UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ESCOLA TÉCNICA SUPERIOR DE ENXEÑARÍA

TRABALLO FIN DE GRAO EN ENXEÑARÍA INFORMÁTICA

DESENVOLVEMENTO DUN SISTEMA PARA A VISUALIZACIÓN WEB DO *TESOURO DO LÉXICO PATRIMONIAL GALEGO E PORTUGUÉS*

Autor:

César Osorio Peláez

Directores do proxecto:

José Varela Pet

Xulio César Sousa Fernández

Santiago de Compostela, Setembro de 2011

ÍNDICE

1. CONTEXTUALIZACIÓN

Os cambios socioeconómicos e tecnolóxicos producidos nas últimas décadas, coa perda do mundo rural tradicional e os seus modos de vida, trouxeron consigo o abandono e a extinción de centos de palabras patrimoniais, ao desapareceren as realidades materiais ás que estaban asociadas. A esta perda do coñecemento léxico non escapan linguas asentadas e recoñecidas nas súas sociedades, como o inglés, o francés, o castelán ou o portugués, e aínda menos aquelas que, como o galego, teñen que vivir en situacións de dificultade. As linguas son un patrimonio inmaterial que hai que preservar. Cada palabra perdida e non rexistrada supón unha perda cultural e tamén a desaparición dun vestixio de grande valor para o estudo da lingua e cultura dunha sociedade.

Consciente da necesidade de contribuír á preservación do léxico patrimonial e da utilidade de contar cun amplo corpus que permita emprender estudos de carácter xeral sobre a variación léxica, practicamente ausentes do ámbito da lingüística galego-portuguesa, o Instituto da Lingua Galega (ILG) emprendeu o proxecto *Tesouro do léxico patrimonial galego e portugués*, unha iniciativa internacional na que participan tamén diversas universidades portuguesas e brasileiras, que ten como obxectivo reunir nunha base de datos todo tipo de fontes que reúnen vocabulario dialectal (monografías sobre a fala dunha localidade, atlas lingüísticos, artigos onomasiolóxicos, etc.)

O sistema web que aquí se presenta foi deseñado e construído no ámbito deste proxecto co fin de dar resposta ás esixencias e requisitos propostos polos investigadores que dirixen o *Tesouro*.

2. OBXECTIVOS

O obxectivo xeral consistiu no desenvolvemento dun sistema web para a explotación do material léxico dialectal recompilado no proxecto *Tesouro do léxico patrimonial galego e portugués* (*Tesouro*). Este obxectivo xeral pódese descompoñer nos seguintes obxectivos concretos:

- Creación dunha base de datos dialectal que permita almacenar, manexar e consultar todo o material dispoñible no corpus do *Tesouro*. A esta base accederase mediante recursos de fácil manexo proporcionados por un servizo web.

- Desenvolvemento dunha aplicación para a explotación eficiente a través da web de toda a información almacenada na base de datos anterior. A interface desta aplicación debe ser de doado manexo, mesmo para usuario sen coñecementos científicos específicos. A información será mostrada ó usuario en formato de texto, multimedia e a través de mapas que representan as referencias xeográficas almacenadas.

A aplicación anterior debe permitir distintos modos de busca e navegación a través do corpus: utilizando lemas, variantes ou os clasificadores semánticos. Ademais debe ofrecer a cartografía coa distribución territorial dos lemas ou das súas variantes, dentro do dominio galego e do dominio portugués, en Portugal e Brasil. Esta

distribución territorial será de tipo administrativo, concellos no caso de Galicia e Portugal e mesorexións para Brasil, cunha extensión xeográfica moito maior. Estes mapas deberán ser interactivos e responder a multitude de eventos do usuario para esclarecer con detalle a información xeográfica do Tesouro.

A visualización interactiva a través de mapas da información almacenada debe de ser áxil e estar baseada na medida do posible en estándares actuais de intercambio de información xeográfica e de visualización de información gráfica. Ademais, será importante que se poida acceder ó sistema dende o maior número de navegadores posible sen a necesidade de instalar software adicional (plug-ins, applets, etc.).

O deseño do sistema debe de ser flexible e escalable e permitir a incorporación futura de información lingüística e cartográfica doutros dominios. O sistema desenvolvido deberá ser implementado con software libre que non leve consigo ningún gasto de licenzas.

3. DESEÑO GLOBAL DO SISTEMA

A arquitectura proposta para o desenvolvemento da aplicación deseñouse para obter un sistema distribuído no que os seus compoñentes estiveran desacoplados, podendo estes incluso estar en máquinas separadas fisicamente. O usuario final poderá acceder ós recursos da aplicación sen percibir a separación que exista entre eles, vendo o conxunto de elementos separados coma un sistema único.

Para permitir unha separación modular e desacoplada escolleuse deseñar o sistema empregando o patrón de deseño Hierarchical-Model-View-Controller (HMVC), ou Modelo Vista Controlador Xerárquico, unha evolución do patrón MVC usado en moitas aplicacións web actuais.

Outro aspecto importante no deseño do sistema será crear unha aplicación web rápida, na que o usuario non se vexa atrapado en longos procesos de espera recuperando a información do servidor. Por isto decidiuse implementar o paradigma SPI *single-page interface* ou *interface de páxina única*, na aplicación web executada no navegador. Este paradigma promove que a aplicación web poida funcionar completamente na mesma páxina, sen ser esta recargada.

Para facer o sistema máis escalable decidiuse separar as súas responsabilidades creando tres compoñentes independentes que se relacionarán entre si para levar a cabo os obxectivos do proxecto. Dous deses compoñentes serán servizos web REST, que grazas á súa arquitectura e simplicidade proporcionarán un acceso eficiente e fácil á información do *Tesouro*, non so ó cliente web a desenvolver, senón a futuras aplicacións, probablemente desenvolvidas por terceiros.

A continuación na **[Figura 1]** amosamos o diagrama UML de despregue da aplicación; este diagrama modela a arquitectura en tempo de execución do sistema e introduce os elementos hardware (nodos) e os compoñentes despregados en cada nodo.

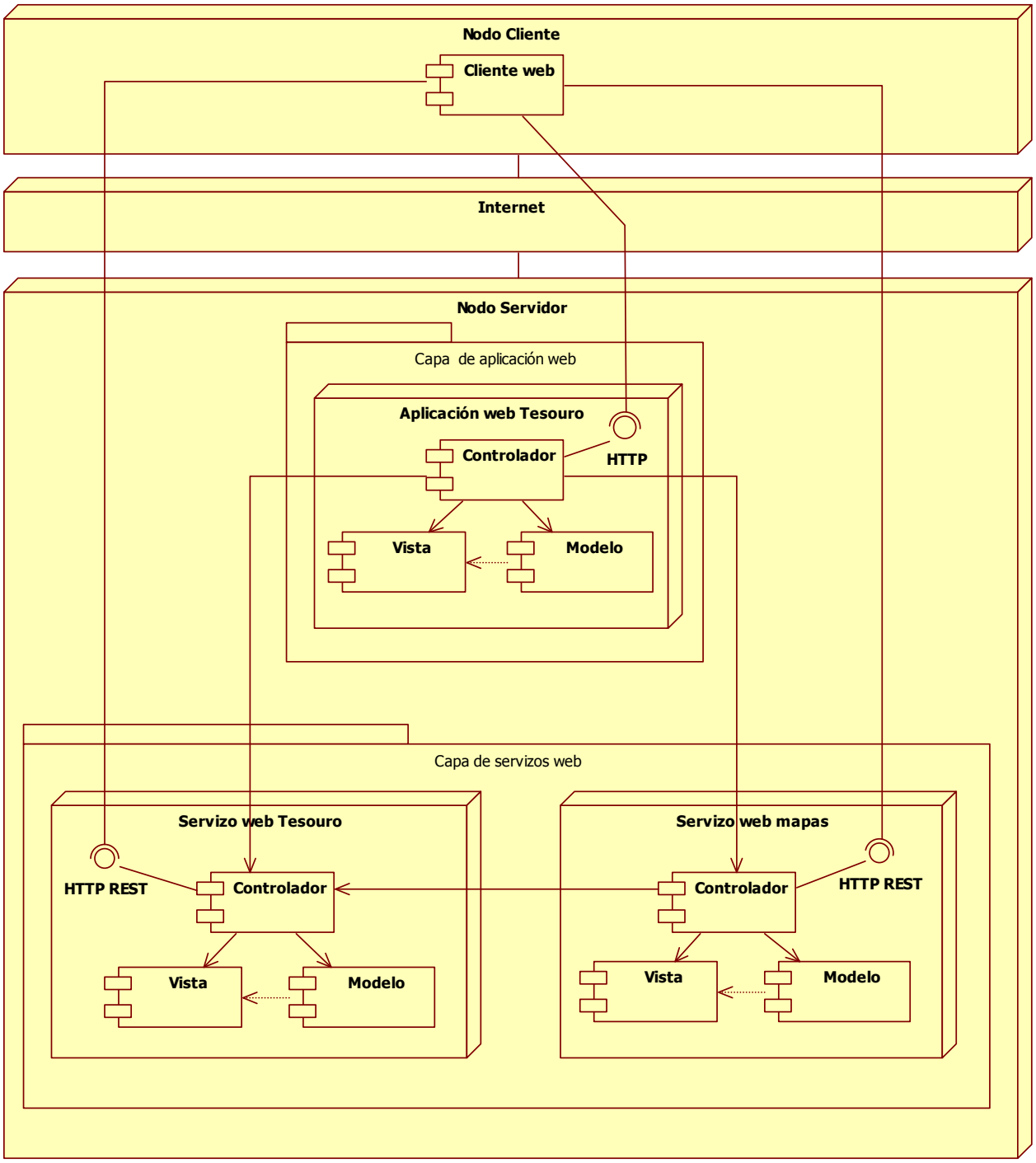

FIGURA 1: DIAGRAMA UML DE DESPREGUE

No Nodo Cliente reside un único compoñente, o cliente web, representado por un navegador web; desde este, o usuario final realizará toda a interacción cos demais compoñentes da aplicación, á parte de servir para probar tódalas funcionalidades implementadas no servidor.

No Nodo Servidor temos dúas capas, a Capa de aplicación e a Capa de servizos web. A primeira posúe un compoñente chamado "Aplicación Web Tesouro" que ten

como responsabilidade xerar a aplicación web HTML/JavaScript que se visualizará e executará no cliente; a comunicación entre este último e o cliente realízase mediante unha interface HTTP. A "Aplicación web Tesouro" comunícase cos controladores do "Servizo web Tesouro" e o "Servizo web mapas" para completar o deseño da aplicación web servida. Esta aplicación web responderá ó paradigma SPI. O código fonte (HTML/JavaScript) da aplicación web SPI será descargado unha única vez por sesión de usuario. Despois o usuario comunicarase co resto dos compoñentes de arquitectura empregando a interface de páxina única (SPI).

A segunda capa presente no Nodo Servidor é a Capa de servizos web, na que se agrupan os distintos servizos web cos que se comunicarán o Cliente web e a capa superior (Capa de aplicación web). O Servizo Web Tesouro será o encargado de explotar a información do *Tesouro*, e proporcionar unha serie de operacións para acceder de forma rápida e sinxela ós contidos almacenados na base de datos do *Tesouro*. O segundo servizo é o Servizo Web de Mapas, que ten como función xerar mapas de divisións administrativas interactivos que serán coloreados no cliente web, empregando os datos proporcionados polo Servizo web Tesouro. Outra das súas funcións será a xeración de mapas de resultados en formatos imprimibles, coma png ou pdf.

A disposición da base de datos do *Tesouro*, así como dos mapas asociados, a través dunha serie de recursos de fácil acceso mediante servizo web REST, será unha baza fundamental para a difusión do *Tesouro*, ademais de ofrecer a posibilidade de creación de futuros recursos ou aplicacións que se nutran da información que ofrece este proxecto. Permitirá tamén, por exemplo, a elaboración de dicionarios dialectais e doutros recursos lexicográficos ou o aproveitamento do corpus para facer comparacións con materiais doutros dominios lingüísticos.

4. SOFTWARE LIBRE NO PROXECTO

Un dos obxectivos e requirimentos principais deste proxecto foi a utilización de software libre para a súa construción e desenvolvemento, o que permitiu ter un custe cero en licenzas de software. Decidiuse liberar o código creado baixo a licenza MIT, xa que é unha licenza que permite a compatibilidade con código existente, e, en xeral, non presenta obstáculos con calquera outra licenza no caso de que se desexe integrar bibliotecas de terceiros.

A linguaxe de programación escollida para a implementación dos distintos subsistemas que residen no servidor foi PHP. Publicada baixo a *PHP License*, a Free Software Foundation considera esta licenza como software libre.

O sistema web esta construído sobre CodeIgniter, un Framework open source escrito en PHP baseado no patrón Modelo Vista Controlador (MVC). O seu obxectivo, igual que o de tódolos frameworks, é axudar os desenvolvedores a crear proxectos dunha forma moito máis rápida e eficiente, sen ter que empezar unha aplicación desde cero, provendo unha serie de ferramentas, librerías e convencións que axilizarán o proceso do traballo, para permitir que os desenvolvedores se centren no desenvolvemento da aplicación. A versión de CodeIgniter enpregada posúe unha licenza

propia similar a Apache/BSD, aínda que na actualizade a licenza foi cambiada a OSL 3.0.

Como sistema de xestión de base de datos (SXBD) do servizo web empregouse PostgreSQL. É un xestor de bases de datos obxecto-relacional que pode ser empregado libremente, está liberado baixo a *PostgreSQL License*, unha licenza libre recollida pola Open Source Initiative.

A maiores das tecnoloxías mencionadas podemos citar as seguintes ferramentas e librerías libres para o desenvolvemento do proxecto. Indícase entre paréntese a licenza libre que empregan:

- **Debian GNU/Linux** (GPL): Sistema operativo libre.
- **jQuery** (MIT): Librería JavaScript.
- **shp2svg** (LGPL): Ferramenta en liña de comandos que transforma arquivos shapefile en SVG.
- **Apache** (Apache License 2.0): Servidor web.
- **Mercurial** (GNU GPL v2): Sistema de control de versións.
- **Netbeans** (CDDL / GPL v2): Entorno de desenvolvemento.
- **cairo rsvg** (GNU GPL v2): Librería Python para a transformación de arquivos SVG.
- **Quantum GIS** (GPL): Sistema de Información Xeográfica de escritorio.

5. ACTIVIDADE PÚBLICA

O *Tesouro do léxico patrimonial galego e portugués* é un proxecto de ámbito internacional, con implicación de universidades de Galicia, Portugal e Brasil, o que garante unha elevada visibilidade e difusión á ferramenta informática que lle serve de base. Do mesmo modo, a calidade científico-técnica do proxecto foi refrendada coa obtención de financiamento externo en convocatorias competitivas[1].

De modo paralelo aos traballos do proxecto e ás reunións internas do grupo de investigación, o *Tesouro* foi presentado en numerosos encontros científicos nacionais e internacionais, mediante contribucións en que a cuestión informática ocupou un papel predominante. Así, ademais de difundir o proxecto, obtívose importante *feedback* da comunidade investigadora nos campos da dialectoloxía ou da lingüística computacional. Unha pequena mostra destas intervencións durante o último ano son:

---

[1] *Tesoro del Léxico Patrimonial Gallego y Portugués. Banco de datos electrónico (corpus gallego)* (subsidiado polo Ministerio de Ciencia e Innovación, 2010-2012) e *Tesouro dialectal português (TEDIPOR)* (Subsidiado pola Fundação para a Ciência e a Tecnologia, 2010-2012).

- Álvarez Pérez, Xosé Afonso, Clarinda Maia et al.: "Tesouro Dialectal Português". Simposio Internacional *Limits and Areas in Dialectology* (Lisboa, outubro 2011).

- Álvarez Pérez, Xosé Afonso & Xulio Sousa (2012): "Presentation of the Tesouro do léxico patrimonial galego e portugués". 7º Congreso da *International Society of Dialectology and Geolinguistics* (Viena, xullo de 2012).

- Brandão, Sílvia (coord.): Mesa redonda *Contribuições do projeto Léxico Patrimonial Galego e Português*, con Silvia F. Brandão (UFRJ), Rosario Álvarez (USC), Vanderci de A. Aguilera (UEL) e Raïssa Ricardo Gillier (CLUL-UL). *II Congresso Internacional de Dialetologia e Sociolinguística. Diversidade Linguística e Políticas de Ensino. Homenagem a Vanderci de Andrade Aguilera.* (Belém, Brasil, setembro de 2012).

- Gillier, Raïssa & Sandra Pereira: "TEDIPOR: Thesaurus of Dialectal Portuguese". Actas do 15th EURALEX International Congress (Oslo, 7-11 Agosto 2012).

- González Seoane, Ernesto (coord.): panel "O Tesouro do léxico patrimonial galego e portugués, aspectos metodolóxicos e aplicacións". X Congreso Internacional da Asociación Internacional de Estudos Galegos (AIEG). (Cardiff, Reino Unido, setembro de 2012). Presentáronse varias comunicacións relativas a esta ferramenta.

Ten especial relevancia o Seminario Internacional *Léxico patrimonial: banco de datos e cartografía*, celebrado en xuño do 2011 polo Instituto da Lingua Galega, con participación de investigadores brasileiros, galegos e portugueses. Nese encontro presentouse publicamente por primeira vez un prototipo da ferramenta informática e comentáronse en detalle todas as súas funcionalidades.

É necesario referir que a actividade pública do Tesouro e da súa ferramenta informática non se restrinxe á comunidade científica, senón que tamén foi referenciado en medios de comunicación xeneralistas, como se constata, por exemplo, en:

- http://www.lavozdegalicia.es/santiago/2011/04/20/0003_201104S20C2997.htm (*Un estudio aborda el léxico patrimonial de Galicia, Brasil y Portugal*)

-http://www.europapress.es/galego/noticia-universidades-galegas-portuguesas-brasileiras-traballan-nunha-base-datos-gratuita-lexicografia-lusofona-20110419182145.html (*Universidades galegas, portuguesas e brasileiras traballan nunha base de datos gratuíta sobre a lexicografía lusófona*)

Por último, cómpre sinalar que este traballo foi merecedor do Premio á Excelencia Lingüística da USC na categoría de proxectos fin de carreira, elixido entre todos os premiados á Calidade Lingüística de tódolos centros da USC.

6. CONCLUSIÓN

Este proxecto fin de carreira, "Desenvolvemento dun sistema para a visualización web do *Tesouro de Léxico Patrimonial Galego e Portugués*", deu como resultado unha ferramenta que está a ser empregada actualmente en universidades de Galicia, Brasil e Portugal[2] e que serve de soporte para a realización dun proxecto internacional de investigación lingüística. Gozou dunha ampla difusión en congresos e encontros de investigadores sobre léxico dialectal e corpus lingüísticos e tamén en medios de comunicación

O proxecto foi avaliado moi positivamente polo tribunal da *Escola Técnica Superior de Enxeñaría* (USC), que o avaliou cunha calificación de sobresaínte (9.5), poñendo en destaque a súa utilidade como exemplo de colaboración entre o ámbito da lingüística e a enxeñaría.

Desde o punto de vista do léxico, esta ferramenta permite ofrecer nunha única plataforma dixital materiais lexicográficos procedentes de fontes moi diversas e heteroxéneas en canto a formato e estrutura. O sistema informático permite a recuperación áxil da información segundo diferentes criterios (variante, lema, clasificador semántico, rexión,...) e inclúe unha aplicación de cartografía automática que axuda ao usuario a localizar a procedencia xeográfica das palabras.

Este recurso é esencial para pór a dispor da comunidade científica e da colectividade en xeral un rico patrimonio inmaterial que está ameazado polo abandono das formas de vida tradicionais e polas dificultades que sofre a lingua galega. O tratamento informático dos datos é clave para a elaboración de traballos lingüísticos baseados nestes materiais

O desenvolvemento do proxecto non finalizou no momento da súa entrega, senón que a día de hoxe séguense a incorporar novas funcionalidades e engadidos. Por outra banda, a publicación da información a través de servizos web REST, facilita e promove a creación de novas aplicacións que exploten a extensa e crecente fonte de información que proporciona o Tesouro, non quedando esta limitada o cliente web desenvolvido para este proxecto, senón a aplicacións desenvolvidas posiblemente por terceiros.

[2] Aínda que a aplicación non foi publicada definitivamente pode consultarse unha versión totalmente funcional no seguinte enlace: http://ilg.usc.es/tesouro/v6/ **user**: mancomun **pass**: tmppassgen75846