Enriquecimiento. Extracción de descriptores

Este apartado contiene las siguientes secciones:

1. Descripción del módulo de extracción de descriptores

La tarea a abordar por el sistema es la extracción de descriptores correspondientes a ROs determinados. Entendemos descriptor como espacio semántico (de granularidad arbitraria) presente en un texto, que puede expresarse léxicamente mediante etiquetas/descriptores que corresponden a sintagmas nominales representativos del descriptor. Los descriptores de grano grueso corresponderán a áreas temáticas, mientras que los descriptores de grano fino se referirán a conceptos más específicos.

El término tópico (o *topic*) en la literatura de PLN se entiende como un espacio semántico (de granularidad arbitraria) incluido en un texto que habitualmente se presenta asociado a la tarea de generación de modelos de tópicos o *topic modelling*. Estos modelos se generan a partir de una colección de textos de forma que se infiere la distribución de tópicos presentes en la colección, así como la distribución de vocabulario asociada a cada tópico. Normalmente, la tarea se centra en extraer tópicos de grano grueso. Por otro lado, la tarea de selección de descriptores o etiquetas que describan estos tópicos es suplementaria y se conoce en la literatura como *topic labelling*.

Otras tareas íntimamente relacionadas con la extracción de descriptores son las siguientes (resumen en tabla 6):

- *Extracción de terminología*: Extraer de un texto la terminología incluida en el mismo, sin realizar ninguna discriminación de acuerdo con su relevancia en el texto. Es decir, el objetivo no es extraer los términos que mejor describan el texto, sino extraer sintagmas que correspondan a unidades terminológicas.
- *Extracción de entidades nombradas*: Identificar en un texto distintos tipos de entidades nombradas como pueden ser nombres de persona, lugares y organizaciones. En esta tarea tampoco se presta atención a la relevancia de cada entidad en el texto.
- *Extracción de palabras clave*: Identificar términos, palabras, entidades que mejor describan el contenido de un texto. No se prefija un nivel de granularidad predeterminado, pero normalmente la tarea se centra en palabras clave específicas y que estén presentes en el texto.
- *Clasificación temática de textos*: Identificar las áreas temáticas o los temas más relevantes en el texto a partir de una lista cerrada de áreas o temas.

	Descriptores grano grueso	Descriptores grano fino	Relevancia descriptiva	Datasets anotados
Topic modelling + topics labelling	V	-	V	-
Extracción terminología	-	V	-	-
Extracción de entidades nombradas	-	V	-	V
Palabras clave	-	V	V	V
Clasificación temática	V	-	V	V

Tabla: Enfoques útiles para la extracción de descriptores.

Dado que el objetivo principal de la extracción de descriptores en el marco de Hércules es la generación de etiquetas descriptoras de los ROs de cara a facilitar la navegación sobre estos por parte de los usuarios, hemos optado de una estrategia mixta que combina técnicas de extracción de terminología, extracción de entidades nombradas, extracción de palabras clave y clasificación temática. Los descriptores extraídos mediante este enfoque automático se combinará, en la medida de lo posible, con los descriptores obtenidos directamente del repositorio mediante APIs o *scraping*.

Se ha determinado que para la generación de etiquetas descriptoras se deben identificar descriptores de diferente granularidad; concretamente, descriptores temáticos y descriptores específicos (Diagrama del sistema propuesto en la imagen 1). Los descriptores temáticos ayudan al usuario a comprender el área de conocimiento a la que pertenece el RO; y los específicos facilitan la exploración de los conceptos específicos más relevantes del RO.

Tanto los descriptores temáticos como los específicos serán enlazados con ítems presentes en las ontologías o taxonomías que se consideren oportunas.

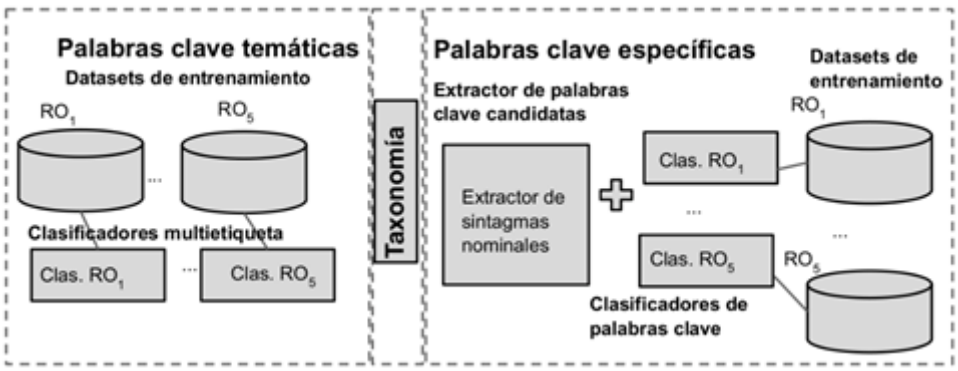


Imagen 1: Diagrama del sistema propuesto.

2. Enriquecimiento. Prototipo y experimentación realizada en el marco del Challenge

Nota: Se incluye información sobre los resultados del Challenge como referencia de algunas técnicas y resultados que se toman como punto de partida en algunas de las tareas de enriquecimiento del proyecto.

Tras analizar la documentación sobre los objetivos del *Challenge*, se consideró que el término "tópico" utilizado hacía referencia tanto a temas generales *subjects areas* (palabras clave -descriptores- temáticas) como a conceptos más específicos (palabras clave -descriptores- específicas). Por ejemplo, se quiere saber si una publicación es relevante al área de "Ingeniería eléctrica y electrónica", y también si incluye conceptos específicos relevantes como "ondas en plasma" o "transistor".

Las palabras clave específicas suelen estar lexicalizadas en el contenido textual de los ROs; las palabras clave temáticas, en cambio, no. Esta circunstancia implica la adopción de diferentes enfoques a la hora de abordar la extracción los dos tipos de palabras clave.

Para abordar la extracción de palabras clave temáticas se han estudiado e implementado dos enfoques:

1. Extracción de descriptores mediante modelos no supervisados de topic labelling.
2. Clasificación multietiqueta supervisada.

Para la extracción de palabras clave específicas se han estudiado dos enfoques:

1. Estrategia supervisada basada en una arquitectura neuronal seq2seq.
2. Estrategia consistente de dos pasos: un primer paso para detección de palabras clave candidatas basado en patrones lingüísticos (utilizados para la extracción de terminología) y NERC, y un segundo paso de cribado de las palabras candidatas basado en un clasificador supervisado booleano.

Los enfoques supervisados se han entrenado sobre datasets (ver tabla 7) de publicaciones y protocolos que incluyen palabras clave temáticas y específicas asignadas manualmente (se ha utilizado el dataset de Krapivin y otros (2009), y se han generado datasets a partir de los repositorios Scopus y Bio-protocol), y se han evaluado según métricas de Precisión, Cobertura y F-score sobre una fracción de cada dataset.

Dataset	**Descripción**	**Nº cat.**	**Nº art.**	**Nº cat. /art.**
Datasets de palabras clave temáticas				
Scop27	Conjunto extraído de Scopus compuesto por pares de abstracts y categorías ASJC de primer nivel.	27	725, 7 K	1,69
Scop262	Conjunto extraído de Scopus compuesto por pares de abstracts y categorías ASJC de segundo nivel.	262	723, 3 K	2,21
Scop5	Conjunto extraído de Scopus compuesto por pares de abstracts y cinco categorías ASJC (seleccionadas aleatoriamente) de primer nivel.	5	580, 5 K	1,29
Bio	Conjunto extraído de Bio-protocol compuesto por pares de artículos completos y categorías de primer nivel de Bio-protocol.	13	3,2 K	1,84
Datasets de palabras clave específicas				
Krapivin	Conjunto extraído de Scopus compuesto por pares de artículos completos y palabras clave.	12,3 K	2,3 K	5,34
Scopus_key	Conjunto extraído de Scopus compuesto por pares de abstracts y palabras clave de granularidad media.	780, 6 K	446, 7 K	5,02
Bio_key	Conjunto extraído de Bio-protocol compuesto por pares de artículos completos y palabras clave.	9,5 K	3,2 K	3,02

Tabla: Datasets construidos para entrenamiento de clasificadores multi-etiqueta (palabras clave temáticas) y de clasificadores booleanos (palabras clave específicas). Se muestra el número total de categorías consideradas en la tercera columna. En la cuarta columna el número total de artículos, y en la quinta el número medio de categorías asignadas a cada artículo.

Describimos, a continuación, cada uno de los enfoques analizados y la evaluación de los resultados:

Extracción de descriptores mediante modelos no supervisados de topic labelling

Para implementar esta estrategia no supervisada se han generado modelos LDA (Blei et al., 2003) para distintas granularidades (100, 200, y 300 topics) a partir de los abstracts del dataset *Scop27*. Para garantizar modelos de gran coherencia se realizaron entrenamientos de 70 iteraciones y 10 pases. Los descriptores y léxico relevante asociado se inspeccionaron de forma manual con el objeto de determinar si los descriptores detectados eran de utilidad para caracterización de los textos, y si la terminología asociada a los descriptores resultaba práctica para la generación de etiquetas. Se observó que, en general, los descriptores detectados no se alineaban de forma notable con "tópicos" interpretables por personas, y que los términos más probables no resultaban útiles como etiquetas o palabras clave para describir el tópico. Estos términos podrían tener cierta utilidad para generar etiquetas más descriptivas, pero habría que aplicar técnicas más sofisticadas de *topic labelling*. Se consideró un enfoque complicado de implementar, que acarreaba cierto riesgo tecnológico, y débilmente competitivo frente a una estrategia supervisada.

Clasificación multietiqueta supervisada

Esta estrategia consiste en abordar la identificación de las palabras clave temáticas de cada documento como una tarea de *clasificación multi-etiqueta*, siendo los documentos los elementos a clasificar y las palabras clave temáticas las etiquetas a asignar. La tarea consiste en identificar, a partir de una lista cerrada de etiquetas, aquellas que son relevantes a cada documento.

La representación de los documentos se hizo de acuerdo a un modelo de bolsa de palabras para el caso de Logistic Regresion (LR) y SVM. Concretamente, el documento se tokenizó para distinguir palabras y signos de puntuación, y posteriormente, a cada palabra se le asignó su valor TFIDF correspondiente. Para eliminar la distorsión ocasionada en la representación por las palabras comunes, se eliminaron las palabras que aparecen en más del 10% de los documentos de la colección a procesar. De esta forma, por cada documento a clasificar se construye un vector cuyas dimensiones son las palabras y los valores los pesos TFIDF. Los clasificadores supervisados multi-etiqueta se alimentan con estos vectores. Para el caso del modelo BERT (Devlin et al., 2018), se utilizó la representación densa y contextual que incluye el modelo para la representación de los textos, y se ajustó el modelo a la tarea de clasificación booleana mediante un proceso de fine-tuning sobre los datos de entrenamiento.

Para implementar los clasificadores multi-etiqueta se optó por una estrategia one-vs-all, es decir, para cada etiqueta se creó un clasificador booleano. Cada clasificador se entrena con ejemplos positivos (textos) correspondiente a los documentos que tienen asignado la etiqueta en cuestión, y con ejemplos negativos (textos) que son todos los demás documentos. Debido al desequilibro entre ejemplos negativos y positivos y el bajo número de ejemplos positivos para algunas etiquetas, se optó por balancear el dataset replicando de forma equilibrada los ejemplos positivos, concretamente aplicando un oversampling aleatorio.

Para entrenar los clasificadores booleanos se utilizaron los clasificadores Logistic Regresion y SVM, usados habitualmente para la clasificación de documentos representados según vectores de palabras. En el caso de BERT, únicamente se generaron clasificadores para el dataset *Scop5* y *Bio* debido al tiempo de entrenamiento que el proceso de fine-tuning requiere.

Estrategia supervisada basada en una arquitectura neuronal seq2seq

Hemos utilizado el dataset *Scopus_key* para entrenar el sistema seq2seq de Meng y otros (2017) ya que éste requiere de un dataset de gran tamaño. Los resultados preliminares que hemos obtenido ofrecen una cobertura de 0.28 para palabras clave específicas presentes en el texto, que aumenta hasta un 0.34 si tenemos en cuenta identificaciones parciales.

Estrategia consistente de dos pasos

La otra estrategia explorada para la identificación de palabras clave específicas consta de dos pasos: En el primer paso se procede a la **de tección de las palabras clave candidatas**, y en el segundo paso las palabras clave candidatas identificadas en el anterior paso son cribadas **mediante un clasificador supervisado** booleano.

Para la detección de candidatos a palabras clave específicas utilizamos una estrategia basada en detección de sintagmas nominales. Para ello se han definido patrones gramaticales, tanto para términos monopolabra (Ej., "nombre") y multipalabra (Ej., "nombre adjetivo", "nombre nombre", ...) que abarcan la mayoría de los sintagmas nominales. Estos patrones se detectan en el texto del artículo una vez se ha anotado con información gramatical. Esta anotación se realiza mediante el procesador lingüistico IXA pipes (Agerri et al., 2014). La anotación se efectúa aplicando al texto de entrada el tokenizador (para distinguir palabras y signos de puntuación), el anotador POS (para asignar las etiquetas POS a cada palabra), y el anotador NERC (para anotar las entidades nombradas multipalabra). Los sintagmas correspondientes a los patrones definidos se extraen de la salida de los tres procesos de IXA pipes, y constituyen las palabras clave candidatas, tanto mono-palabra como multipalabra. De las palabras clave multipalabra también extraemos sub-sintagmas nominales, a no ser que sean entidades nombradas.

- **Frecuencia**: Frecuencia del candidato en el artículo.
- **Posición**: Posición relativa de la primera ocurrencia del candidato en el artículo.
- **Frecuencia en colección de dominio abierto**: Frecuencia del candidato en una colección de noticias.
- **Frecuencia en colección de publicaciones científicas**: Frecuencia del candidato en una colección de artículos científicos.
- **Incluido en título**: Valor booleano que indica si el candidato está incluido en el título o no.
- **Incluido en abstract**: Valor booleano que indica si el candidato está incluido en el abstract o no.
- **Termhood** (sólo monopalabra): Valor que indica el grado de especificidad y relevancia del candidato en el artículo medido mediante TFIDF o Loglikelihood ratio.
- **Nivel de anidación** (sólo monopalabra): Porcentaje de ocurrencias en las que aparece anidado en un candidato multipalabra.
- **Unithood** (sólo multipalabra): Valor que indica la coherencia del término multipalabra medido mediante la mediada de asociación Loglikelihood ratio.
- **Similitudes basadas en embeddings**: Similitud semántica (coseno) entre el embedding del candidato y el embedding construido sobre la unificación del título y el abstract como un solo texto. Se han utilizado embeddings estáticos (FastText) y contextuales (Flair).

Tabla: Atributos (*features*) para caracterizar cada palabra clave candidata.

En el siguiente paso se toma como punto de partida la lista de palabras clave candidatas identificadas en el paso anterior. El objetivo de este paso es eliminar de esta lista las palabras clave candidatas menos probables a ser palabras clave. Efectuamos esta criba como un proceso de clasificación booleana, y utilizamos para ello un clasificador supervisado previamente entrenado con datasets que incluyen documentos y sus respectivas palabras clave. Para la clasificación supervisada representamos las palabras clave candidatas mediante un vector que incluye los atributos mostrados en la tabla 8. Al existir un gran desequilibrio entre el número de ejemplos positivos y de negativos (la mayoría de los candidatos extraídos del texto no son palabras clave), se aplicó un *oversampling* aleatorio sobre los ejemplos de entrenamiento. Se experimentó con los algoritmos Logistic Regresion (LR) y Gradient Boosting Descent (GB).

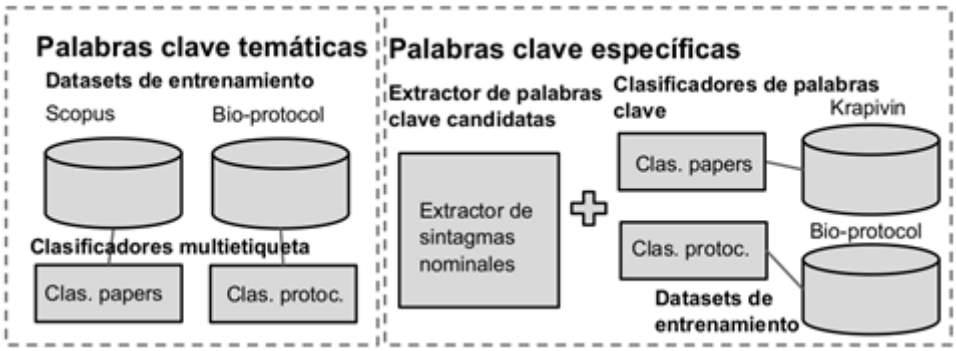


Imagen 2: Diagrama del prototipo presentado al Challenge.

Evaluación del sistema del Challenge

Los clasificadores, tanto de palabras clave temáticas como de específicas, se han evaluado según las métricas de Precisión, Cobertura y F-score sobre una fracción (20%) de cada dataset. El 80% restante se ha utilizado para entrenar cada clasificador.

Palabras clave temáticas

Para el caso de las publicaciones científicas, como era esperable, se obtienen mejores resultados (ver tabla 9) con el repertorio de palabras clave temáticas más genéricas (*Scop27*). En los tres casos (*Scop262*, *Scop27*, *Scop5*) las coberturas obtenidas son altas, sobre todo en los clasificadores LR. Los valores de las precisiones son más bajos, pero tras una inspección manual de los falsos positivos, hemos podido averiguar que muchos de ellos serían clasificaciones correctas pero que no constaban como tal en la referencia. La referencia es fruto de una anotación manual donde en algunos casos no quedan recogidas todas las categorías relevantes al documento. Teniendo en cuenta este aspecto, hemos considerado que, en el caso de las publicaciones, los clasificadores LR son más apropiados para implementar el módulo de extracción, ya que ofrecen una mayor cobertura. Los resultados de los modelos BERT, si bien se limitan a un dataset con un número limitado de categorías, son notablemente superiores a los de LR y SVM en términos de precisión y F-score, por lo que los clasificadores BERT también pueden resultar aptos para su integración en el módulo de extracción.

Los resultados de los clasificadores de protocolos (*Bio*) ofrecen mejores precisiones que los de publicaciones, y en este caso los clasificadores LR también ofrecen mejor cobertura que los SVM, por lo que son éstos los que se han integrado en el módulo de extracción. En este caso BERT no parece ser competitivo, seguramente por el tamaño limitado de *Bio* y por la limitación de BERT a representar documentos largos, como es el caso de los documentos de la colección de *Bio*.

	Scop262		Scop27		Scop5			Bio		
	LR	SVM	LR	SVM	LR	SVM	BERT	LR	SVM	BERT
Precisión	0,21	0,3	0,41	0,44	0,17	0,33	0,58	0,66	0,71	0,52
Cobertura	0,71	0,48	0,83	0,72	0,7	0,42	0,44	0,61	0,57	0,6
F-score	0,31	**0,36**	0,53	**0,54**	0,27	0,37	**0,50**	**0,63**	**0,63**	0,53

Tabla: Resultados de la clasificación multi-etiqueta para identificación de palabras clave temáticas.

Palabras clave específicas

Los resultados (ver tabla 10) ponen de relieve la dificultad de la tarea, sobre todo si nos fijamos en los valores de precisión obtenidos, menores o iguales a 0,1 en todos los casos. Los clasificadores consiguen identificar (valores altos de cobertura) las palabras clave incluidas en el gold-standard, pero añaden un número significativo de candidatos no definidos como tales. Tras una inspección manual de estos falsos positivos, pudimos observar que muchos de ellos eran palabras clave apropiadas. En el gold-standard sólo se establecen entre 3 y 5 palabras clave por artículo, y no se siguen criterios claros para hacer esa selección, lo que dificulta el aprendizaje de su detección. Es por ello que, en esta tarea, parece preferible fijarse en medidas de cobertura a un determinado cutoff, como por ejemplo R@10 (cobertura sobre el top 10). Según R@10 y F-score, los clasificadores GB ofrecen los mejores resultados -también frente a los de la arquitectura seq2seq previamente mencionada-, por lo que serán éstos los que se integren en el módulo de extracción de palabras clave específicas. Debido a la baja precisión de los resultados, este proceso debería venir acompañado de una revisión manual final.

	Mono				Multi			
	Krapivin		*Bio_key*		*Krapivin*		*Bio_key*	
	LR	GB	LR	GB	LR	GB	LR	GB
Precisión	0,01	0,02	0,02	0,05	0,07	0,10	0,02	0,04
Cobertura	0,89	0,88	0,89	0,89	0,82	0,82	0,85	0,82
F-score	0,02	0,04	0,04	0,09	0,13	0,18	0,04	0,07
R@10	0,21	0,23	0,47	0,64	0,61	0,65	0,58	0,67

Tabla: Resultados de la clasificación para identificación de palabras clave específicas.

Además de las evaluaciones automáticas mostradas en las tablas 9 y 10, también se realizó una evaluación cualitativa efectuada mediante inspecciones manuales de los resultados. Para ello utilizamos como conjuntos de testeo las colecciones de ROs establecidas en el *Challenge* para tal fin. Las colecciones correspondían a tres tipos de ROs (publicaciones científicas, protocolos experimentales, y proyectos de GitHub) y se evaluó un único sistema por cada tipo de RO:

- Artículos científicos: Extractor de palabras clave temáticas basado en *Scop262* y *LR*, y extractor de palabras clave específicas basado en *Krapivin* y *GB*.
- Protocolos experimentales: Extractor de palabras clave temáticas basado en *Scop262* y *LR*, y extractor de palabras clave específicas basado en *GB* y *Bio_key*.
- Proyectos GitHub: Extractor de palabras clave específicas basado en extracción de sintagmas nominales y NERC.

Describimos, a continuación, los resultados de la inspección manual por cada tipo de RO:

- Artículos científicos.
    - El sistema propone palabras clave específicas monopalabra y multipalabra, que coinciden con las introducidas manualmente ofreciendo buena cobertura, siempre que la palabra clave específica sea relevante en el texto. En los casos en que no se ha encontrado, la palabra clave manual no está en el texto o sólo está una vez.
    - En el caso de la precisión, la propuesta de palabras clave específicas incluye bastantes más que las introducidas manualmente por los usuarios, si bien se han observado como pertinentes hasta un umbral de 0,55 en monopalabra y 0,5 en multipalabra. Más allá de ajustes de los algoritmos o de sus umbrales, esto sugiere que un aspecto crítico de usabilidad será la facilidad con que el interfaz permita aceptar o descartar palabras clave propuestas.
    - En cuanto a las palabras clave temáticas, la cobertura y precisión observadas son altas.
- Protocolos experimentales.
    - Como en el caso anterior, el sistema propone palabras clave específicas monopalabra y multipalabra, que coinciden con las introducidas manualmente ofreciendo una cobertura alta, siempre que la palabra clave específica esté en el texto.
    - En cuanto a la precisión, la propuesta de palabras clave se comporta del mismo modo descrito en el caso anterior.
    - En cuanto a las palabras clave temáticas, la cobertura y precisión son muy altas, si bien en este caso sólo se reconoce el primer nivel de la taxonomía, ya que no había ejemplos suficientes para entrenar con el segundo nivel.
- Proyectos GitHub.
    - No se han podido aplicar las mismas técnicas que en los casos anteriores por no disponer de un dataset de entrenamiento apropiado (que habrá que construir) y porque la naturaleza de los repositorios de GitHub va a necesitar una aproximación específica para obtener textos con los que poder operar. Las palabras claves específicas no han tenido, en general, ni la cobertura ni la precisión adecuadas.

3. Enriquecimiento. Análisis para la extracción de descriptores temáticos

El proceso de reclamación (*claim*) de un RO debe pasar por un proceso de enriquecimiento que le incorpore un conjunto de descriptores. Estos descriptores serán, en primer lugar, específicos y se tratará de palabras o conjuntos de palabras reconocidos en el texto y alineados, si es posible, con entidades de ontologías existentes.

En segundo lugar, el enriquecimiento podría recuperar o asignar descriptores temáticos al RO, que no tendrían por qué encontrarse en el contenido textual del RO. Consideramos que estos descriptores tendrían un interés menor que los específicos en cuanto a definir el contenido del RO, por su mayor granularidad, pero que, sin embargo, aportarían valor en el conjunto del proyecto Hércules, particularmente en los Métodos de Análisis (MA)

Es decir, que un RO estuviera categorizado en una o más áreas de conocimiento, que podrían haber sido recuperadas desde la fuente original o asignadas mediante un clasificador supervisado, no aportaría demasiado al propio investigador en la gestión de su trayectoria investigadora; pero sí lo haría en un conjunto amplio de ROs pertenecientes al Sistema Universitario Español como el que aspira a gestionarse mediante el proyecto MA.

Esta clasificación de los ROs habilitaría o facilitaría, al menos, las explotaciones de datos previstas en MA, particularmente las que se refiriesen a las comparaciones entre entidades y al análisis de la producción científica nacional o regional; y también sería una parte relevante del proceso de recomendación en el asistente de configuración de equipos de proyecto.

Además, la clasificación temática obtenida podría alinearse con categorizaciones exigidas en procedimientos administrativos de la Universidad Española, como por ejemplo la Nomenclatura de Ciencia y Tecnología de la UNESCO, ofreciendo a los investigadores un asistente que les podría ahorrar parte del trabajo administrativo.

Por lo tanto, una de las funcionalidades del extractor de tópicos será la identificación de descriptores temáticos relevantes a un RO determinado. Entendemos como descriptor un término que denote un área temática concreta.

De cara a la implementación de la extracción de descriptores temáticos relevantes en un RO determinado consideramos que hay dos cuestiones claves que merecen un análisis pormenorizado:

1. Selección de la taxonomía base de referencia que guiará la extracción de descriptores temáticos.
2. Método para la extracción de los descriptores temáticos.

Por un lado, de cara a garantizar un acceso coherente a los diferentes tipos de ROs es conveniente que la extracción de descriptores temáticos se haga de acuerdo con una taxonomía de descriptores unificada. Además, esta taxonomía común debe contener descriptores que sean fácilmente interpretables por el usuario y que le faciliten la tarea de acceso y navegación de los ROs. Por otro lado, la implementación del proceso de extracción de descriptores temáticos para todos los tipos de RO no resulta una tarea trivial, ya que puede requerir la combinación del uso de APIs de repositorios de ROs y clasificadores supervisados entrenados ad-hoc.

En el capítulo 3.1 presentamos un análisis de las diferentes taxonomías de áreas temáticas que pueden ser apropiadas para su uso en el proceso de extracción de descriptores temáticos. En el capítulo 3.2 ponemos el foco en el problema de la implementación del proceso de extracción analizando distintas estrategias.

3.1 Análisis de taxonomías

En esta sección se presentan y describen brevemente las iniciativas más relevantes identificadas durante el proceso de análisis de taxonomías existentes para la representación de áreas de conocimiento en el dominio de la investigación, haciendo mención igualmente a las posibilidades de obtener un dataset de entrenamiento ya etiquetado con la taxonomía en cuestión.

Además, se hará una breve comparación de las diferentes taxonomías analizadas de manera que se pueda determinar cómo se podrían reutilizar en la tarea de extracción de descriptores temáticos de los diferentes Research Objects (ROs) para estandarizar el espacio de salida de estos y poder integrarlos posteriormente dentro de la ontología unificada del proyecto HERCULES.

3.1.1 Publicaciones o artículos científicos

A continuación, se describen las taxonomías encontradas para Research Objects (ROs) de tipo publicación o artículo científico.

ASJC de Scopus

La base de datos bibliográfica Scopus utiliza ASJC (All Science Journal Classification) como taxonomía para clasificar sus registros.

Dispone de una taxonomía basada en 5 áreas temáticas (Physical Sciences, Health Sciences, Social Sciences, Life Sciences y Multidisciplinary) con una clasificación de primer nivel de 27 entradas[1] y un segundo nivel de 307 entradas[2].

Dataset de entrenamiento

Ya disponemos de un dataset basado en esta taxonomía porque es el que se ha utilizadó en el *challenge* en la tarea de extracción de descriptores temáticos. Se podría utilizar ya para realizar una categorización automática de ROs, siempre que se tuviera acceso a su contenido.

LCC de DOAJ

La iniciativa Directory of Open Access Journals (DOAJ) utiliza tanto palabras clave de texto libre como términos de clasificaciones formales para representar las áreas de conocimiento sobre cualquier registro de su base de datos.

La clasificación LCC tiene 21 categorías de primer nivel[3] y 228 de segundo nivel[4]. Cada categoría LLC tiene categorías muy generales hasta el nivel 2. Por ejemplo, a segundo nivel tenemos 1 clase referente a Mathematics (QA Mathematics), pero en Scopus-ASJC hay 5: Computational Theory and Mathematics, Applied Mathematics, Computational Mathematics, Discrete Mathematics and Combinatorics, Mathematical Physics.

La fuente de datos de DOAJ devuelve etiquetas <dc:subject> (del vocabulario Dublin Core) cuyo contenido tiene como prefijo el esquema formal utilizado para clasificar la publicación en cuestión, excepto en el caso de la Clasificación de la Biblioteca del Congreso o LCC, en la que se indica dicha clasificación mediante un atributo adicional en la etiqueta:

<dc:subject>SOME_NONLCC_FORMAL_CLASSIFICATION:term</dc:subject>

<dc:subject xsi:type="dcterms:LCC">Term</dc:subject>

Dataset de entrenamiento

DOAJ permite descargar un dataset de registros con metadatos sobre las publicaciones que tiene en su base de datos, pero no el texto completo de la publicación. Para descargar el texto de cada artículo habría que implementar un crawler o herramienta de web scraping.

Clasificación de la plataforma de datos de la BNE

Las clasificaciones temáticas de la BNE (Biblioteca Nacional Española) son genéricas y planas, ya que por ejemplo hay:

- 676 categorías en “materia_simple” y 30.480 de ellas no tienen una categoría madre,
- 040 categorías en “subencabezado general” y 4.023 categorías están sin categoría madre,
- 368 categorías en “subencabezamientos” con 353 categorías sin categoría madre.

Dataset de entrenamiento

No disponible, aunque se puede descargar un dataset de registros con metadatos sobre las publicaciones que tiene en su base de datos, pero no el texto completo de la publicación. Para descargar el texto de cada artículo habría que implementar un crawler o herramienta de web scrapping.

Nomenclatura de Ciencia y Tecnología de la UNESCO

Existen dos clasificaciones: la Nomenclatura de Ciencia y Tecnología de la UNESCO[5] y el Tesauro de la UNESCO[6], ambas con 3 niveles.

La primera clasificación es la que se suele utilizar en la universidad española y contiene 2.505 categorías (24 de ellas a primer nivel y 248 a segundo):

1. CAMPOS (24). Se refieren a las áreas más generales y están identificados por los 2 primeros dígitos del código. Cada campo comprende varias disciplinas.
2. DISCIPLINAS (248). Suponen una descripción general de grupos de especialidades en Ciencia y Tecnología. Se identifican con los 4 primeros dígitos del código (coincidiendo los dos primeros con los del campo al que pertenecen).
3. SUBDISCIPLINAS (2505). Son las entradas más específicas de la nomenclatura y representan las actividades que se realizan dentro de una disciplina. Están codificadas con 6 dígitos (coincidiendo los cuatro primeros con la disciplina a la que pertenecen).

Parece ser que hay muchas categorías que están obsoletas, ya que la comunidad científica suele tener dificultad para encuadrar sus áreas de investigación actuales en las disciplinas de segundo nivel y las subdisciplinas de tercer nivel.

Dataset de entrenamiento

Tal vez podríamos conseguir un dataset anotado de algún repositorio de la universidad española, pero sería difícil que llegase con una cantidad suficiente de ejemplos para todas las categorías de nivel 3.

Clasificaciones de la Web of Science

Web of Science (WoS) es una plataforma on-line de Clarivate Analytics que contiene bases de datos de información bibliográfica[7] y recursos de análisis de la información que permiten evaluar y analizar el rendimiento de la investigación. Su finalidad no es proporcionar el texto completo de los documentos, sino proporcionar herramientas de análisis que permitan valorar su calidad científica.

Dispone de 16 esquemas de clasificación[8] en áreas de investigación, de los cuales 12 se basan en la asignación de datos a sistemas externos de clasificación. Los esquemas externos se desarrollan en asociación con los organismos de evaluación de la investigación en una determinada región y suelen basarse en clasificaciones de revistas o en el mapeo de categorías ya existentes en WoS. Además, existen otros 4 esquemas de clasificación exclusivos de WoS:

- Citation Topics[9] (WoS-CT). Posee 3 niveles (macro, meso y micro) con diferente número de categorías: 10 (macro), 326 (meso) y 2.444 (micro). Con WoS-CT los documentos se etiquetan manualmente en categorías macro y meso según su contenido. Los micro se asignan algorítmicamente con el descriptor más significativo para el documento obtenido con un algoritmo (desarrollado por CWTS Leiden[10]) que se basa en procesar las citas de los documentos y no su contenido.
- WoS Research Areas[11] (WoS-RA). Contiene un primer nivel de 252 áreas temáticas para ciencias, ciencias sociales y artes y humanidades. Áreas muy amplias como física y ciencia de los materiales están representadas por subáreas más pequeñas de segundo nivel, pero no hemos encontrado cuántas categorías hay en dicho nivel. WoS-RA se utiliza en Science Citation Index Expanded Journals, Social Sciences Citation Index Journals y Arts & Humanities Citation Index Journals.
- Essential Science Indicators[12] (WoS-ESI). Contiene un primer nivel de 22 áreas temáticas. Se utiliza para clasificar 11.728 revistas de Web of Science Core Collection (solamente Science Citation Index Expanded y Social Sciences Citation Index). Con WoS-ESI cada revista solamente se puede asignar a una de las 22 áreas de investigación. Revistas como Science y Nature se clasifican excepcionalmente como multidisciplinares, ya que publican investigaciones en muchos campos diferentes. Los artículos publicados en estas revistas multidisciplinares se asignan a un área temática u otra en función de la representación de las revistas citadas.
- Institutional Profiles Research Areas[13] (WoS-IPRA). Contiene un primer nivel de 6 áreas temáticas y un segundo nivel con 267 subáreas. WoS-IPRA es utilizada por Clarivate Analytics para elaborar perfiles institucionales de las principales universidades e instituciones de investigación del mundo.

Dataset de entrenamiento

Hemos comprobado que en el contexto del proyecto HERCULES tenemos acceso a alguna de las APIs de Web of Science[14]. Haciendo llamadas a la API se podría generar un dataset de registros con metadatos sobre las publicaciones, pero no el texto completo de la publicación. Sin embargo, se podría utilizar el abstract para la extracción de descriptores temáticos.

3.1.2 Taxonomías para protocolos

Dos de los repositorios de protocolos experimentales más utilizados por la comunidad científica y que son mencionados en el pliego de HERCULES son Bio-protocol.org y Protocol-exchange.

Bio-protocol.org[15] usa descriptores específicos de texto libre y dispone de una categorización de hasta 3 niveles.

Por su parte, Protocol-exchange[16] sí dispone de una categorización de un único nivel con unos 107 términos[17]. Cada protocolo está asignado al menos a uno de sus “subject terms”. Si algún autor no encuentra una categoría adecuada a su protocolo en el formulario de registro correspondiente, puede ponerse en contacto con protocol.exchange@nature.com para estudiar la posibilidad de ampliar la lista existente.

Existen otras iniciativas como Protocols.io[18] que ofrece una API[19] para consultar su repositorio de protocolos, pero aparentemente no dispone de ninguna clasificación para los mismos ni almacena ningún metadato sobre áreas de conocimiento.

Dataset de entrenamiento

Se podría implementar un crawler o herramienta de web scraping para generar un dataset a partir de la web de Protocol-exchange (con algo más de variedad temática) o desde la web de bio-protocols (bio-ciencias).

3.1.3 Taxonomías para proyectos de código

Dos de los repositorios de código más utilizados por la comunidad científica y que son mencionados en el pliego de HERCULES son GitHub[20] y Bitbucket[21].

GitHub solamente usa descriptores específicos de texto libre que denomina “topics”[22] para clasificar y realizar búsquedas de repositorios de código. Este metadato y otros pueden ser recuperados a través de su API[23].

Bitbucket parece que no utiliza ninguna clasificación. En la descripción de su API[24] tampoco aparece ningún metadato sobre un repositorio que se pueda interpretar como elemento de clasificación.

Sin embargo, en SourceForge[25] sí parece que se utiliza una clasificación de varios niveles para el catálogo de aplicaciones que ofrece, dado que en su directorio permite realizar búsquedas por categorías[26]. No hemos encontrado documentación sobre el número de niveles ni sobre el número de categorías en cada nivel.

Dataset de entrenamiento

Se podría implementar un crawler o herramienta de web scraping para generar un dataset basado en el catálogo de SourceForge.

3.1.4 Comparación de taxonomías

En la siguiente tabla se comparan las diferentes taxonomías analizadas teniendo en cuenta sus características principales en cuanto a la calidad de sus términos y la posibilidad de obtener un dataset de entrenamiento, el número de niveles que posee, el número de categorías por cada nivel, fechas asociadas (tiempo de vida) y uso que se está haciendo de las mismas.

Taxonomía	**Características**	**# Niveles**	**# Categorías**	**Fechas**	**Uso en apps**
ASJC - Scopus	· Clasificación significativa de categorías. · Dataset disponible para publicaciones y protocolos.	2	Nivel 1: 27 Nivel 2: 307	2020	Scopus y otras aplicaciones de Elsevier
LCC - DOAJ	· Categorías muy generales hasta el nivel 2. · Dataset de publicaciones a elaborar mediante web scrapping.	5	Nivel 1: 21 Nivel 2: 228	2021	DOAJ
Clasificación de la plataforma de datos de la BNE	· Categorías muy genéricas y planas. · Dataset de publicaciones a elaborar y quizá de baja calidad.	¿3?	Difícil de saber en concreto.	¿?	datos.bne.es
Nomenclatura de Ciencia y Tecnología de la UNESCO	· Categorías obsoletas. · Tradición en la universidad española. · Dataset de publicaciones anotado quizá de algún repositorio de la universidad española.	3	Nivel 1: 24 Nivel 2: 248	1973-1988-2021	CVN-FECYT
WoS-CT	· Categorización completa y actualizada. · Dataset de publicaciones con abstract y metadatos de categorías si se solicita acceso a InCites API.	3	Nivel 1: 10 Nivel 2: 326	2019-2021	WoS y otras de Clarivate Analytics
WoS-RA	· Categorización completa y actualizada de un nivel. · Dataset de publicaciones con abstract y metadatos de categorías.	1	Nivel 1: 252	?-2021	WoS y otras de Clarivate Analytics
WoS-ESI	· Categorización actualizada, pero con pocos términos. · Dataset de publicaciones con abstract y metadatos de categorías.	1	Nivel 1: 22	2010-2021	WoS y otras de Clarivate Analytics
WoS-IPRA	· Categorización completa y actualizada. · Dataset de publicaciones con abstract y metadatos de categorías.	2	Nivel 1: 6 Nivel 2: 267	2009-2021	WoS y otras de Clarivate Analytics
		1	Nivel 1: 107		

Clasificación Protocol-exchange	· Categorización abierta de un nivel, muy centrada en ciertos dominios científicos. · Dataset de protocolos a elaborar mediante web scraping.			2019-2021	Protocol Exchange (repositorio abierto de Nature)
Clasificación SourceForge	· Categorización muy extensa de la que se desconoce su estructura. · Dataset de protocolos a elaborar mediante web scraping.	?	?	1999-2021	SourceForge

3.1.5 Conclusiones sobre la comparación de taxonomías

De todas las taxonomías analizadas y comparadas, las más apropiadas serían las que tienen una buena cobertura, es decir, un número significativo no muy elevado de categorías en al menos dos niveles, para todas las áreas de investigación y para todos los tipos de ROs descritos en el proyecto HERCULES.

Tal y como se plantea en el proyecto, lo habitual es que los distintos tipos de RO estén relacionados entre sí formando un RO compuesto, por lo que también sería apropiado que se utilizara una misma taxonomía para todos ellos.

Por número no muy elevado entendemos aquél que ronde los 200-300 términos, de manera que sea abordable su utilización con un clasificador supervisado, uno de los enfoques que se plantean en el proceso de extracción de descriptores temáticos de ROs.

Y por significativo entendemos aquél que cubra con cierta concreción un número suficiente de categorías de segundo nivel para cada una de las áreas de investigación de primer nivel.

Dicho lo anterior, las candidatas más apropiadas serían la ASJC de Scopus, la LCC de DOAJ, la Nomenclatura de Ciencia y Tecnología de la UNESCO y la clasificación IPRA de WoS.

La ASJC de Scopus tiene la ventaja de que ha sido utilizada en el Challenge del proyecto HERCULES y ya disponemos, por tanto, de un dataset anotado para su utilización. Sin embargo, se generaron dudas respecto a su cobertura para todas las áreas de investigación, dando lugar al análisis que se presenta en este documento.

La Nomenclatura de Ciencia y Tecnología de la UNESCO parece que genera problemas respecto a la actualidad de sus términos, dado que la comunidad científica así lo ha expresado de manera informal en diferentes foros.

Por último, si además tenemos en cuenta el volumen de ROs existentes en las fuentes de datos y la cantidad de bases de datos de diferentes dominios científicos, la fuente de datos de WoS es mayor que el de DOAJ.

Todo ello permite concluir que la clasificación IPRA de WoS podría ser la opción más interesante para estandarizar el espacio de salida de descriptores temáticos de las áreas de conocimiento de los ROs y sería la que se integraría en la ROH (Red de Ontologías de HERCULES).

3.2 Enfoques para extracción de descriptores temáticos de ROs

La tarea de extracción de descriptores temáticos propuesta consiste en, dada una taxonomía de áreas temáticas y un RO, identificar las categorías temáticas que son relevantes al RO. Para abordar esta tarea proponemos tres enfoques:

- Enfoque basado en API de repositorio: Los descriptores temáticos se obtienen directamente del API del repositorio donde está alojado el RO.
- Enfoque basado en scraping: Los descriptores temáticos se obtienen del HTML del repositorio correspondiente al RO.
- Enfoque basado en clasificadores supervisados: Los descriptores temáticos se obtienen mediante un clasificador supervisado previamente entrenado con ejemplos.

Describimos en los siguientes capítulos los tres enfoques propuestos, y la propuesta final para abordar la extracción de descriptores temáticas para todos los tipos de ROs.

3.2.1 Extracción de descriptores temáticos a partir de APIs

A continuación, se describe someramente el enfoque que se utilizaría para la extracción de descriptores temáticos al reclamar una publicación que estuviera almacenada en este tipo de fuentes de datos.

Dado que en la sección 1.5 del presente documento se ha concluido que la clasificación WoS-IPRA es la más apropiada y sería la que se integraría en la ROH (Red de Ontologías de HERCULES), la fuente principal para la reclamación de ROs será la API de la Web of Science.

APIs de la Web of Science

Hemos realizado consultas a la Web of Science API Expanded y se pueden obtener metadatos sobre las publicaciones, entre los que se encuentran las áreas de conocimiento, pero no está claro a qué esquema pertenecen los resultados que devuelve:

<category_info>

<headings count="1">

<heading>Science & Technology</heading>

</headings>

<subheadings count="2">

<subheading>Physical Sciences</subheading>

<subheading>Technology</subheading>

</subheadings>

<subjects count="6">

<subject ascatype="traditional" code="EA">Chemistry, Analytical</subject>

<subject ascatype="traditional" code="IQ">Engineering, Electrical & Electronic</subject>

<subject ascatype="traditional" code="OA">Instruments & Instrumentation</subject>

<subject ascatype="extended">Chemistry</subject>

<subject ascatype="extended">Engineering</subject>

<subject ascatype="extended">Instruments & Instrumentation</subject>

</subjects>

</category_info>

Aunque no hemos encontrado confirmación en la documentación oficial de las APIs de WoS, parece que los “subject” con “ascatype=traditional” se refieren al esquema WoS-IPRA y los que tienen “ascatype=extended” usan indistintamente los esquemas WoS-ESI y WoS-RA.

Tras realizar una búsqueda en la web y encontrar lo que parece ser una versión más reciente de WoS-RA[27] en la que aparecen los valores contenidos en los “subject” con “ascatype=traditional”, pero con 2 categorías más (254), y otra clasificación en áreas de investigación parecidas a las de IPRA[28], parece que los valores de los elementos “subheading” se corresponden con las 5 “research areas” de la columna derecha de la siguiente tabla:

IPRA	**Research Areas (Categories / Classification)**
Arts & Humanities (28)	Arts & Humanities (14 sub)
Clinical, Pre-Clinical & Health (47)	Life Sciences & Biomedicine (76 sub)
Engineering & Technology (51)	Physical Sciences (17 sub)
Life Sciences (53)	Social Sciences (25 sub)
Physical Sciences (38)	Technology (21 sub)
Social Sciences (49)	

En cualquier caso, para evitar tener que realizar una alineación de las diferentes clasificaciones y dado que se ha concluido en la sección 1.5 del presente documento que la clasificación WoS-IPRA es la más adecuada, se utilizaría el contenido de los elementos “subject” con “ascatype=traditional” como descriptores temáticos del RO, pudiendo así representar como URIs del grafo de conocimiento de HERCULES basado en la ROH.

Otras fuentes de datos vía API

Si se reclama un RO almacenado en otra fuente de datos, se utilizarán los metadatos que ofrezca dicha fuente de datos para RO en la fuente de datos principal de WoS y se utilizarían los descriptores temáticos extraídos de ella.

De esta manera, se utilizará un único espacio de salida estandarizado de áreas de conocimiento y se podrían representar sus descriptores temáticos como una URI del grafo de conocimiento de HERCULES basado en la ROH.

Si el RO no se encuentra en la WoS, se aplicará la propuesta descrita en el apartado 2.4 del presente documento. Se descarta reutilizar los metadatos relativos a descriptores temáticos ofrecidos por otra fuente de datos diferente a WoS para no incurrir en el sobrecoste innecesario de realizar una alineación de cada descriptor de dicha fuente (texto libre o cualquier otra clasificación) con un término de la clasificación WoS-IPRA.

3.2.2 Extracción de descriptores temáticos vía web scraping

Dos de las fuentes identificadas para protocolos, Protocol Exchange y bio-protocol, carecen de un API del que recuperar la información del RO reclamado por el usuario. De partida, esto obliga a realizar un proceso de scraping para obtener la información del RO en estas fuentes, lo que se aprovechará para obtener los descriptores temáticos de las dos fuentes indicadas, siempre que mantengan las condiciones y permisos de reutilización actuales. En Protocol Exchange las condiciones dependen de cada protocolo, mientras que todo el contenido de bio-protocols es Open Access.

En el desarrollo se implementará un API que permitirá definir y mantener el scraping de estos (y potencialmente otros) sitios web, asociando los elementos HTML con los metadatos a recuperar del RO, por ejemplo: abstract, background, keywords, procedure, categories, etc.

En particular, las categorías recuperadas estarían alineadas con la taxonomía propuesta para los descriptores temáticos, bien porque ya tenían una correspondencia (por ejemplo, “Analytical chemistry” de Protocol Exchange se alinearía con “CHEMISTRY, ANALYTICAL” de IPRA) o bien porque se ha extendido la taxonomía para alojar categorías propias de los protocolos (por ejemplo, “Systems Biology” de bio-protocols, que no se alinearía con ninguna categoría de IPRA).

3.2.3 Extracción de descriptores temáticos mediante un clasificador supervisado

La identificación de descriptores temáticos relevantes a un RO puede abordarse mediante un enfoque basado en clasificadores supervisados. Los clasificadores serían de tipo multi-etiqueta y se entrenarían a partir de datasets de entrenamiento que incluirían ejemplos de ROs anotados con los correspondientes descriptores temáticos. La representación vectorial de cada ROs requerida por el algoritmo de aprendizaje se hará de acuerdo a un determinado paradigma de representación textual. El clasificador entrenado de esa forma sería capaz de identificar los descriptores temáticos de ROs no presentes en el entrenamiento, siempre y cuando su estructura textual no variase significativamente respecto a la de los ejemplos utilizados en el entrenamiento.

Uno de los aspectos clave a la hora de valorar la viabilidad de un enfoque supervisado es la disponibilidad de datasets de entrenamiento. En este caso necesitaríamos conjuntos de ROs con sus correspondientes descriptores temáticos, y disponer de un número suficiente de ejemplos de ROs por cada descriptor temático.

Según el análisis presentado el capítulo 1, podríamos generar datasets para los siguientes tipos de ROs y taxonomías:

- Posibles datasets para **Papers** (uno de los siguientes):
  - Scopus-ASJC.
  - DOAJ-LCC.
  - WoS-RA.
  - WoS-IPRA.
- Posibles datasets para **Protocolos** (uno de los siguientes):
  - Protocol-exchange.
  - Bio-protocol.
- **Proyectos código**:
  - SourceForge

Para generar clasificadores que siguieran una única taxonomía tendríamos que, de alguna manera, armonizar las taxonomías asociadas a los diferentes datasets en una única taxonomía unificada. Por otro lado, también habría que analizar – mediante experimentación – qué estrategia proporciona los mejores resultados; entrenar un clasificador por cada tipo de RO, o entrenar un clasificador único a partir de la unión de los datasets de los diferentes tipos de RO. Desde un punto de vista de ingeniería del software sería más apropiado un único clasificador, pero debido a las diferencias entre las características de los textos de los distintos tipos de RO, esta estrategia podría no ser la que mejores resultados diera.

Otro aspecto que será determinado mediante distintos experimentos será la granularidad de categorías que los clasificadores supervisados puedan ofrecer. Cuantas más categorías se establecen en el entrenamiento más difícil resulta el proceso de aprendizaje, por lo que habrá que buscar un compromiso entre este número y la tasa de acierto deseable.

En principio, el enfoque basado en clasificadores supervisados se utilizará de manera complementaria a los enfoques basados en API y scraping. Computacionalmente es el más costoso y también el más limitado en cuanto a tasa de acierto.

4. Conclusiones de primeros experimentos, participación en el Challenge, análisis de extracción de descriptores temáticos, y próximos trabajos

De los resultados obtenidos en los primeros experimentos y en el *Challenge* podemos concluir que la diferenciación de dos tipos de tópicos (descriptores temáticos y específicos) es viable y eficaz a la hora de ofrecer caracterizaciones de los ROs más fácilmente interpretables por el usuario.

Hemos comprobado que las estrategias basadas en aprendizaje automático supervisado son viables, y que ofrecen buenos resultados para los dominios de publicaciones científicas y protocolos experimentales. El trabajo realizado también supone un punto de partida sólido para extender la estrategia supervisada al dominio de proyectos GitHub.

De todas formas, de acuerdo con los primeros resultados, es esperable que no se va a alcanzar una precisión del 100% en la detección de descriptores, por lo que se vislumbra que este proceso deberá ser asistido manualmente y combinado con enfoques basados en recuperar los descriptores directamente de los repositorios. Las correcciones realizadas por los usuarios se recogerán para enriquecer los datasets de entrenamiento, y así poder reentrenar clasificadores de mayor precisión y cobertura.

Se vislumbra que la estrategia adecuada para abordar la tarea de extracción debería ser una estrategia híbrida que combine los distintos tipos de enfoque descritos en el punto 3.3. Se dará prioridad al enfoque basado en API, por ser el más robusto y el que ofrece los resultados más precisos. El enfoque basado en scraping también ofrece resultados precisos, pero resulta sensible a cambios en la maquetación HTML del repositorio de donde se recuperan los descriptores. Por esa razón, será el segundo enfoque prioritario. En tercer lugar, se hará uso del enfoque basado en clasificadores automáticos. Las distintas combinaciones de los enfoques vendrán determinadas por el tipo de RO a tratar, ya que la viabilidad, o necesidad, de cada uno de ellos viene condicionada por el tipo de RO. Mostramos, a continuación, un resumen de la estrategia propuesta para extraer descriptores temáticos de todos los tipos de ROs:

- **Papers**: Se tomarán los descriptores temáticos (IPRA) de WoS en caso de que el paper esté alojado en WoS. En caso contrario, se optará por procesar el paper con el clasificador supervisado.
- **Protocolos**: Se tomarán los descriptores temáticos de los repositorios bio-protocol o protocol-exchange mediante el enfoque basado en scraping, en caso de que el protocolo esté disponible en uno de estos repositorios. En caso contrario, se hará uso del clasificador supervisado.
- **Referencias**. Se tomarán los descriptores temáticos del RO referenciado.
- **Anotaciones**. Se tomarán los descriptores temáticos del RO anotado.
- **Proyectos de código**: Se tomarán los descriptores temáticos del clasificador supervisado.

Listamos y describimos, a continuación, los aspectos que, tras los primeros experimentos y análisis, todavía permanecen abiertos y que se pretenden abordar durante el proyecto:

Evaluación de modelos de lenguaje

En los experimentos realizados en el marco del *Challenge* se entrenó y evaluó un clasificador de descriptores temáticos basado en BERT. Por limitaciones de tiempo el clasificador se entrenó sobre un dataset que únicamente abarcaba un repertorio limitado de categorías temáticas. Los resultados obtenidos superaban a los de los clasificadores basados en el paradigma de bolsa-de-palabras.

Con objeto de verificar si los clasificadores basados en BERT (o modelos de lenguaje similares como roBERTa) ofrecen los mejores resultados sobre datasets más amplios y de distintos tipos de RO, se van a realizar una serie de experimentos. También se cuantificarán los consumos de CPU/GPU y memoria, para sopesarlos con los resultados de las evaluaciones.

Evaluación de arquitecturas seq2seq para extracción de descriptores específicos

En los experimentos previos se entrenó y evaluó un extractor de descriptores específicos basado en una arquitectura neuronal seq2seq para artículos científicos. Los resultados obtenidos no eran mejores que los ofrecidos por los sistemas alternativos evaluados, al menos para el dominio de los artículos científicos.

Se realizarán experimentos adicionales sobre otros tipos de RO, para ver si en esos dominios este enfoque puede ser una alternativa robusta. Al igual que en el caso de modelos basados en BERT, se cuantificarán los consumos de CPU/GPU y memoria, para sopesarlos con los resultados de las evaluaciones.

Generación de datasets para otros tipos de ROs e idiomas

En el *Challenge* se puso el foco en tres tipos de ROs (artículos científicos, protocolos experimentales, y proyectos GitHub) y se experimentaron enfoques supervisados para dos tipos de RO (artículos científicos y protocolos experimentales) para los cuales se crearon datasets de entrenamiento (tanto para extracción de palabras temáticas como específicas).

Se analizará en qué medida se pueden reutilizar los datasets (o incluso los mismos clasificadores) creados para entrenar clasificadores orientados a tres tipos adicionales de ROs: proyectos GitHub, referencias bibliográficas, y anotaciones. En caso de ser necesario, se generarán nuevos datasets para esos tres tipos de ROs. Para ello se analizarán diferentes repositorios, y estrategias, tanto automáticas como semiautomáticas que requieran supervisión manual.

Por otro lado, se estudiarán estrategias para poder entrenar clasificadores bilingües, de forma que además de textos en inglés, también sean capaces de procesar textos en castellano. Para abordar esa tarea se han identificado dos estrategias:

- Traducción de los datasets en inglés: Mediante esta estrategia los textos del dataset en inglés se traducen al castellano mediante un sistema de traducción automática. Para ello se utilizará una variante, adaptada al dominio abierto, del sistema presentado por Elhuyar en la competición Biomedical Translation Task del WMT20 donde consiguió la segunda mejor puntuación.
- Uso de modelos multilingües de lenguaje: Los modelos de lenguaje multilingües pre-entrenados permiten entrenar un clasificador a partir de un dataset en un idioma que luego puede aplicarse sobre otros idiomas (incluidos en el modelo multilingüe pre-entrenado).

Ontologías o taxonomías de referencia para la extracción y enlazado de descriptores

Los clasificadores supervisados entrenados en el contexto del *Challenge* aprenden los descriptores presentes en los datasets. En el caso de los datasets generados para los ROs de artículos científicos y protocolos experimentales las palabras temáticas seguían las taxonomías ASJC y la propia de Bio-protocol, respectivamente. Los descriptores específicos, en cambio, no se correspondían con ninguna taxonomía.

Para los descriptores temáticos necesitaremos datasets de entrenamiento que cuente con un volumen suficiente de ROs anotados manualmente contra una taxonomía reconocida, tal y como se ha explicado en el capítulo 3.3. Esta taxonomía se publicará como datos enlazados (Linked Data) en formato SKOS como uno de los resultados del proyecto Hércules.

En cuanto a los descriptores específicos, una vez extraídos las enlazaremos con espacios de datos enlazados genéricos y temáticos, como:

- DBpedia, por ejemplo Word processor.
- MESH Medical Subject Heading. TopicalDescriptor, Term y Concept. Por ejemplo, Anti-Bacterial Agents o Antibiotics.
- CSO Computer Science Ontology, por ejemplo, word processing.

Estableceremos una relación, que el administrador del sistema podrá modificar más adelante, entre los ítems de la taxonomía (los descriptores temáticos que se identificarán en cada RO) y los espacios de datos temáticos. Consideramos que, en principio, los descriptores específicos de todos los tipos de ROs serán enlazables a DBpedia, por su carácter general, pero el enlazado con espacios temáticos más concreto se decidirá en función de los descriptores temáticos identificados. Así, si el RO se clasifica en categorías como “Drug Discovery”, “Pharmacology” o “General Medicine”, podemos establecer que sus descriptores específicos se enlazarán con MESH.

Por el contrario, si se clasifica en “Computer Science Applications”, “Human-Computer Interaction” y “Artificial Intelligence”, podemos decidir que se enlacen con CSO.

El proceso de enlazado no será sólo mediante string-matching, sino que desambiguará el descriptor específico comparando los contextos del RO (el resto de descriptores) con el del término a enlazar, usando:

- Términos relacionados.
  - CSO: superTopicOf, contributesTo, relatedEquivalent, sameAs.
  - MESH: concept, preferredConceot, allowableQualifier, preferredTerm, broaderDescriptor.
- Términos y contenido.
  - DBPedia: description (presencia de descriptores específicos), subject, sameAs (en dominios ajenos a las versiones de idiomas de DBpedia), isPrimaryTopicOf, wikiPageRedirects, wikiPageWikiLink, genre, primaryTopic.

Además, el proceso de enlazado tendrá en cuenta las acciones que otros usuarios hayan podido hacer en relación con la aceptación o no de propuestas anteriores.

Del mismo modo que en el caso de la propuesta de descriptores, también para este caso de uso se convierte el interfaz en un componente muy importante, ya que debe permitir al usuario comprobar y verificar con rapidez y sencillez si el enlazado propuesto es o no correcto.

Combinación de enfoques para la extracción de descriptores

La estrategia que se propone para abordar la tarea de extracción de descriptores será una estrategia híbrida que combinará los distintos tipos de enfoques. Se dará prioridad al enfoque basado en API, por ser el más robusto y el que ofrece los resultados más precisos. El enfoque basado en scraping también ofrece resultados precisos, pero resulta sensible a cambios en la maquetación HTML del repositorio de donde se recuperan los descriptores. Por esa razón, será el segundo enfoque prioritario. En tercer lugar, se hará uso del enfoque basado en clasificadores automáticos.

Se deberán implementar los diferentes enfoques, así como su combinación. Las distintas combinaciones de estos enfoques vendrán determinadas por el tipo de RO a tratar, ya que la viabilidad, o necesidad, de cada uno de ellos viene condicionada por el tipo de RO. Mostramos, a continuación, un resumen de la estrategia propuesta para extraer descriptores temáticos de todos los tipos de ROs:

- **Papers**: Se tomarán los descriptores temáticos (IPRA) de WoS en caso de que el paper esté alojado en WoS. En caso contrario, se optará por procesar el paper con el clasificador supervisado.
- **Protocolos**: Se tomarán los descriptores temáticos de los repositorios bio-protocol o protocol-exchange mediante el enfoque basado en scraping, en caso de que el protocolo esté disponible en uno de estos repositorios. En caso contrario, se hará uso del clasificador supervisado.
- **Referencias**. Se tomarán los descriptores temáticos del RO referenciado.
- **Anotaciones**. Se tomarán los descriptores temáticos del RO anotado.
- **Proyectos de código**: Se tomarán los descriptores temáticos del clasificador supervisado.

Flujo e interfaces del enriquecimiento

Introducción

Este documento describe el flujo e interfaces del proceso de enriquecimiento de ED, que añade áreas temáticas (descriptores temáticos) y tópicos específicos (descriptores específicos) a los ROs, sean estos recuperados desde fuentes externas o introducidos a mano por el investigador.

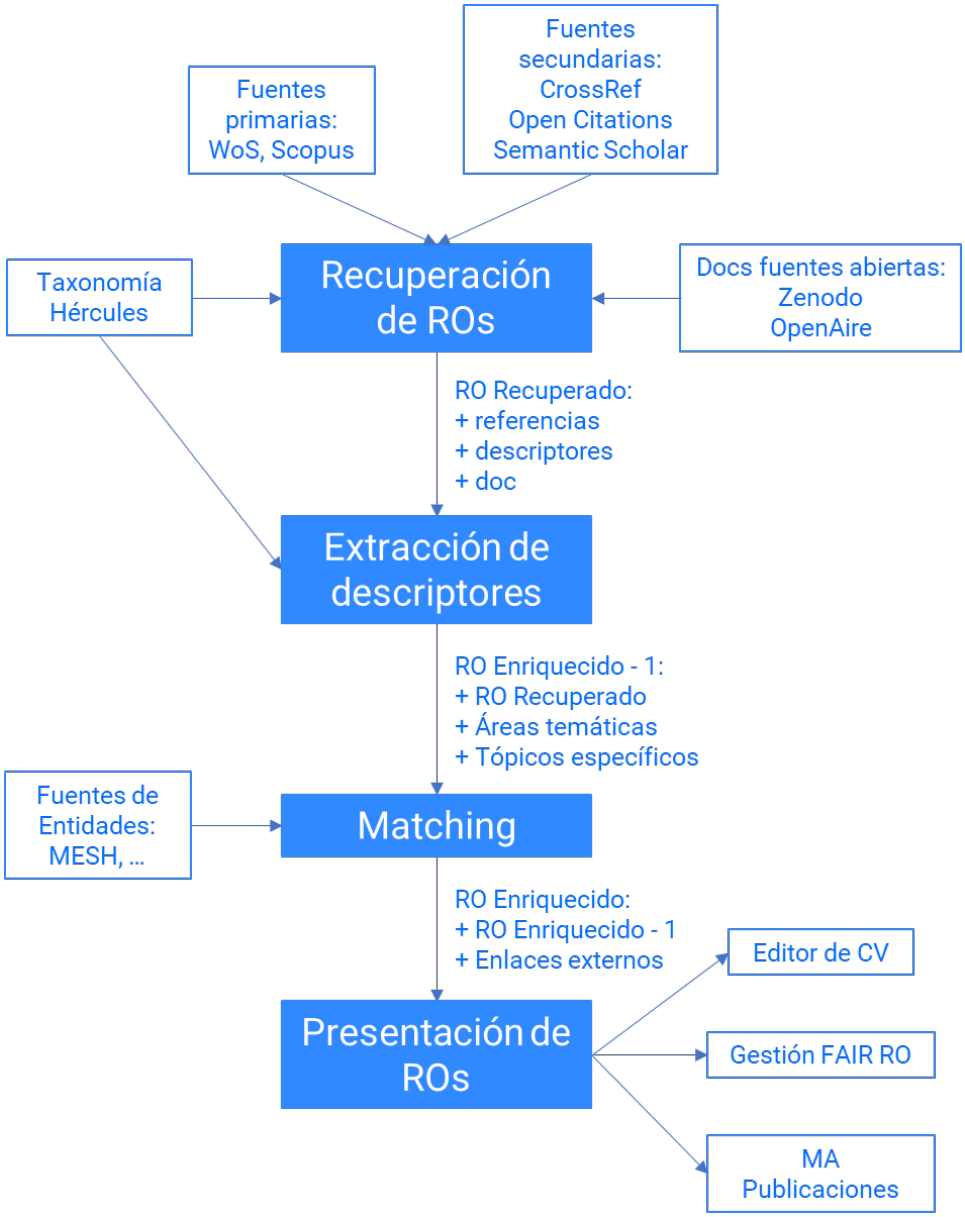
Los objetivos del proceso son:

- Enriquecer la información de los ítems recuperados desde fuentes externas de investigación científica
- Utilizar los descriptores recuperados para potenciar la experiencia de búsqueda, recuperación y consulta de la información
- Utilizar los descriptores recuperados para explicar e ilustrar la similitud

El proceso de enriquecimiento tiene los siguientes pasos:

- Extracción de descriptores temáticos, alineados con la taxonomía unificada
- Extracción de descriptores específicos
- Matching de los descriptores específicos con entidades definidas en fuentes externas.
- Presentación de descriptores al usuario para su gestión, en 2 interfaces distintos:
  - Edición CV para los ROs correspondientes a la norma CVN (contenidos en el apartado de publicaciones científicas)
  - Gestión FAIR RO (sprint 3), para todos los tipos de ROs.

El flujo del proceso corresponde al siguiente diagrama:



Extracción de descriptores

El proceso de extracción de descriptores trabaja sobre los ROs obtenidos desde fuentes externas, que cuentan con sus metadatos, referencias y citas, descriptores de las fuentes externas (palabras clave y categorías) y el enlace al documento, en el caso de que lo tengan.

Como salida obtenemos un RO enriquecido con áreas temáticas (descriptores temáticos) y tópicos específicos (descriptores específicos), generados con los algoritmos de enriquecimiento.

Taxonomías para descriptores temáticos

Los descriptores temáticos extraídos para los ROs se corresponderán con los ítems de la taxonomía unificada para Hércules

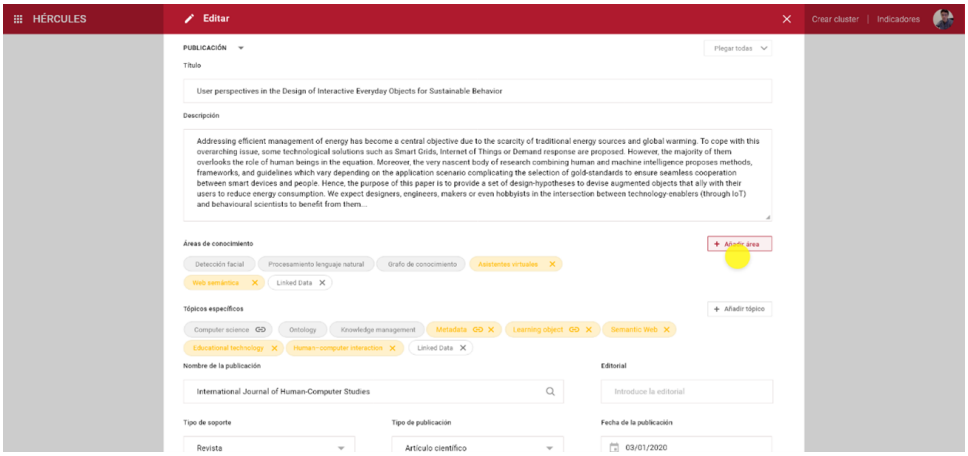
- Las fuentes consultadas en la elaboración de la taxonomía están descritas en el Análisis de taxonomías (Confluence).
- El proceso de unificación y su resultado se puede consultar en Taxonomía unificada de descriptores temáticos para Hércules (Confluence).
- La taxonomía unificada está alineada con otras taxonomías:
  - Para los papers y ROs genéricos: ASJC-Scopus + arXiv + MESH-Pubmed + WoS-JCR.
  - Para usarse en procesos de exportación y/o carga (alineación en curso): UNESCO, CVN-FECYT y UMU.
  - Para bio-protocolos: Bio-protocol.org
  - Para proyectos código: Sourceforge.

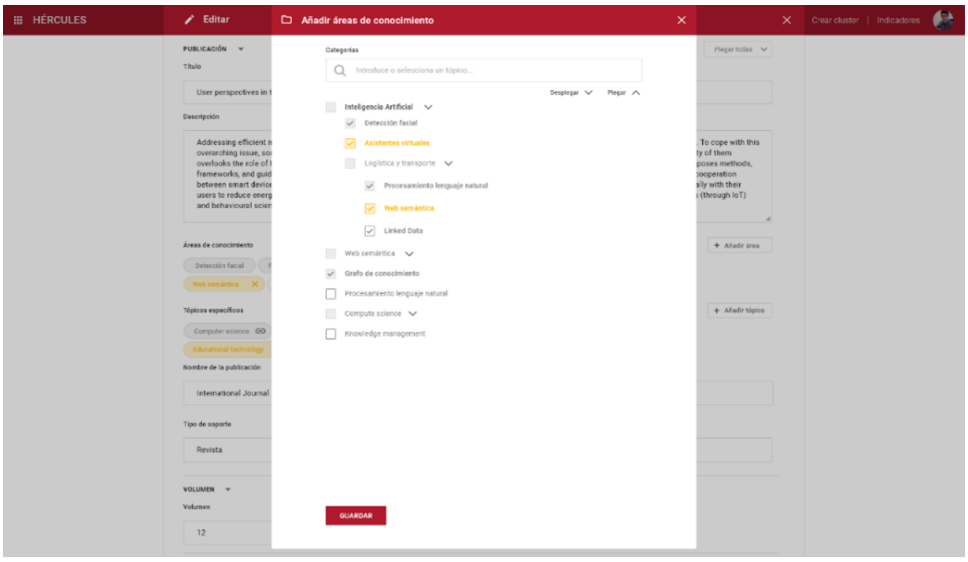
Extracción en ROs de fuentes externas

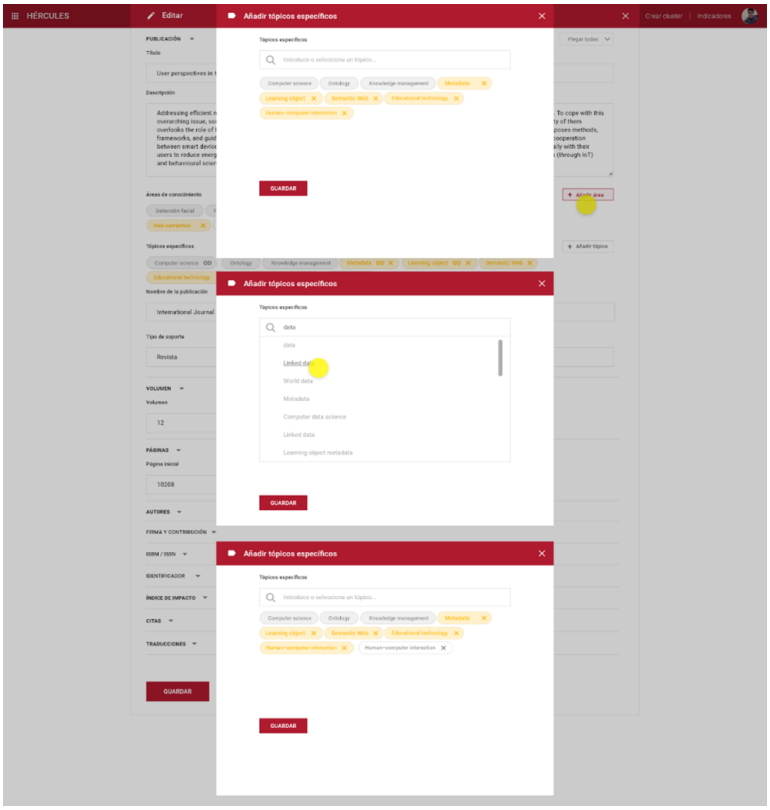
El proceso de extracción de descriptores comienza tras recuperar la información de ROs desde fuentes externas de información. Se describe a continuación el caso de recuperación de publicaciones científicas, que será similar de otros ROs, como los de código, bio-protocolos, datasets, etc.

- Fuentes primarias (WoS y Scopus) + documento open (Zenodo, OpenAire)
  - Obtención de descriptores temáticos:
    - Se recuperan de las fuentes primarias.
    - Los descriptores recuperados se mapean con la taxonomía unificada.
    - Extracción de descriptores temáticos adicionales si aportan valor, tras validación con los usuarios.
  - Obtención de descriptores específicos:
    - Se recuperan de las fuentes primarias.
    - Extracción descriptores específicos adicionales. Se espera una aportación de valor más clara que en el caso de los temáticos adicionales
  - En el siguiente sprint se podría añadir Open Aire como fuente primaria, ya que está indexando el contenido del proyecto Recolecta de FECYT.
- Fuentes secundarias (CrossRef, Open Citations, Semantic Scholar) + documento open (Zenodo, OpenAire)
  - Ítems obtenidos como referencias y citas (corresponden al tipo de RO Referencia).
  - Entre la información recuperada de estos ítems no hay descriptores temáticos, sólo en algunos casos hay específicos.
  - Todos los ítems de fuentes secundarias se envían al proceso de extracción para obtener descriptores temáticos y específicos.

Al usuario se le mostrarán los tópicos recuperados en gris, sin opción a su eliminación, y los descriptores adicionales, que podrá descartar, en naranja.





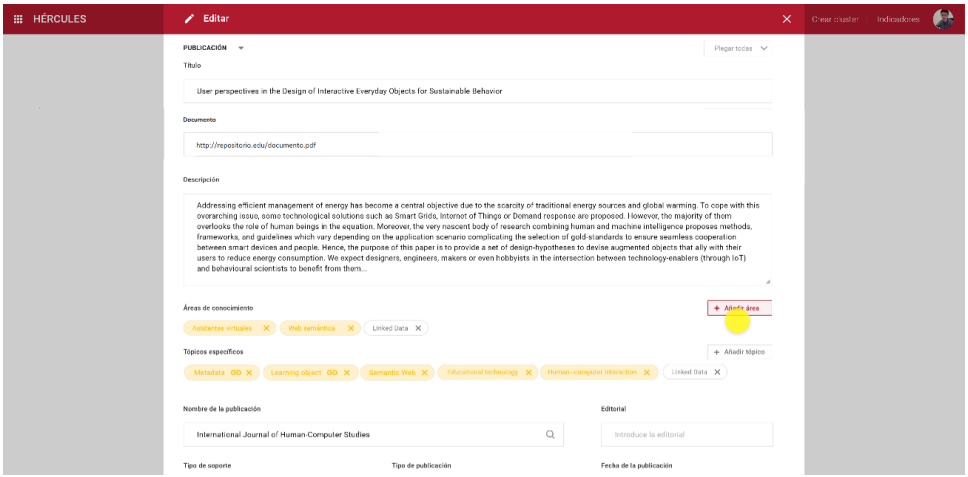


También podrá añadir otros descriptores adicionales, tanto áreas temáticas como tópicos específicos, según se muestra en los siguientes ejemplos.

ROs introducidos por el usuario

Este sería un caso un poco diferente al descrito en el flujo presentado al principio del documento, en el que el primer paso, de Recuperación de ROs, no existe y la información del RO introducido por el usuario se envía directamente al paso de Extracción de descriptores.

El proceso se invocará automáticamente en la edición del RO, cuando el usuario haya introducido información suficiente, que será: título, documento (opcional) y resumen-abstract. Como resultado, el usuario obtendrá unos descriptores temáticos y específicos (en naranja), que podrá rechazar, y también podrá añadir otros mediante las acciones de añadir área y añadir tópico.



ROs procesados masivamente

El proceso de carga inicial más habitual incluirá habitualmente la Recuperación de ROs. En algún caso podría suceder que esta carga contase ya con información suficiente de los ROs y sólo hiciera falta su enriquecimiento desde el paso de “Extracción de descriptores”. Este proceso también sería posible.

Otro supuesto de procesamiento masivo sería el que se produciría tras la importación de un CV en formato CVN. En este caso los ROs importados pasarían por el paso de “Recuperación de ROs”, para completar su información antes de continuar con la “Extracción de descriptores”.

Matching

El proceso de matching actúa sobre los descriptores específicos propuestos por el Enriquecimiento, no sobre el RO. Las características generales del proceso son:

- El administrador define las fuentes externas de entidades con las que intentar el matching de descriptores específicos.
- Entre las posibles fuentes de entidades podemos tener:
  - Fuentes Linked Open Data con punto SPARQL de consulta (p.e. MESH)
  - Fuentes con API de búsqueda.
  - Datasets descargables.
- El usuario investigador selecciona las fuentes de entidades que quiere utilizar para hacer el matching.
- El sistema propone una o más entidades externas con las que enlazar cada descriptor que tenga un matching.

Configuración y Funcionamiento

Cada fuente de entidades externas tendrá un microservicio que se encargará de los procesos de interrogación y de presentar una propuesta de match para un descriptor, en el caso de que la encuentre.

Para el caso de una fuente Linked Data con punto de interrogación SPARQL se configurará del siguiente modo (con el ejemplo de MESH):

- Tipos de recursos en los que buscar el descriptor (mesh:Descriptor, mesh:Concept, mesh:Term) con orden de preferencia
- Propiedades de los recursos en las que buscar el descriptor por cada tipo de recurso (rdfs:label para Descriptor y Concept; mesh: altLabel y mesh:sortVersion para Term)
- Propiedades de los recursos a considerar en el proceso de desambiguación, mediante la presencia de otros descriptores en:
  - Propiedades de los recursos por tipo de recurso (mesh:annotation para Descriptor; mesh:scopeNote para Concept)
  - Propiedades de los recursos que apunten a otros recursos relacionados en los que buscar otros descriptores identificados , por tipo de recurso (mesh:broaderDescriptor y mesh:concept para Descriptor; mesh:narrowerConcept y mesh:relatedConcept para Concept)

Esta sería una solución escalable a otras fuentes externas SPARQL mediante configuración.

Con independencia de la fuente externa, el proceso de matching se encarga de localizar las entidades externas con las que enlazar, con los siguientes pasos:

- Búsqueda de descriptores candidatos, con coincidencia exacta y aproximada, ordenados por tipo y ranking de coincidencia.
- Generación de un ranking de candidatos, preferentemente con una única propuesta
- Presentación al usuario de uno o más enlaces con entidades externas para cada descriptor específico enlazado.
- El usuario podrá eliminar los enlaces en el interfaz (ver apartado siguiente).

Posibles fuentes de matching:

- SPARQL Endpoint, terminología salud, medicina

  https://id.nlm.nih.gov/mesh/query

```
SELECT ?d ?label
FROM <http://id.nlm.nih.gov/mesh>
WHERE {
  {?d a meshv:Descriptor} UNION {?d a meshv:Concept} .
  ?d rdfs:label ?label .
  FILTER(REGEX(?label,'anti-bacterial','i'))
}
ORDER BY ?d
```

- USGS Thesaurus. API. “Topics and methods of scientific study carried out by USGS, with product types, scientific disciplines, geologic time, and types of institutional structure and activities. Broad and shallow, used to help people find scientific information”.

  https://apps.usgs.gov/thesaurus/tab-term.html

  https://apps.usgs.gov/thesaurus/search-pattern.php?text=structural%20geology

- SAGE Terminology Service. SPARQL Endpoint. “The SAGE Social Science Thesaurus is a multidisciplinary vocabulary of the most important concepts in the social sciences”. En inglés.

  http://concepts.sagepub.com:6081/dataset.html?tab=query&ds=/skosmos

  https://concepts.sagepub.com/social-science/concept/welfare_reform

- SPARQL Enpoint. Ver descripción en About - AGROVOC Thesaurus - Organizations - "FAO catalog"

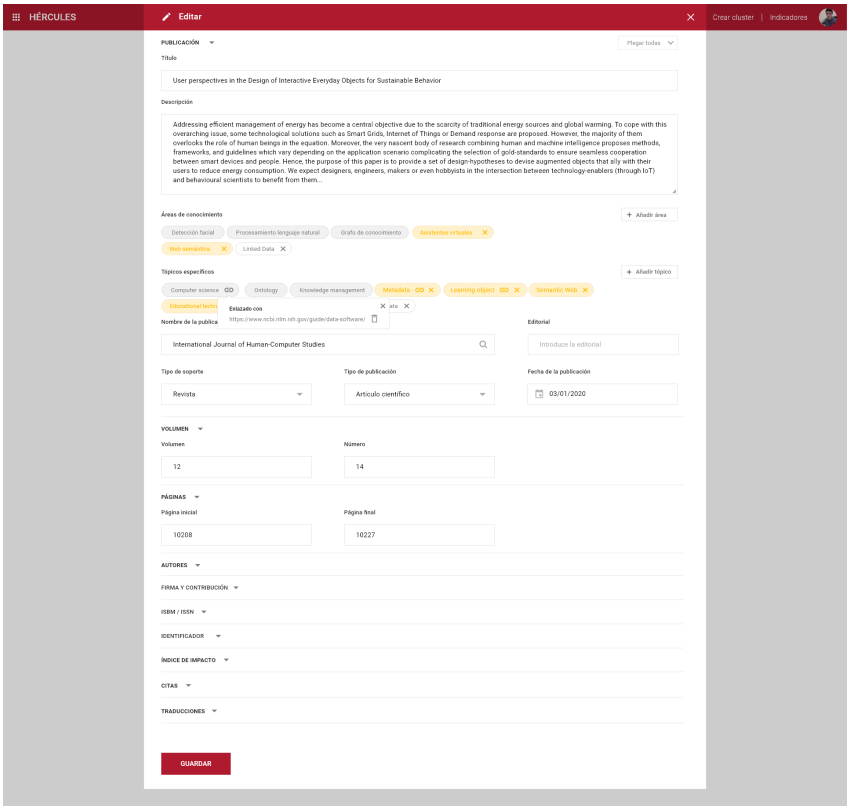
  https://agrovoc.fao.org/sparql

  http://aims.fao.org/aos/agrovoc/c_765.html (Bacteria)

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT distinct ?sub ?label WHERE {
  ?sub ?pred ?obj .
  ?sub <http://www.w3.org/2004/02/skos/core#prefLabel> ?label .
  FILTER(REGEX(?label,'mammal','i'))
} LIMIT 100
```

Presentación y uso del matching

El usuario podrá eliminar los enlaces propuestos en la edición del CV y en la Gestión de FAIR RO, como se muestra en la siguiente imagen.



Las entidades resultantes del matching se presentan y usan del siguiente modo:

- Descriptores temáticos visualmente enlazados hacia la entidad externa
  - Enlace hacia la web externa
  - Presentación integrada de información del repositorio externo (p.e. pharmacologicalAction, indexerConsiderAlso)
- Explotación de contextos utilizando las relaciones de la entidad externa
  - Mejora de la divulgación en MA con información contextualizada.
- Enlace(s) en los datos de la entidad (Linked Data).
- Priorización en el orden de descriptores propuestos por el Enriquecimiento (tras validación con los usuarios).
- Posible uso en la explicabilidad del algoritmo de similitud.

Implementación de Matching (entity linking)

- No es un proceso crítico, por lo que se desarrollará como un proceso offline con cola de procesamiento y reintentos para que no bloquee ni ralentice el funcionamiento online del sistema (gestión de CV y de ROs).
- Enlazado de descriptores específicos de ámbito biomédico con MESH, según lo descrito en el apartado anterior de Configuración y Funcionamiento.
  - P.e.: https://id.nlm.nih.gov/mesh/D012343.html
  - Recuperación de información de padres e hijos del concepto conectado.
- Uso del API de UMLS para conectar los términos MESH con la terminología SNOMED CT (pendiente de probar).
  - Ver en https://documentation.uts.nlm.nih.gov/rest/source-asserted-identifiers/
  - Recuperación de información de padres e hijos del concepto SNOMED conectado.

Enriquecimiento. Experimentos y resultados

Preparación de datasets

Se ha definido una taxonomía unificada que cubre las diferentes áreas de conocimiento y los diferentes tipos de de RO:

- Papers (ASJC + Arxiv+ Pubmed) + protocolos (Bio-protocol) + proyectos código (Sourceforge).

Para la generación de datasets de entrenamiento se han utilizado las siguientes fuentes:

- Protocolos: Bio-protocol.
- Proyectos código: Sourceforge.
- Papers: Arxiv + dataset interno (autores españoles).

Dataset de ROs de bioprotocolos

Se ha creado un dataset mediante scraping de las páginas de bio-protocol.org. Hemos recuperado 6.473 protocolos, obteniendo los siguientes metadatos para cada uno de ellos: título, autores, DOI, publicación, resumen, palabras clave, background, materiales, equipamiento, software, procedimiento, análisis de datos, notas, recetas, reconocimientos, declaración de intereses, ética, referencias y categorías.

Category	Total in dataset	Train	Dev	test
Biochemistry	1009	705	154	150
Biophysics	43	32	4	7
Cancer_Biology	320	226	49	45
Cell_Biology	1401	972	213	216
Developmental_Biology	172	130	14	28
Environmental_science	11	7	2	2
Immunology	435	297	71	67
Microbiology	895	601	153	141
Molecular_Biology	733	498	103	132
Neuroscience	441	296	78	67
Plant_Science	754	542	98	114
Stem_Cell	162	113	24	25
Systems_Biology	97	67	18	12
Total	6473	4486	981	1006
Total examples	3489	2435	525	529

Tabla 1: Dataset de bioprotocolos.

Dataset de descripciones de proyectos de código

Se ha creado un dataset que incluye descripciones de proyectos de código enlazados con los correspondientes descriptores temáticos a partir de sourceforge.net. En un primer paso, se han descargado los nombres de los proyectos alojados en sourceforge mediante scraping a partir del directorio https://sourceforge.net/directory. Posteriormente, se han obtenido los metadatos (que incluyen los descriptores temáticos) de cada nombre utilizando el API sourceforge.net/projects/codeblocks.). Se muestra, a continuación, las estadísticas para los descriptores de primer nivel:

Category	**Total in dataset**	**Train**	**Dev**	**test**
Multimedia	7112	4970	1080	1062
Desktop Environment	4539	3180	684	675
Social sciences	393	273	62	58
Mobile	1356	956	205	195
Sociology	309	232	33	44
Blockchain	120	81	19	20
Scientific/Engineering	6977	4912	982	1083
Software Development	9055	6352	1343	1360
Internet	7993	5566	1223	1204
Formats and Protocols	4713	3344	666	703
Religion and Philosophy	553	389	85	79
Database	4900	3442	738	720
Security	4684	3300	692	692
Communications	6047	4191	928	928
Games/Entertainment	5482	3860	812	810
Office/Business	5623	3984	808	831
System	7888	5500	1176	1212
Education	4719	3292	726	701
Terminals	1061	737	162	162
Text Editors	4504	3180	661	663
Printing	830	567	135	128
Total	88858	62308	13220	13330
Total examples	57687	40381	8653	8653

Tabla 2: Dataset de proyectos de código.

Dataset de artículos científicos

Se está creando un dataset de artículos científicos con un doble objetivo: servir de entrenamiento para la extracción de descriptores y utilizarlo para obtener datos que permitan realizar pruebas de carga. Para la elaboración del dataset nos hemos encontrado con múltiples problemas legales y restricciones en el uso de fuentes de datos como WoS, Dialnet y otras. Por ello, se están utilizando las siguientes fuentes de datos disponibles para ello y que igualmente condicionan la elaboración de una taxonomía unificada de descriptores temáticos:

- ArXiv.
- PubMed.
- Dataset interno creado a partir de Scopus en el marco de un proyecto del Plan de Impulso de la Lengua y utilizado en el Challenge.

ASJC de Scopus y arXiv utilizan descriptores diferentes en sus clasificaciones, por lo que ha sido necesario hacer una fusión entre los dos sistemas de clasificación. También se ha realizado el mismo proceso de integración con los descriptores de la clasificación MESH de PubMed que cuelgan de la entrada “Medicine”, dado que investigadores consultados en dicho dominio así lo han sugerido. Por lo tanto, la taxonomía fusionada es la resultante de la única aproximación factible encontrada. La taxonomía resultante tiene tres niveles de granularidad:

- Nivel 1: 27 descriptores
- Nivel 2: 392 descriptores
- Nivel 3: 167 descriptores

Category	Total in dataset	Train	Dev	Test
Agricultural and Biological Sciences	13,128	9106	2007	2015
Arts and Humanities	6,857	4804	1060	992
Biochemistry, Genetics and Molecular Biology	15,506	10880	2303	2323
Business, Management and Accounting	6,896	4789	1050	1057
Chemical Engineering	8,277	5790	1195	1292
Chemistry	13,180	9204	1925	2051
Computer Science	13,637	9573	2036	2028
Decision Sciences	5,344	3742	798	804
Dentistry	2,970	2087	454	429
Earth and Planetary Sciences	6,429	4493	974	962
Economics, Econometrics and Finance	6,771	4776	994	1001
Energy	6,476	4552	923	1001
Engineering	14,885	10379	2191	2315
Environmental Science	10,350	7206	1541	1603
Health Professions	5,107	3608	786	713
Immunology and Microbiology	7,494	5237	1164	1093
Materials Science	9,531	6717	1367	1447
Mathematics	12,415	8699	1873	1843
Medicine	24,845	17442	3731	3671
Multidisciplinary	4,344	3059	633	652
Neuroscience	6,391	4480	971	940
Nursing	5,365	3725	812	828
Pharmacology, Toxicology and Pharmaceutics	6,122	4257	946	918
Physics and Astronomy	14,277	9992	2181	2104
Psychology	6,412	4495	968	948
Social Sciences	11,434	8054	1701	1678
Veterinary	5,043	3569	771	703
Total	123,965	86,775	18,594	18,594
Total examples	249,486	174,715	37,355	37,411

Tabla 3. Dataset de papers (nivel 1 de granularidad).

Entrenamiento y evaluación de clasificadores multi-etiqueta para identificación de descriptores temáticos

A partir del dataset presentado anteriormente se han entrenado clasificadores multi-etiqueta utilizando diferentes estrategias de representación textual y diferentes algoritmos. Por un lado, se ha analizado una representación textual no densa basada en bolsa-de-palabras, y por otro, una representación densa basada en embeddings contextuales.

Para implementar la clasificación basada en la representación vectorial no densa se han estudiado los algoritmos Logistic Regression y SVM. Los valores de los vectores se computan en base al estadístico TFIDF, y la clasificación multi-etiqueta se aborda mediante la estrategia one-vs-all y oversampling aleatorio en cada clasificador booleano.

El clasificador basado en la representación densa se ha implementado utilizando modelos de lenguaje neuronal pre-entrenados (BERT, BART, Electra...) que se han afinado (*fine-tuned)* a la tarea multi-etiqueta.

Se muestran, a continuación, los resultados (según las métricas de Precision, Recall, y F-score) obtenidos en los experimentos:

Overall results (macro avg)	**P**	**R**	**F**
LR	**0.66**	**0.59**	**0.62**
SVM	0.59	0.46	0.52
Bert-base Binary Classifiers (oversample)			
Bert-base - (Zuhaitz-TF) - oversample max	0.67	0.56	0.60
Bert-base - (Transformers-Pytorch) - oversample max	0.68	0.56	0.61
Electra-base - (Transformers-Pytorch) - oversample max	0.69	0.54	0.58
RoBerta-base - (Transformers-Pytorch) - oversample max	**0.68**	**0.61**	**0.64**
RoBerta-large - (Transformers-Pytorch) - oversample max	**0.65**	**0.65**	**0.65**
Bert-large-cased - (Transformers-Pytorch) - oversample max	0.63	0.59	0.60
Electra-large - (Transformers-Pytorch) - oversample max	0.65	0.48	0.53
BigBird-base-(2048) - (Transformers-Pytorch) - oversample max	0.66	0.58	0.61

Tabla 4: Resultados para bioprotocolos.

Overall results (macro avg)	**P**	**R**	**F**
LR	**0.49**	**0.60**	**0.53**
SVM	0.53	0.52	0.51
Bert-base Binary Classifiers (oversample)	0.65	0.49	0.55
Bert-base (512) - (Transformers-Pytorch) - oversample max	**0.68**	**0.55**	**0.61**
RoBerta-base - (Transformers-Pytorch) - oversample max	0.66	0.56	0.60
RoBerta-large - (Transformers-Pytorch) - oversample max	**0.67**	**0.59**	**0.62**
Bert-large-cased (256) - (Transformers-Pytorch) - oversample max	0.66	0.57	0.61

Tabla 5: Resultados para proyectos de código.

Granularity	**System**	**P**	**R**	**F**
L0 (5 label)	LR	0.82	0.82	0.82
	SVM	0.77	0.86	0.81
	BERT	**0.92**	**0.92**	**0.92**
L1 (27 label)	LR	0.61	0.86	0.72
	SVM	0.75	0.79	0.77
	BERT	**0.93**	**0.92**	**0.92**
L2 (344 label)	LR	0.41	0.86	0.55
	SVM	0.68	0.72	0.7
	BERT	0.89	0.84	0.86
	mBERT (test on Spanish papers)	**0.94**	**0.93**	**0.94**

Tabla 6: Resultados para papers con distintas granularidades de la taxonomía. También se incluyen resultados de evaluación multilingüe mostrando resultados para papers en castellano (mBERT).

Entrenamiento y evaluación de sistema para identificación de descriptores específicos

Se ha experimentado con un enfoque de extracción consistente de dos componentes: extractor de sintagmas nominales para identificar los candidatos a descriptor específico, y clasificador supervisado para cribar los candidatos. Para implementar el clasificador supervisado se han analizado dos enfoques. La estructura del sistema estudiado sería la siguiente:

- Extractor de sintagmas nominales: Procesador lingüistico de spaCy + patrones morfosintácticos.
- Clasificador supervisado:
  - Basado en atributos (Gradient Boosting): Frecuencia, posición, frecuencia en colección de dominio abierto, incluido en título?, incluido en abstract?,. …
  - Basado en un modelo BERT ajustado a la tarea.

Se muestran, a continuación, los resultados obtenidos sobre el dataset de papers Krapivin descrito en el capítulo 2.

	Gradient Boosting			**Bert (pair-sentence classification)**		
Score type	**All**	**Single-word**	**Multi-word**	**All**	**Single-word**	**Multi-word**
R@5	0.35	0.50	0.62	0.31	0.43	0.38
R@10	0.53	0.65	0.77	0.44	0.52	0.54
R@15	0.62	0.75	0.83	0.52	0.58	0.62
R@20	0.70	0.82	0.89	0.57	0.62	0.67

Tabla 7: Resultados para extracción de descriptores específicos sobre papers (dataset *Krapivin*). Métrica de evaluación, recall at X.

Próximos trabajos:

- Integración de matching (enlazado de entidades) en el proceso de extracción de descriptores específicos. Ver apartado de Matching en Flujo e interfaces del enriquecimiento.

Enriquecimiento. Identificación de ROs similares (en elaboración)

Experimentos y resultados

Se han creado los siguientes datasets a partir de los datasets descritos en el anterior capítulo:

- Papers: 300.000 abstracts.
- Protocolos: 3.489 protocolos.
- Proyectos código: 57.687 fichas.

La evaluación de los diferentes sistemas estudiados ha consistido en analizar manualmente los rankings de documentos similares devueltos por cada sistema para un grupo de 20 documentos test determinado inicialmente para cada tipo de RO (papers, protocolos, proyectos de código). Se han anotado manualmente los cinco documentos más similares de cada ranking determinando si son similares o no, para así calcular la precisión en ese corte.

Para la extracción de los rankings de documentos similares, a partir de los datasets, para cada documento test, se han analizado dos enfoques.

1. Similitud basada en representación Bag-of-Words, ponderación tf-idf, y distancia coseno:
   - Filtrado de vocabulario muy frecuente (> 10% docs) y poco frecuente (df<4).
2. Similitud basada en representación densa (embeddings) y ajustada a la tarea STS (Semantic Text Similarity), y distancia coseno:
   - Embeddings estimados por Bi-encoders neuronales basados en BERT y ajustados a tarea STS con más de 1000 M de tuplas de ejemplos.

Se muestran en la siguiente tabla los resultados obtenidos:

	BoW	Bi-encoder MiniLM
P@5	0.55	0.93

Tabla 1: Resultados (Precision at 5) para rankings de paper similares

Se listan, a continuación, los próximos trabajos a abordar (aspectos comentados en la reunión de presentación del sprint):

- Evaluación de representación textual basada en selección de descriptores y uso de grafos.
- Evaluación translingüe.
- Evaluación de protocolos y proyectos de código.
- Experimentación sobre selección de descriptores relevantes a relación de similitud:
    - Enfoque 1: Intersección de descriptores de textos A y B.
    - Enfoque 2: Descriptores más cercanos al embedding promedio de textos A y B.