GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
UNIVERSIDAD DE MURCIA
FONDO EUROPEO DE DESARROLLO REGIONAL (FEDER)
Unión Europea
Una manera de hacer Europa

OAI-PMH, Carga de Datos y descubrimiento

Álvaro Palacios
RIAM Intelearning Lab – GNOSS
alvaropalacios@gnoss.com

HĒRCULES

gnoss
Deusto
Universidad de Deusto
Deustuko Unibertsitatea
University of Deusto

Introducción

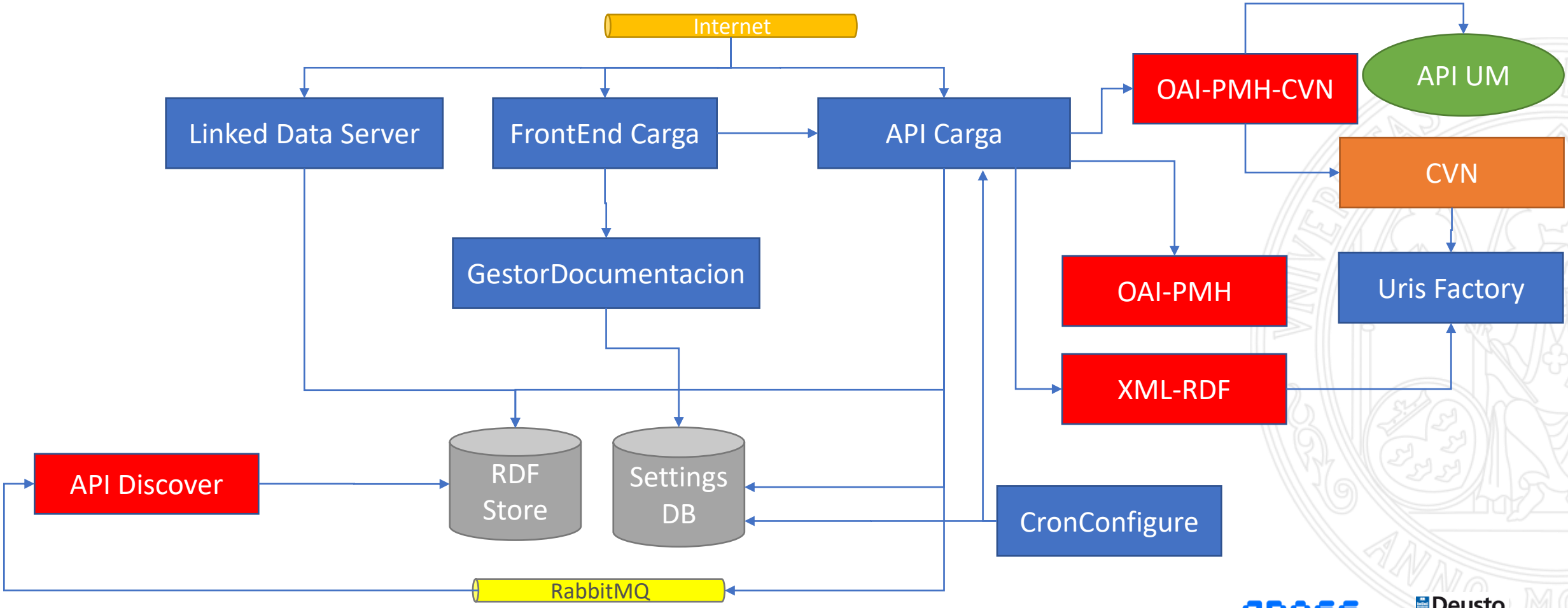
- Arquitectura ASIO
- OAI-PMH
- OAI-PMH en ASIO
- Conversor XML RDF
- Descubrimiento en ASIO
- Reconciliación de entidades
- Descubrimiento de enlaces
- Descubrimiento de equivalencias
- Conclusión

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
UNIVERSIDAD DE MURCIA
FONDO EUROPEO DE DESARROLLO REGIONAL (FEDER)
Una manera de hacer Europa
Unión Europea

Hércules ASIO. Arquitectura ASIO

gnoss
Deusto
Universidad de Deusto
Deustuko Unibertsitatea
University of Deusto

Arquitectura ASIO



Arquitectura ASIO

- API Carga: Contiene los procesos ETL (Extract, Transform and Load) necesarios para la carga de datos. Junto con los procesos de gestión de repositorios, validaciones y sincronizaciones.
- API Discover: Servicio encargado del descubrimiento: reconciliación de entidades, descubrimiento de enlaces y detección de equivalencias.
- CronConfigure: Es un api para la gestión y configuración del programado de tareas, tanto de ejecución recurrente como ejecución única sobre los repositorios configurados.
- OAI-PMH: Servicio recolector de metadatos.
- XML-RDF: Tras generar los XML con OAI-PMH, los transforma en RDF.
- FrontEndCarga: Constituye el interfaz Web de administración de las cargas de datos en la plataforma Hércules ASIO.

Arquitectura ASIO

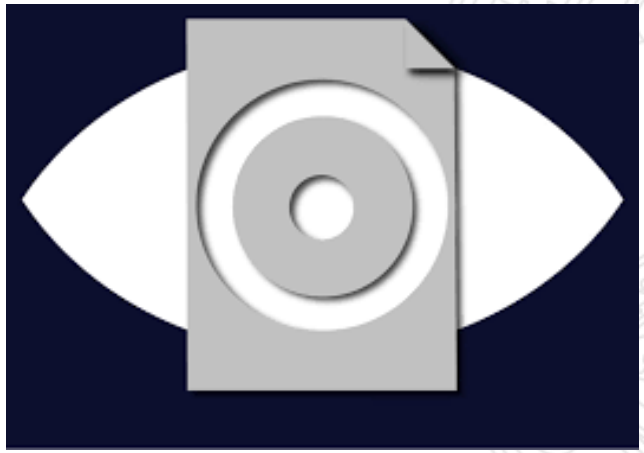
- Gestor Documentación: Permite publicar páginas webs que informen del nodo ASIO de la universidad, con contenido estático y dinámico, obtenido éste último del API de consulta o del SPARQL Endpoint
- Identity Server: Encargado de la securización mediante tokens para los APIs que forman el proyecto.
- URIs Factory: Es el encargado de generar las URIs de todas las entidades existentes en ASIO.
- LinkedDataServer: Proporciona el servicio de datos enlazados de Hércules ASIO, cumpliendo la recomendación Linked Data Platform .

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
UNIVERSIDAD DE MURCIA
FONDO EUROPEO DE DESARROLLO REGIONAL (FEDER)
Una manera de hacer Europa
Unión Europea

Hércules ASIO. OAI-PMH.

OAI-PMH

- Protocolo: https://www.openarchives.org/OAI/openarchivesprotocol.html
- OAI-PMH -*Open Archives Initiative Protocol for Metadata Harvesting*- es un protocolo para la transmisión de metadatos por internet.
- La versión actual es la 2.0, creada en 2002.



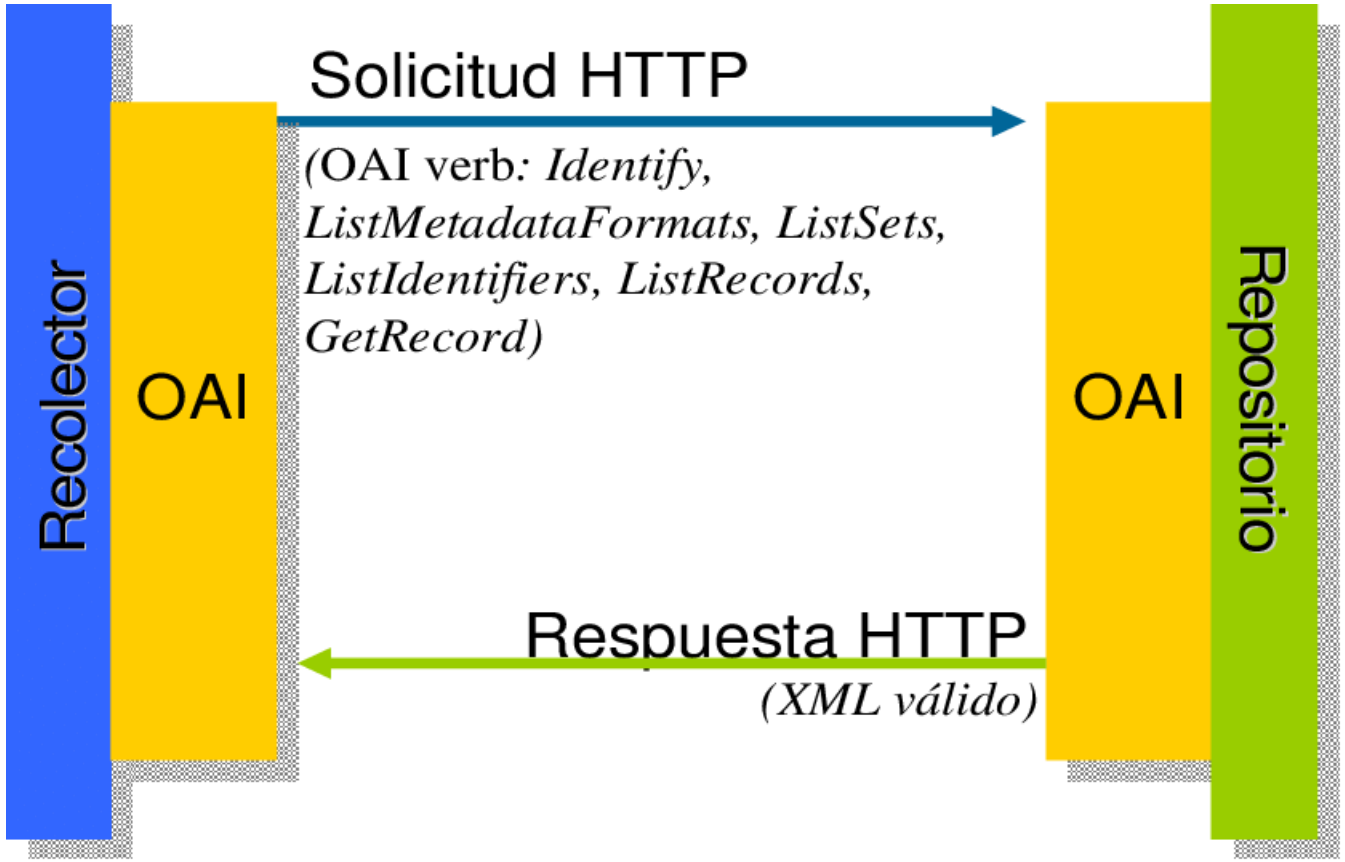
OAI-PMH

El Protocolo OAI-PMH presenta las siguientes características:

1. Su funcionamiento se basa en una arquitectura cliente-servidor en la que un servicio recolector de metadatos pide información a un proveedor de datos.
2. Las peticiones se expresan en HTTP, utilizando únicamente los métodos GET o POST.
3. Todas las respuestas deben ser documentos XML bien formados codificados en UTF-8.
4. Fechas y tiempo se codifican mediante la ISO 8601 y se expresan en UTC.
5. Soporta la difusión de registros en diversos formatos de metadatos.
6. Tiene control de flujo.
7. Cuando hay un error o una excepción los repositorios deben indicarlos distinguiéndolos de los códigos de estado HTTP por incluir uno o más elementos de error en la respuesta.

OAI-PMH

Protocolo de peticiones:



OAI-PMH

Protocolo de peticiones:

El servicio OAI-PMH expone 6 peticiones distintas:

1. *Identify*, para obtener información sobre el servidor.
2. *ListSets*, para obtener registros pertenecientes a una clase determinada creada por el servidor.
3. *ListMetadataFormats*, para obtener la lista de los formatos bibliográficos usados por el servidor.
4. *ListIdentifiers*, para obtener encabezamientos.
5. *GetRecord*, para obtener un registro determinado.
6. *ListRecords*, para obtener registros completos.

OAI-PMH. Identify.

Permite recuperar información sobre un repositorio. Los repositorios también **pueden** emplear el verbo Identificar para devolver información descriptiva adicional.

Argumentos**:**

No tiene.

OAI-PMH. Identify.

Respuesta:

```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<OAI-PMH xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"
xmlns="http://www.openarchives.org/OAI/2.0/">
    <responseDate>2021-02-16T12:12:27Z</responseDate>
    <request verb="Identify">http://herc-as-front-desa.atica.um.es/oai-pmh-
    cvn/OAI_PMH</request>
    <Identify>
        <repositoryName>OAI_PMH_CVN</repositoryName>
        <baseURL>http://herc-as-front-desa.atica.um.es/oai-pmh-cvn/OAI_PMH</baseURL>
        <protocolVersion>2.0</protocolVersion>
        <adminEmail>test@domain.ch</adminEmail>
        <earliestDatestamp>1987-02-16T00:00:00Z</earliestDatestamp>
        <deletedRecord>no</deletedRecord>
        <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
    </Identify>
</OAI-PMH>
```

OAI-PMH. ListSets.

Recupera la estructura establecida de un repositorio, adecuado para el consumo selectivo.

Argumentos:

- **resumptionToken**: Token de flujo de control.

OAI-PMH. ListSets.

Respuesta:

```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<OAI-PMH xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"
xmlns="http://www.openarchives.org/OAI/2.0/">
  <responseDate>2021-02-16T12:18:26Z</responseDate>
  <request verb="ListSets">http://herc-as-front-desa.atica.um.es/oai-pmh-
cvn/OAI_PMH</request>
  <ListSets>
    <set>
      <setSpec>cvn</setSpec>
      <setName>Currículum Vítae Normalizado</setName>
      <setDescription>Currículum Vítae Normalizado</setDescription>
    </set>
  </ListSets>
</OAI-PMH>
```

OAI-PMH. ListMetadataFormats.

Recupera los formatos de metadatos disponibles en un repositorio.

Argumentos

- **identifier (opcional)**: Especifica el identificador único del elemento para el que se solicitan los formatos de metadatos disponibles.

OAI-PMH. ListMetadataFormats.

Respuesta:

```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<OAI-PMH xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"
xmlns="http://www.openarchives.org/OAI/2.0/">
  <responseDate>2021-02-16T12:20:50Z</responseDate>
  <request verb="ListMetadataFormats">http://herc-as-front-desa.atica.um.es/oai-
pmh-cvn/OAI_PMH</request>
  <ListMetadataFormats>
    <metadataFormat>
      <metadataPrefix>rdf</metadataPrefix>
    </metadataFormat>
  </ListMetadataFormats>
</OAI-PMH>
```

OAI-PMH. ListIdentifiers.

Permite recuperar encabezados en lugar de registros. Según el repositorio, un encabezado puede tener estado eliminado si se eliminó un registro que coincida con los argumentos.

Argumentos

- **from (opcional)**: Especifica un límite inferior para la recolección selectiva basada en fecha.
- **until (opcional)**: Especifica un límite superior para la selección selectiva basada en fecha.
- **metadataPrefix**: Especifica que los encabezados deben devolverse solo si el formato de los metadatos coincide con el metadataPrefix.
- **set (opcional)**: Especifica los criterios establecidos para la recolección selectiva.
- **resumptionToken (opcional)**: Token de control de flujo.

OAI-PMH. ListIdentifiers.

Respuesta:

```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<OAI-PMH xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd" xmlns="http://www.openarchives.org/OAI/2.0/">
  <responseDate>2021-02-16T12:23:01Z</responseDate>
  <request verb="ListIdentifiers" metadataPrefix="rdf">http://herc-as-front-desa.atica.um.es/oai-pmh-cvn/OAI_PMH</request>
  <ListIdentifiers>
    <header>
      <identifier>1</identifier>
      <datestamp>2020-11-09T10:43:04Z</datestamp>
      <setSpec>cvn</setSpec>
    </header>
    <header>
      <identifier>2</identifier>
      <datestamp>2020-11-10T10:43:04Z</datestamp>
      <setSpec>cvn</setSpec>
    </header>
    …
    <header>
      <identifier>8</identifier>
      <datestamp>2020-11-16T10:43:04Z</datestamp>
      <setSpec>cvn</setSpec>
    </header>
  </ListIdentifiers>
</OAI-PMH>
```

OAI-PMH. GetRecord.

Recupera un registro de metadatos individual de un repositorio.

Argumentos

- **identifier**. Especifica el ID único del elemento en el repositorio desde el que se debe difundir el registro.

- **metadataPrefix**. Especifica el metadataPrefix del formato que debe incluirse en la parte de metadatos de un registro devuelto.

OAI-PMH. GetRecord.

Respuesta:

```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<OAI-PMH xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd" xmlns="http://www.openarchives.org/OAI/2.0/">
  <responseDate>2021-02-16T12:25:55Z</responseDate>
  <request verb="GetRecord" identifier="1" metadataPrefix="rdf">http://herc-as-front-desa.atica.um.es/oai-pmh-cvn/OAI_PMH</request>
  <GetRecord>
    <record>
      <header>
        <identifier>1</identifier>
        <datestamp>2020-11-09T10:46:10Z</datestamp>
        <setSpec>cvn</setSpec>
      </header>
      <metadata>
        <rdf:RDF xmlns:bibo="http://purl.org/roh/mirror/bibo#" xmlns:foaf="http://purl.org/roh/mirror/foaf#"
xmlns:obobfo="http://purl.org/roh/mirror/obo/bfo#" xmlns:oboro="http://purl.org/roh/mirror/obo/ro#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:roh="http://purl.org/roh#" xmlns:rohes="http://purl.org/rohes#"
xmlns:vcard="http://purl.org/roh/mirror/vcard#" xmlns:vivo="http://purl.org/roh/mirror/vivo#">
          <rdf:Description rdf:about="http://graph.um.es/res/person/1a4174b2-df39-4b7b-9304-0de2b3bf33a9">
            <foaf:name rdf:datatype="http://www.w3.org/2001/XMLSchema#string">PEDRO MIGUEL RUIZ MARTINEZ</foaf:name>
            <roh:correspondingAuthorOf rdf:resource="http://graph.um.es/res/book/18496e8a-d583-497c-a686-7c9d2dd69340" />
            <roh:participates rdf:resource="http://graph.um.es/res/conference/INFORMATIK+2005" />
          </rdf:Description>
        </rdf:RDF>
      </metadata>
    </record>
  </GetRecord>
</OAI-PMH>
```

OAI-PMH. ListRecords.

Es una combinación de ListIdentifiers y GetRecords. Obtiene un listado con todos los records solicitados y sus metadatos

Argumentos:

- **from (opcional)**: Especifica un límite inferior para la recolección selectiva basada en fecha.
- **until (opcional)**: Especifica un límite superior para la selección selectiva basada en fecha.
- **metadataPrefix**: Especifica que los encabezados deben devolverse solo si el formato de los metadatos coincide con el metadataPrefix.
- **set (opcional)**: Especifica los criterios establecidos para la recolección selectiva.
- **resumptionToken**: Token de control de flujo.

OAI-PMH. ListIdentifiers.

Respuesta:

```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<OAI-PMH xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd" xmlns="http://www.openarchives.org/OAI/2.0/">
  <responseDate>2021-02-16T12:23:01Z</responseDate>
  <request verb="ListRecords" metadataPrefix="rdf">http://herc-as-front-desa.atica.um.es/oai-pmh-cvn/OAI_PMH</request>
  <ListRecords>
    <record>
      <header>
        <identifier>1</identifier>
        <datestamp>2020-11-09T10:43:04Z</datestamp>
        <setSpec>cvn</setSpec>
      </header>
      <metadata>
        <rdf:RDF xmlns:bibo="http://purl.org/roh/mirror/bibo#" xmlns:foaf="http://purl.org/roh/mirror/foaf#"
xmlns:obobfo="http://purl.org/roh/mirror/obo/bfo#" xmlns:oboro="http://purl.org/roh/mirror/obo/ro#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:roh="http://purl.org/roh#" xmlns:rohes="http://purl.org/rohes#"
xmlns:vcard="http://purl.org/roh/mirror/vcard#" xmlns:vivo="http://purl.org/roh/mirror/vivo#">
          <rdf:Description rdf:about="http://graph.um.es/res/person/1a4174b2-df39-4b7b-9304-0de2b3bf33a9">
            <foaf:name rdf:datatype="http://www.w3.org/2001/XMLSchema#string">PEDRO MIGUEL RUIZ MARTINEZ</foaf:name>
            <roh:correspondingAuthorOf rdf:resource="http://graph.um.es/res/book/18496e8a-d583-497c-a686-7c9d2dd69340" />
            <roh:participates rdf:resource="http://graph.um.es/res/conference/INFORMATIK+2005" />
          </rdf:Description>
        </rdf:RDF>
      </metadata>
    </record>
    …
  </ListRecords>
</OAI-PMH>
```

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
UNIVERSIDAD DE MURCIA
FONDO EUROPEO DE DESARROLLO REGIONAL (FEDER)
Una manera de hacer Europa
Unión Europea

Hércules ASIO. OAI-PMH en ASIO

gnoss
Deusto
Universidad de Deusto
Deustuko Unibertsitatea
University of Deusto

OAI-PMH EN ASIO

- En ASIO se pueden gestionar los repositorios con el API Carga, dentro del FrontEndCarga existen pantallas para realizar estas gestiones: https://herc-as-front-desa.atica.um.es/carga-web/RepositoryConfig
- Estos repositorios se utilizan en las sincronizaciones:
  - Se listan los identificadores con ListIdentifiers.
  - Se obtienen los MetadataFormats disponibles en el repositorio.
  - Se obtienen los records de forma individual.
  - Se llama al servicio conversor XML-RDF si es necesario.
  - Se manda al API Carga el RDF en formato ROH

OAI-PMH EN ASIO

- En estos momentos hay configurados 3 repositorios:
- **CRIS from Radbound University, NL:**

  Repositorio con datos de la Universidad de Radbound (Holanda) (https://oamemtfp.uci.ru.nl/metis-oaipmh-endpoint/OAIHandler)
- **CVN_OAI_PMH:**

  Repositorio con 8 cvs de prueba de la Universidad de Murcia (http://herc-as-front-desa.atica.um.es/oai-pmh-cvn/OAI_PMH)
- **OAI-PMH-XML:**

  Repositorio con varios XML de prueba (http://herc-as-front-desa.atica.um.es/oai-pmh-xml/OAI_PMH)

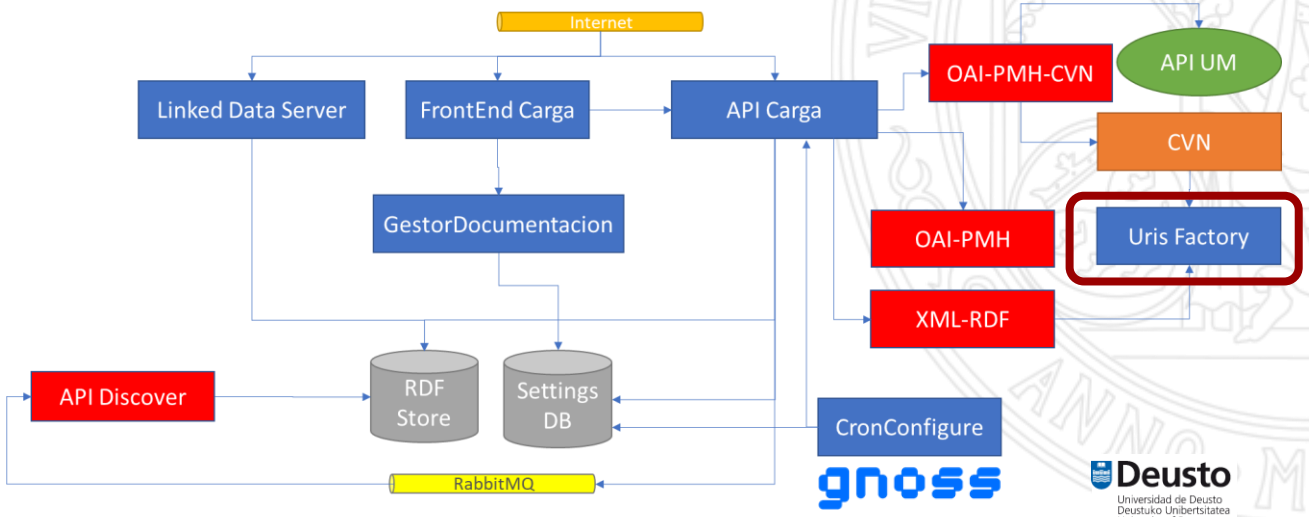
GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
UNIVERSIDAD DE MURCIA
FONDO EUROPEO DE DESARROLLO REGIONAL (FEDER)
Unión Europea
Una manera de hacer Europa

Hércules ASIO. Conversor XML-RDF.

Conversor XML-ROH TODO ORGNAIZAR.

- Los servicios OAI-PMH pueden devolver los metadatos directamente en formato RDF ROH, en cuyo caso no es necesario aplicar el conversor (caso del OAI-PMH con los cvs de la Universidad de Murcia).
- Cuando esto no es así (por ejemplo se quiere reutilizar un repositorio ya existente), es necesario utilizar este conversor para transformar el XML en un RDF ROH.
- Este servicio interactúa con el servicio UrisFactory para la creación de las URLs de las entidades.

- Este servicio cuenta con dos métodos:
  - El primero mostrará un listado con los formatos admitidos.
  - El segundo, mediante el formato y el XML a transformar, creará el RDF pertinente.



Conversor XML-ROH. appsettings.json

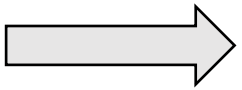
```
{
  "Logging": {
    "LogLevel": {
      "Default": "Information",
      "Microsoft": "Warning",
      "Microsoft.Hosting.Lifetime": "Information"
    }
  },
  "AllowedHosts": "*",
  "UrlUrisFactory": "http://herc-as-front-desa.atica.um.es/uris/"
}
```

Conversor XML-ROH. Ficheros de configuración TOML. (1/2)

- Un fichero TOML es un formato de archivo de configuración que tiene como función mapear datos de forma sencilla.
- Su nombre viene de “Tom’s Obvious, Minimal Language”.
- Es de código abierto y tiene actualizaciones continuas.

Formato	Estándar formal	Estándar flexible	Fuertemente tipado	Implementación fácil	Legible	Permite comentarios
TOML	Sí	No	Sí	Sí	Sí	Sí
JSON	Sí	No	Sí	Sí	Sí	No
YAML	Sí	No	Sí	No	Sí	Sí
INI	No	Sí	No	Sí	Sí	Sí

[nombre de la sección]
clave = “valor” # Comentario



[Persona]
nombre = “Álvaro” # Nombre de la persona

Conversor XML-ROH. Ficheros de configuración TOML. (2/2)

- Actualmente, TOML va por la versión 1.0.0
- TOML está implementado en más de 40 lenguajes de programación, de los cuales se incluye C#, JavaScript, Python, PHP etc.
- Algunas de sus características son:
  - Distingue entre mayúsculas y minúsculas.
  - El documento tiene que estar codificado en UTF-8.
  - La extensión del archivo resultante es .toml
- Web oficial: https://toml.io/en/
- Documentación: https://toml.io/en/v1.0.0

Conversor XML-ROH. Métodos.

http://herc-as-front-desa.atica.um.es/conversor_xml_rdf/swagger/index.html

- ConfigurationFilesList: Muestra la lista de configuraciones que tiene establecida el servicio.

- Convert: Permite convertir los ficheros XML a RDF.
  - *pType*: Indica el fichero de configuración a utilizar.
  - *pXmlFile*: Selección del XML a convertir.

Conversor XML-ROH. Configuración TOML. Entity (1/2).

Contiene la información de la entidad a tratar.

- *rdftype* (string) → Tipo de la entidad en la ontología.
- *rdftypeproperty* (string) → Nodo del cual habrá que obtener el tipo de la entidad.
- *id* (string) → Atributo identificador del nodo.
- *nameSpace* (string) → Espacio de nombre del nodo.
- *source* (string) → Nodo al que corresponde en el XML.
- *property* (string) → Propiedad a la que hay que acceder y se encuentra en un nodo padre.
- *datatype* (string) → Tipo de dato de la propiedad que hay que acceder.
- *mappingrdftype* (Mapping[]) → Mapa para obtener el tipo.
- *properties* (Property[]) → Lista de propiedades.
- *subentities* (Subentity[]) → Lista de subentidades.

Conversor XML-ROH. Configuración TOML. Entity (2/2).

```
[[entities]]
rdftype = "http://purl.org/roh/mirror/foaf#Person"
nameSpace = "https://www.openaire.eu/cerif-profile/1.1/"
source = "ns:Person"
id = "@id"
```

*Con “@ + nombre_atributo_del_nodo” se accede a dicho atributo.

Conversor XML-ROH. Configuración TOML. Mapping.

Se utiliza para indicar si una entidad es un subtipo de una herencia.

- *nameSpace* (string) → Espacio de nombre del nodo.
- *source* (string) → Contenido del nodo del XML.
- *target* (string) → Subtipo de la entidad.

```
[[entities.mappingrdftype]]
nameSpace = "https://www.openaire.eu/cerif-profile/vocab/COAR_Publication_Types"
source = "http://purl.org/coar/resource_type/c_6501"
target = "http://purl.org/roh/mirror/bibo#AcademicArticle"
```

Conversor XML-ROH. Configuración TOML. Property.

Indica si la entidad tiene una propiedad.

- *property* (string) → Tipo de la propiedad de la ontología.
- *source* (string) → Nodo al que corresponde en el XML.
- datatype (string) → Tipo de la propiedad.

[[entities.properties]]

property = “http://purl.org/roh/mirror/foaf#name”

source = “ns:PersonName”

Conversor XML-ROH. Configuración TOML. Subentity.

Propiedad que apunta a otra entidad.

- *property* (string) → Tipo de la propiedad a la que apunta.
- *inverseProperty* (string) → Tipo de la propiedad inversa.
- *entities* (Entity[]) → Lista de entidades.

[[entities.subentities]]

property = "http://purl.org/roh/mirror/vivo#relatedBy"

inverseProperty = http://purl.org/roh/mirror/vivo#relates

Demo: http://herc-as-front-desa.atica.um.es/conversor_xml_rdf/swagger/index.html

https://oamemtfp.uci.ru.nl/metis-oaipmh-endpoint/OAIHandler?verb=GetRecord&identifier=oai%3ametis.ru.nl%3aProjects

Conversor XML-ROH. Demo

XML de un proyecto: https://oamemtfp.uci.ru.nl/metis-oaipmh-endpoint/OAIHandler?verb=GetRecord&identifier=oai%3ametis.ru.nl%3aProjects%2fIGMD+6+-+Paediatrics&metadataPrefix=oai_cerif_openaire

Conversor: https://localhost:44313/swagger/index.html

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
UNIVERSIDAD DE MURCIA
FONDO EUROPEO DE DESARROLLO REGIONAL (FEDER)
Una manera de hacer Europa
Unión Europea

Hércules ASIO. Descubrimiento en ASIO.

gnoss
Deusto
Universidad de Deusto
Deustuko Unibertsitatea
University of Deusto

Descubrimiento en ASIO

Una vez cargados los datos desde el SGI, se ejecutan los siguientes procesos de descubrimiento:

- ❑ Reconciliación.
- ❑ Descubrimiento de enlaces.
- ❑ Detección de equivalencias.

Descubrimiento en ASIO

Una vez cargados los datos desde el SGI, se ejecutan los siguientes procesos de descubrimiento:

- Reconciliación.
  - Evita la duplicación de entidades mediante un conjunto de reglas.
  - Toma decisiones autónomas si la evaluación de las reglas supera un umbral.
  - Solicita la validación del usuario si la evaluación queda en un rango de duda.
  - Utiliza datos obtenidos en el descubrimiento de enlaces, desde fuentes externas (ORCID, DOI, etc.) y desde Unidata.
- Descubrimiento de enlaces.
- Detección de equivalencias.

Descubrimiento en ASIO

Una vez cargados los datos desde el SGI, se ejecutan los siguientes procesos de descubrimiento:

- Reconciliación.
- Descubrimiento de enlaces.
  - Obtención de identificadores.
  - Enriquecimiento con enlaces a fuentes externas y/o Unidata.
  - Información para la reconciliación.
  - Proceso con ejecución continua que no se ejecuta sólo en el proceso de carga.
- Detección de equivalencias.

Descubrimiento en ASIO

Una vez cargados los datos desde el SGI, se ejecutan los siguientes procesos de descubrimiento:

- ❑ Reconciliación.
- ❑ Descubrimiento de enlaces.
- ❑ Detección de equivalencias.
  - ❑ Obtención de enlaces a entidades de otros nodos ASIO.
  - ❑ Uso del nodo Unidata.
  - ❑ Información para la reconciliación.

Descubrimiento en ASIO

En resumen, las funciones de el API Descubrimiento, que son parte del proceso de carga, se dividen en 3 grupos:

- Reconciliación de entidades. Evita la duplicación de entidades, detecta las entidades ya cargadas en el nodo ASIO y evita que se carguen entidades que ya están cargadas con otra URI.
- Descubrimiento de enlaces. Genera enlaces hacia datasets externos (incluidos los de otros datasets ASIO a través del nodo Unidata), puede incorporar datos en ASIO y ofrece información de ayuda en la reconciliación de entidades.
- Detección de equivalencias. Detecta equivalencias con los datos cargados en el nodo Unidata.

Los 3 grupos de funciones actúan en el proceso de descubrimiento para todos los datos a cargar en ASIO.

Descubrimiento en ASIO

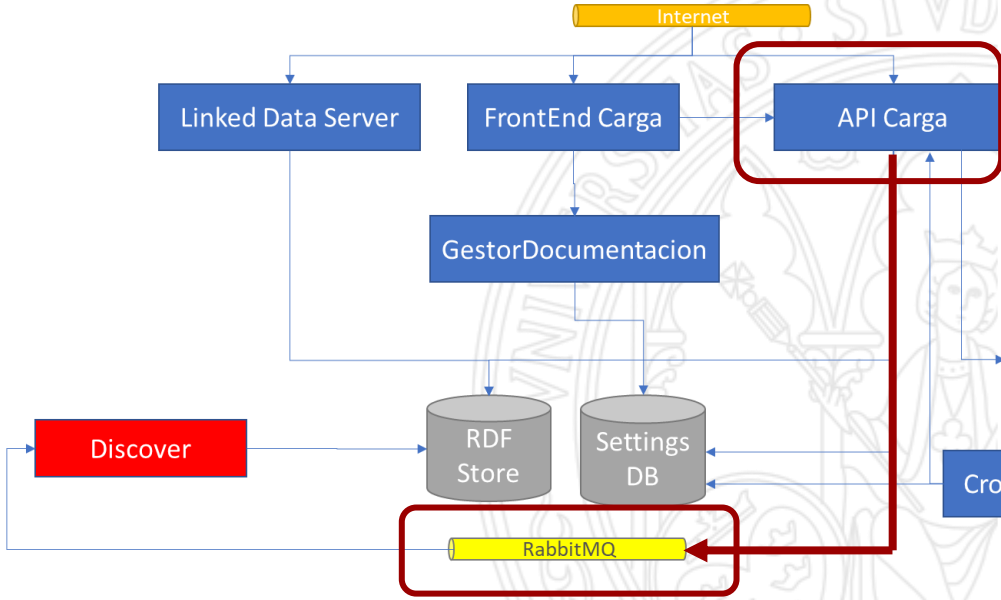
El servicio de descubrimiento tiene dos procesos diferenciados:

- Proceso general de descubrimiento. Este proceso se ejecuta cada vez que se carga un RDF en el sistema.
- Enriquecimiento continuo. Este proceso está ejecutándose continuamente aplicando el descubrimiento de enlaces a los datos que ya están cargados en el sistema.

Descubrimiento en ASIO. Proceso general de descubrimiento.

Este proceso se ejecuta cada vez que se carga un RDF a través del API de Carga con los datos provenientes de un repositorio OAI-PMH transformados a RDF con el servicio XML-RDF.

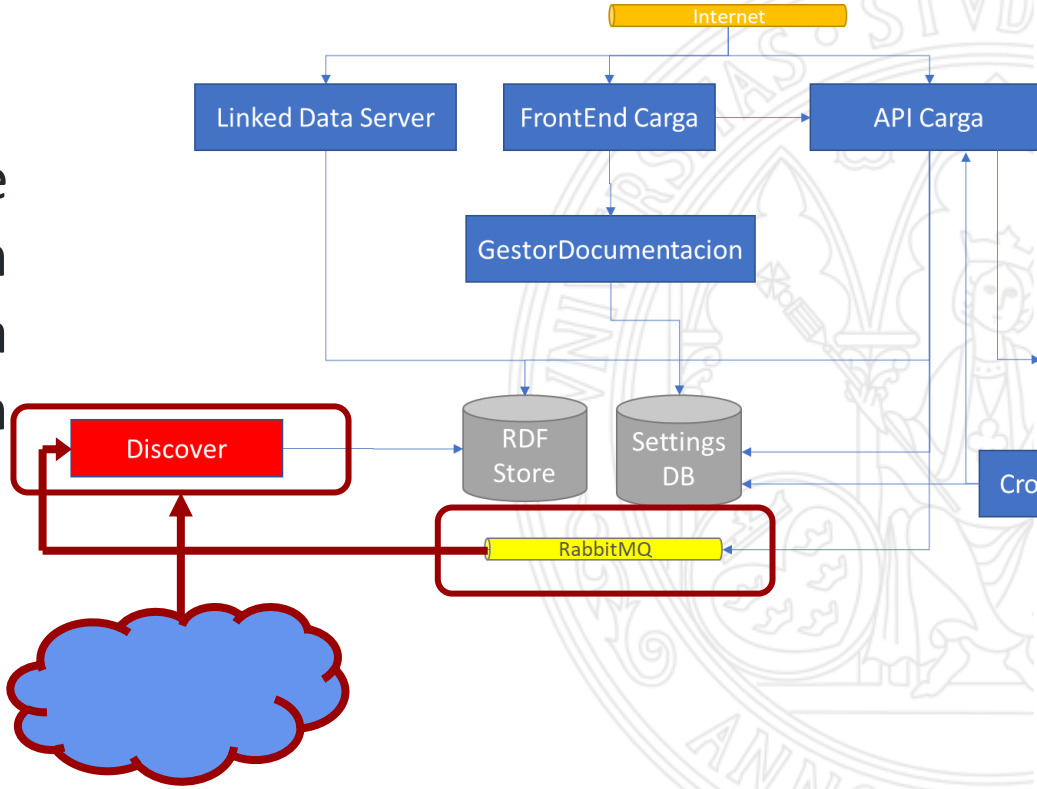
El API de Carga añade los RDFs a una cola de RabbitMQ.



Descubrimiento en ASIO. Proceso general de descubrimiento.

El Servicio de Descubrimiento procesa cada RDF transformado, leído desde la cola de Rabbit.

El Servicio trabaja con cada RDF en memoria y le aplica los tres procesos de forma iterativa, que van enriqueciendo/modificando el RDF, hasta que llega una iteración en la que no se detecta ningún enriquecimiento/modificación adicional.



Descubrimiento en ASIO. Proceso general de descubrimiento.

Como resultado de la aplicación del proceso de descubrimiento pueden suceder dos cosas:

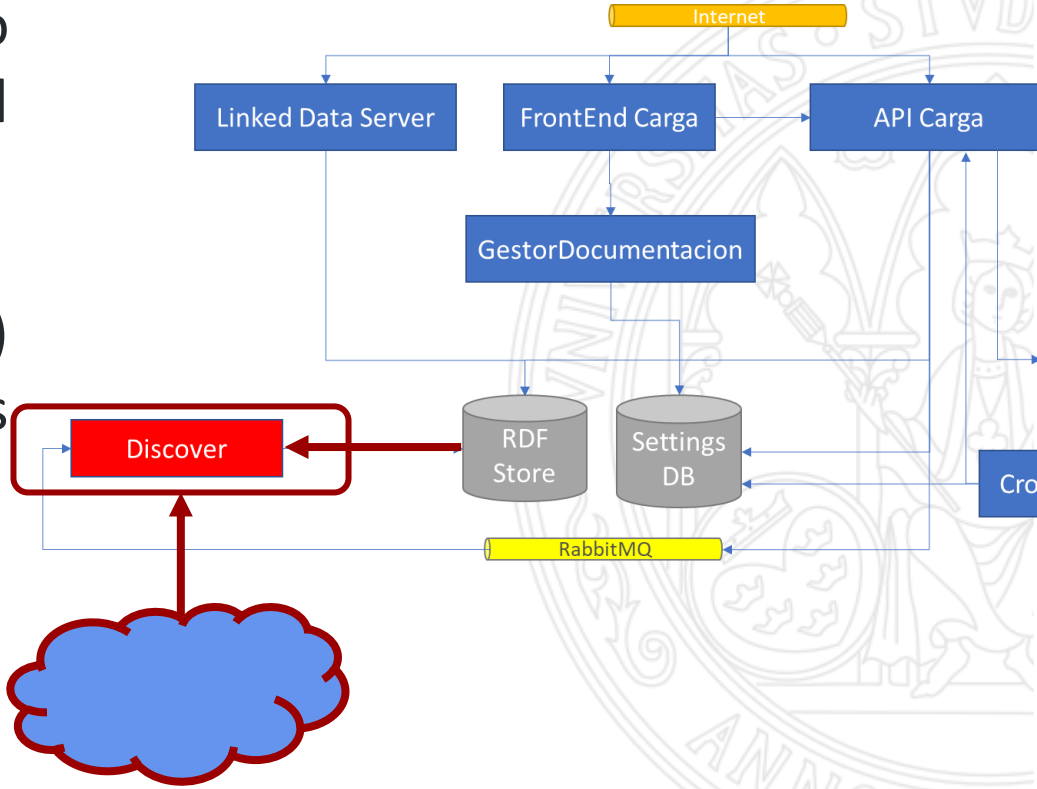
- ❑ Si no hay ningún problema de desambiguación con el proceso de reconciliación se envía el RDF al RDF Store.
- ❑ Si hay algún problema de desambiguación el RDF no se publica y se queda a la espera de que se resuelva el problema de desambiguación de forma manual.

Ejemplo de problema de desambiguación:
https://herc-as-front-desa.atica.um.es/carga-web/Job/274?repository_id=14e66ce5-7bbf-44f4-8ad4-e4019f34edad

Descubrimiento en ASIO. Enriquecimiento continuo.

El enriquecimiento continuo se ejecuta con una periodicidad establecida, aplicando el descubrimiento de enlaces a los datos que ya están cargados en el grafo.

El objetivo es detectar y añadir enlaces (descubrir) que hayan sido introducidos en los sistemas externos y en Unidata posteriormente a su alta en el grafo.



GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
UNIVERSIDAD DE MURCIA
FONDO EUROPEO DE DESARROLLO REGIONAL (FEDER)
Una manera de hacer Europa
Unión Europea

Hércules ASIO. Reconciliación de entidades.

gnoss
Deusto
Universidad de Deusto
Deustuko Unibertsitatea
University of Deusto

Reconciliación de entidades.

El proceso de reconciliación tiene como entrada un RDF generado a partir de un XML proveniente de la sincronización con un repositorio OAI-PMH.

El proceso de reconciliación de entidades estará apoyado por el proceso de descubrimiento de enlaces y de equivalencias.

Para llevara cabo la reconciliación de entidades se utilizan las configuraciones establecidas en el fichero config/reconciliationConfig.json. En el que se especifican las reglas para llevar a cabo la desambiguación entre entidades.

Reconciliación de entidades. Configuraciones para la reconciliación

En el fichero config/reconciliationConfig.json se configuran las reglas por tipo de entidad (rdf:type) y se tiene en cuenta la herencia de clases.

Cada configuración tiene 3 propiedades:

1. rdfType: Clase a la que afecta la configuración.
2. identifiers: Propiedades utilizadas como identificador.
3. properties: Propiedades para detectar el grado de igualdad de las entidades.

```
{
        "rdfType": "http://purl.org/roh/mirror/foaf#Person",
        "identifiers": [
                "http://purl.org/roh/mirror/vivo#identifier",
                …
        ],
        "properties": [
        {
                …
        },
        {
                …
        }
}
```

Reconciliación de entidades. Configuraciones para la reconciliación

Dentro de cada uno de los elementos de la propiedad ‘properties’ se establecen varios parámetros para detectar el grado de igualdad de las entidades:

1. property: Propiedad a tener en cuanta en al reconciliación ('@@@' implica un 'salto' y '?' implica que puede ser cualquier propiedad)
2. mandatory: Indica si el cumplimineto de esa propiedad es condición necesaria para considerar a dos entidades la misma
3. inverse: Si vale 'false' se buscan los valores de las propiedades utilizando la entidad como sujeto, si vale 'true' se buscan los valores de las propiedades utilizando la entidad como objeto

Reconciliación de entidades. Configuraciones para la reconciliación

4. type: Es el tipo de igualdad que se debe cumplir
   - 0 (equals): Misma entidad o mismo valor de la propiedad
   - 1 (ignoreCaseSensitive): Mismo valor de la propiedad (ignorando mayúsculas y minúsculas)
   - 2(name): Uso del algoritmo para nombres (para nombres de personas)
   - 3(title): Uso del algoritmo para títulos (para títulos de documentos por ejemplo)

5. maxNumWordsTitle: En el caso de que type tenga como valor '3' implica el número de palabras que debe tener el título para considerar el máximo valor de igualdad
6. scorePositive: Score positivo que se da a la relación cuando se da una coincidencia.
7. scoreNegative: Score negativo que se da a la relación cuando no se da una coincidencia y ambas entidad tienen algún valor para esa propiedad.

Reconciliación de entidades. Configuraciones para la reconciliación

A continuación se muestra un fragmento para la configuración para las entidades de tipo 'http://purl.org/roh/mirror/foaf#Person'

```
{
    "rdfType": "http://purl.org/roh/mirror/foaf#Person", //Entidad a la que afecta
    "identifiers": [//Listado de propiedades que actúan como identificadores
      "http://purl.org/roh/mirror/vivo#identifier",
      "http://purl.org/roh/mirror/vivo#eRACommonsId",
      "http://purl.org/roh/mirror/vivo#researcherId",
      "http://purl.org/roh#ORCID"
    ],
    "properties": [
      {
        "property": "http://purl.org/roh/mirror/foaf#name", //Propiedad
        "mandatory": true,            //Obligatoria
        "inverse": false,             //Propiedad directa
        "type": 2,                    //Algoritmo de nombres
        "scorePositive": 0.89,        //Score positivo
        "scoreNegative": null         //Score negativo
      },
```

Reconciliación de entidades. Configuraciones para la reconciliación

```
    {
      "property": "http://purl.org/roh#participates", //Propiedad
      "mandatory": false,      //No obligatoria
      "inverse": false,                  //Propiedad directa
      "type": 0,                         //Igualdad por valor
      "scorePositive": 0.5,    //Score positivo
      "scoreNegative": null    //Score negativo
    },
    {
      "property": "http://purl.org/roh/mirror/foaf#mbox", //Propiedad
      "mandatory": false,                //No obligatoria
      "inverse": false,                  //Propiedad directa
      "type": 0,                         //Igualdad por valor
      "scorePositive": 0.9,    //Score positivo
      "scoreNegative": 0.025   //Score negativo
    },
    {
      "property": "http://purl.org/roh/mirror/bibo#authorList@@@?", //Propiedad con salto y con variable
      "mandatory": false,                //No obligatoria
      "inverse": true,                   //Propiedad inversa
      "type": 0,                         //Igualdad por valor
      "scorePositive": 0.9,    //Score positivo
      "scoreNegative": null    //Score negativo
    }
  ]
}
```

Reconciliación de entidades. Flujo

1. Se lee el RDF y se obtienen todas las entidades para realizar la reconciliación.

2. Para cada una de las entidades se hace una consulta al grafo RDF del RDF Store para obtener posibles candidatos para la reconciliación con las propiedades que estén configuradas para llevar a cabo la desambiguación.
    1. Se busca si existe alguna entidad con la misma URI.
    2. Si no se ha encontrado la entidad y la entidad cuenta con algún identificador, se intenta reconciliar la entidad a través de alguno de sus identificadores. Si existe alguna entidad cargada en el nodo ASIO que comparta identificador se considera la misma entidad.
    3. Si no se ha encontrado la entidad, se buscan similitudes con entidades ya cargadas. Para cada tipo de entidad utilizan las reglas establecidas en el fichero reconciliationConfig.json.

Reconciliación de entidades. Flujo

3. Una vez obtenidos todos los candidatos se aplican las reglas de cálculo de reconciliación para obtener las entidades finales ya cargadas. Este punto además es apoyado por el descubrimiento de enlaces

4. En función del resultado obtenido se realiza una de las siguientes acciones:
   - Si para alguna entidad hay más de un candidato que supere el umbral máximo o el umbral mínimo, se agregará el RDF a una BBDD junto con todos los datos necesarios para una revisión manual.
   - Si para alguna entidad sólo se obtiene un candidato que supere el umbral máximo, se modificará la URL de la entidad en el RDF a cargar por la URL de la entidad encontrada.
   - Se obtienen las entidades principales del RDF y se eliminan todos los triples que haya en el grafo RDF en los que aparezcan como sujeto u objeto.
   - Se eliminan todos los triples de las entidades cuyo sujeto y predicado estén en el RDF a cargar y estén marcados como monovaluados según la especificación de la ontología.
   - Se vuelcan los triples al grafo RDF.

Reconciliación de entidades. Algoritmo de nombres

• Se ha optado por una medida basada en conjuntos de caracteres, usando n-gramas y obteniendo el coeficiente de Jaccard. Los aspectos a considerar son:

- Reordenar la cadena de nombre + apellidos si aparece una coma. Por ejemplo, “Pérez Lara, Ángel” a “Ángel Pérez Lara”.
- Dividir el nombre y apellidos en sus palabras, retirando stop words (de, del, la) y guiones.
- Considerar la puntuación de las palabras con un coeficiente de Jaccard por encima de 0,5. Si no se supera, el índice resultante sería 0.
- Otorgar un peso fijo de 0,5 al reconocimiento de una inicial (“Eduardo” y “E.”).
- La puntuación de una palabra será 0 si no aparece en el orden adecuado.

Reconciliación de entidades. Algoritmo de nombres

Para "Ángel Pérez Lara" podríamos obtener los siguientes candidatos:

Superan el corte de 0,5:

- Ángela Pérez Lara: 0,90
- A. Pérez Lara: 0,83
- Miguel Pérez Lara: 0,67
- Ángel Pérez Rodríguez: 0,67
- Miguel Ángel Pérez Lara: 0,625
- Ángel Pedro Pérez Laras: 0,58

NO superan el corte de 0,5:

- Ángel Pedro Pérez Talavera: 0,46
- Ángel Pedro Pérez Calatayud: 0,46
- Ángel Yoset Lara Pérez: 0,42

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
UNIVERSIDAD DE MURCIA
FONDO EUROPEO DE DESARROLLO REGIONAL (FEDER)
Una manera de hacer Europa
Unión Europea

Hércules ASIO. Descubrimiento de enlaces.

gnoss
Deusto
Universidad de Deusto
Deustuko Unibertsitatea
University of Deusto

Descubrimiento de enlaces

Se buscan en los diferente APIs las coincidencias de publicaciones +autores.

El nombre del autor o de la publicación, por sí solos, no son suficientes para considerarlos la misma entidad. Se utilizan en el caso de que exista una relación entre los nombres de ambas entidades.

Hay diferencias en lo que se puede hacer con cada una de las APIs externas, por lo que no todas se usan del mismo modo.

Descubrimiento de enlaces. Procedencia

Los datos de fuentes externas se cargan junto con unos triples que indican su procedencia. El triple con el dato obtenido desde una fuente externa se carga en un grafo que indica su procedencia. Por ejemplo, un código ORCID recuperado desde DBLP:

Grafo: http://graph.um.es/graph/sgi

```
roh:res/researcher/id1 roh:ORCID "00000".

roh:res/agent/idAgente1
    a prov:SoftwareAgent;
    foaf:name "Algoritmo Hércules ASIO";
```

Grafo: http://graph.um.es/graph/dblp

```
roh:res/researcher/id1 prov:wasUsedBy _:bnode1.
_:bnode1
    a prov:Activity;
    rdf:predicate roh:ORCID;
    rdf:object "00000";
    prov:startedAtTime "2020-04-25T01:30:00Z"^^xsd:dateTime;
    prov:endedAtTime "2012-04-25T03:40:00Z"^^xsd:dateTime;
    prov:wasAssociatedWith roh:res/agent/idAgente1;
    prov:wasAssociatedWith roh:res/organization/dblp;
```

Descubrimiento de enlaces. Fuentes externas

Crossref (https://www.crossref.org/)

Crossref makes research outputs easy to find, cite, link, assess, and reuse. A not-for-profit membership organization that exists to make scholarly communications better.

Dentro de discover se hacen llamadas al método del API 'https://api.crossref.org/works?query.author={nombre_autor}&rows=200' y se obtienen las publicaciones que tienen como autor el parámetro pasado.

Descubrimiento de enlaces. Fuentes externas

DBLP Computer Science Bibliography (https://dblp.org/)

The dblp computer science bibliography is the on-line reference for bibliographic information on major computer science publications. It has evolved from an early small experimental web server to a popular open-data service for the whole computer science community.

Dentro de discover se hacen llamadas al método del API 'https://dblp.org/search/author/api?q={nombre_autor}&h=5' y se obtienen los autores junto con sus publicaciones, a continuación, se llama al método 'https://dblp.org/pid/{id_autor}' con los identificadores de DBLP de los autores para obtener más metadatos.

Descubrimiento de enlaces. Fuentes externas

DOAJ (https://doaj.org/)

The DOAJ (Directory of Open Access Journals) was launched in 2003 with 300 open access journals. Today, this independent database contains over 15 000 peer-reviewed open access journals covering all areas of science, technology, medicine, social sciences, arts and humanities. Open access journals from all countries and in all languages are welcome to apply for inclusion.

Dentro de discover se hacen llamadas a los métodos del API 'https://doaj.org/api/v2/search/articles/title:{nombre_documento}' y 'https://doaj.org/api/v2/search/journals/title:{nombre_documento}' y se obtienen las publicaciones junto con sus autores.

Descubrimiento de enlaces. Fuentes externas

ORCID (https://orcid.org/)

ORCID is a nonprofit organization helping create a world in which all who participate in research, scholarship and innovation are uniquely identified and connected to their contributions and affiliations, across disciplines, borders, and time.

Dentro de discover se hacen llamadas al método del API 'https://pub.orcid.org/v3.0/expanded-search?q={nombre_autor}&rows=5' para obtener los identificadores de los autores y posteriormente se hacen llamadas a los métodos del API 'https://pub.orcid.org/v3.0/{id_autor}/person' y 'https://pub.orcid.org/v3.0/{id_autor}/works para obtener metadatos de los autores y sus obras respectivamente.

Descubrimiento de enlaces. Fuentes externas

PubMed (https://pubmed.ncbi.nlm.nih.gov/)

PubMed® comprises more than 30 million citations for biomedical literature from MEDLINE, life science journals, and online books.

Dentro de discover se hacen llamadas al método del API 'https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?term={nombre_documento}&field=title&sort=relevance&retmax=10' con el que obtenemo los IDs de los documentos y posteriormente se hacen llamadas al método del API 'https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id={id_documento}&retmode=xml' con los identificadores de los documentos y obtenemos los documetnos junto con sus autores.

Descubrimiento de enlaces. Fuentes externas

Recolecta (https://recolecta.fecyt.es/)

RECOLECTA, o Recolector de Ciencia Abierta, es el agregador nacional de repositorios de acceso abierto. En esta plataforma se agrupan a todas las infraestructuras digitales españolas en las que se publican y/o depositan resultados de investigación en acceso abierto.

Dentro de discover se hacen llamadas al método del API 'https://buscador.recolecta.fecyt.es/buscador-recolecta?search_api_fulltext={nombre_documento}' y se obtienen las publicaciones junto con sus autores.

Descubrimiento de enlaces. Fuentes externas

Scopus (https://www.scopus.com/)

Scopus is the largest abstract and citation database of peer-reviewed literature: scientific journals, books and conference proceedings. Delivering a comprehensive overview of the world's research output in the fields of science, technology, medicine, social sciences, and arts and humanities, Scopus features smart tools to track, analyze and visualize research.

Dentro de discover se hacen llamadas al método del API 'https://api.elsevier.com/content/search/scopus?query=TITLE({nombre_documento})&view=COMPLETE&apiKey={ScopusApiKey}&httpAccept=application/xml' y se obtienen las publicaciones junto con sus autores, posteriormente se llama al método del API 'https://api.elsevier.com/content/author/author_id/{id_autor}?apiKey={ScopusApiKey} para obtener metadatos de los autores.

Descubrimiento de enlaces. Fuentes externas

Web of Science (http://wos.fecyt.es/)

FECYT provides access to Web of Science, the world's largest publisher-neutral citation index and research intelligence platform.

Dentro de discover se hacen llamadas al método del API 'http://search.webofknowledge.com/esti/wokmws/ws/WokSearch' con los nombres de las publicaciones y se obtienen las publicaciones junto con sus autores.

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
UNIVERSIDAD DE MURCIA
FONDO EUROPEO DE DESARROLLO REGIONAL (FEDER)
Una manera de hacer Europa
Unión Europea

Hércules ASIO. Descubrimiento de equivalencias.

gnoss
Deusto
Universidad de Deusto
Deustuko Unibertsitatea
University of Deusto

Detección de equivalencias

Este proceso utiliza la información del nodo Unidata para detectar equivalencias de entidades entre el nodo del SGI y Unidata.

Detectara las equivalencias de entidades y añadirá los triples ‘SameAs’ a las entidades para que estén relacionadas entre los nodos SGI y Unidata.

Enriquece el nodo del SGI pudiendo obtener identificadores de Unidata

Apoya el proceso de reconciliación tras enriquecer los datos de las entidades

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
UNIVERSIDAD DE MURCIA
FONDO EUROPEO DE DESARROLLO REGIONAL (FEDER)
Una manera de hacer Europa
Unión Europea

Hércules ASIO. Flujo de carga de datos.

Flujo de carga de datos

Para realizar una carga hay que realizar los siguientes pasos:

1. Localizar o crear un repositorio OAI-PMH
    1. Si está ya creado habría que configurar el conversor XML-ROH para que transforme los datos a RDF ROH.
    2. Si se crea uno nuevo se puede reutilizar alguna configuración del XML-ROH o que devuelva directamente RDF en formato ROH
2. Una vez creado y configurado el repositorio configurar la frecuencia de las actualizaciones. https://herc-as-front-desa.atica.um.es/carga-web/RepositoryConfig
3. Esperar a que se realiza la carga e intervenir si existen problemas de desambiguación https://herc-as-front-desa.atica.um.es/carga-web/Job/274?repository_id=14e66ce5-7bbf-44f4-8ad4-e4019f34edad

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
UNIVERSIDAD DE MURCIA
FONDO EUROPEO DE DESARROLLO REGIONAL (FEDER)
Una manera de hacer Europa
Unión Europea

GRACIAS

gnoss
Deusto
Universidad de Deusto
Deustuko Unibertsitatea
University of Deusto