A codificación e estruturación dos documentos

Os documentos que se incorporan ao CORGA codifícanse segundo o estándar XML (*eXtensible Markup Language*) co fin de incrementar as posibilidades de recuperación de información e, ademais, garantir a permanencia no tempo. Esta codificación implica un deseño e estruturación que dá conta da disposición interna característica de cada un dos grandes tipos de textos (xornal, teatro, ensaio etc.). Así, exemplificando cun texto xornalístico, esta estruturación permite considerar un xornal un único documento que está organizado en noticias, distribuídas en *seccións*, as cales, á súa vez, conteñen obrigatoriamente un *corpo* e opcionalmente un *titular*, *resumo* e/ou *pé de foto*. A maiores, cada un destes elementos está constituído por parágrafos (texto comprendido entre dous puntos e á parte), e estes son segmentados en oracións (secuencia textual separada do resto do texto por un signo forte de puntuación). Naturalmente, o xornal posúe ademais unha cabeceira complexa na que se recollen os datos bibliográficos e unha cabeceira específica por noticia onde se inclúen os datos relativos a autor, sección e áreas temáticas que correspondan.

Esta disposición detallada habilita a posibilidade de que, no sistema de recuperación de información a través da opción Buscar en, unha vez indexados os documentos, se poidan realizar consultas sobre a totalidade do documento (*noticia*, seguindo co exemplo) ou sobre unha unidade estrutural concreta (para o tipo de documento xornal: *titular, resumo, pé de foto* ou *corpo*).

A maiores das unidades estruturais xa citadas que dan conta da organización interna das noticias que poden conformar xornais, revistas e blogs (*corpo, titular, resumo* e *pé de foto*), tivéronse en conta na estruturación interna outras que adoitan aparecer no tipo de documento *libro*. Comparten, así mesmo, outra característica: todas son opcionais, pois o único elemento constitutivo obrigatorio é *corpo*. Trátase de *prólogo* (engloba as distintas denominacións: *presentación, limiar, introdución, preámbulo, prefacio* etc.), *apéndice* (inclúe *epílogo, agradecementos, glosario* etc.), *dedicatoria*, *cita* (acolle só aquelas que os autores empregan totalmente illadas de calquera outro texto, e en ningún caso aparecen delimitadas as citas textuais usadas en ensaios como xustificación do que se di ou refire), *encabezamento* e *nota* (menos as referencias bibliográficas que non se indexan).

Todos os elementos estruturais descompóñense en parágrafos, e estes, á súa vez, en oracións, sendo esta última a unidade sobre a que se realizan as buscas no sistema de consultas e, loxicamente, tamén á que pertence o segmento que se recupera que se ofrece en concordancias, aínda que logo se poida ampliar o contexto a dúas oracións anteriores e dúas posteriores.

Cómpre subliñar que os textos introducidos no CORGA non conservan o formato: cursivas, grosas ou emprego de versais como elemento marcador desaparecen cando se prepara o documento para a súa inclusión no corpus. Tampouco se mantén a paxinación. Débese ter en conta que unha vez que un texto se converte en documento electrónico deixa de ter importancia a localización por páxina do orixinal, cuxo mantemento suporía cortar unidades, pois as localizacións fanse doutros xeitos. Non se cita xa a ocorrencia dunha forma dada na páxina *n* de tal obra, senón que a referencia se fai pola aparición da forma *x* na obra *z* que aparece no corpus, no noso caso, CORGA.

A codificación que se lles aplica aos textos que se introducen no CORGA inclúe, á parte da estruturación nas unidades anteriores segundo corresponda, unha serie de etiquetas que achegan información de diverso tipo. Pode verse unha descrición das mesmas na pestana Guía premendo en *Relación de etiquetas empregadas na codificación do corpus*.