Sistema de Informação Processual para a Provedoria de Justiça

Marco Fernandes, Miguel Alho, Pedro Almeida,
Joaquim Arnaldo Martins, Joaquim Sousa Pinto, Hélder Zagalo

***Resumo* – Este artigo apresenta um Sistema de Informação Processual (SIP), desenvolvido pela Universidade de Aveiro para a Provedoria de Justiça. O sistema desenvolvido permite a pesquisa, recolha, representação e anotação de processos previamente digitalizados. No artigo, é descrita a estratégia adoptada no armazenamento e manipulação dos documentos digitalizados, na pesquisa de informação proveniente de múltiplos repositórios e no desenvolvimento de um sistema de anotações para documentos web.**

***Abstract* –This article presents the project Sistema de Informação Processual, developed in the University of Aveiro for the Provedoria de Justiça. The system allows searching, retrieving, presenting and annotating the digital processes. In the paper, we describe the strategy adopted in the storage and manipulation of the digital documents, the search of information in multiple repositories and the development of a parallel system for the annotation of web documents.**

I. INTRODUÇÃO

Na sociedade de informação em que vivemos actualmente, as bibliotecas digitais assumem um papel de crescente relevo. As suas aplicações abrangem vários conteúdos, desde o papel (digitalização de documentos por uma organização/instituição), vídeo, fotografia, etc.

Com a sua implementação, pretende-se facilitar o processo de obtenção de informação por parte dos interessados, sendo este realizado de uma forma rápida e simples, em vários pontos de acesso. Por outro lado, procura-se criar um conjunto de facilidades, nomeadamente a facilidade de gestão e de organização, que dificilmente se poderia obter com os repositórios físicos originais (estejam eles em papel, vídeo analógico, etc.).

Apesar de oferecer um conjunto muito mais vasto e flexível de funcionalidades, as bibliotecas digitais colocam também vários problemas para quem as desenvolve. A segurança e a privacidade de informação obrigam a preocupações renovadas. Com efeito, numa lógica de acesso em vários pontos de uma rede (intranet ou, especialmente, internet) é imperativo reduzir ao mínimo as vulnerabilidades de segurança dos sistemas.

A decisão quanto ao formato de armazenamento é também um problema de grande relevo. Esta escolha depende de vários parâmetros, como o espaço ocupado, a perda de qualidade, o custo associado, o tempo de vida previsto para o formato e as ferramentas disponíveis para a manipulação dos documentos.

Requisitos

A Provedoria de Justiça tem, nos seus arquivos, mais de três milhões de documentos relativos aos processos existentes. Como é de fácil compreensão, a tarefa de obter um desses documentos para consulta pode ser inglória. Além disso, a gestão deste gigantesco repositório é um trabalho complexo.

Assim, a Universidade de Aveiro tinha de desenvolver um sistema que permitisse consultar os processos digitalizados e a informação existente na base de dados. Para facilitar o processo de consulta, deveriam ser desenvolvidas diversas funcionalidades de pesquisa. Finalmente, para permitir uma documentação mais completa e uma pesquisa mais orientada, deveria ser desenvolvido um sistema de anotação dos processos.

A aplicação web a desenvolver seria destinada à rede intranet da Provedoria, logo o número máximo de utilizadores simultâneos a que esta teria de responder ficaria abaixo de uma centena.

Modelo de 3 Camadas

Muitas aplicações actuais baseiam-se numa tecnologia cliente/servidor de duas camadas. Neste modelo, os dados são armazenados num servidor centralizado, que responde aos pedidos dos clientes que lhe acedem. Embora a gestão da informação seja simplificada, este modelo é muito pouco flexível e de difícil alteração posterior.

O desenvolvimento de uma aplicação complexa, seja ela web ou desktop, requer uma organização flexível e eficiente dos vários componentes que a constituem. O modelo mais comum para atingir este objectivo é composto por três camadas (*3-tier model*) [1]: dados, lógica e apresentação. Com esta arquitectura, conseguem-se melhorias substanciais na escalabilidade, robustez e reutilização do sistema. Alguns exemplos de bibliotecas digitais que usam este modelo são a biblioteca Alexandria [2] e o DSpace [3].

O acesso aos dados por parte da camada de apresentação é absolutamente transparente. Além disso, qualquer alteração na política de armazenamento e de acesso aos dados pode ser feita apenas nos componentes que lhe dizem respeito, sem ser necessário reconstruir todos os outros.

Por outro lado, o âmbito da aplicação pode ser alargado, já que outra qualquer aplicação pode ser construída sobre as camadas lógica e de dados já existentes.

Finalmente, com o acesso controlado aos dados, todos os detalhes de armazenamento e acesso aos mesmos podem ser encapsulados, aumentando consideravelmente a segurança do sistema.

Sistemas de Anotação

Os sistemas de anotação têm aplicações muito vastas e podem ser de grande utilidade. A grande funcionalidade destes sistemas é permitir inserir informações sobre o documento que se visualiza sem alterar o mesmo. Este conceito não é novo e algumas das mais populares aplicações de manipulação de documentos têm já ferramentas para inserir e manipular anotações: aplicações da família Microsoft Office [4], Adobe Acrobat [5], etc.

Embora esta funcionalidade esteja já bem implementada em aplicações desktop, ferramentas equivalentes para anotar documentos web estão ainda pouco desenvolvidas. Alguns exemplos são o Annotation Engine [6] e o MemoBook Notes [7].

No caso particular da Provedoria de Justiça, é de grande utilidade um mecanismo que complemente a informação disponível na base de dados ou nos documentos digitalizados. Estas anotações desde que feitas por pessoal qualificado, como é o caso em questão, são muito importantes para ajudar na classificação e posterior procura dos documentos, de uma forma mais selectiva e eficaz. Além disso permitem ir classificando os documentos de uma forma incremental, à medida que vão utilizados ou reutilizados, pois é impensável ter classificados em tempo razoável a enorme quantidade de documentos existente.

Num cenário intranet, como é o da Provedoria, uma aplicação deste tipo deve armazenar as anotações num repositório centralizado. Assim, ficam automaticamente disponíveis para os restantes utilizadores (se for essa a intenção), independentemente do ponto de acesso à rede.

V. ARQUITECTURA

Na Figura 1 está representada a arquitectura adoptada para o SIP. Como se pode verificar, o sistema foi desenvolvido de forma modular, o que permite efectuar alterações apenas nos componentes necessários.

Para aceder aos repositórios de informação, foram desenvolvidas três bibliotecas: ImageLibrary, DataLibrary e CNote. Sobre essas bibliotecas, dois web services foram criados – SIPws (geral) e espiritUs_ws (para o subsistema de anotações) – para disponibilizar os métodos necessários.

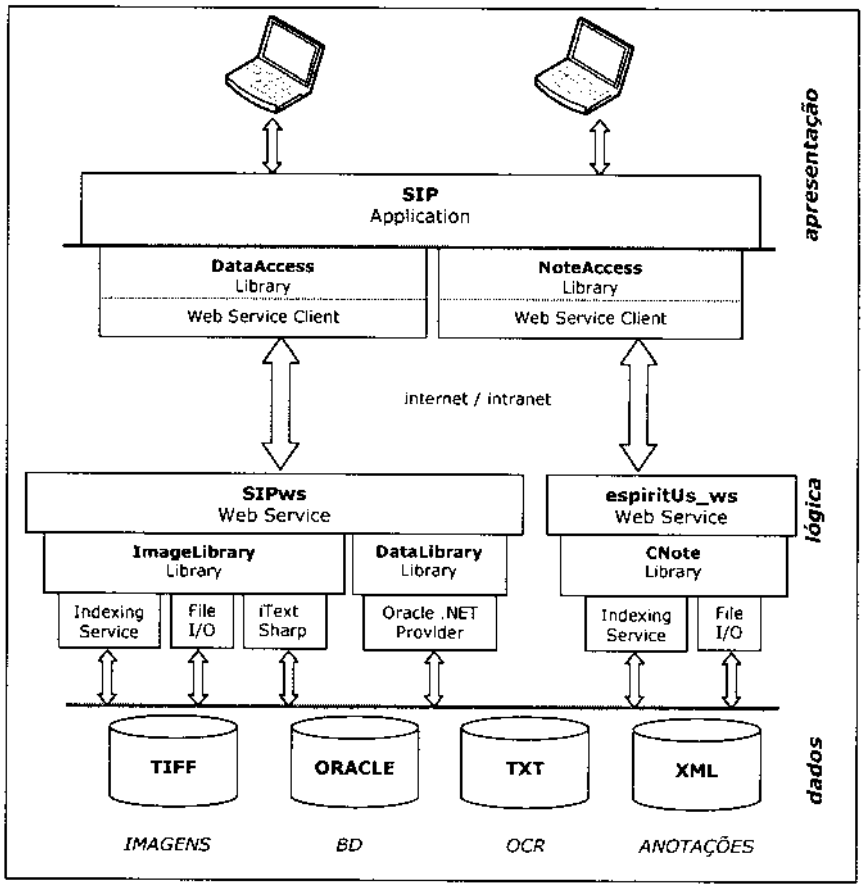

Figura 1 - Arquitectura geral do SIP

Finalmente, a aplicação web (SIP) acede de forma transparente aos dados utilizando os módulos DataAccess e NoteAccess, que por sua vez invocam os web services. Estes dois módulos permitem isolar a aplicação dos detalhes de acesso aos dados.

A. Camada de Dados

O sistema desenvolvido tem de manipular vários tipos de informação: os dados dos processos existentes na base de dados (Oracle) da Provedoria de Justiça, os documentos digitalizados desses mesmos processos, o texto extraído desses documentos por OCR (*Optical Character Recognition*) e as anotações efectuadas pelos utilizadores.

A.1 Repositórios de Informação

Na Provedoria de Justiça existia já uma base de dados com informação relativa aos processos existentes. Para implementar o Sistema de Informação Processual, foram criados três novos repositórios de informação: documentos digitalizados, documentos de texto extraído dos processos e documentos XML com as anotações aos processos.

A.2 Documentos Digitalizados

A escolha de um formato para armazenamento dos documentos digitalizados foi feita tendo em conta a conservação de uma alta qualidade de imagem sem exigir demasiada capacidade de armazenamento. Depois de uma análise dos vários formatos existentes, foi definido que os documentos seriam digitalizados para o formato TIFF (*Tag Image File Format*) [8], com compressão CCITT (*Comité Consultatif International Télèphonique et Télègraphique*) Grupo 4 e a uma resolução de 300 dpi. O formato é bastante eficaz na conservação da informação, permitindo quer a visualização quer a impressão de alta qualidade e ainda a possibilidade da transcodificação para outros formatos. Visto que os documentos são (tipicamente) a preto e branco, o tamanho final do ficheiro é bastante reduzido (cerca de 40KB) apesar das suas dimensões (2481x3512).

A distinção das páginas é feita através do nome do ficheiro que inclui a informação do número da página e também o número e ano do processo. Esta designação única em conjunto com o serviço de indexação permite que a estrutura dos directórios no repositório seja arbitrária, já que o serviço encarrega-se de determinar a sua localização, tornando o acesso aos documentos transparente. O serviço de indexação foi implementado utilizando o Indexing Service da Microsoft.

O Indexing Service (IS) [9] é um serviço do Windows 2000 e versões posteriores que extrai conteúdo de documentos e cria um catálogo para permitir uma pesquisa mais rápida e eficiente. Por defeito, o IS filtra documentos Office, HTML, mensagens MIME e ficheiros de texto e indexa informação específica (autor, conteúdo, etc.). Para todos os restantes ficheiros, apenas são filtradas propriedades genéricas (data, nome e localização do ficheiro, etc.)

A.3 Documentos OCR

O texto digitalizado dos documentos dos processos é armazenado em ficheiros de texto simples. O formato é universalmente aceite e armazenado no mais reduzido espaço possível. Assim, a sua transferência é bastante rápida. Tal como no caso dos documentos digitalizados (imagens), o nome do ficheiro serve para identificar a que página de determinado processo pertence o texto.

A.4 Anotações

Ao contrário do que acontece nos repositórios anteriores, não é utilizado apenas o nome do ficheiro para determinar a que processo (e, eventualmente, a que documento) pertence um conjunto de anotações. Como se explicará na secção B.3, um ficheiro de anotações fica associado ao URL (*Uniform Resource Locator*) do documento exibido no *browser* do utilizador. Concretamente, a estrutura de directórios e o nome dos ficheiros no repositório reflectem o endereço de um determinado documento web.

O armazenamento das anotações é feito no formato XML. O XML (Extensible Markup Language) [10] é um formato de texto simples e flexível derivado do SGML (Standard Generalized Markup Language) [11]. Este formato permite armazenar informação em blocos indentificados com marcadores. A sua utilização traz grandes vantagens, nomeadamente: flexibilidade estrutural, nomeação dos marcadores de forma personalizada e intuitiva e existência cada vez mais relevante de ferramentas de manipulação de XML.

Um documento XML de anotações tem um formato semelhante ao que se apresenta de seguida.

```
<?xml?>
<notes url="">
 <query txt="">
  <note id="" userid="" private="" date="">
  </note>
 </query>
</notes>
```

Esta estrutura de informação armazena o URL do documento consultado, a identificação do utilizador, a indicação se a anotação é privada ou pública, a data (e hora) em que foi realizada a notação e a anotação propriamente dita.

Como este componente de anotações ainda está em fase de desenvolvimento, optou-se por utilizar o XML como formato de armazenamento das anotações devido à sua flexibilidade. Se, numa fase posterior, se pretender extender a funcionalidade deste componente, nomeadamente ao permitir criar *threads* de anotações (encadeamento), facilmente se alterará a estrutura XML de

suporte ao armazenamento da informação, ao contrário do que sucederia com bases de dados relacionais.

B. Camada Lógica

A camada lógica é responsável por implementar políticas de acesso e manipulação dos dados. Nesta camada, implementou-se um conjunto de funcionalidades que visam aumentar a *performance* do sistema e a transformação de dados. Como forma de isolar a camada lógica da camada de apresentação, utilizaram-se Web Services.

B.1 Web Services

Os Web Services [12] são uma das principais ferramentas para implementar arquitecturas distribuídas em aplicações web. São unidades individuais de código, que permitem a partilha de dados entre aplicações em diferentes máquinas, mesmo em plataformas distintas. Além disso, o seu formato de transferência de informação é o XML, cuja utilização, como se viu na secção anterior, tem grandes vantagens.

B.2 Manipulação de Documentos Digitalizados

O formato TIFF não é suportado para representação de imagens nos *browsers* típicos. Assim, para permitir a apresentação dos processos via web, é necessário um processo intermédio.

A solução adoptada podia ter passado pela instalação de um *plug-in* (por exemplo, o Quicktime [13] da Apple) que permitisse a visualização de imagens do tipo TIFF. Embora esta solução funcione, exige que todos os utilizadores possuam o componente. Além disso, o tempo de carregamento do *plug-in* torna a experiência de navegação na aplicação pouco agradável.

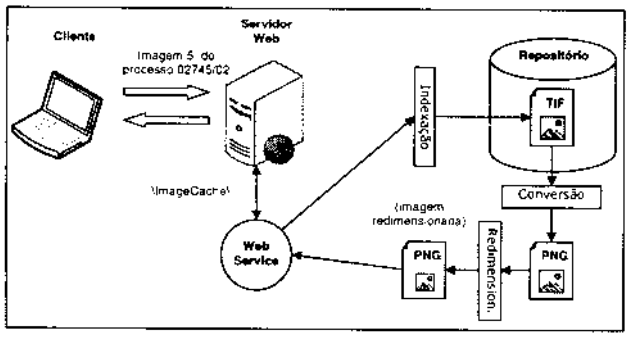

Figura 2 - Consulta de uma página de um processo

Assim, optou-se por efectuar, *on-demand* e em tempo real, a conversão dos documentos TIFF para o formato PNG (*Portable Network Graphics*) [14]. Este é um dos formatos suportados pelos *browsers* com uma das melhores relações qualidade/tamanho. Além da conversão, é feito ainda um redimensionamento do tamanho para 680x962. Desta forma, a informação transmitida entre a camada de dados e a de apresentação é mais reduzida. Por outro lado, um redimensionamento prévio da imagem para o tamanho em que vai ser exibido torna-a mais legível do que quando o mesmo é feito pelos browsers. O processo é demonstrado na Figura 2.

Como se pode observar na Figura 2, cada documento convertido fica armazenado numa cache de imagens de tamanho configurável. Desta forma, se o mesmo ou outro utilizador voltar a requisitar uma imagem já existente na cache, o processo de conversão e redimensionamento não é feito. Quando a cache fica preenchida, o sistema determina a imagem mais antiga e remove-a. Tendo em conta que, de acordo com os elementos da Provedoria, num determinado período temporal existe um padrão que indica a consulta sistemática dos mesmos processos, a utilização de uma cache (mesmo que de tamanho modesto) revela-se uma mais-valia para o sistema.

Quanto aos processos digitalizados, o sistema de informação processual apresenta ainda outra funcionalidade – a impressão completa ou parcial de um processo. Para atingir esse objectivo, foi criada uma biblioteca que reúne todos os documentos solicitados (TIFF) num único PDF (*Portable Document Format*) [15], que é posteriormente enviado para o cliente. Em cada página, para além da imagem digitalizada, pode ser ainda colocado um cabeçalho com o número da página e do processo. O processo está exemplificado na Figura 3.

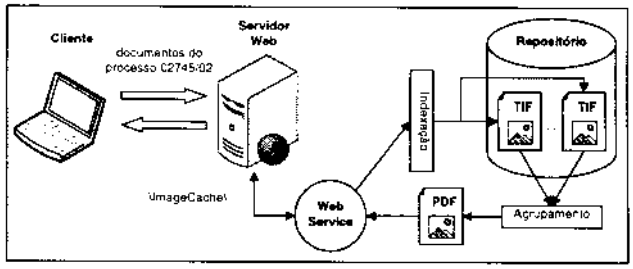

Figura 3 - Criação de PDF para impressão

Tal como no caso dos PNG, os documentos PDF são armazenados numa cache do sistema (no mesmo espaço físico que a de PNG). Contudo, esta cache terá, previsivelmente, uma utilização menos eficaz que a anterior, já que, para cada processo, cada utilizador pode gerar uma colecção diferente de documentos.

B.3 Gestão das Anotações

O sistema de anotação desenvolvido armazenada as anotações de acordo com o URL da página que está a ser visualizada. Por exemplo, se o sistema for utilizado para anotar a página http://www.ieeta.pt/A/B/pagina.htm, a anotação é armazenada na directoria [Repositorio]\A\B\ com o nome pagina.htm.xml.

No caso do sistema da Provedoria, para visualizar a ficha do processo 1234 de 2001, é utilizado um endereço semelhante a http://[*Servidor*]/processo.aspx?n=01234/01. Caso fosse este o endereço utilizador para realizar as anotações, estas seriam armazenadas no repositório de anotações no ficheiro processo.aspx.xml no directório

[Repositorio]\[Servidor]. A informação que aparece a seguir ao ponto de interrogação seria armazenada no elemento *query* do documento XML, conforme apresentado na secção A.4.

Esta forma de armazenar as anotações, embora genérica, tornaria o sistema composto por ficheiros em reduzido número mas de grande dimensão, já que todas as anotações a fichas de processos seriam armazenadas neste ficheiro. Posteriormente, utilizar-se-ia o elemento *query* para determinar quais as anotações de um determinado processo. Estes ficheiros, de grande dimensão, poderiam levantar problemas no processo de pesquisa e manipulação de anotações.

Para ultrapassar este problema, foi desenvolvido um filtro ISAPI (*Internet Information Application Protocol Interface*) [16], que se encarrega de transformar o URL de apresentação dos processos para o seguinte formato: http://[*Servidor*]/processos/2001/1234. Ao anotar esta página, é armazenado um ficheiro no directório [Repositorio]\[*Servidor*]\documentos\2001\1234.xml no servidor de anotações. Deste modo, as anotações relativas a cada processo ficam armazenadas em ficheiros separados.

B.4 Funcionalidades de Pesquisa

A pesquisa incide, essencialmente, na base de dados, no texto OCR e no texto anotado. Quanto à base de dados, foram escolhidos alguns campos da mesma, considerados relevantes pelos elementos da Provedoria, para efectuar pesquisas. Quanto à pesquisa no texto OCR e nas anotações foi utilizado o IS, embora de formas distintas. No caso dos processos digitalizados em TIFF, é utilizado o serviço de indexação para determinar quais os ficheiros (através do seu nome) associados a um determinado processo.

No caso do texto associado a esses documentos (OCR), a metodologia seguida é um pouco diferente. Com efeito, estes documentos são essencialmente utilizados para efectuar pesquisas por texto livre no seu conteúdo, sendo o número do processo determinado posteriormente a partir do nome do ficheiro.

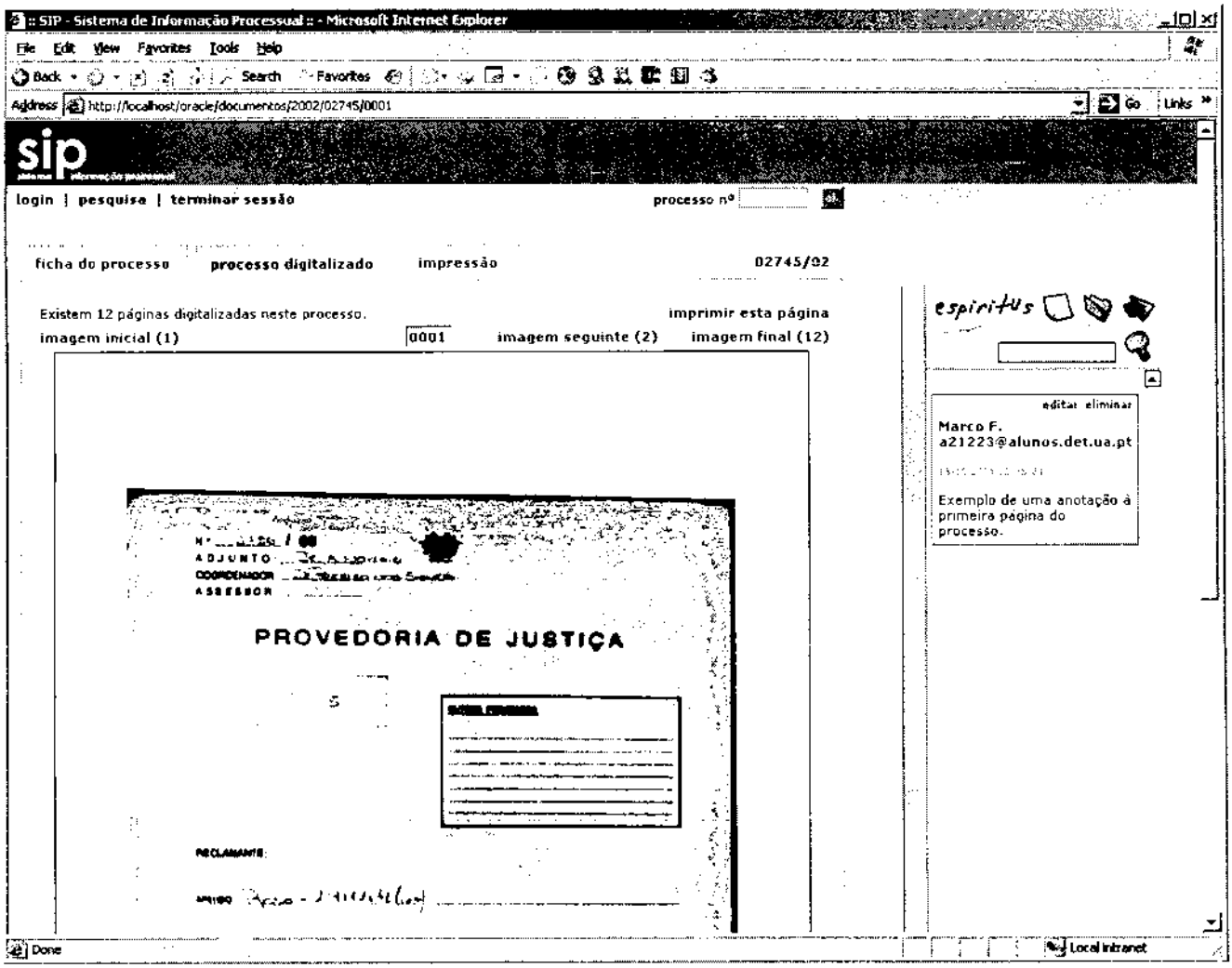

Figura 4 - Interface do SIP

Como o IS não tem, por defeito, um filtro para documentos XML, foi necessário instalar um filtro XML (IFilter , da Quilogic [17]) para indexar as anotações.

C. Camada de Apresentação

Como se pode observar na Figura 4, a informação é disponibilizada aos utilizadores sob a forma de páginas web. Existe uma página de pesquisa onde se podem procurar processos pelos critérios descritos acima. Para cada processo, é disponibilizada a ficha do mesmo (proveniente da base de dados) e é permitido consultar todas as páginas digitalizadas (no formato PNG). Os utilizadores podem igualmente gerar PDFs de todo o processo ou de apenas de algumas páginas para salvaguardar localmente ou imprimir.

Quanto às anotações, foi desenvolvido um componente que se pode colocar em qualquer página onde se pretenda que existam anotações (no caso da Provedoria, foi colocado na página dos processos). Cada utilizador, depois de autenticar-se, pode fazer anotações a um processo no seu todo ou a uma página em concreto. Esta anotação pode ser pública (livre para consulta) ou privada (visível apenas para o autor). Pode ainda obter a listagem de todas as suas anotações e efectuar uma pesquisa nas anotações existentes (suas e de outros, quando públicas).

VI. CONCLUSÕES

O sistema desenvolvido cumpre as funcionalidades essenciais exigidas, integrando na página web informação de fontes heterogéneas.

Encontrou-se uma solução que permite conservar os documentos digitalizados num formato com alta resolução – o TIFF. Tanto a conversão para PNG como o processo de junção de várias imagens num PDF são feitas de forma relativamente célere. Além disso, a existência de uma cache de documentos torna o sistema notoriamente mais eficiente.

Por outro lado, graças à utilização do sistema de indexação, o acesso aos repositórios é feito de forma completamente transparente para a aplicação, já que esta não necessita de saber a localização de cada documento.

Finalmente, a arquitectura do sistema desenvolvido é completamente modular, pelo que alterações posteriores podem ser executadas de forma eficaz.

REFERÊNCIAS

[1] Jupitermedia Corporation, *three-tier – Webopedia.com*, 2004, http://winplanet.webopedia.com/TERM/T/three_tier.html

[2] University of California, *Alexandria Digital Library Project*, 2004, www.alexandria.ucsb.edu

[3] MIT Libraries & Hewlett-Packard, *DSpace Federation*, 2003, http://dspace.org

[4] Microsoft Corporation, *Microsoft Office System Informação de Produto*, 2004, http://www.microsoft.com/portugal/office

[5] Adobe Systems, *Adobe Acrobat 6.0 Professional*, 2004, http://www.adobe.com/products/acrobatpro/overview.html

[6] Berkman Center for Internet & Society – Harvard, *Annotation Engine*,2004, http://cyber.law.harvard.edu/projects/annotate.html

[7] Obout Inc., *MemoBook Notes*, 2003, http://www.memobook.com

[8] O'Reilly & Associates, Inc, *GFF Format Summary: TIFF*, 1996, http://netghost.narod.ru/gff/graphics/summary/tiff.htm

[9] Microsoft Corporation, *What is Indexing Service*, 2003, http://msdn.microsoft.com/library/en-us/indexsrv/html/ixintro_0311.asp

[10] World Wide Web Consortium, *Extensible Markup Language (XML) 1.0 (Third Edition)*, 2004, http://www.w3.org/TR/REC-xml

[11] ISO 8879:1986, *Standard Generalized Markup Language (SGML)*, 2001, http://www.iso.org

[12] World Wide Web Consortium, *Web Services*, 2004, http://www.w3.org/TR/ws-arch

[13] Apple Computer, *Apple – Quicktime*, 2004, http://www.apple.com/quicktime

[14] Greg Roelofs, PNG (Portable Network Graphics) Home Site, 2004, http://www.libpng.org/pub/png

[15] Adobe Systems, *What is Adobe PDF?*, 2004, http://www.adobe.com/products/acrobat/adobepdf.html

[16] Microsoft Corporation, *ISAPI Filter Overview*, 2004, http://msdn.microsoft.com/library/en-us/iissdk/iis/isapi_filter_overview.asp

[17] Quilogic, *XML IFilter for indexing XML files*, 2003, http://www.quilogic.cc/ifilter.htm