Enriquecimiento de entidades de Wikidata mediante un modelo de descomposición y mapeado de categorías de Wikipedia

Tomás Saorín[1], Juan-Antonio Pastor-Sánchez[2]

[1] https://orcid.org/0000-0001-9448-0866 – Universidad de Murcia (España) – tsp@um.es
[2] https://orcid.org/0000-0002-1677-1059 – Universidad de Murcia (España) – pastor@um.es

Resumen: El objetivo de este trabajo es explorar la relación entre las categorías asignadas a los artículos de Wikipedia con la descripción y metadatos generados en Wikidata. Las categorías de Wikipedia, desde el punto de vista de la organización del conocimiento, la categorización/clasificación, representan un esfuerzo único en la historia de estructuración del conocimiento histórico, cultural, científico y factual realizado de forma colaborativa y masiva. La comunidad de editores, al mismo tiempo que elabora contenidos (artículos), trabaja en su organización conforme a un esquema de conceptos (categorías), el cual evoluciona orgánicamente junto al contenido. Se plantea utilizar la categorización de artículos de Wikipedia para enriquecer la descripción de entidades en Wikidata. Para ello se propone procesar los literales de las categorías mediante técnicas de procesamiento de lenguaje natural (PLN) estableciendo patrones que permitan identificar tanto propiedades como entidades o valores con los que construir declaraciones para una entidad. La secuencia de operaciones propuesta sería el siguiente: 1) Selección de un conjunto coherente de categorías, 2) Establecimiento de patrones de procesamiento de literales y asignación a propiedades y elementos de Wikidata, 3) Creación de declaraciones con cualificadores para cada categoría procesada y 4) Programación de bots para el procesamiento automático de categorías, enriquecimiento y validación de las descripciones de elementos de Wikidata. La propuesta recogida en este trabajo se centra en el uso de diferentes propiedades y entidades de Wikidata para desarrollar el punto 3. Aunque el sistema de categorías de Wikipedia está sometido a discusión y responde a necesidades coyunturales de organización y navegación de contenidos, consideramos que se trata de un trabajo directo de los editores de la enciclopedia en el que se materializa conocimiento en relación con los artículos. El mecanismo descrito para codificar el resultado de un reconocimiento de entidades y conceptos en las categorías, pese a poder ser ya operativo, mejoraría mediante la creación de una propiedad específica “scope os topic combined” o bien “used with property”, cuyo rango estuviera restringido a propiedades, y que sólo pudiera usarse como cualificador de una declaración con la propiedad P971. De esta forma el significado y el uso estaría claramente establecido en la ontología de Wikidata. La automatización de un proceso regular que a partir de las categorías asignadas a cada artículo, en cualquier idioma, enriquezca y valide las declaraciones de cada elemento, constituiría un refuerzo sin duda efectivo, aprovechando una dinámica de edición – la asignación y creación de categorías – que ya existe. Además serviría para avanzar en tener un esquema de conceptos de más calidad, al especificarse el significado de las categorías que suponen una composición de varios términos y que en realidad resuelven necesidades descriptivas por otros medios. También serviría para detectar redundancias e inconsistencias a varios niveles.