Logroño, diciembre 2021

Hércules edma. Flujo e interfaces del proceso de enriquecimiento

Proyecto Hércules ed - ma

Contenido

[1 Introducción 3](#_Toc90500992)

[2 Extracción de descriptores 5](#_Toc90500993)

[2.1 Taxonomías para descriptores temáticos 5](#_Toc90500994)

[2.2 Extracción en ROs de fuentes externas 5](#_Toc90500995)

[2.3 ROs introducidos por el usuario 8](#_Toc90500996)

[2.4 ROs procesados masivamente 8](#_Toc90500997)

[3 Matching 9](#_Toc90500998)

[3.1 Configuración y Funcionamiento 9](#_Toc90500999)

[3.2 Presentación y uso del matching 11](#_Toc90501000)

# 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.

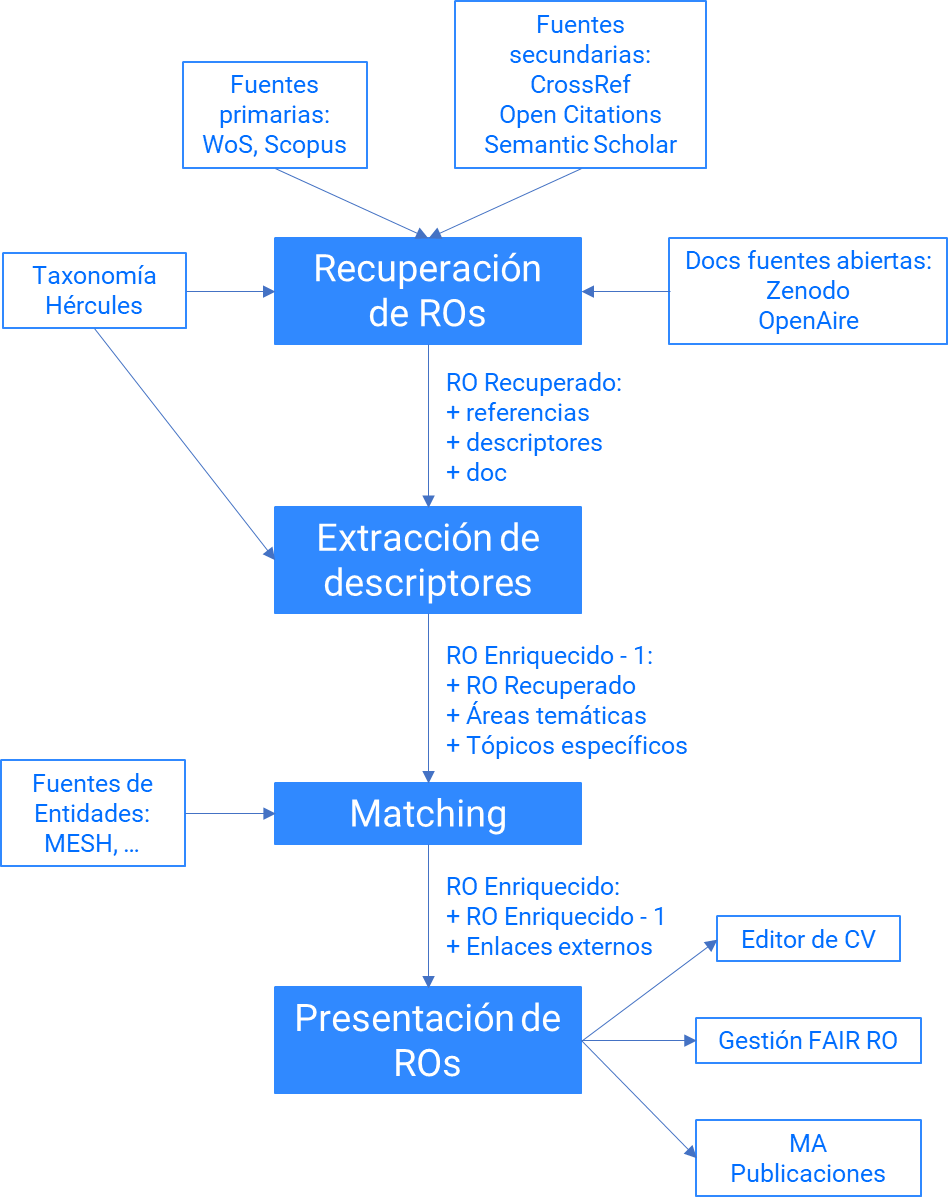
Los objetivos del proceso son:

* 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 (enlazar con Confluence)
* 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).](https://confluence.um.es/confluence/pages/viewpage.action?pageId=222822823)
* El proceso de unificación y su resultado se puede consultar en [Taxonomía unificada de descriptores temáticos para Hércules (Confluence).](https://confluence.um.es/confluence/pages/viewpage.action?pageId=306774073)
* 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.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

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

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

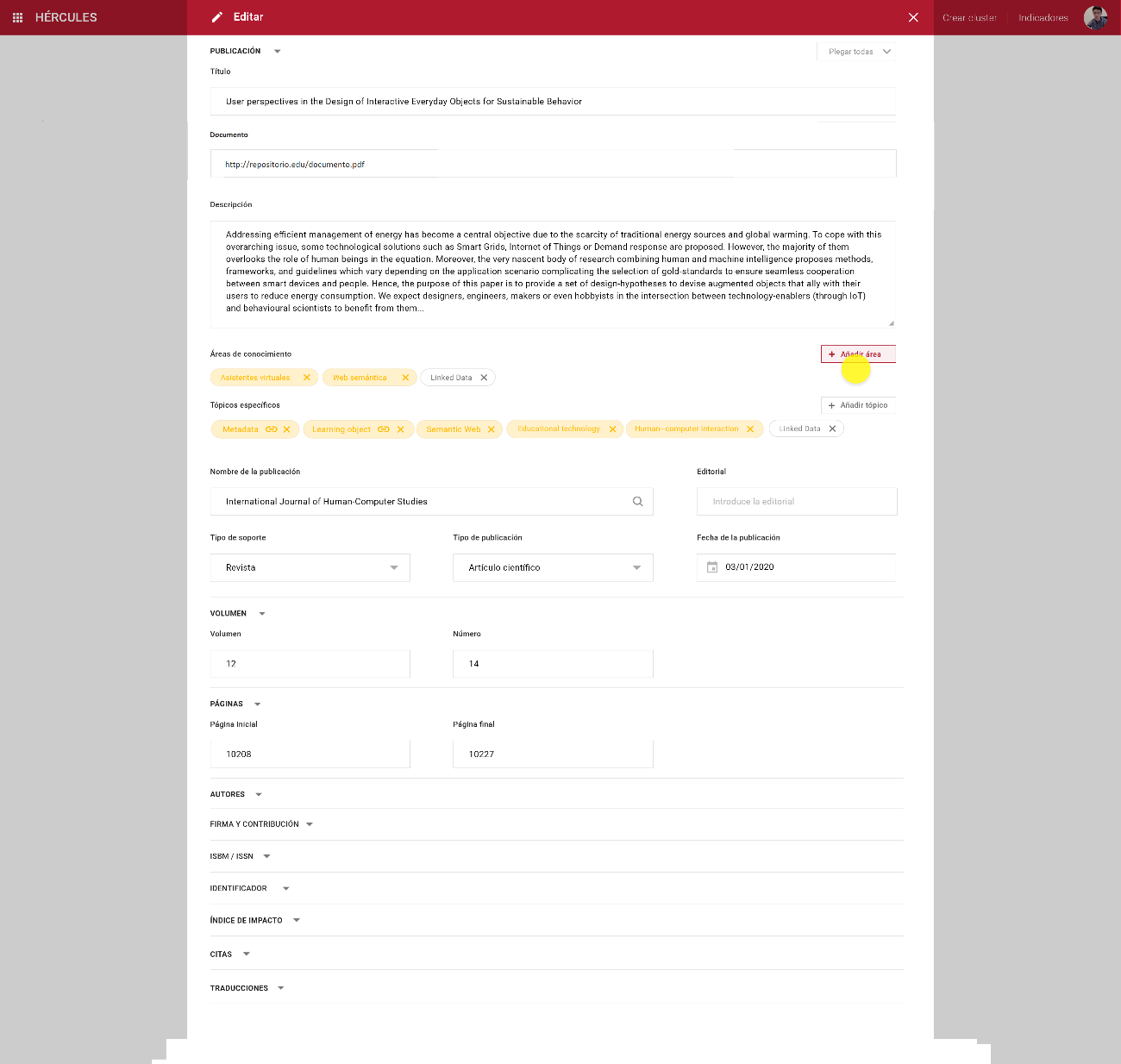
Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

## 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:

* MESH. 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>

* AGROVOC. SPARQL Enpoint. Ver descripción en [About - AGROVOC Thesaurus - Organizations - "FAO catalog"](https://data.apps.fao.org/catalog/organization/about/agrovoc)

<https://agrovoc.fao.org/sparql>

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

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

## 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.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamenteLas 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).