos. Isso introduz um alto grau de liberdade para utilização desses elementos estruturais. Entretanto, se usada dessa forma, a ordenação hierárquica força a decisão de importantes atributos diferenciadores para serem colocados no topo das hierarquias (exemplo: “privado -> fotos -> 2005” vs. “2005 -> fotos -> privado”). Essa priorização precisa ser feita de ante-mão e não pode ser facilmente modificada mais tarde.

2.2.2.4 Tesauros

É uma lista de termos e suas definições que padroniza a utilização de palavras para indexação. Além das definições, um tesauro fornece relacionamentos entre os termos. Estes relacionamentos podem ser do tipo hierárquico, associativo e de

equivalência (sinônimos). Segundo a American National Standards Organization / ANSI, tesauros é:

> um vocabulário controlado organizado segundo uma ordem conhecida e estruturado de modo a disponibilizar claramente os relacionamentos de equivalência, associação, hierárquicos e homônimos existentes entre termos. Esses relacionamentos devem ser evidenciados através de identificadores padronizados para os relacionamentos. (...) O objetivo básico de um tesauro é facilitar a recuperação e obter consistência na indexação de documentos escritos. (ANSI/NISO Monolingual Thesaurus Standard, 2005)

De acordo com Breitman (2010), podemos definir um tesauro como uma taxonomia acrescida de um conjunto de relacionamentos semânticos (equivalência, hierárquicos e associação) entre seus termos.

Note que os tipos de relacionamentos entre os termos de um tesauro são finitos e bem definidos. Esse conjunto é útil na criação de vocabulários controlados, mas não é suficiente para modelar outros aspectos do mundo real. Muitas vezes é necessário relacionar conceitos utilizando-se relacionamentos do tipo parte-de, membro-conjunto, fase-processo, lugar-região, material-objeto, causa-efeito, entre muitos outros.

Atualmente o tesauro mais utilizado é o WordNet[12]. Desenvolvido por especialistas em cognição, o WordNet é um banco de dados léxico para língua inglesa com mais de 150.000 termos.

[12] Disponível em: http://wordnetweb.princeton.edu/perl/webwn Acessado em: 20/01/2014

2.2.3 Linguagem Livre (*por atribuição*)

A linguagem livre é aquela adotada sem critérios, ou seja, o indexador opta por termos que julga adequados à representação temática, sem que estes ocorram necessariamente no documento ou mesmo em um vocabulário controlado. A indexação é dita livre porque não existem limitações quanto aos termos a serem empregados.

2.2.3.1 *Tags*

Um bom exemplo da aplicação da linguagem livre são as *tags* amplamente utilizadas na internet em sistemas como Del.icio.us[13] e Flickr[14]. Nesses sistemas, as *tags* são atribuídas a URLs, fotos, vídeos, documentos, etc, pelo próprio usuário como uma busca invertida, isto é, para que o próprio possa encontrar esses objetos-digitais mais tarde. De acordo com Shirky:

> *Tags* são simples etiquetas para URLs, selecionadas para ajudar o usuário a encontrar essas URLs. *Tags* tem o efeito adicional de agrupar URLs relacionadas. Não existe um conjunto definido de categorias ou escolhas oficiais pré-aprovadas. Você pode usar palavras, acrônimos, números, o que quer que faça sentido para você, sem se preocupar com as necessidades, interesses ou requisitos de outra pessoa. (SHIRKY, 2005)

Atualmente, a base para o acesso a informação na internet é a busca livre por palavras-chave. As *tags* seguem portanto, um princípio complementar ao da busca: escolher uma combinação de palavras-chave, específicas o suficiente para que se possa distinguir o objeto em seu contexto, mas generalistas o suficiente para facilitar sua busca (recuperação) mais tarde.

Para entender o sucesso das *tags* e o modo como são atribuídas, precisamos compreender os processos de **categorização** e **classificação** a partir de uma perspectiva cognitiva.

---

[13] Disponível em: https://delicious.com Acessado em: 20/01/2014

[14] Disponível em: http://www.flickr.com Acessado em: 20/01/2014

Categorização

Categorização é a ferramenta cognitiva fundamental para compreensão do mundo. De acordo com Jacob (2004, p. 515-540), “a categorização fragmenta nossa experiência do mundo em grupos ou categorias cujos membros compartilham similaridades perceptíveis em um dados contexto”. A definição desses conjuntos de “coisas” são a base do processo cognitivo e da comunicação. Categorias são definidas por (1) intenção (definição das propriedades compartilhadas pelos membros da categoria), (2) extensão (o conjunto de todos os membros pertencentes àquela categoria) e (3) relação com outras categorias.

Categorização não é apenas um processo de reconhecimento de características e propriedades, mas também um processo criativo. Somos capazes de construir rapidamente categorias *adhoc*, como: “10 coisas para levar na viagem de férias”. De acordo com Sinha (2005), a força das *tags* está relacionada com essa facilidade com que categorias são ativadas diante de um objeto qualquer. Escrever as primeiras categorias que vem à mente, sem a preocupação de ser exato ou escolher o termo “correto”, é algo que podemos fazer sem muito esforço.

Classificação

Classificação é baseada na categorização, mas introduz ferramentas adicionais. De acordo com Jacob ela “envolve a associação ordenada e sistemática de cada entidade (ou membro) a uma e apenas uma classe em um sistema de classes mutuamente excludentes e não sobrepostas.” Enquanto a categorização é uma tarefa mental cotidiana, a classificação, por outro lado, envolve a decisão das propriedades relevantes de uma entidade (ou membro) em relação com uma conceitualização externa pré-existente (um vocabulário controlado por exemplo). Quanto maior e mais bem definida (portanto, mais forte) essa conceitualização, mais difícil é essa decisão.

No mundo digital, não apenas categorizamos os objetos, mas procuramos otimizar sua “encontrabilidade”. Temos que considerar não apenas a categoria mais adequada, mas também, como provavelmente vamos procurar pelo item no futuro. Essas duas questões muitas vezes geram respostas conflitantes, complexificando o processo de decisão e, consequentemente, de indexação.

De acordo com Sinha:

> Não se trata de uma simples categorização *adhoc* – colocar o objeto (ou entidade) em uma categoria, qualquer categoria que venha a mente. Temos que considerar o esquema de categorias como um todo. Meu esquema está desequilibrado? Tenho muitos itens em uma categoria e poucos em outra? Se coloco todos os itens em uma única categoria nunca serei capaz de encontrar algo. Preciso de uma nova categoria para esse item? Esse item cabe nesse esquema? (SINHA, 2005)

Todas essas questões levam ao que Sinha nomeou "*post activation analysis paralysis*" (paralisia pós ativação conceitual), que é quando o usuário fica em dúvida quanto ao termo mais adequado para indexação de uma entidade ou objeto. A decisão errada pode dificultar a futura recuperação do objeto indexado no esquema de categorias em questão.

O *taggeamento* livre elimina boa parte do problema descrito anteriormente, utilizando a forma mais simples de arquitetura da informação: livre associação de termos com entidades. Cognitivamente, captura associações com categorias e propriedades em um nível subjetivo, sem a necessidade de se remeter a um "esquema maior", vocabulário controlado ou outro meio específico para organização da informação.

Em resumo, as *tags* trazem exatamente as propriedades significativas de diferenciação e relevância que um item tem sob a perspectiva de um usuário específico em uma dada situação, de uma forma leve e de fácil obtensão.

2.2.3.2 Folksonomias

Embora as *tags* sejam utilizadas para adicionar marcadores de relevância subjetiva e pessoal, o principal ponto para popularização das *tags*, é o fato de sistemas colaborativos de anotação permitirem aos usuários o compartilhamento de suas *tags* com uma comunidade. Sistemas como Del.icio.us e Flickr, permitem não apenas o registro e marcação de entidades (objetos-digitais) para posterior recuperação, mas fazem com que a informação produzida seja disponibilizada para outros usuários, habilitando múltiplos pontos de acesso semânticos para navegação nos conteúdos.

De acordo com Quintarelli:

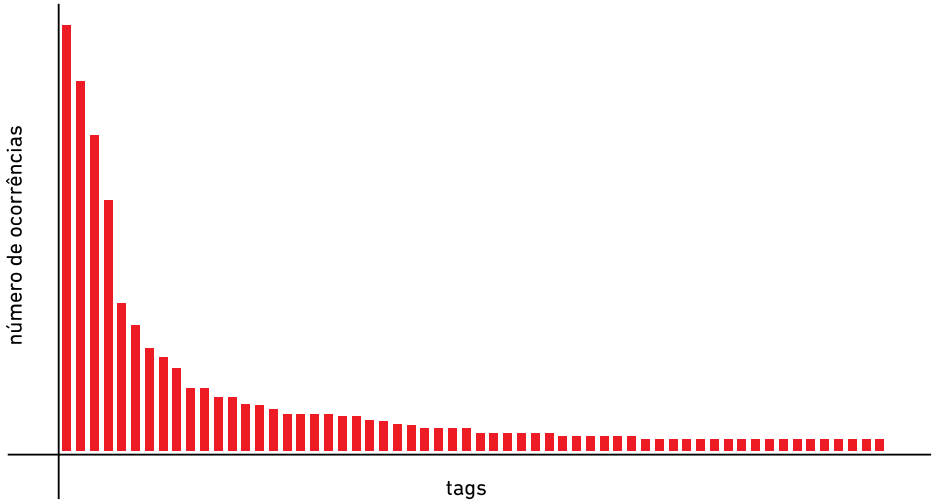
> Uma folksonomia é uma classificação gerada pelo usuário, que emerge de um consenso colaborativo (*bottom-up*). Uma fusão dos termos *folks* e taxonomia, o primeiro uso do termo folksonomia é atribuído a Thomas Vander Wal. Taxonomia vem de *taxis* e *nomos. Taxis* significa classificação. *Nomos* (ou *nomia*) significa gerenciamento. *Folks* significa pessoas. (QUINTARELLI, 2005)

A força real das *tags* livres surge, portanto, do processo colaborativo: sistemas de anotações colaborativas dão suporte para que uma comunidade de usuários faça anotações em objetos-digitais de forma social e transparente. Isso dá para cada usuário a consciência tanto das suas *tags* pessoais como das *tags* de outros membros da comunidade. Por oferecer um retorno imediato (tanto no nível pessoal como social), padrões de utilização das *tags* vão emergindo com o tempo.

A Cauda Longa

Vários estudos empíricos com folksonomias tem confirmado que a distribuição das *tags* tende a seguir uma lei geral – uma pequena quantidade de *tags* é usada muitas vezes e a grande maioria das *tags* é utilizada raramente. Esse padrão se mantém tanto no nível individual como coletivo.

Figura 14, A típica distribuição das *tags* em cauda longa, em uma folksonomia



Além disso, tem se demonstrado que essa distribuição das *tags* se estabiliza com o passar do tempo. Em uma folksonomia isso é geralmente considerado um bom sinal, indicando que existe um consenso no julgamento das entidades e no vocabulário

utilizado – ou pelo menos que o mecanismo para sugestão de *tags* está funcionando bem.

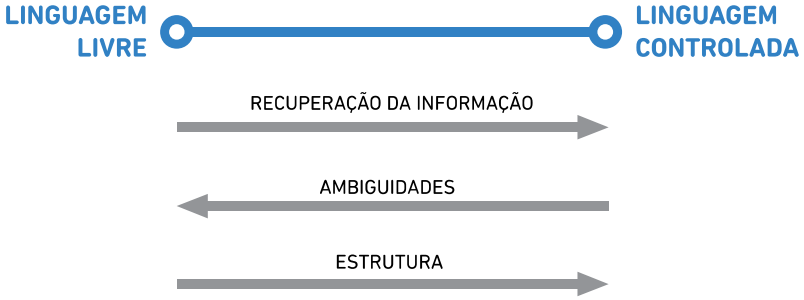
Claro que a introdução de um mecanismo simplificado e não controlado tem problemas estruturais inerentes:

- Propriedades típicas da linguagem como sinonímia (grafias diferentes com o mesmo significado), homonímia e polissemia (mesma grafia com significados diferentes) causam inconsistências nas anotações.
- Além disso, erros de digitação, capitalização inconsistente (“design” vs. “Design”) ou pluralização (“gato” vs. “gatos”) podem acidentalmente levar ao uso de múltiplas anotações com a mesma intencionalidade.
- As pessoas categorizam em um nível básico. Se apresentadas a uma foto de um labrador, anotariam “cachorro”, mas não “animal” ou “mamífero” (conceitos de ordem macro – não informativos). Da mesma forma, as anotações são aplicadas neste nível básico, adicionando as diferenças relevantes, e deixando de fora o óbvio (anotar uma foto de um gato com “animal” é pouco comum, mas pode ser valioso para uma futura busca e recuperação).

2.3 **Linguagem Livre x Linguagem Controlada**

As linguagens livre e controlada, com suas respectivas singularidades, apresentam vantagens e desvantagens de uso. Em linhas gerais, pode-se dizer que a primeira apresenta a linguagem mais próxima do usuário, porém ao mesmo tempo, ambiguidade entre termos. A segunda, por sua vez, distancia-se da linguagem habitual do usuário, entretanto estabelece maior controle linguístico e semântico. O esquema abaixo ilustra essas diferenças.

Figura 15, Linguagem Livre x Linguagem Controlada.



Como vimos anteriormente, a linguagem livre trabalha com a estrutura mais simples de arquitetura da informação, oferencendo maior agilidade para a indexação. Entretanto, sua falta de estrutura impossibilita o tratamento dos problemas típicos da linguagem (polissemia e sinonímia), impactando na qualidade da recuperação da informação.

Para escolha entre uma ou outra linguagem, devem ser analisados os objetivos do projeto, a área de conhecimento dos documentos e as características dos usuários a serem atendidos. De acordo com Nóbrega (2011, p.32):

> A experiência do indexador deve nortear a escolha dos termos em cada uma das linguagens. Estudos sobre as abordagens livre e controlada para recuperação da informação demonstram que a combinação dos dois modelos usados em conjunto oferece maior recuperação. (NÓBREGA, 2011)

Dessa forma, considerando as singularidades de um projeto de indexação, faz-se importante uma análise da metodologia já utilizada pelos pesquisadores do Museu da Pessoa, a fim de compreender o processo de trabalho já implantado. Como exemplo, vamos analisar a indexação realizada para o website[15] do projeto "Memórias do Comércio de São Paulo". Embora esse projeto não atribua *tags* aos trechos de vídeo (apenas ao vídeo completo), a análise da metodologia serve de base para compreensão do processo e definição de uma nova estratégia de indexação. Segundo Thompson (1978):

> (...)não há nenhum modelo claramente estabelecido para se seguir, de modo que é importante usar um sistema que admita modificações à luz da experiência. E, acima de tudo, ele deve ser projetado para ajudar e não para substituir a imaginação, a compreensão e a intuição humanas. (THOMPSON, 1978)

2.4 **O caso do projeto "Memórias do Comércio de São Paulo"**

O projeto "Memórias do Comércio de São Paulo", reúne aproximadamente 230 depoimentos. Esses depoimentos foram indexados com *tags* (linguagem livre) por diversos pesquisadores ao longo de quase 10 anos (o projeto iniciou-se em 1994). Sendo assim, como era de se esperar de um processo colaborativo tão longo, foi necessária uma normatização das *tags*, consolidando um vocabulário controlado com 73 termos (linguagem controlada).

---

15 Disponível em: http://www.memoriasdocomerciosp.museudapessoa.net Acessado em: 20/01/2014

Para organizar os depoimentos e facilitar o acesso dos usuários aos vídeos, os termos foram agrupados manualmente nas seguintes categorias: Origens, Tipo de Comércio, Período, Temas e Sonhos. A definição dessas categorias foi resultado de um processo iterativo onde novas categorias eram criadas ou suprimidas, buscando um balanço na quantidade de vídeos por categoria. Para organização dos depoimentos e facilitação do acesso dos usuários aos vídeos. As telas abaixo ilustram o resultado desse processo, onde as categorias são representadas na interface como menus *dropdown.*

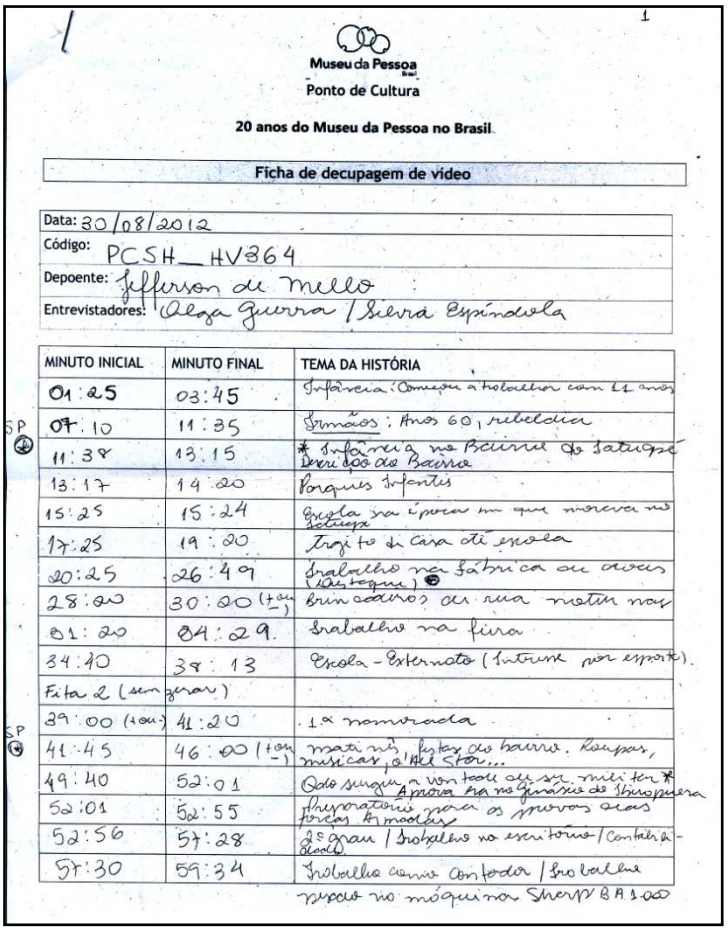
Figura 16, Telas do website “Memórias do Comércio de São Paulo”.

Cada uma dessas categorias reúne um conjunto de *tags*. Selecionando uma das *tags* no menu *dropdown,* filtra-se a matriz de quadrados que apresenta então apenas os depoentes indexados com aquela *tag.* Clicando sobre o depoente o usuário é direcionado para a página com o vídeo editado. Repare que as outras *tags* associadas aquele depoente aparecem à direita do vídeo, permitindo uma navegação transversal (um mesmo depoente pode estar presente em mais de uma categoria).

2.5 **Estratégia para indexação de trechos dos depoimentos**

Para criar relacionamentos entre os trechos de vídeo, atendendo as recomendações do *OHDA*, foi preciso definir uma nova estratégia de indexação com base na análise descrita anteriormente. Conforme observado, o Museu da Pessoa utiliza uma abordagem híbrida para indexação dos depoimentos. Em uma primeira fase, os pesquisadores utilizam a linguagem livre para atribuir *tags* às entrevistas. Mesmo havendo um roteiro de perguntas, seria pouco útil a definição prévia de um vocabulário controlado dada a natureza imprevisível dos depoimentos. Além disso, nesse projeto, a liguagem livre oferece a agilidade necessária para indexação dos segmentos de vídeo. Os pesquisadores realizam anotações durante as entrevistas para marcar os trechos mais relevantes do depoimento.

Figura 17, Ficha de decupagem de vídeo (primeira página) do Museu da Pessoa.



Museu da Pessoa
Ponto de Cultura

20 anos do Museu da Pessoa no Brasil

Ficha de decupagem de vídeo

Data: 30/08/2012
Código: PCSH_HV864
Depoente: Jefferson de Mello
Entrevistadores: Olga Guerra / Silvia Espindola

MINUTO INICIAL	MINUTO FINAL	TEMA DA HISTÓRIA
01:25	03:45	Infância: Começa a trabalhar com 11 anos
07:10	11:35	Irmãos: Anos 60, rebeldia
11:38	13:15	* Infância no Bairro de Jabaquara Descrição do Bairro
13:17	14:20	Parques Infantis
15:25	15:24	Escola na época em que morava no Jabaquara
17:25	19:20	Trajeto de Casa até escola
20:25	26:49	Trabalho na fábrica de ovos (Jabaquara)
28:20	30:20 (+ ou -)	brincadeiras na rua motin nas
01:20	04:29	Trabalho na feira
34:40	38:13	Escola - Externato (Interno por opção)
Fita 2 (sem gravar)		
39:00 (+ ou -)	41:20	1ª namorada
41:45	46:00 (+ ou -)	mais nós, festas do bairro. Roupas, musicas, o Rei Star...
49:40	52:01	Após sempre a vontade de ser militar * Ingressou no Ginásio de Ibirapuera Alistamento nas forças Armadas
52:01	52:55	
52:56	57:28	2º grau / Trabalho no escritorio / Contador - Brasil
57:30	59:34	Trabalha como Contador / Trabalha

Em um segundo momento, quando as fichas de decupagem são efetivamente cadastradas no sistema, erros de digitação, pluralização, sinônimos, siglas, etc; são padronizados a fim de garantir a consistência da indexação. O resultado desse processo é a consolidação de um vocabulário controlado.

Nesse projeto, vamos indexar não apenas um trecho, mas vários trechos do vídeo. A partir do mesmo video bruto (matriz hospedada no YouTube) serão indexados diversos segmentos de vídeo.

Figura 18, Seleção de diversos trechos de um depoimento em vídeo.

Sendo assim, podemos esperar um importante aumento no número de *tags*. Essa nova abordagem, multiplica a quantidade de entidades (trechos de vídeo)

indexadas por depoente. Ao invés de um único vídeo, ou único trecho indexado, cada depoimento pode conter diversas histórias. Consequentemente, multiplicam-se as *tags* cadastradas no sistema. Além disso, com o vídeo completo disponível para consulta, novos trechos podem ser futuramente indexados.

Nesse projeto vamos utilizar uma taxonomia simples, isto é, sem subcategorias, para agrupar *tags* que pertencem a um mesmo tema, a fim de facilitar a navegação do usuário através desse extenso e dinâmico vocabulário.

2.6 **Design de Informação Computacional**

Como descrito anteriormente, projetos de história oral podem coletar centenas de depoimentos. A análise posterior desse material adiciona constantemente novas anotações (*tags*). Essa quantidade de informação torna difícil sua compreensão global. O problema aumenta pela natureza mutante (contante modificação) dos dados, resultado das novas anotações adicionadas, ou das anotações antigas refinadas. Essa quantidade de informações demanda novas ferramentas, e sua complexidade requer uma consideração extra no que diz respeito a sua representação visual, afim de destacar hierarquias, revelar padrões e apresentar relacões entre os dados através de múltiplas dimensões. De acordo com Vassão:

> Uma possível resposta é que o “complexo” é aquilo que está além da nossa compreensão. Outra possível resposta é que o “complexo” é um conjunto de coisas “simples”, percebidas como algo complexo apenas pela acumulação de simplicidades muito numerosas. (...) Por outro lado, uma maneira de compreender-se a Complexidade é como um conjunto muito grande, muito extenso, de coisas muito simples – a “complicação” da complexidade, ou seja, nossa dificuldade de compreendê-la, é apenas conseqüência do acúmulo de entidades muito numerosas. Como diriam alguns, a Complexidade é um conjunto de simplicidades. (VASSÃO, 2010, p.24)

De acordo com Ben Fry, para lidar corretamente com problemas de visualização de dados complexos, diversos campos precisam ser conectados como parte de um processo unificado. Combinando-se as habilidades necessárias em um único, claramente documentado campo, o processo se torna mais acessível para aqueles que tem apenas uma parte do conhecimento. Designers podem aprender a ciência da computação necessária para lidar com grandes quantidades de informação, ou estatísticos podem comunicar suas informações de forma mais efetiva através do entendimento de princípios básicos para representação de dados.

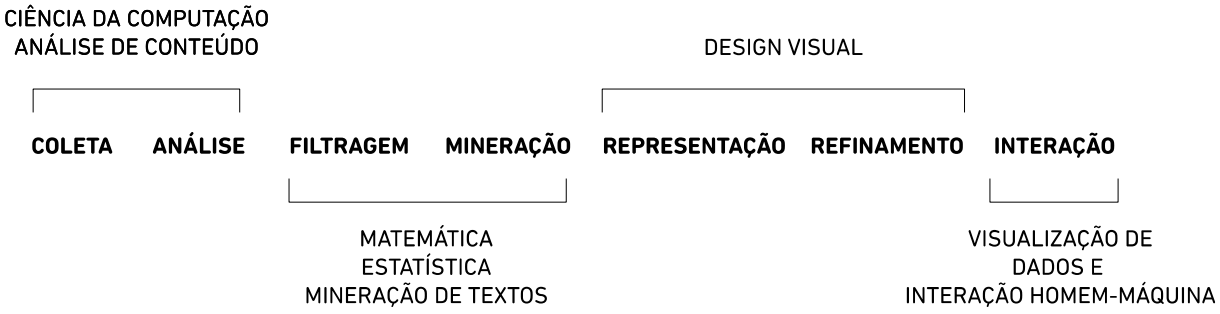
Os métodos não são novos, mas seu isolamento em campos individuais tem impedido que sejam utilizados como um todo, uma vez que é raro uma única pessoa ter conhecimento em todas as áreas envolvidas.

Numa tentativa de adquirir um melhor entendimento das informações, campos como "visualização da informação", "mineração de dados", “design visual” e "análise de conteúdo" são empregados, cada um resolvendo uma parte isolada de um problema específico, mas falhando num sentido mais amplo. De acordo com Ben Fry:

> Parte do problema com as abordagens individuais para lidar com dados é que a separação dos campos acaba levando cada pessoa a resolver uma parcela isolada do problema, e ao longo do processo, algo é perdido em cada transição de uma etapa para a seguinte. É como um telefone sem-fio, onde cada etapa do processo subtrai aspectos da questão inicial a ser considerada. (FRY, 2004)

Como solução para este problema, Fry propõe que esses campos independentes sejam agrupados como parte de um processo unificado entitulado "Design de Informação Computacional", uma metodologia combinada para exploração, análise, e representação de dados complexos.

Figura 19, Etapas do processo de Design de Informação Computacional



Coleta

Diz respeito a coleta dos dados, no caso deste projeto, corresponde a etapa de indexação dos trechos de vídeo com suas *tags*.

Análise

Adiciona alguma estrutura às *tags* cadastradas, organizando-as em categorias.

Filtragem

Limpeza e manutenção das *tags*.

Mineração

Aplicação de métodos para identificar padrões e tendências em um contexto matemático.

Representação

Definição da forma gráfica para representação básica dos dados.

Refinamento

Melhorias na representação básica a fim de torná-la mais clara e atraente.

Interação

Adição de métodos para manipulação dos dados ou controles dos elementos visíveis.

Para o desenvolvimento do protótipo no próximo capítulo, vou percorrer o processo de Design de Informação Computacional, destacando a interdepedência entre as etapas desde a coleta de dados até a interação com a informação.

3 **PROTÓTIPO**

3.1 **O caso do projeto “Memórias da Vila Madalena”**

Para aplicar o processo descrito no capítulo anterior, escolhi o projeto "Memórias da Vila Madalena", do Museu da Pessoa. Desde o início da pesquisa estive em contato com a equipe do Museu da Pessoa para entender sua metodologia de trabalho e avaliar qual seria um bom projeto que fornecesse subsídios reais para o desenvolvimento e teste do sistema aqui proposto. Inicialmente me interessei pelo projeto Imigrantes (que já havia sido transcrito e indexado), mas nem todos os vídeos estavam disponíveis em formato digital. Em uma das visitas ao Museu da Pessoa fui apresentado ao projeto "Memórias da Vila Madalena". As entrevistas haviam sido gravadas poucos dias antes e havia a disponibilidade imediata dos vídeos brutos em formato digital. Nada havia sido ainda transcrito ou indexado.

Além da parceria firmada com o Museu da Pessoa, essa escolha tem fundamento prático. Esse é um projeto pequeno com entrevistas de aproximadamente 30 minutos, o que viabiliza o processo de indexação em um curto espaço de tempo. São 12 entrevistas coletadas no dia 26/08/2012, na Vila Madalena, São Paulo. Homens e mulheres de diversas idades deram seu depoimento de história de vida. As entrevistas seguem um roteiro cronológico focado na história de vida de cada pessoa com ênfase nas memórias da Vila Madalena.

Figura 20, Os depoentes do projeto “Memórias da Vila Madalena”

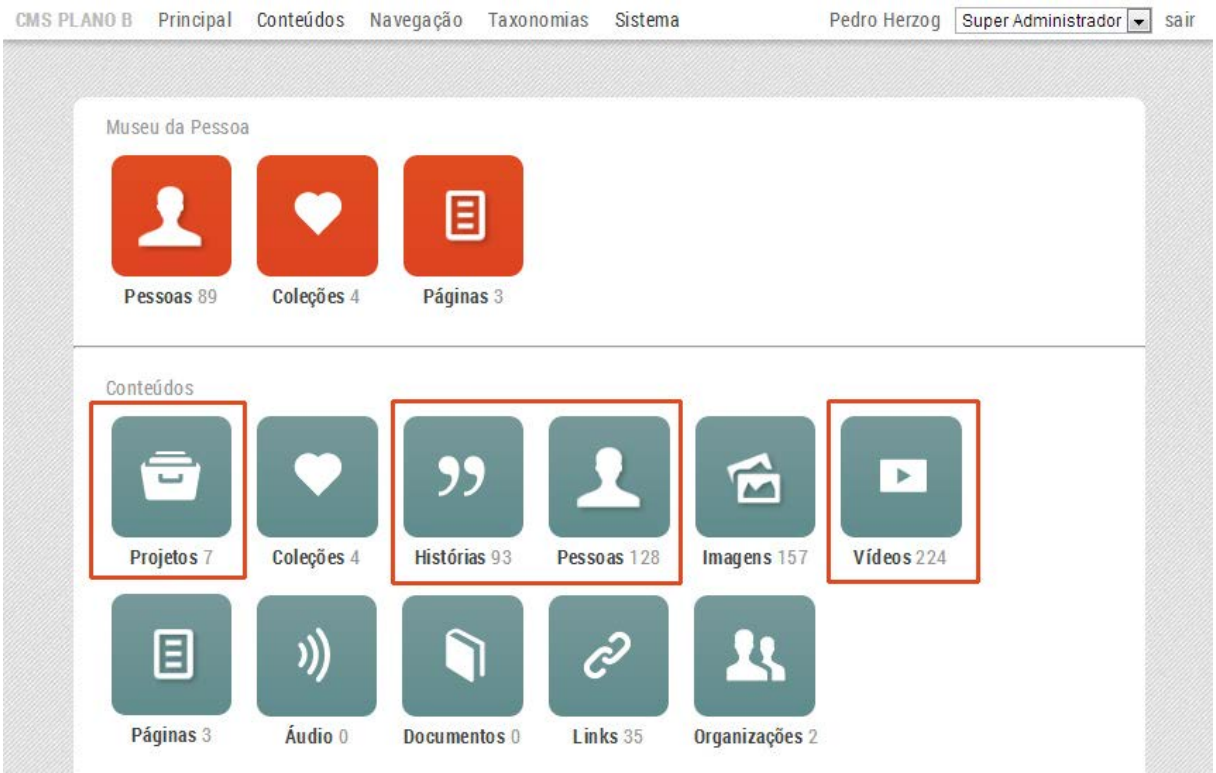
3.2 **A plataforma**

Para realizar a indexação dos trechos de vídeo, vou utilizar um *Content Management System* (*CMS*) próprio[16]. A plataforma Shiro 1.0, ainda em fase beta, é um *CMS* desenvolvido no *framework Code Igniter*[17] (PHP/MySQL). É uma plataforma flexível para cadastro, relacionamento e anotação (*taggeamento*) de entidades, que se adapta às necessidades de cada projeto. O pesquisador do Museu utiliza o ambiente administrativo da plataforma para indexar os trechos de vídeo.

3.2.1 Entidades

Para este projeto de História Oral, as seguintes entidades foram utilizadas: (1) projetos, (2) histórias, (3) pessoas e (4) vídeos.

Figura 21, Tela principal do CMS Shiro 1.0 beta

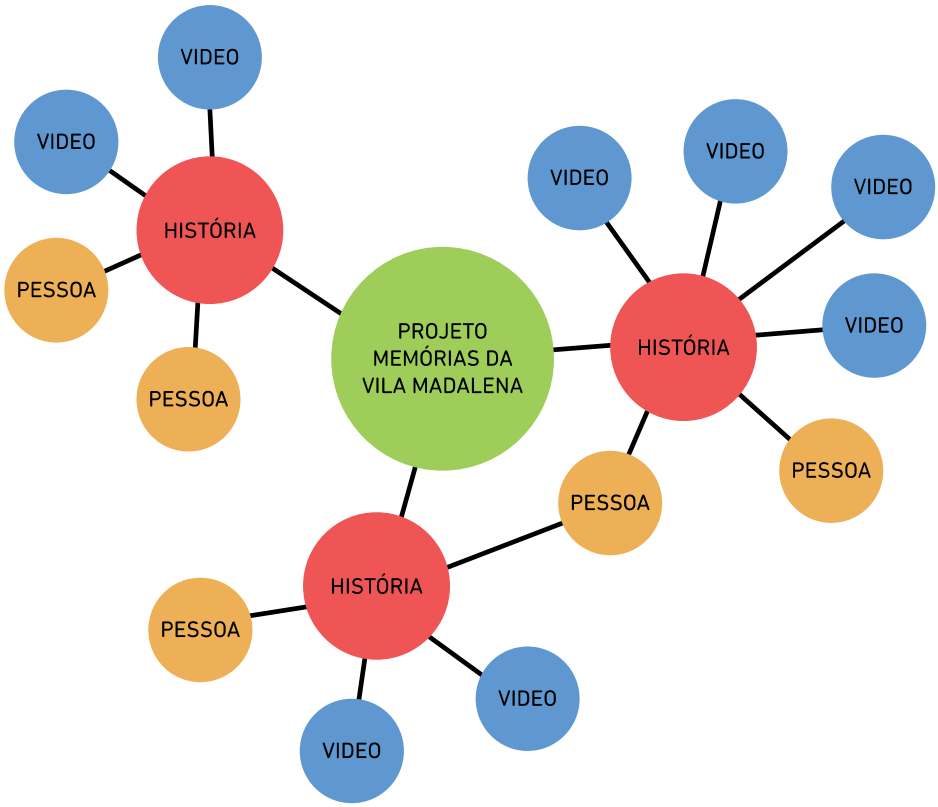


[16] Desenvolvido com Sergio Boiteux e equipe (na Plano B Design)

[17] CodeIgniter é uma estrutura *open-source* (código-aberto) para desenvolvimento rápido de *websites* dinâmicos em PHP. Disponível em: http://ellislab.com/codeigniter Acessado em: 20/01/2014

Essas entidades podem se relacionar livremente de acordo com os objetivos do projeto. Para indexação dos segmentos de vídeo, vamos agrupá-los conforme o diagrama a seguir:

Figura 22, Diagrama de relacionamentos entre entidades no CMS.

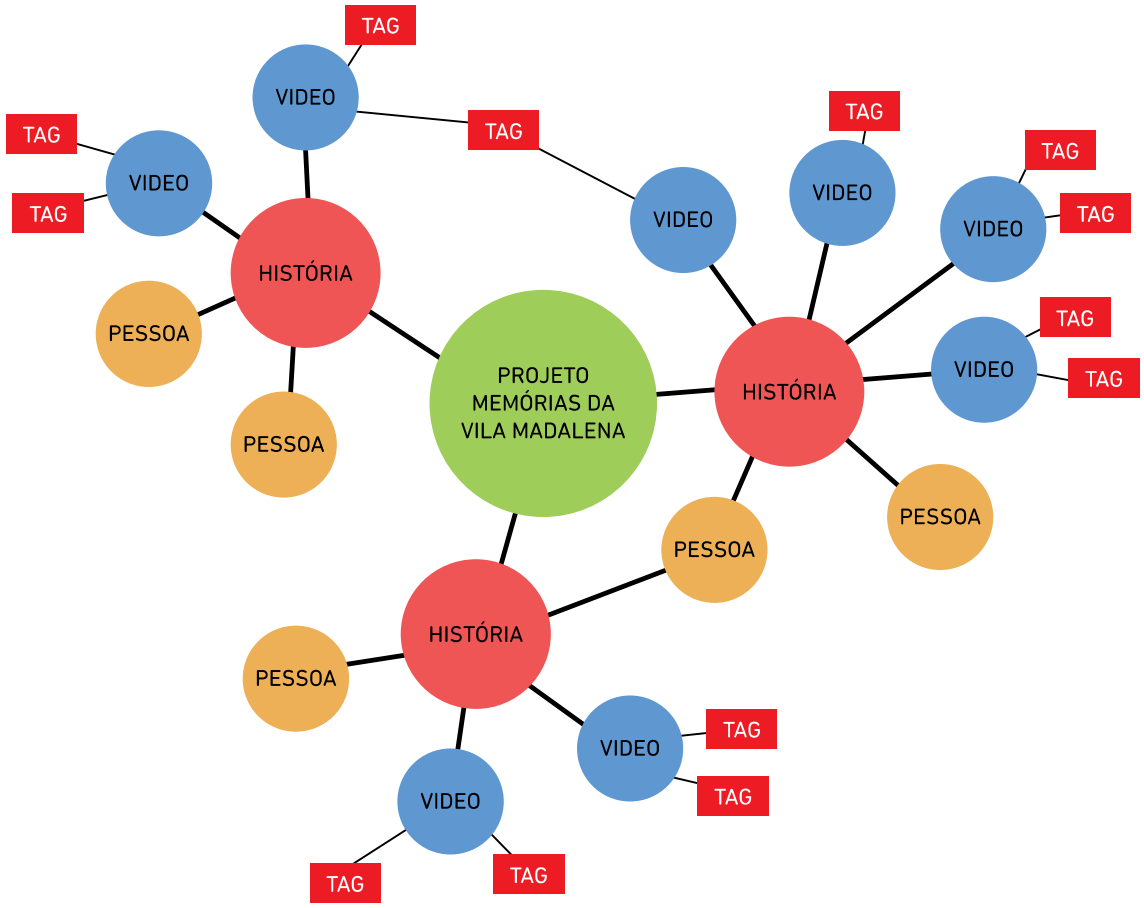


A entidade “projeto” reúne todas as “histórias” (no diagrama aparecem apenas 3 histórias, mas neste caso teremos os 12 depoimentos). As entidades “histórias”, reúnem “pessoas” (depoente e entrevistador) e “vídeos”. Cada vídeo corresponde a um trecho da entrevista.

3.2.2 Anotações ou *Tags*

Embora a plataforma permita que todas as entidades recebam *tags*, neste projeto vamos atribuir *tags* apenas as entidades “vídeos”.

Figura 23, Diagrama de relacionamento entre vídeos e *tags*.



Um vídeo pode receber mais de uma *tag*. Uma mesma *tag* pode ser compartilhada por diferentes vídeos.

3.2.3 Taxonomias

A plataforma Shiro permite a construção de taxonomias para diversas aplicações. Neste projeto, vamos criar uma taxonomia para agrupar as *tags* cadastradas em temas ou categorias. Dessa forma, cada tema contém uma lista de *tags* relacionadas. Embora o sistema permita, esse projeto não utiliza sub-categorias na taxonomia.

3.3 **Etapas do processo**

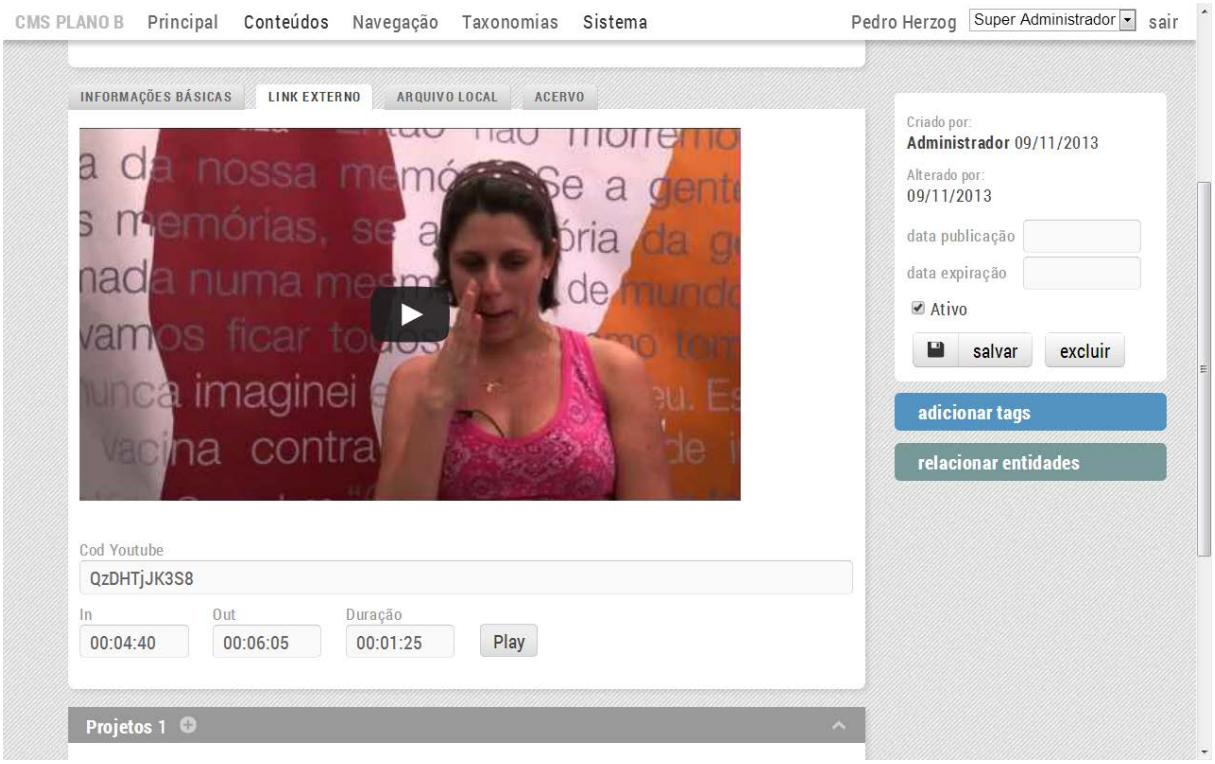
3.3.1 Coleta

Os 12 vídeos coletados pela equipe do Museu da Pessoa são cadastrados na plataforma *CMS*. Como este é um projeto pequeno, vamos utilizar os servidores de mídia do YouTube para reprodução dos vídeos. Para projetos maiores, outras soluções de hospedagem e *streaming* de vídeos podem ser necessárias para garantir uma boa performance. O sistema também pode funcionar com arquivos locais, onde o acesso a internet é restrito.

Uma vez cadastrados no YouTube, podemos acessar o vídeo a partir da plataforma *CMS*.

Na tela de cadastro de vídeo, além do campo para o código do YouTube, o sistema possibilita a marcação de um início (*in*) e um fim (*out*) para definir o trecho a ser indexado.

Figura 24, Tela para cadastro de vídeo.

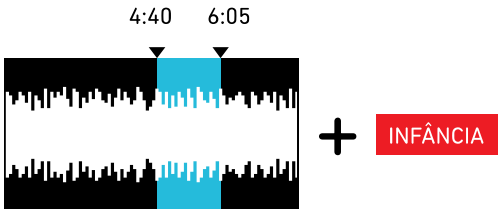


A partir desses campos o sistema gera um código com a URL no padrão especificado pelo YouTube.

```
<iframe width=”560” height=”315”
src=”http://www.youtube.com/v/W-NCDovAWB8?start=04:40&end=06:05&autoplay=1
&rel=0&showinfo=0&autohide=1&version=3” frameborder=”0” allowfullscreen>
</iframe>
```

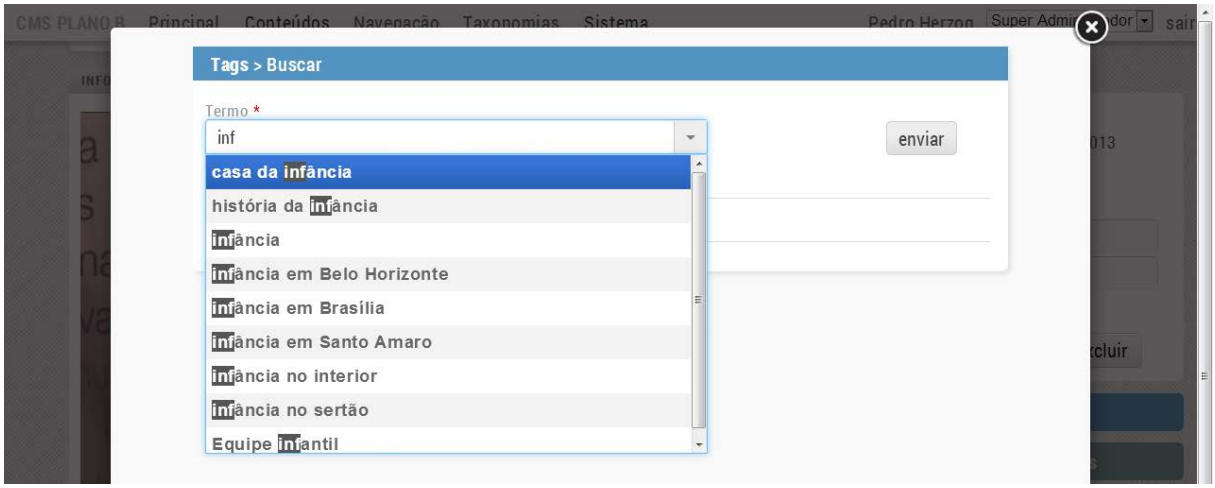
Definido o trecho, o pesquisador adiciona uma anotação livre ou *tag* para identificar a história ou o assunto e facilitar sua recuperação.

Figura 25, O trecho é indexado com uma ou mais *tags.*



A plataforma *CMS* utiliza um formulário *autofill* que apresenta as *tags* já cadastradas no sistema para que não sejam re-cadastradas desnecessariamente. Além disso, o *autofill* ajuda a evitar erros de digitação, pluralização e outras inconsistências no cadastro das *tags.*

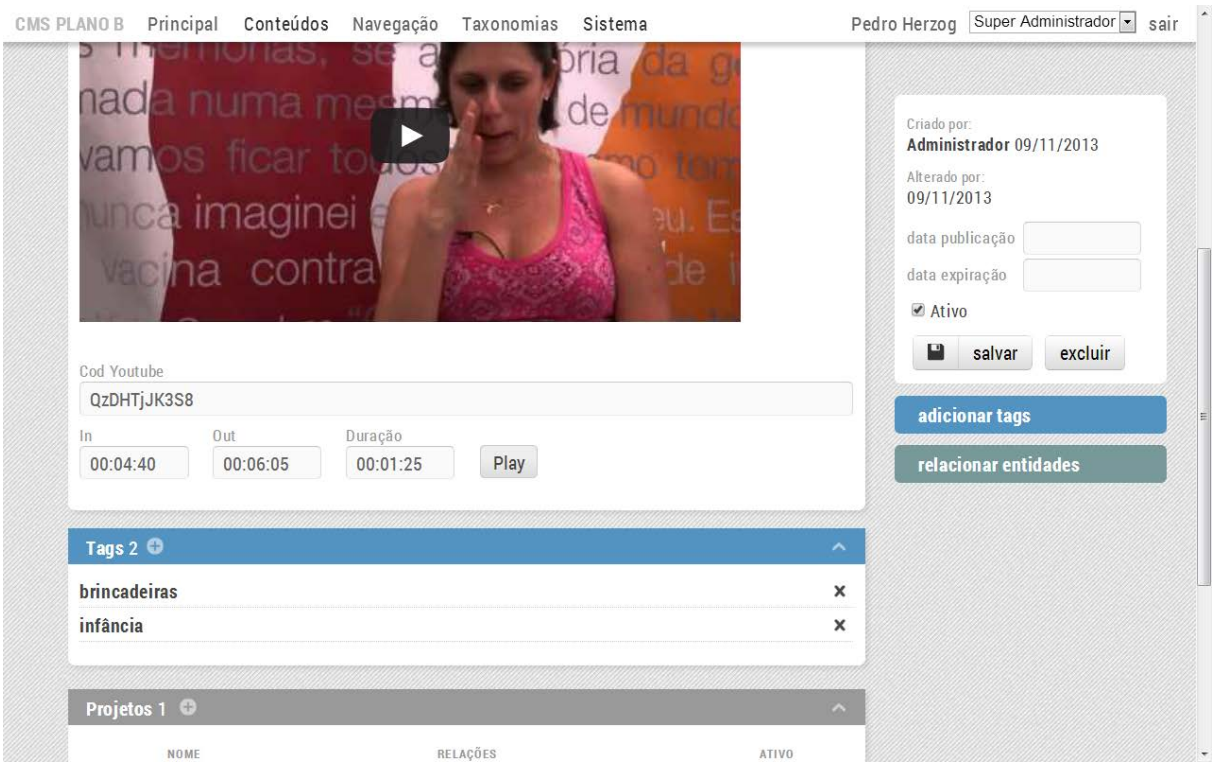
Figura 26, Formulário *autofill* para cadastro de *tags.*



Se a *tag* escolhida pelo pesquisador não estiver presente na lista oferecida pelo *autofill,* a nova *tag* pode ser imediatamente cadastrada.

O pesquisador pode utilizar mais de uma *tag* para indexar um trecho de vídeo. Em uma conversa sobre “infância”, por exemplo, a depoente falou de muitas “brincadeiras”. Estas duas tags podem ser empregadas, ampliando as chances de acessarem aquele trecho a partir de uma busca. As *tags* aumentam o grau de recuperação dos trechos de vídeo.

Figura 27, *Tags* relacionadas ao trecho de vídeo indexado.



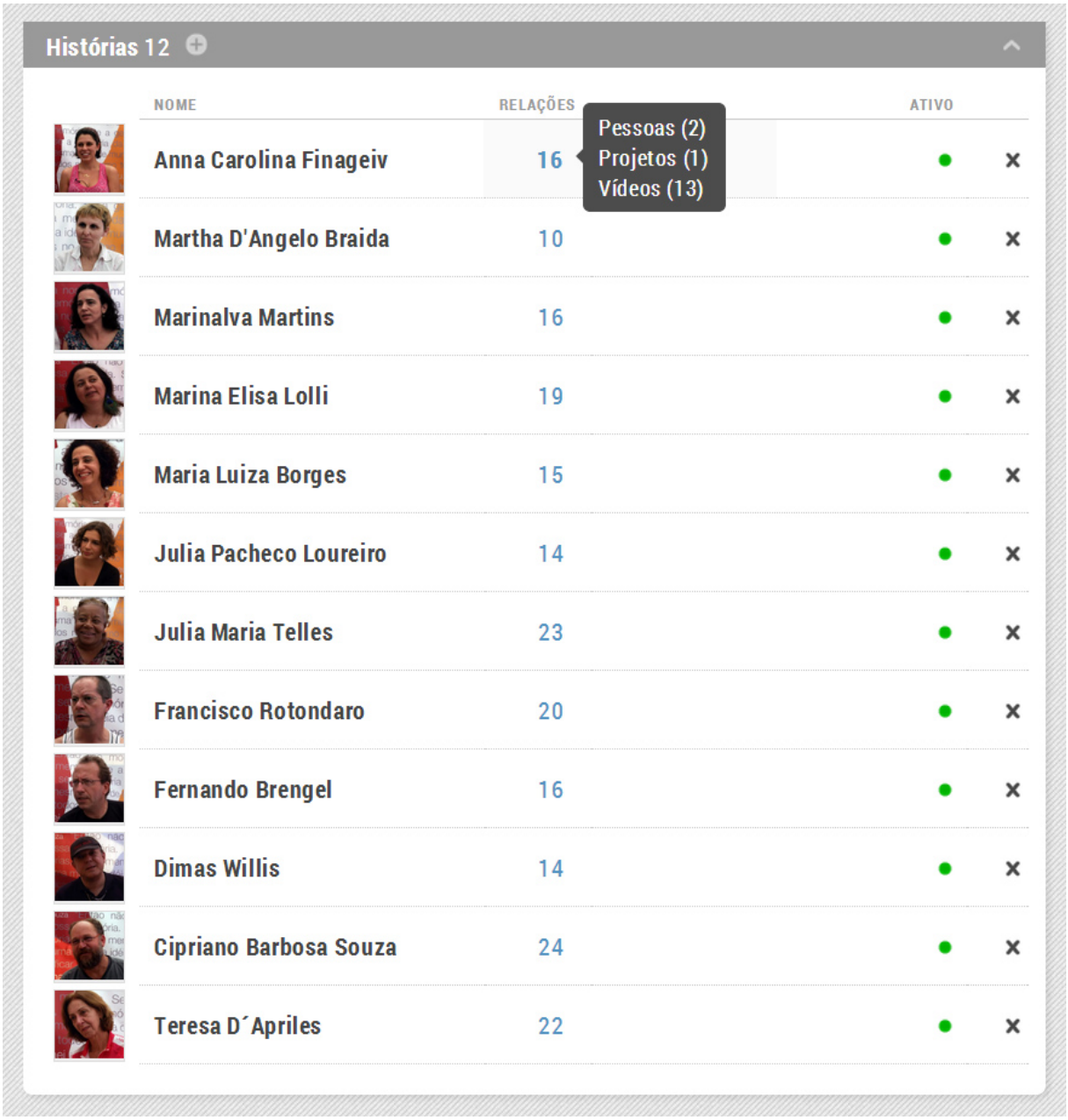
Verena Alberti chama a atenção para a importância de uma avaliação quantitativa dos termos empregados, a fim de equilibrar a recuperação dos itens indexados:

> Uma etapa fundamental no desenvolvimento de base de dados é a definição dos descritores (ou temas) que darão conta dos assuntos tratados nas entrevistas. Novamente é necessária a criação de uma subtabela para uniformizar todos os descritores, que devem ser suficientemente, mas não excessivamente, abrangentes. Por exemplo, de pouco adianta criar um descritor ao qual será relacionada apenas uma entrevista do acervo, mas também não é produtivo criar um descritor ao qual estão vinculadas todas as entrevistas sem exceção. (ALBERTI, 2005 p. 141)

Como resultado de uma indexação preliminar, apresentamos abaixo as 12 histórias cadastradas no sistema. Os números presentes na coluna “relações” referem-se as entidades (“pessoas”, “vídeos” e “projetos”) relacionadas àquela história. Neste caso, toda história tem sempre 1 projeto e 2 pessoas relacionadas

(depoente e entrevistador), portanto o número restante refere-se aos trechos de vídeos indexados.

Figura 28, Os doze depoentes e os trechos de vídeo indexados.



3.3.2 Análise

A plataforma *CMS* permite a criação de taxonomias para diversas aplicações. Neste caso, vamos criar uma taxonomia para agrupar as *tags* em categorias ou temas.

A definição desses temas está relacionada com os objetivos do projeto e tem base no próprio roteiro de perguntas. No caso do projeto "Memórias da Vila Madalena", diversos assuntos foram abordados: origem, família, infância, escola, juventude, amor, universidade, etc. O roteiro não é fixo, dependendo da história que está sendo contada novas perguntas são formuladas e outras nem são realizadas. Abaixo uma lista com alguns dos termos cadastrados para ilustrar a heterogeneidade das *tags* livres:

Figura 29, Lista de *tags.* Veja no anexo I a lista completa das *tags* cadastradas na etapa anterior.

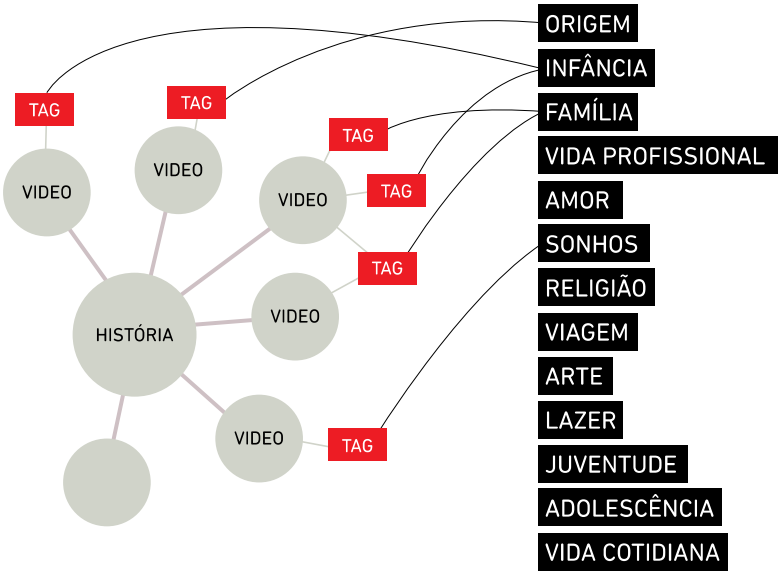
A partir de uma análise rápida do universo de *tags* cadastradas, definimos os seguintes temas como ponto de partida:

Figura 30, Lista preliminar de temas.

Figura 31, Cadastro da taxonomia “Vila Madalena” com os temas preliminares.

A plataforma permite o relacionamento das *tags* livres com essas estruturas taxonômicas.

Figura 32, Diagrama de relacionamento entre *tags* e temas.



Ao longo do processo, o pesquisador pode avaliar se os temas escolhidos devem ser mantidos, renomeados ou excluídos sem o prejuízo das *tags* cadastradas. Novos temas poderão ser criados e outros podem ser mesclados. Os critérios para definição dos temas variam de projeto para projeto.

Como resultado final deste processo, os temas são apresentados na interface como pontos de acesso para navegação do usuário.

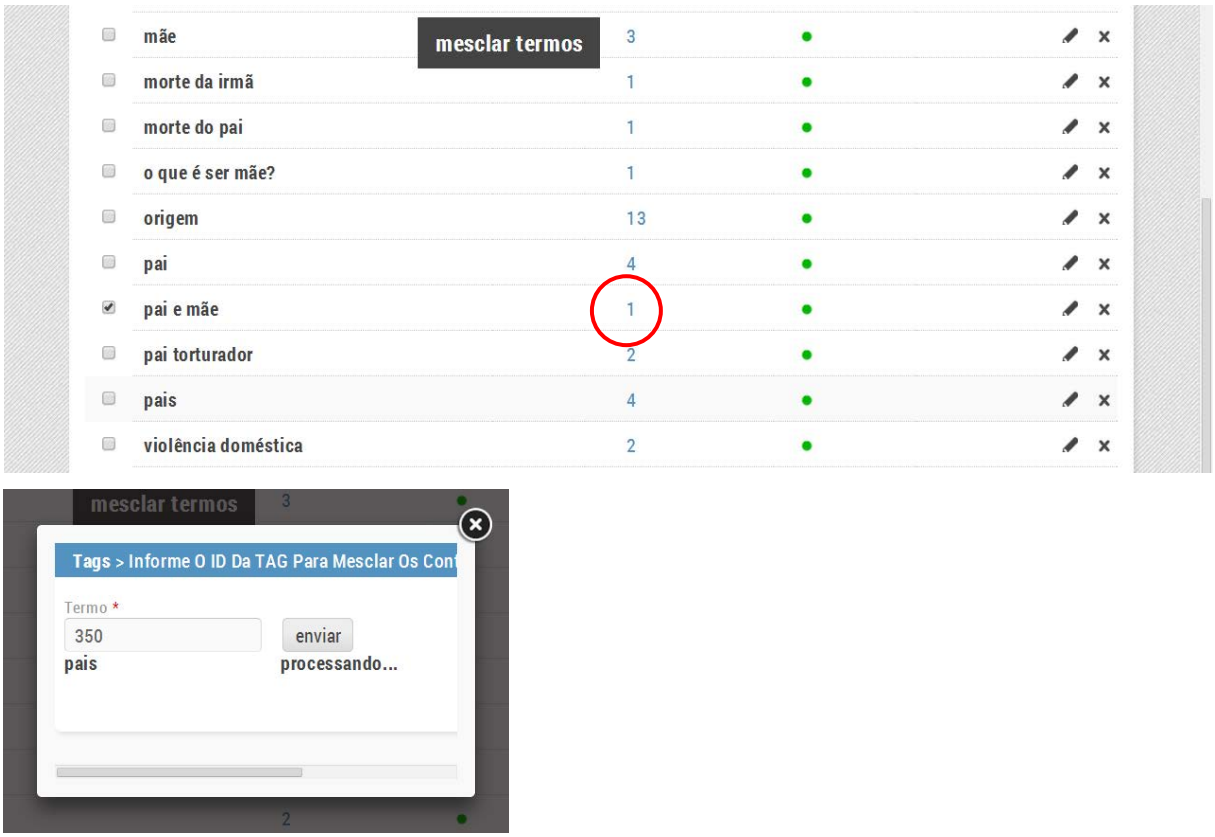
3.3.3 Filtragem

Nessa etapa os pesquisadores realizam a gestão das *tags*. A indexação por *tags* carrega as limitações descritas no capítulo anterior. Embora tenhamos utilizado o *autofill* na etapa de coleta, propriedades típicas da linguagem como sinonímia (grafias diferentes com o mesmo significado) e polissemia (mesma grafia com significados diferentes) podem causar inconsistências nas anotações. Além disso, erros de digitação, capitalização inconsistente ou pluralização podem acidentalmente levar ao uso de múltiplas anotações com a mesma intencionalidade, impactando na análise quantitativa das *tags*. Por tudo isso, é importante um monitoramento constante das *tags* cadastradas.

Quando listamos as *tags* relacionadas ao tema “família”, identificamos os termos “pais” e “pai e mãe”. Entendendo que são *tags* com o mesmo significado, vamos migrar o termo “pai e mãe”, que só tem uma relação, para o termo “pais”.

Na plataforma CMS, para mesclar *tags*, basta selecionar as *tags* que serão mescladas e informar o ID da *tag* de destino.

Figura 33, Mesclando o termo “pai e mãe” com “pais”.



No exemplo acima, mesclamos os termos “pai e mãe” e “pais”. Repare que as relações são preservadas. A *tag* “pais” herdou a relação antes atribuída a *tag* “pai e mãe”, por isso, o número de relações passou de 4 para 5.

Figura 34, Depois de mescladas, as tags preservam as relações criadas anteriormente.

3.3.4 Mineração

A plataforma oferece uma lista com todas as *tags* cadastradas no sistema. Como o projeto “Memórias da Vila Madalena” compartilha o mesma base de dados de outros projetos do Museu da Pessoa, a listagem apresentada na figura inclui *tags* não utilizadas ou relacionadas com temas da taxonomia “Vila Madalena”.

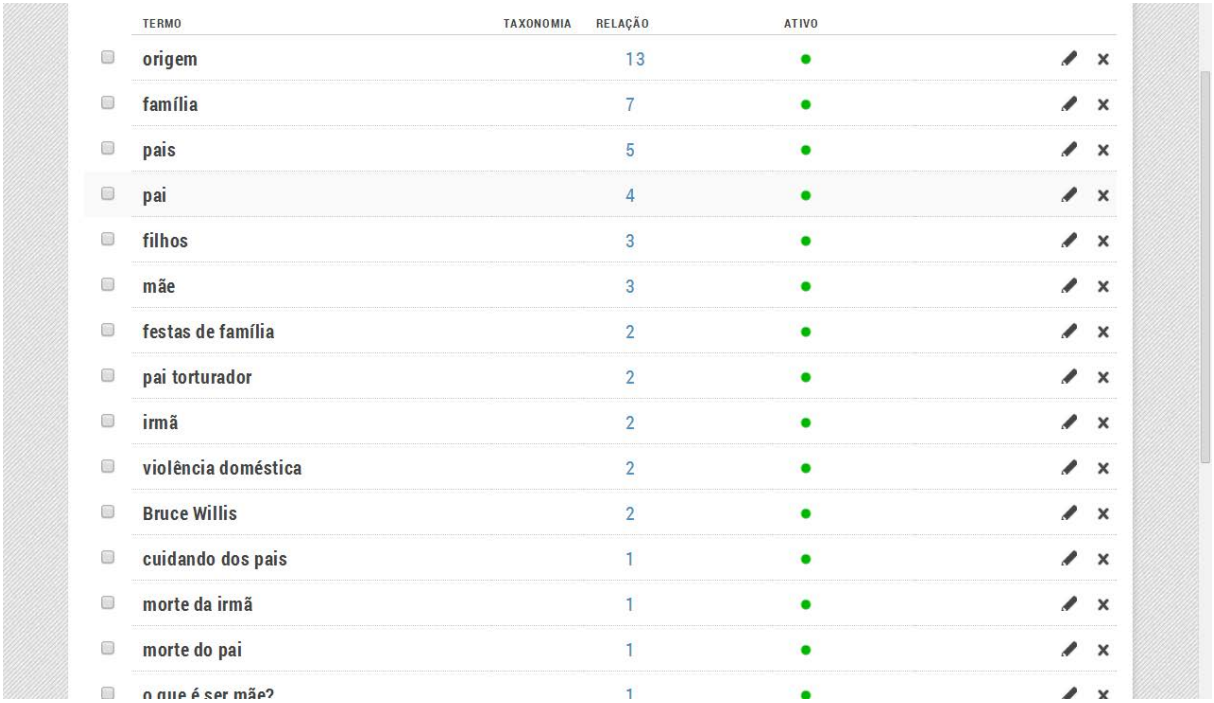
Figura 35, Listagem das *tags* cadastradas, ordenadas pelo número de relações.

Figura 36, A cauda longa com as *tags* da taxonomia "Vila Madalena".

De qualquer forma, nessa etapa já é possível confirmar uma tendência no processo de indexação livre (ou *taggeamento*): poucos termos são utilizados muitas vezes, e a grande maioria de termos é empregado poucas vezes, ou mesmo uma única vez.

Quando listamos apenas as *tags* relacionadas ao tema “família” da taxonomia “Vila Madalena”, a tendência se repete.

Figura 37, Listagem das tags relacionadas ao tema “família” da taxonomia “Vila Madalena”.



TERMO	TAXONOMIA	RELAÇÃO	ATIVO
origem		13	
família		7	
pais		5	
pai		4	
filhos		3	
mãe		3	
festas de família		2	
pai torturador		2	
irmã		2	
violência doméstica		2	
Bruce Willis		2	
cuidando dos pais		1	
morte da irmã		1	
morte do pai		1	
o que é ser mãe?		1	

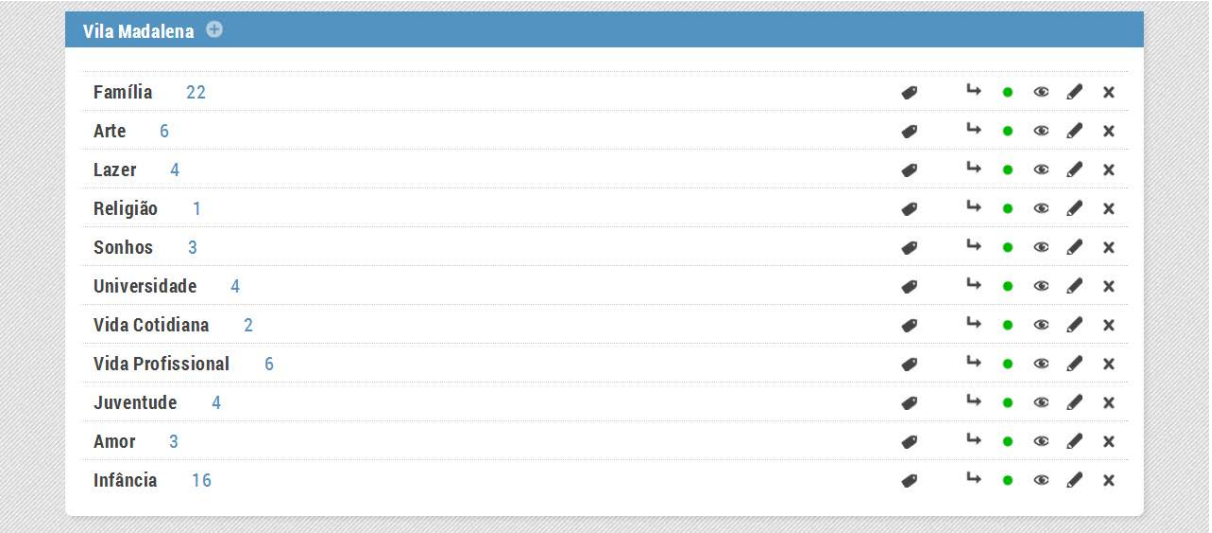
Também nessa etapa são contabilizadas as *tags* distribuídas pelos temas na taxonomia “Vila Madalena”.

Figura 38, Número de *tags* relacionadas aos temas da taxonomia.

Família	24
Arte	4
Lazer	3
Religião	1
Sonhos	2
Universidade	4
Viagem	1
Vida Cotidiana	2
Vida Profissional	6
Juventude	3
Adolescência	1
Amor	3
Infância	16

Na figura acima, as categorias “religião”, “adolescência” e “viagem” tem apenas uma *tag* relacionada, não justificando a existência desses temas. A tag “adolescência em SP” foi relacionada com o tema “juventude”, a tag “viagem” foi relacionada com o tema “lazer”, a tag “religião” foi utilizada na indexação de 2 vídeos, justificando sua permanência. Este é um critério arbitrário.

Figura 39, Resultado da reorganização dos temas da taxonomia “Vila Madalena”.



Repare que um tema com muitas *tags* **não é necessariamente** o tema mais povoado de vídeos. Um tema com poucas *tags* muito utilizadas pode ser mais povoado do que um outro com muitas *tags* empregadas poucas vezes.

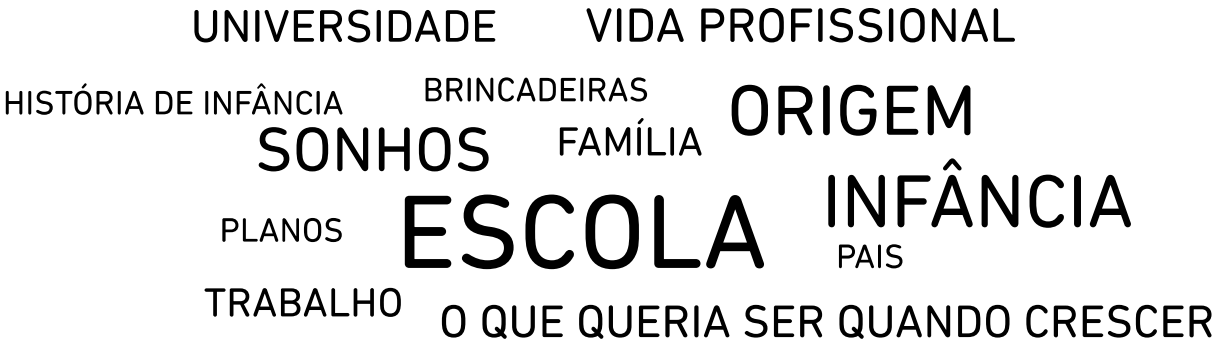
3.3.5 Representação

Nesta etapa vamos tratar da representação básica das *tags* e taxonomias.

3.3.5.1 *Tags*

É freqüente o uso das nuvens de *tags* para representação de um universo de *tags.* Nesse tipo de representação, os termos mais empregados aparecem com maior peso visual e destaque.

Figura 40, Nuvem de *tags.*



As nuvens de *tags,* são frequentemente utilizadas para navegação: clicando em uma das *tags* o usuário é direcionado para uma página contendo todas as entidades em que aquela *tag* foi atribuída.
De acordo com Stefaner[18], embora as nuvens de *tags* representem definitivamente um dos mais bem sucedidos exemplos de visualização de dados nos últimos tempos, elas tem algumas limitações:

- Nuvens de *tags* não são suficientes para navegação efetiva pela cauda longa: Com o tempo, emerge um padrão onde teremos algumas *tags* dominantes (a "cabeça") e um grande número de *tags* raramente utilizadas (a "cauda longa"). Enquanto as tags populares da "cabeça" permanecem estáticas, em destaque, caracterizando os tópicos mais abrangentes (mais utilizados), a "cauda longa" contém os termos mais específicos. A nuvem de *tags* prioriza visualmente a "cabeça". Entretanto, tanto para busca como para exploração, o acesso a cauda longa é vital, já que é onde está contida a informação que diferencia o objeto indexado dos demais.

- Para a representação de um grande número de termos, uma nuvem de *tags* pode ficar um tanto confusa. Nesse tipo de representação, as *tags* menos utilizadas são em geral eliminadas para garantir a legibilidade dos termos apresentados. Além disso, a variação de peso não oferece um julgamento preciso das quantidades. O próprio tamanho das palavras (número de letras) influencia nossa percepção. No exemplo acima, qual dos termos foi empregado mais vezes: UNIVERSIDADE ou O QUE QUERIA SER QUANDO CRESCER? A pequena diferença no corpo das letras não é suficiente para uma resposta segura.

- As nuvens de *tags* não representam a dinâmica das *tags*, isto é, não leva em consideração as últimas *tags* adicionadas. Para resolver esse problema, Chirag Mehta[19] implementou um *slider* temporal. Entretanto, outro problema ficou evidente: Nuvens de *tags* não funcionam muito bem com animação ou

---

18 Stefaner, M. 2007 Visual Tools for the Socio-Semantic Web (p.55)

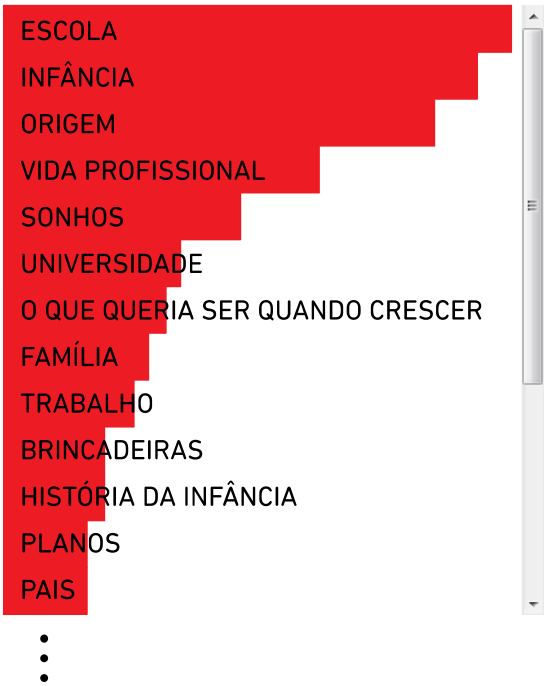
19 Disponível em: http://chir.ag/projects/preztags/ Acessado em: 20/01/2014

transições suaves em função do tempo. Como a posição de cada tag é definida pelo tamanho e posição da tag anterior (no caso da ordenação alfabética), os termos começam a “pular” quando as *tags* são escalonadas, ou quando *tags* aparecem ou desaparecem, gerando uma leitura confusa.

Figura 41, Clássica representação da nuvem de *tags*. Os termos estão ordenados alfabeticamente, os termos maiores são os termos mais utilizados, e os termos em cinza claro, representam os termos mais antigos.

Para representação de todos os termos cadastrados vou utilizar o gráfico da cauda longa apresentado na etapa anterior. Mas para facilitar a leitura das *tags,* o gráfico é rotacionado 90º.

Figura 42, O gráfico da cauda longa rotacionado 90º.



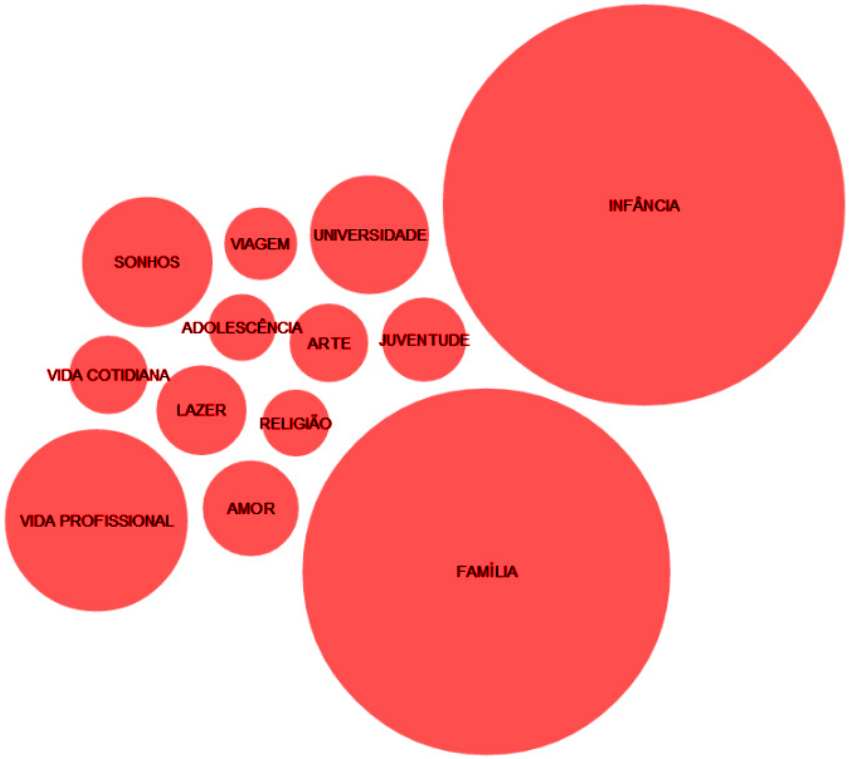
Cada um dos termos é apresentado dentro de uma barra. O tamanho da barra traduz o número de ocorrências de cada *tag.* Dessa forma, fica fácil dizer que a *tag* UNIVERSIDADE foi empregada mais vezes que O QUE QUERIA SER QUANDO CRESCER. A barra de rolagem vertical permite o acesso aos termos menos empregados, no final da cauda longa.

3.3.5.2 Taxonomia

Como vimos anteriormente, em nuvens de *tags,* os termos são ordenados alfabeticamente ou por tamanho – seria interessante se os termos que estão relacionados a um mesmo campo semântico fossem apresentados juntos. Alguns desses relacionamentos podem ser deduzidos automaticamente, pela observação do uso das *tags:* algumas *tags* são frequentemente utilizadas em conjunto (por exemplo: “Vida Profissional” e “Trabalho”). Para lidar com essa questão vamos organizar as *tags* em uma taxonomia.

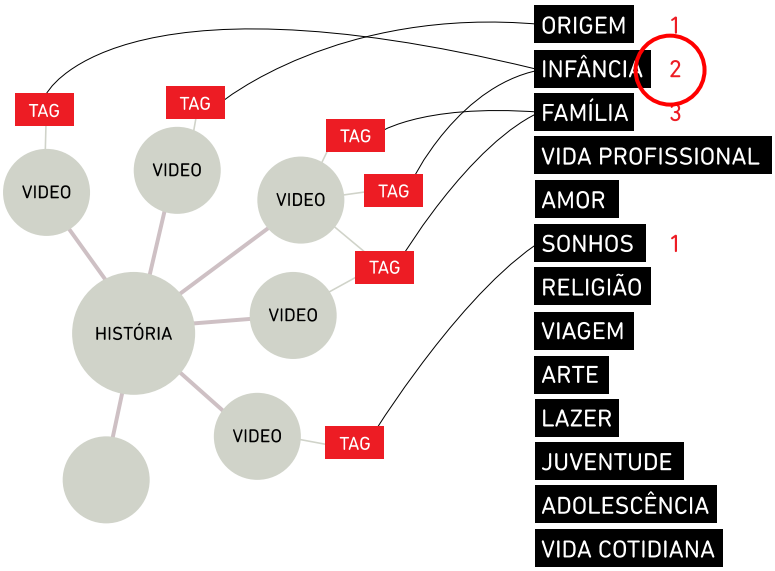
Para representação de taxonomias é comum a utilização de menus e árvores hierárquicas. Mas como a taxonomia criada não tem sub-categorias, neste projeto vamos representar os temas como conjuntos de *tags*. Esses conjuntos funcionam como campos semânticos, agregando as *tags* associadas àquele tema, e são representados por círculos. A área do círculo corresponde a quantidade de vídeos relacionados a *tags* relacionadas aquele tema na taxonomia.

Figura 43, Representação básica dos temas da taxonomia.



Embora o tema FAMÍLIA tenha 24 *tags* e o tema INFÂNCIA apenas 16, as *tags* relacionadas a categoria INFÂNCIA foram empregadas mais vezes. Por isso o conjunto INFÂNCIA aparece maior nessa representação.

Figura 44, Relacionamento do trecho de vídeo com a *tag*, e da *tag* com o tema.



Este gráfico ilustra a lógica para quantificação da ocorrência de videos por tema. Embora na figura, o tema “família” tenha apenas 2 *tags* relacionadas, uma delas foi empregada em 2 vídeos, e por isso totalizam-se 3 ocorrências nesse tema.

3.3.5.3 Navegação básica

Selecionando-se o tema “família”, o sistema apresenta os vídeos com as *tags* relacionadas àquele tema. Os vídeos são ordenados randomicamente em uma matriz de quadrados que rola horizontalmente e se adapta ao formato da tela.

Figura 45, A matriz de vídeos.

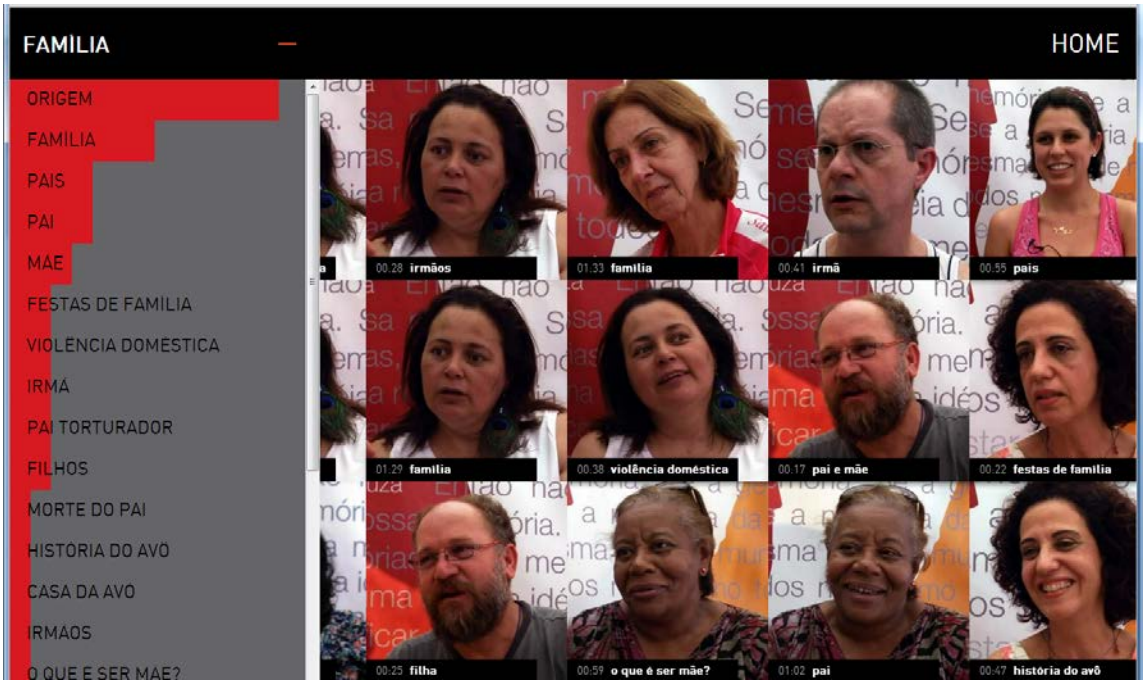


Figura 46, A tela do depoente.

Selecionando-se um dos videos, a interface apresenta o trecho escolhido com as outras *tags* relacionadas ao mesmo depoente, assim como vídeos de outros depoentes com as mesma *tag.*

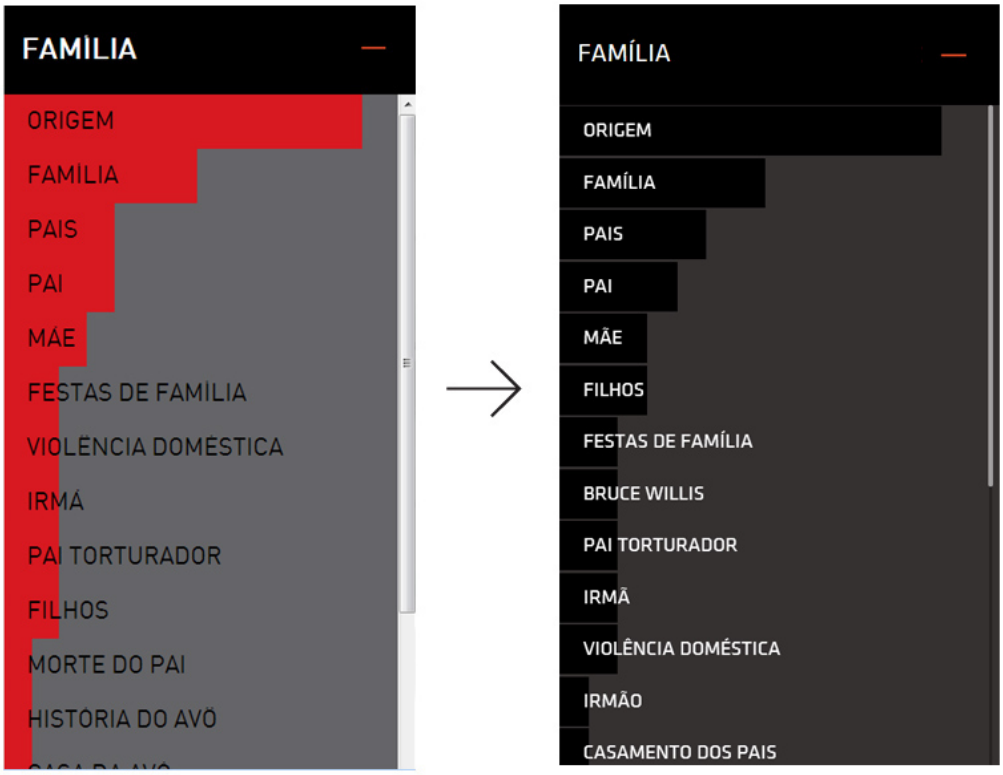
3.3.6 Refinamento

Nessa etapa, o resultado da etapa anterior passa por melhorias no aspecto visual, a fim de torná-lo mais claro e envolvente/atraente.

3.3.6.1 *Tags*

A cauda longa vertical, assim como outros elementos da interface (barras de rolagem, tipografia, cores, etc...), sofreu pequenos ajustes.

Figura 47, Melhorias no visual da cauda longa vertical.



Os gráficos acima foram editados a partir de capturas de tela. Novas *tags* foram cadastradas ao longo dessa etapa, por isso observam-se algumas diferenças na listagem de termos.

Para solucionar problemas de acentuação, a tipografia foi alterada e reduzida, mantendo a espessura das barras para não perder superfície de ativação (área clicável). Foi criada uma linha de separação entre as barras a fim de torná-las mais claras. O contraste entre as barras e o fundo foi reduzido, as barras perderam a cor (a cor será usada para indicar a seleção) e a tipografia passou de preto para branco.

3.3.6.2 Taxonomia

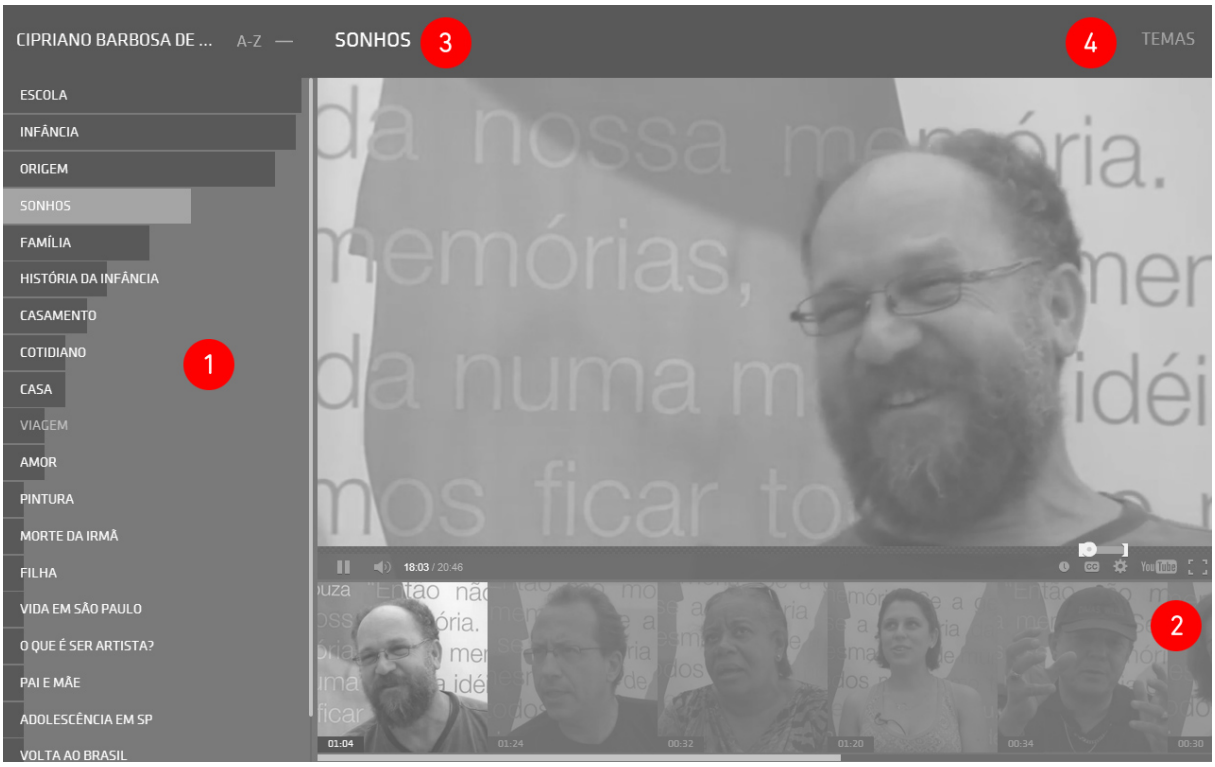
Figura 48, Melhorias visuais na representação dos temas.

Além de ajustes na cor, tanto dos círculos quanto da tipografia, nessa fase foi implementado no código valores mínimo e máximo para a área dos círculos. Esses valores são definidos em função do número de círculos (temas) e do tamanho da tela. Para este projeto, definimos como 25 pixels de raio o valor mínimo para representação de um tema, a fim de garantir uma área clicável confortável.

3.3.6.3 Navegação

A presente proposta pretende oferecer ao usuário uma navegação que permita o acesso aos trechos de vídeo de um mesmo depoente, assim como o relacionamento entre entrevistas, isto é, que o usuário possa escolher entre ver outros assuntos da mesma entrevista, ou outros depoentes falando do mesmo assunto.

Figura 49, Opções de navegação a partir da tela do depoente.



1. Outros trechos deste mesmo depoente;
2. Trechos de outros depoentes indexados com a mesma *tag*;
3. Acesso a matriz de trechos de depoimentos daquele tema;
4. Acesso aos temas (tela principal).

Figura 50, A matriz de vídeos depois do refinamento.

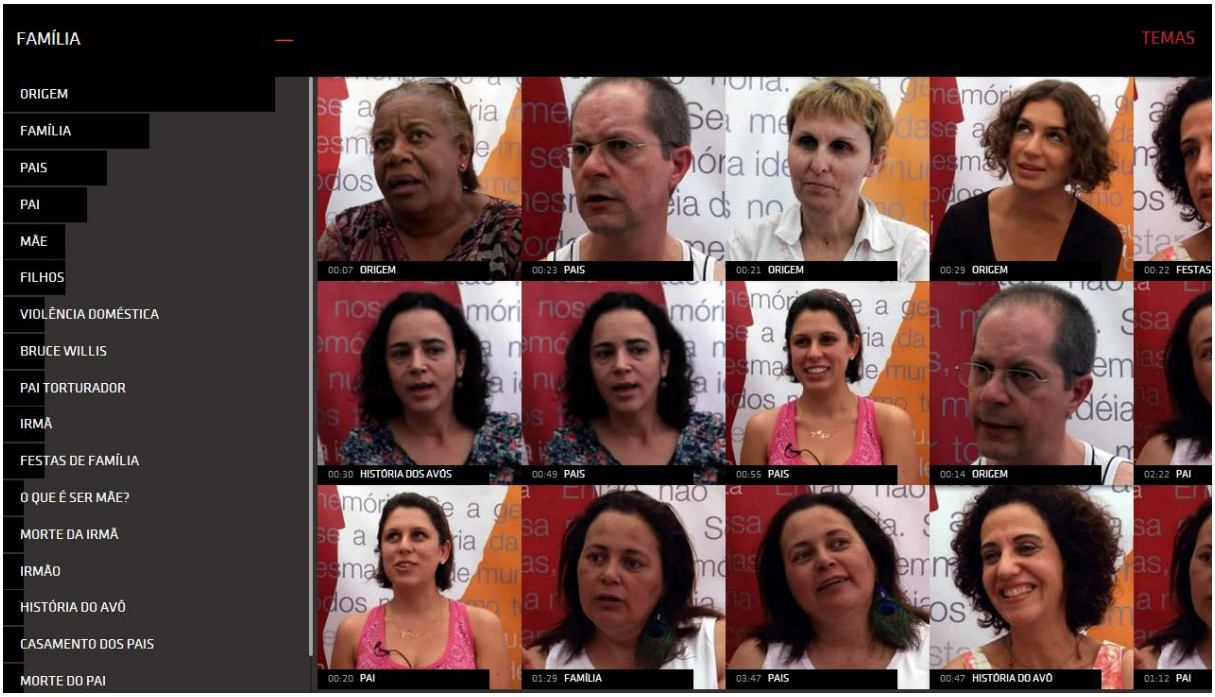


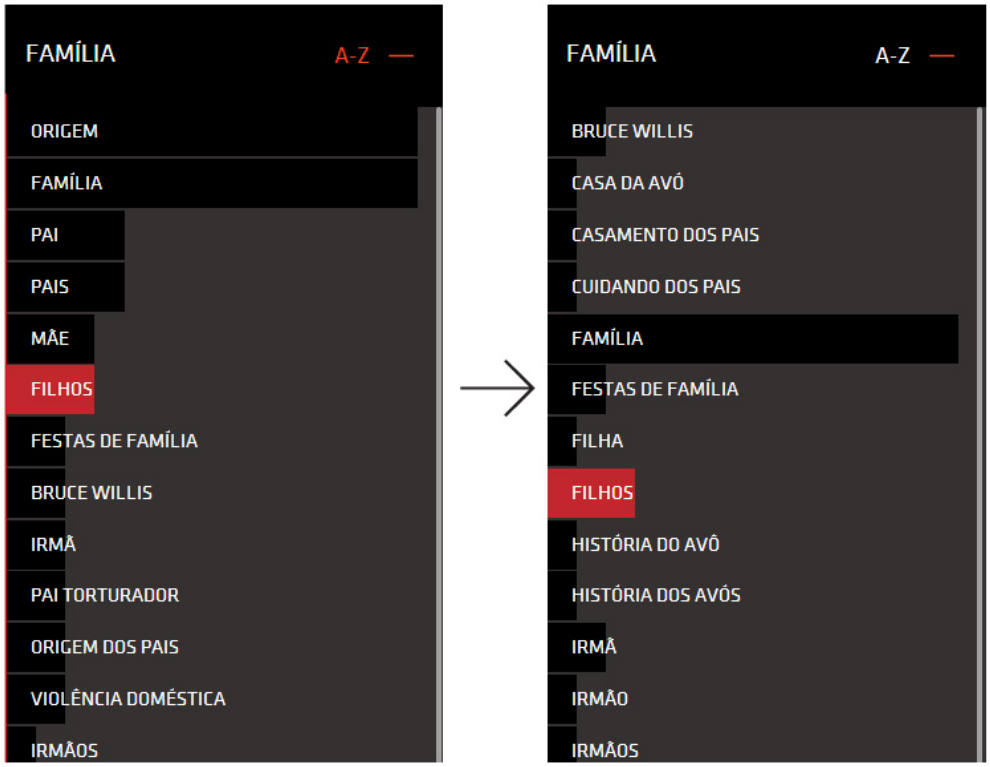
Figura 51, A tela do depoente depois do refinamento.

3.3.7 Interação

Essa etapa diz respeito aos métodos adicionados à etapa anterior para otimizar a navegação, ordenação e controle dos elementos.

No topo da barra lateral destinada para a cauda longa vertical adicionamos um controle para ordenação alfabética das *tags*.

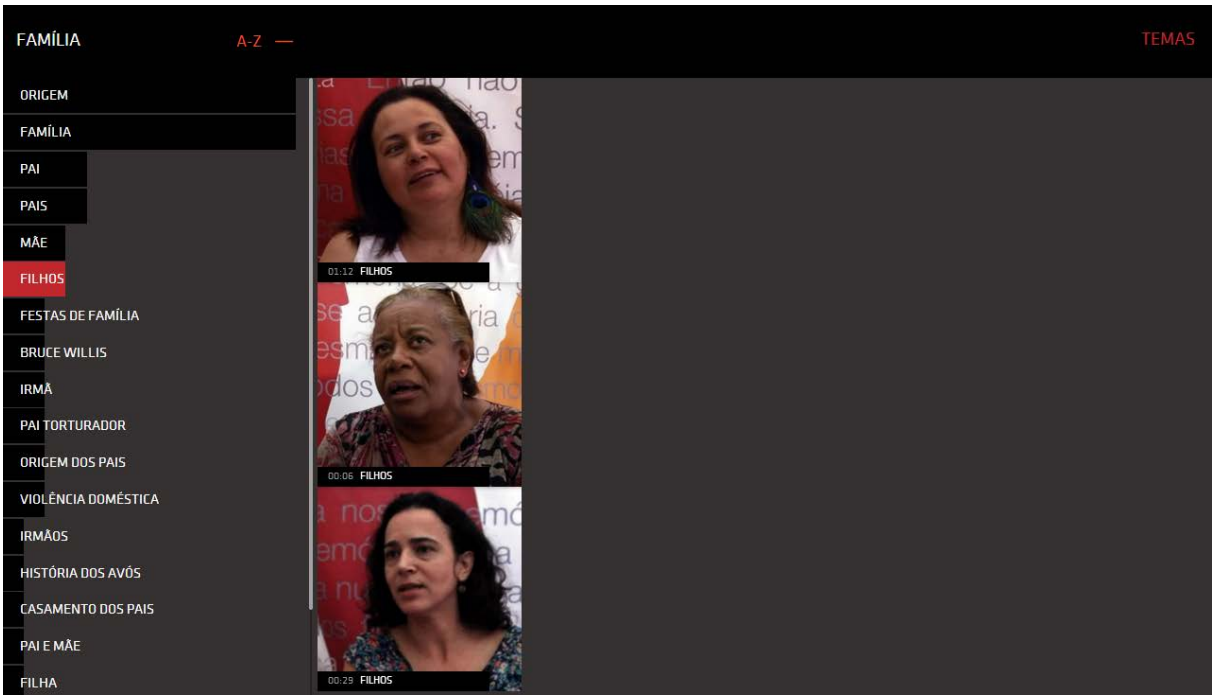
Figura 52, A ordenação alfabética dos termos.



A ordenação alfabética permite que o usuário encontre tags específicas sem ter que percorrer toda a cauda longa.

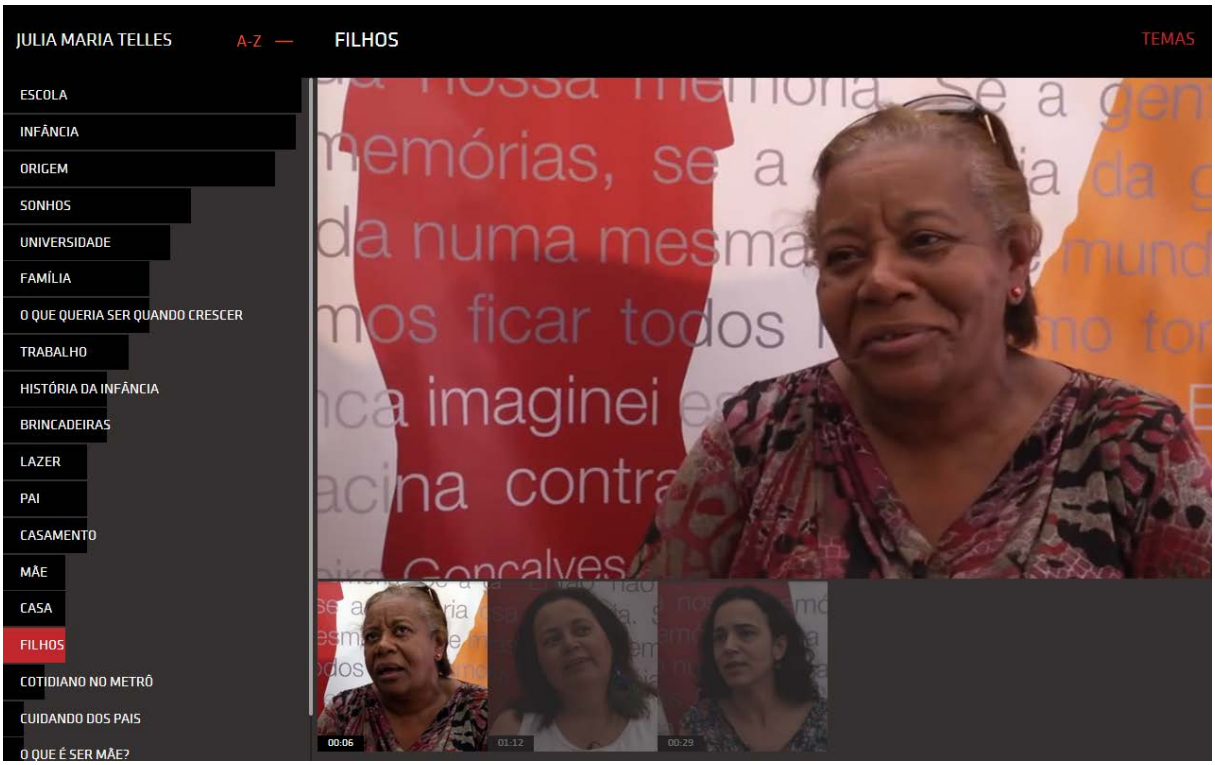
Clicando na *tag*, o usuário filtra a matriz de vídeos.

Figura 53, Matriz de vídeos filtrada pela *tag* “filhos”.



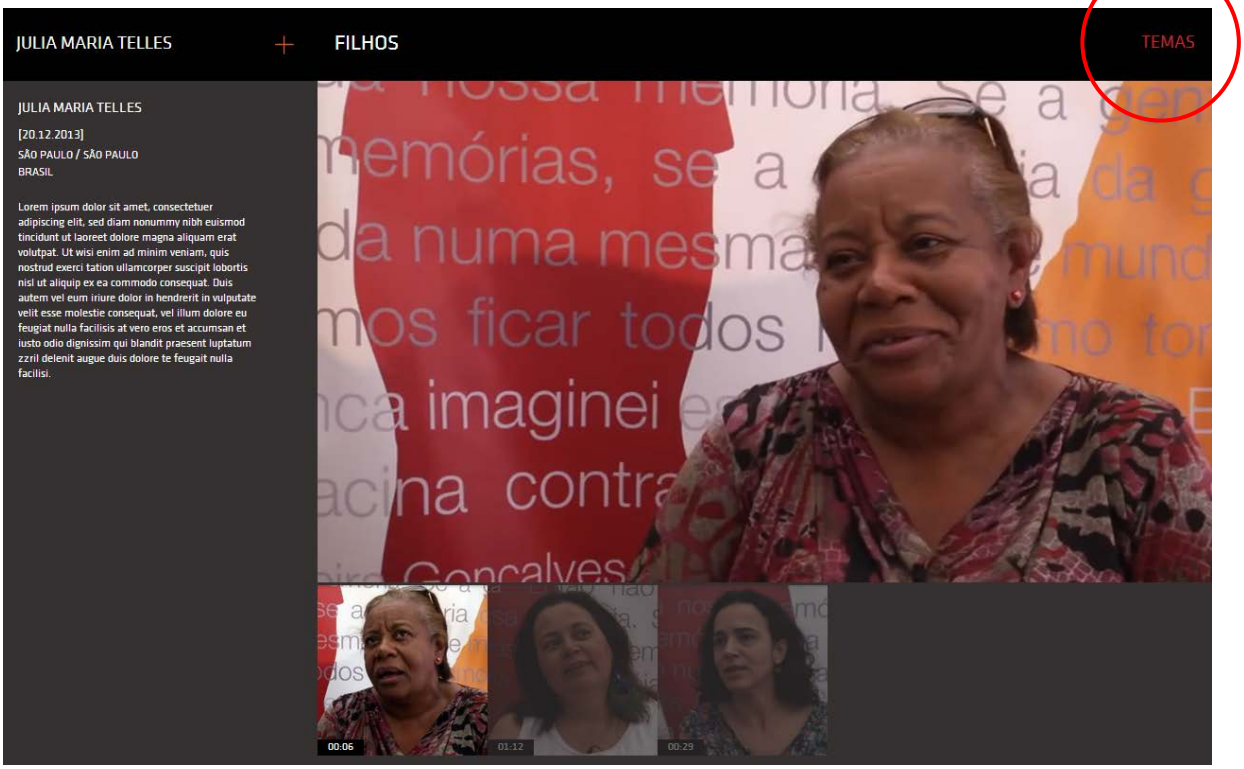
Clicando sobre o vídeo, o usuário é direcionado para a tela daquele depoente. A lista de *tags* agora só apresenta termos relacionados aquele depoente.

Figura 54. Tela do depoente com lista de tags “aberta”.



Repare que na barra superior ao invés do nome da categoria ou tema, agora consta o nome da pessoa. Após 3 segundos, o menu lateral desliza horizontalmente, revelando informações básicas do depoente. Para acessar o menu, o usuário pode clicar sobre o sinal de “+”.

Figura 55. Tela do depoente com lista de *tags* “fechada”.



No canto superior direito está sempre disponível um retorno para a tela inicial, onde são apresentados os temas. Abaixo do vídeo, uma barra horizontal apresenta os outros trechos indexados com a mesma *tag*.

Figura 56, Acima do vídeo são disponibilizadas as *tags*, permitindo um retorno para a matriz de vídeos.

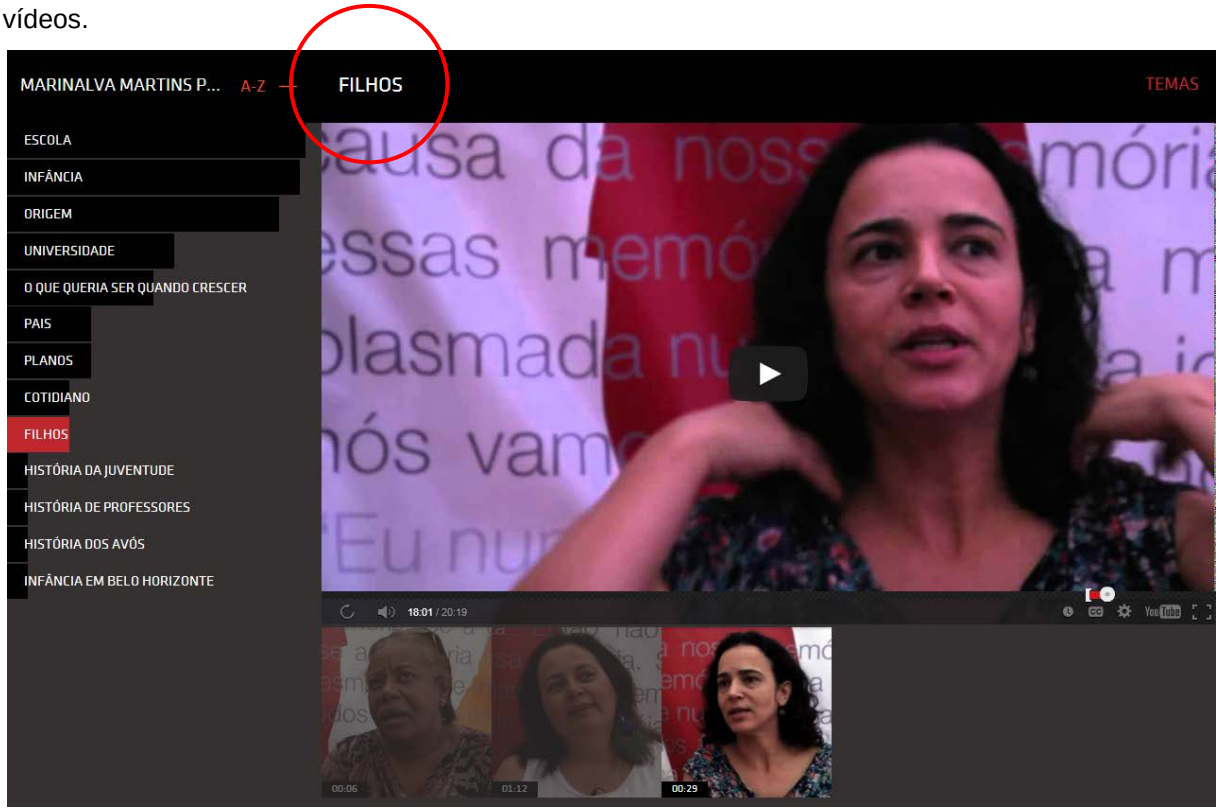
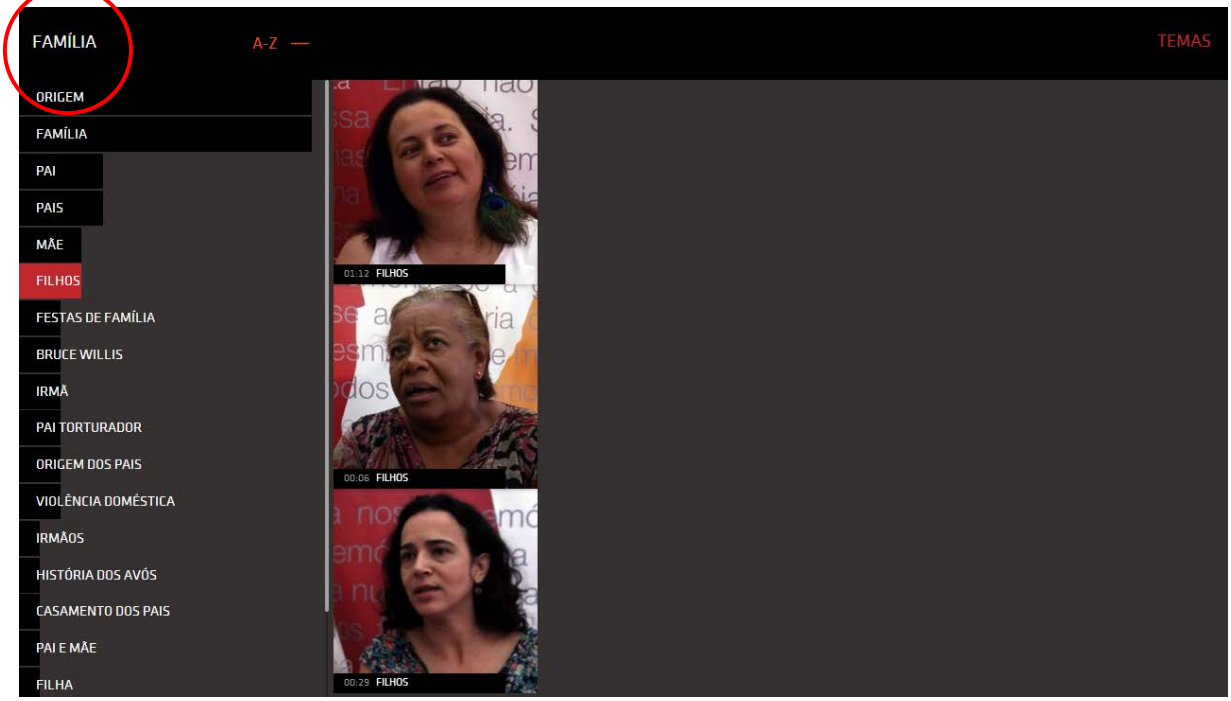
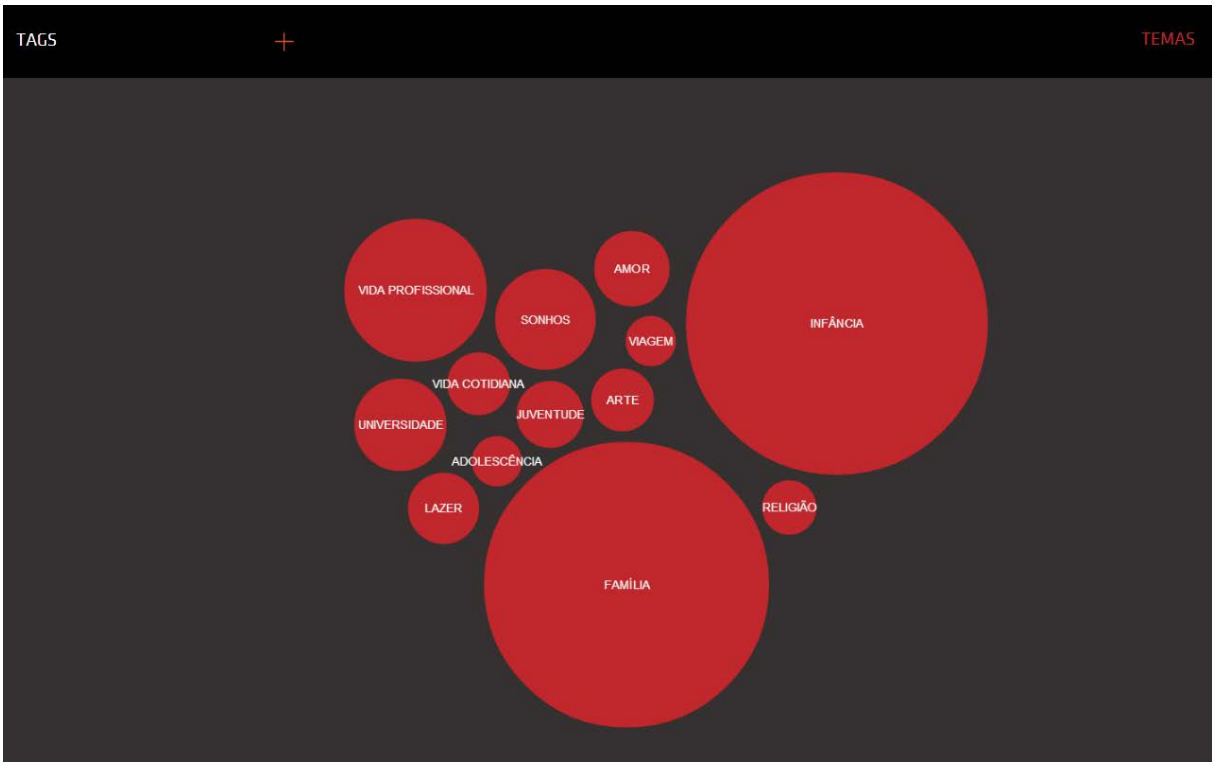


Figura 57, De volta a matriz de vídeos filtrada pela *tag* “filhos”, o usuário pode agora ver todas as *tags* relacionadas ao tema original, neste caso “família”.



Na tela principal, os círculos são posicionados randomicamente e convergem para o centro da tela utilizando um algoritmo de *circle packing*[20].

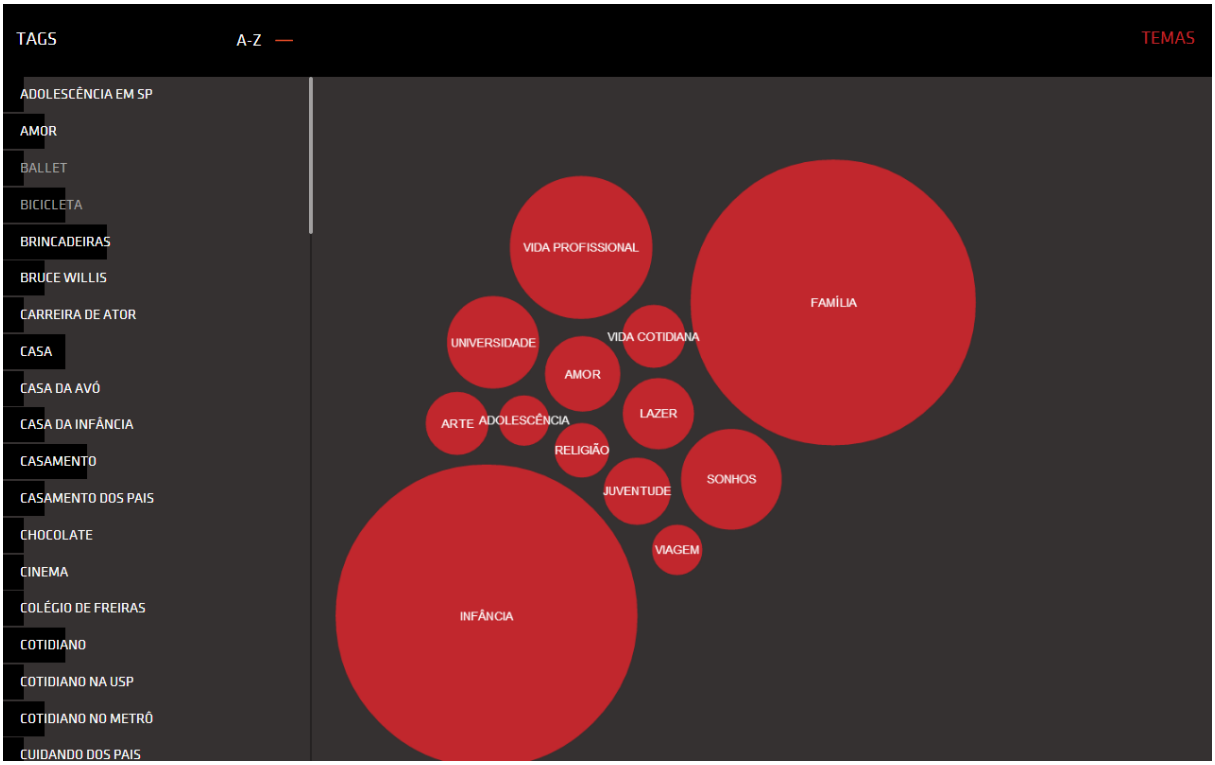
Figura 58, Tela principal.



[20] Cf. Wolfram Alpha, um *circle packing* é um arranjo de círculos dentro de uma determinada região, de forma que não ocorra superposição de dois círculos e alguns (ou todos) sejam mutuamente tangentes. Disponível em: http://mathworld.wolfram.com/CirclePacking.html Acessado em: 20/01/2014

Nessa tela, o menu lateral lista todas as *tags* independente da categoria.

Figura 59, Tela principal com a lista de *tags* “aberta” ordenada alfabeticamente.



As novas *tags* cadastradas aparecem nessa listagem em um tom de cinza rebaixado, indicando que essas *tags* ainda não foram relacionadas a nenhum tema. Dessa forma ajudamos os pesquisadores a identificar esses termos e providenciar o relacionamento com o tema mais indicado ou mesmo criar um novo tema. Se esse relacionamento não é feito, a *tag* aparece apenas na listagem completa da tela inicial ou na tela do depoente, não figurando nas listas das categorias.

Na imagem acima, podemos ver que o termo “ballet” está marcado em cinza. Clicando no termo o sistema apresenta os trechos de vídeo indexados com a tag “ballet” (neste exemplo, apenas um trecho de vídeo).

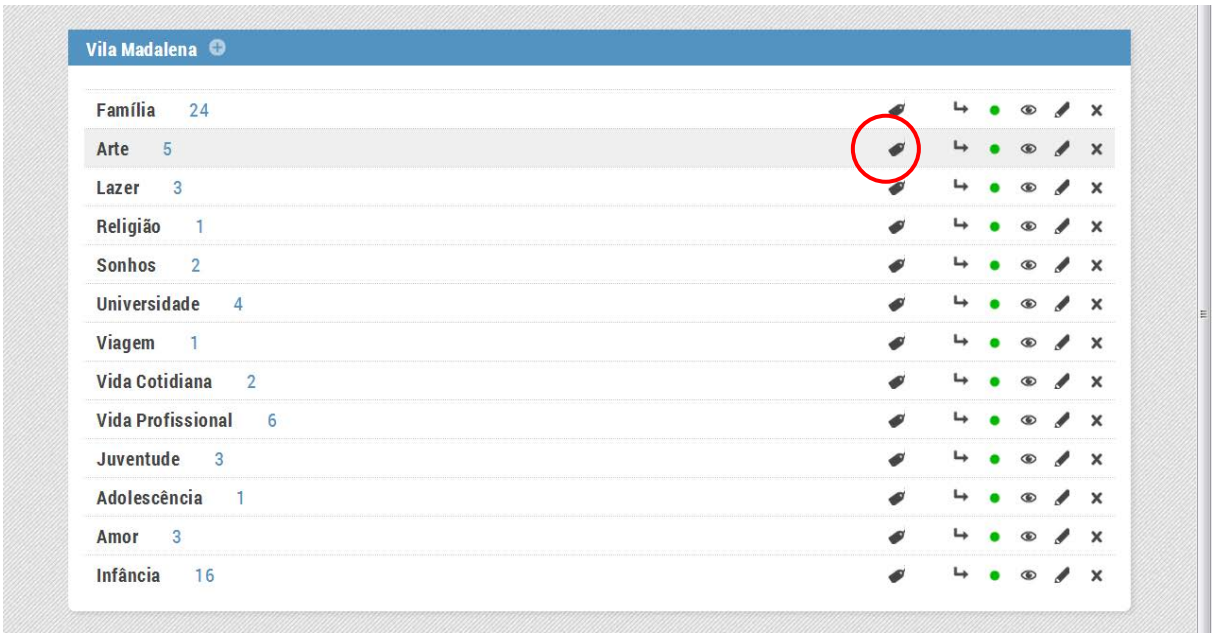
Figura 60, Resultado da filtragem.

Clicando no vídeo podemos ver que o mesmo foi indexado com as *tags* “sonhos” (relacionada ao tema “sonhos”) e “música” (relacionada ao tema “arte”).

Figura 61, Diversas *tags* relacionadas a um mesmo trecho.

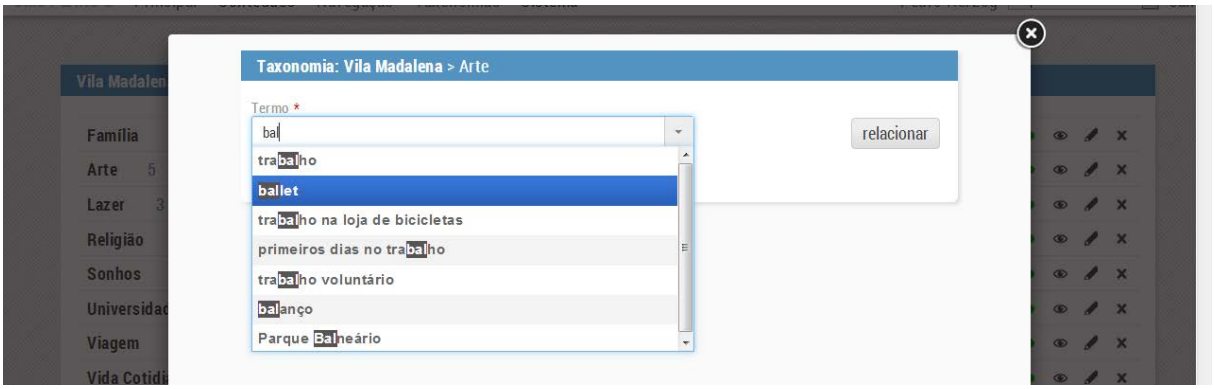
Para relacionar o termo “ballet” com o tema “arte”, o pesquisador acessa a taxonomia Vila Madalena na área administrativa.

Figura 62, Relacionando tags aos temas.



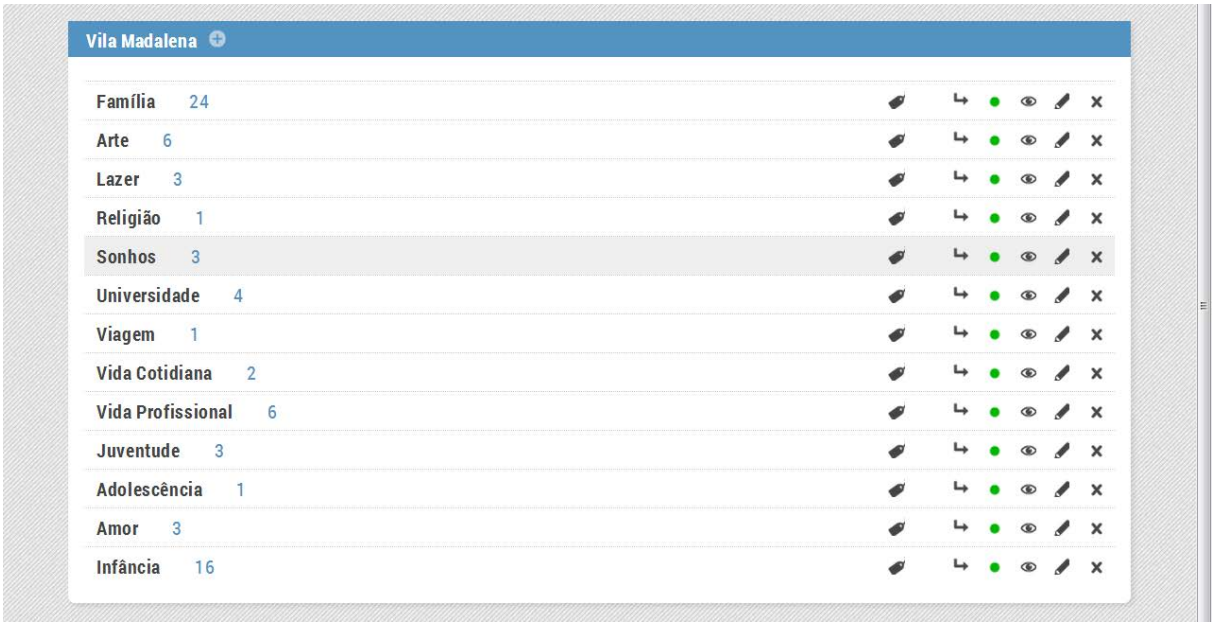
Para realizar o relacionamento da *tag*, o pesquisador utiliza o mesmo formulário *autofill* utilizado no início do processo.

Figura 63, *Autofill*.



Relacionamos o termo “ballet” com os temas “arte” e “sonhos”. Repare que o número ao lado dessas categorias foi acrescido de 1 (uma) *tag.*

Figura 64, A taxonomia atualizada.



Uma vez criado o relacionamento, o termo passa a aparecer normalmente nas listas de *tags* das categorias.

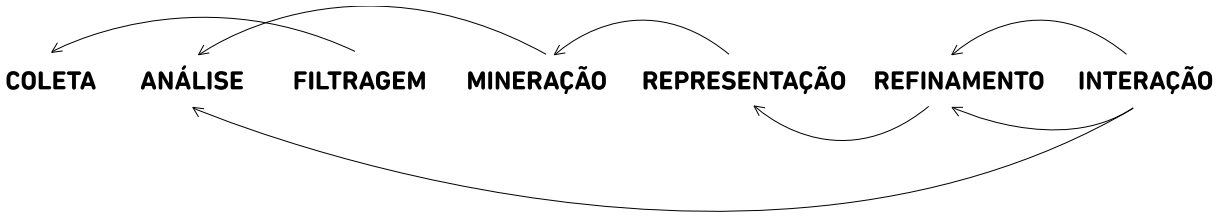
Figura 65, A lista atualizada.

Com os vídeos completos disponíveis para consulta no YouTube, os pesquisadores do Museu da Pessoa podem sempre acessá-los para indexar novos trechos, corrigir eventuais erros de indexação, ou apenas atribuir novas *tags* a trechos já selecionados. O resultado dessas novas indexações é representado na cauda longa, oferecendo um *feedback* imediato para os pesquisadores. Dessa forma, o sistema oferece aos pesquisadores novas formas para acompanhar o processo de indexação a fim de garantir um equilíbrio da quantidade de vídeos por tema.

4 **CONCLUSÃO**

O sistema aqui descrito, demonstra a possibilidade de se habilitar pontos de acesso para os trechos de vídeo, criando relacionamentos entre as entrevistas. O protótipo oferece uma ferramenta para os pesquisadores visualizarem o impacto do seu trabalho (como indexadores) na interface oferecida para o usuário final. Dessa forma, fica evidente a interdependência entre as etapas do processo de Design de Informação Computacional proposto por Fry.

Figura 66, A interdependência entre as etapas do processo.



A Interação depende diretamente das categorias criadas na etapa de Análise. A etapa de Filtragem depende da forma de indexação utilizada na etapa de Coleta. As soluções de Interação são decorrentes da etapa de Refinamento e Representação, que por sua vez está intimamente ligada aos dados adquiridos na etapa de Mineração.

A interface proposta atende as recomendações para a prática de história oral elencadas na introdução deste trabalho. Além de oferecer acesso direto aos segmentos de vídeo, a criação de categorias a partir da análise das anotações livres realizadas pelos pesquisadores, permite o relacionamento temático entre os diversos trechos de vídeo.

Não sendo o foco desse trabalho, outro aspecto importante a destacar, é que esse sistema modifica o fluxo do processamento das entrevistas. A transcrição dos depoimentos é uma das etapas mais longas em um projeto de história oral, sendo frequentemente associada aos atrasos na publicação dos resultados. A abordagem aqui proposta, não depende da transcrição dos depoimentos para o início do processo de indexação e, consequentemente, disponibilização dos vídeos.

Além disso, no sistema aqui proposto, a edição dos vídeos pode ser realizada pelos próprios pesquisadores, minimizando os gastos com ilha de edição.

Ao longo do processo de indexação foram identificadas algumas possíveis melhorias na interface:

- Além das duas possibilidades de ordenação das *tags* na cauda longa (por incidência ou alfabética) o sistema poderia oferecer uma busca com *autofill* (semelhante a solução aplicada no ambiente adminstrativo) para facilitar o acesso daquele usuário que já sabe o que está buscando.
- Ainda na cauda longa, o botão (A-Z) que ativa e desativa a ordenação alfabética não representa claramente sua função. Ao invés de uma simples mudança na cor, seria mais evidente se fossem utilizados diferentes ícones para representar os tipos de ordenação.
- Embora os círculos na tela inicial ofereçam uma visão geral dos temas, a navegação entre temas poderia ser facilitada através de um menu *dropdown* no canto superior direito, sem que o usuário tivesse que voltar toda hora para a página inicial. Outras soluções de interface podem ser exploradas nesse ponto.
- Na cauda longa, além da diferenciação de cor utilizada para representar as *tags* que ainda não foram relacionadas a um tema, podem ser utilizadas outras codificações de cor para representação, por exemplo, das *tags* mais recentes, ou mais acessadas.
- Os *thumbs* (imagens quadradas) utilizados para representação dos trechos de vídeo poderiam ser gerados automaticamente. Neste protótipo, ocorrem *thumbs* repetidos, porque estes foram carregados manualmente, isto é, para cada trecho de vídeo foi feito o *upload* de uma imagem.
- Pode ser melhor explorado o recurso de *hover* (*mouseover*) para evidenciar pontos de ativação em diversos elementos desse protótipo. Entretanto, a utilização desse recurso está relacionada com o tipo de interação pretendida. No caso de uma superfície *touchscreen,* por exemplo, esse recurso acaba limitado pela própria natureza do dispositivo (tablets ou TVs touchcreens).

Além disso, podem ser incorporadas ao processo metodologias para automatização (ainda que parcial) da indexação. Ferramentas para extração automática de termos, por exemplo, podem ser úteis quando temos um grande volume de depoimentos já transcritos.

Embora esse projeto tenha sido desenvolvido para atender às necessidades específicas do projeto Memórias da Vila Madalena, a estratégia de indexação e a interface aqui descritas oferecem flexibilidade para lidar com coleções maiores. Sendo assim, como desdobramento desse trabalho, podemos aplicar o mesmo sistema para indexação de outros projetos do Museu da Pessoa, ou mesmo para outros projetos de história oral.

REFERÊNCIAS

ALBERTI, V. **Manual de história oral** – 3. ed. – Rio de Janeiro: Editora FGV, 2005.

AMERICAN NATIONAL STANDARDS ORGANIZATION. **ANSI NISO Z 39.19**: Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. Bethesda: NISO Press, 2005.

ANTONIOU, G.; HARMELEN, F. **A Semantic Web Primer**. Cambridge Massachusetts: MIT Press, 2008.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific American**, v.284, n. 5, p. 34-43, 2010.

BREITMAN, K. K. **Web Semântica: a internet do futuro**. Rio de Janeiro: LTC, 2010.

BÓNIS, F. **Béla Bartók. His Life in Pictures and Documents**. Budapest: Kossuth Printing House, 1981.

FRY, B. **Computational Informational Design.** Cambridge Massachusetts: MIT Press, 2004.

JACOB, E. K. Classification and categorization: a difference that makes a difference. **Library Trends**, 2004.

KRIPPENDORFF, K. **Content Analisys: An Introduction to Its Methodology** (2nd Edition). Thousand Oaks, CA: Sage Publications, 2012.

LAMBERT, D.; FRISCH M. **Meaningful access to audio and video passages: A two- tiered approach for annotation, navigation, and cross-referencing within and across oral history interviews.** in: Oral History in the Digital Age. Institute of Library and Museum Services, 2012.

Disponível em: <http://ohda.matrix.msu.edu/2012/06/meaningful-access-to-audio-and-video-passages-2/> Acessado em: 20/01/2014.

LANCASTER, F. W. **Indexação e resumos:** teoria e prática. Brasília: Briquet de Lemos, 2004.

VIEIRA, S. B. **Indexação automática e manual: revisão de literatura.** Brasília: UnB, 1988.

MAZÉ, E. A. **Metadata: best practices for oral history access and preservation**. in:Oral history in the digital age. Institute of Library and Museum Services, 2012. Disponível em: <http://ohda.matrix.msu.edu/2012/06/metadata/> Acessado em: 20/01/2014

MEIHY, J. **Manual de História Oral**. São Paulo: Ed. Loyola, 1996.

NÓBREGA, D. L. **Indexação de artigos de periódicos em Ciência da Informação: elaboração de politica de indexação para a base ABCDM**. Brasília: UnB, 2011.

QUINTARELLI, E. **Folksonomies: Power to the people**, 2005. Disponível em: <http://www-dimat.unipv.it/biblio/isko/doc/folksonomies.htm#overview>. Acessado em: 20/01/2014

STEFANER, M. **Visual Tool for the Socio-Semantic Web**. Potsdam: University of Apllied Sciences, 2007.

SHIRKY, C. **Ontology is Overrated: Categories, Links and Tags**, 2005. Disponível em: <http://www.shirky.com/writings/ontology_overrated.html>. Acessado em: 20/01/2014.

SINHA, R. **A Cognitive Analysis of Tagging**, 2005. Disponível em: <http://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging/>. Acessado em: 20/01/2014

TEBEAU, M. **Case Study: Visualizing Oral History,** in: Oral History in the Digital Age, Washington, D.C.: Institute of Museum and Library Services, 2012. Disponível em: <http://ohda.matrix.msu.edu/2012/06/visualizing-oral-history/>. Acessado em: 20/01/2014

VASSÃO, C. A. **Metadesign: ferramentas, estratégias e ética para a complexidade**. São Paulo: Blucher, 2010.

ANEXO - Lista de *tags* utilizadas no protótipo

ADOLESCÊNCIA EM SP
AMOR
BALLET
BICICLETA
BRINCADEIRAS
BRUCE WILLIS
CARREIRA DE ATOR
CASA
CASA DA AVÓ
CASA DA INFÂNCIA
CASAMENTO
CASAMENTO DOS PAIS
CHOCOLATE
CINEMA
COLÉGIO DE FREIRAS
COTIDIANO
COTIDIANO NA USP
COTIDIANO NO METRÔ
CUIDANDO DOS PAIS
CURSO DE ATOR
DANÇA
ESCOLA
FACULDADE
FAMÍLIA
FESTAS DE FAMÍLIA
FILHA
FILHOS
GOSTOS PESSOAIS
HISTÓRIA
HISTÓRIA DA ESCOLA
HISTÓRIA DA INFÂNCIA
HISTÓRIA DA JUVENTUDE
HISTÓRIA DE PROFESSORES
HISTÓRIA DO AVÔ
HISTÓRIA DOS AVÓS
HISTÓRIAS DE BICICLETA
INFÂNCIA
INFÂNCIA EM BELO HORIZONTE
INFÂNCIA EM BRASÍLIA
INFÂNCIA EM SANTO AMARO
INFÂNCIA NO INTERIOR
INFÂNCIA NO SERTÃO
IRMÃ
IRMÃO
IRMÃOS
JUVENTUDE
LAZER
LITERATURA
MÃE
MATEMÁTICA
MORTE DA IRMÃ
MORTE DO PAI
MÚSICA
O QUE É SER ARTISTA?
O QUE É SER MÃE?
O QUE É SER PROFESSOR?
O QUE QUERIA SER QUANDO CR...
ORIGEM
ORIGEM DOS PAIS
PAI
PAI E MÃE
PAI TORTURADOR
PAIS
PINTURA
PLANOS
PRIMEIROS DIAS NO TRABALHO
PROFESSOR
REENCONTRO DE AMIGOS
RELIGIÃO
SAIA NA NOITE
SONHOS
TRABALHO
TRABALHO NA LOJA DE BICICLETAS
TRABALHO VOLUNTÁRIO
UNIVERSIDADE
VIAGEM
VIDA EM SÃO PAULO
VIDA PROFISSIONAL
VIOLÊNCIA DOMÉSTICA
VOLTA AO BRASIL
VOLTA PARA SÃO PAULO