

EGEO: La estructura geográfica de una base de conocimiento

Ignacio González, Sergi Cervell, Josep M. Merenciano, David Trotzig y Joan Vaqué
{igonzalez, scervell, jmmerenciano, dtrotzig, jvaque}@planeta-actimedia.es

Departamento de Lingüística Computacional
Banco de Contenidos
Planeta Actimedia S.A.

Resumen

En la presente contribución se describe la Estructura Geográfica (EGEO) que forma parte del banco de contenidos (BCP) desarrollado en el grupo Planeta. Dicha descripción incide tanto en las bases formales y constructivas del EGEO en sí, como en su interrelación funcional y de diseño con la base de conocimiento en general.

1. Preliminares

La resolución de la localización geográfica es una necesidad para cualquier sistema gestor de información. Para garantizar su funcionalidad, todo sistema de este género debiera cumplir ciertos requisitos:

- a) Una base toponímica sólida, rica en contenidos e internamente consistente.
- b) Una organización estructural que refleje las relaciones entre las entidades físicas y políticas que componen el sistema, a la par que garantice su imbricación con las demás estructuras de conocimiento de la ontología.
- c) Una ductilidad apropiada para el tratamiento tanto de las formas canónicas como de las variantes toponímicas.

Para atender tales requisitos se ha desarrollado una Estructura Geográfica (EGEO), en el marco general de su ontología lingüística, la Base de Conocimientos (BDCon).

2. El conjunto de la BDCon

La BDCon también incluye un árbol tipológico (ATIP), donde las entidades léxicas son clasificadas en una estructura que establece relaciones de hiponimia e hiperonimia. En el contexto de la BDCon, el ATIP tiene distintas funcionalidades, entre ellas: el estudio sistemático de la desambiguación semántica y de la sinonimia, y la estructuración e indexación

de materiales enciclopédicos. En la actualidad, el ATIP recoge más de 170.000 acepciones en 110.000 nodos.

La categorización temática es gestionada por la estructura arbórea ATEM. Las referencias temporales por la estructura ECRON y ciertas categorías, atributos y metadatos, por la estructura de soporte ASOP.

3. Inserción del EGEO en la BDCon

Todos los nodos presentes en el EGEO están ligados al árbol tipológico (ATIP) de la BDCon; p.ej.: el nodo del EGEO físico “Etna” está relacionado con el nodo ATIP “volcán”, así como el nodo del EGEO político “España” lo está con el nodo ATIP “país”.

Del mismo modo, los antropónimos, acontecimientos y demás elementos léxicos gestionados por la BDCon tienen su correspondiente relación con el EGEO.

Si consideramos las propias estructuras internas del EGEO, que describiremos a continuación, es el ligamen de todas estas estructuras de conocimiento lo que proporciona sentido pleno a la clasificación de las entidades toponímicas del BCP. Lo cual se traduce en múltiples niveles y posibilidades, tanto de organización de la información como de búsqueda de la misma. P.ej.: el sistema es capaz de organizar y permitir la recuperación de la información por parámetros como “municipio” (descriptor tipológico ATIP), “Venezuela” (nodo geográfico EGEO) y “división administrativa de segundo orden” (metadato ASOP).

4. Estructura y uso referencial del EGEO

El EGEO está formado por más de 317.000 topónimos, y su función primera consiste en organizar dichos topónimos. Su segunda funcionalidad consiste en servir de

referente a otros elementos de la BDCon y permitir clasificar, según su contexto geográfico, los distintos tipos de entidades léxicas contenidos en ella; esto es, en atención al marco físico y político al que tales entidades están significativamente relacionadas. En su fase actual, dichas entidades son básicamente nombres propios (antropónimos, marcas comerciales, acontecimientos, instituciones, obras, etc.).

El EGEO consta de tres estructuras de tipo arbóreo, basadas en relaciones jerárquicas de precesión (“incluye a”) y sucesión (“pertenece a”) entre sus nodos: se trata del EGEO físico (donde se clasifican las entidades léxicas referentes a entidades geofísicas: continentes, océanos, mares, montañas, ríos...), el EGEO político (que estructura las entidades geopolíticas: países, comarcas, distritos, municipios...) y el EGEO histórico (recoge las entidades geopolíticas ya extintas: p.ej., el Imperio romano y la URSS).

El nexo entre los tres árboles se cifra en un entramado de relaciones, tendidas discrecionalmente entre los árboles EGEO físico y EGEO político, y desde el EGEO histórico hacia los EGEOs físico y político. Se desglosan del siguiente modo:

a) Relaciones de “totalmente contenido en” y “parcialmente contenido en”. Desde el EGEO político al EGEO físico (p.ej., Alemania está totalmente contenida en la región de Europa central) y viceversa (p.ej., el archipiélago de Hawai está contenido totalmente en Estados Unidos), y desde el EGEO histórico hacia el físico (Checoslovaquia estaba totalmente contenida en la región Europa central).

b) Relación “Coincide con”. Desde el EGEO histórico hacia el EGEO político. P.ej.: la URSS coincidía con –el conjunto de– Rusia, Ucrania, Bielorrusia, Moldavia, etc.

Se establecerán tantas relaciones de “parcialmente comprendido en” y “coincide con” como precise el nodo desde el que se tienda dicha relación (p.ej. completando el caso anterior Checoslovaquia coincide con –el conjunto de– las actuales repúblicas Checa y Eslovaca).

c) No hay que olvidar que con las relaciones anteriores interactúan las jerarquías arbóreas y las propias

relaciones establecidas en la asignación de un topónimo. P.ej. podemos saber directamente cuáles son los topónimos asignados a un nodo de una estructura cualquiera de la BDCon, como los ríos (ATIP) de Alemania (EGEO político) y también, indirectamente, a través de la herencia arbórea y de las relaciones entre árboles, establecer los ligámenes entre la colina (ATIP) de Kamparkalns (EGEO físico), sita en Lituania (EGEO político), con respecto al continente europeo (uno de los nodos superiores del EGEO físico).

En otro orden de cosas, el EGEO no es menos importante como referente de las formas canónicas de la toponimia del BCP, y como gestor de sus diferentes transcripciones y variantes.

5. Funcionamiento de la interfaz

La interfaz de búsqueda de topónimos permite consultar las estructuras de la BDCon, tanto por separado como en búsquedas cruzadas.

Una vez abierta la interfaz, la pantalla se divide en dos grandes zonas: a la izquierda, la zona de visualización y manipulación de árboles, a la derecha la zona de selección de consultas. Una ventana inferior mostrará el resultado de la búsqueda solicitada. Y sobre cualquier topónimo la interfaz muestra toda la información que contiene la base.

6. Desarrollos futuros

Es posible seguir desarrollando tanto nuevas utilidades como el tratamiento de los datos internos de la base de topónimos. En el primer aspecto, este sistema puede servir como base para la implementación de un reconocedor de nombres propios (topónimos) en texto; o para la implementación de sistemas que impliquen la generación o gestión de léxico cartográfico. En referencia a su estructura interna, queda la puerta abierta al desarrollo de nuevas relaciones entre los diversos árboles EGEO, y entre éstos y el conjunto de la BDCon; y la posible migración de los datos de las actuales estructuras arbóreas a grafos e incluso a puntos de un sistema estructurado de mapas.