
**INSTITUTO TECNOLÓGICO Y DE ESTUDIOS
SUPERIORES DE MONTERREY
CAMPUS MONTERREY**

INFERENCIA DE REDES DE COLABORACIÓN

GReCo

CONTENIDO

- ❑ **Introducción**
 - ❑ Inferencia de afinidades
 - ❑ Objetivo
- ❑ **Contexto**
- ❑ **Modelo de Solución**
- ❑ **Resultados**

INTRODUCCION

- ✘ Las bibliotecas digitales son un conjunto de recursos electrónicos y capacidades técnicas asociadas para crear, buscar y utilizar información, en este sentido el contenido de las bibliotecas digitales incluye gran variedad de datos y metadatos que describen diversos aspectos de sus contenidos.

La inferencia de afinidades en colecciones de documentos pertenecientes a bibliotecas digitales involucra el determinar mediante ciertas técnicas la similitud de contenido en los documentos y poder cuantificar esa similitud.

- ✘ El objetivo es poder representar dicha afinidad en términos de una red de colaboración, red capaz de expresar el grado de similitud entre documentos.

CONTEXTO

El desarrollo de este trabajo se basa en las colecciones de documentos provenientes de los repositorios institucionales de tres miembros de RABiD:

- + CIRIA (UDLAP).
- + Phronesis (ITESM).
- + Redalyc (UAEM).

Por ser parte de RABiD las características de estos repositorios son:

- + Pertenecen a una red abierta por medio de la cual puedan compartirse colecciones y servicios.
- + Son proveedores de datos de la iniciativa de archivos abiertos (OAI), facilitando la recuperación del contenido de documentos bajo el protocolo estándar para cosecha de metadatos (OAI-PMH).

-
- × El protocolo OAI-PMH hace uso de los metadatos Dublin Core (DC), es el formato recomendado por OAI usado para describir el contenido, nombre, título y otras características de los recursos disponibles en una biblioteca digital.
 - × La figura 1 muestra los metadatos: <dc:title>, <dc:description> y <dc:creator> .

Estos metadatos almacenan información del contenido del documento para un registro típico de un documento digital tomada de la colección de Phronesis.


```

<record>
<header>
<identifier>ITESMMTY200042</identifier>
<datestamp>2002-01-14</datestamp>
<setSpec>MTY</setSpec>
</header>
<metadata>
<oai_dc:dc
  xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:title>
Identificación de las tecnologías claves que tienen mayor impacto en la administración de la
cadena de proveeduría, en la aplicación del concepto de empresa extendida
</dc:title>
<dc:identifier>http://copernico.mty.itesm.mx/phronesis/mty/busqueda/bajarOAI.cgi?filename=ITESMM
TY199913.ps</dc:identifier>
<dc:date>2000-12-12</dc:date>
<dc:description>
El objetivo de esta investigación fue la identificación de las tecnologías de información que son
claves y tienen mayor impacto en la empresa de manufactura en la habilitación de los procesos
llevados a cabo en la administración de la cadena de proveeduría en el concepto de empresa
extendida.
En el desarrollo del estudio se describe un modelo que facilita la definición e identificación de
los procesos y tecnologías utilizadas en la administración de la cadena de proveeduría (Supply
Chain Management - SCM) bajo el concepto de empresa extendida. De tal forma, que se
identifican tecnologías desde una perspectiva generalizada y no "productos comerciales" como
soluciones de mercado, ya que estos pueden ser variados e integrar tecnologías de distintas
formas.
</dc:description>
<dc:creator> José Vladimir Burgos Aguilar</dc:creator>
<dc:language>en</dc:language>
<dc:subject>Innovación y competitividad Innovación y tecnologías de información y de
comunicaciones</dc:subject>
<dc:publisher>Instituto Tecnológico y de Estudios Superiores de Monterrey</dc:publisher>
<dc:publisher>ITESM</dc:publisher>
<dc:publisher>Mexico</dc:publisher>
<dc:format>application/pdf</dc:format>
<dc:source>ITESMMTY200042.pdf</dc:source>
<dc:contributor> </dc:contributor>
</oai_dc:dc>
</metadata>
</record>

```

Figura 1. Ejemplo de un registro típico de una tesis digital

MODELO DE SOLUCIÓN

- ✘ La figura 2 describe la metodología que consta de tres etapas:

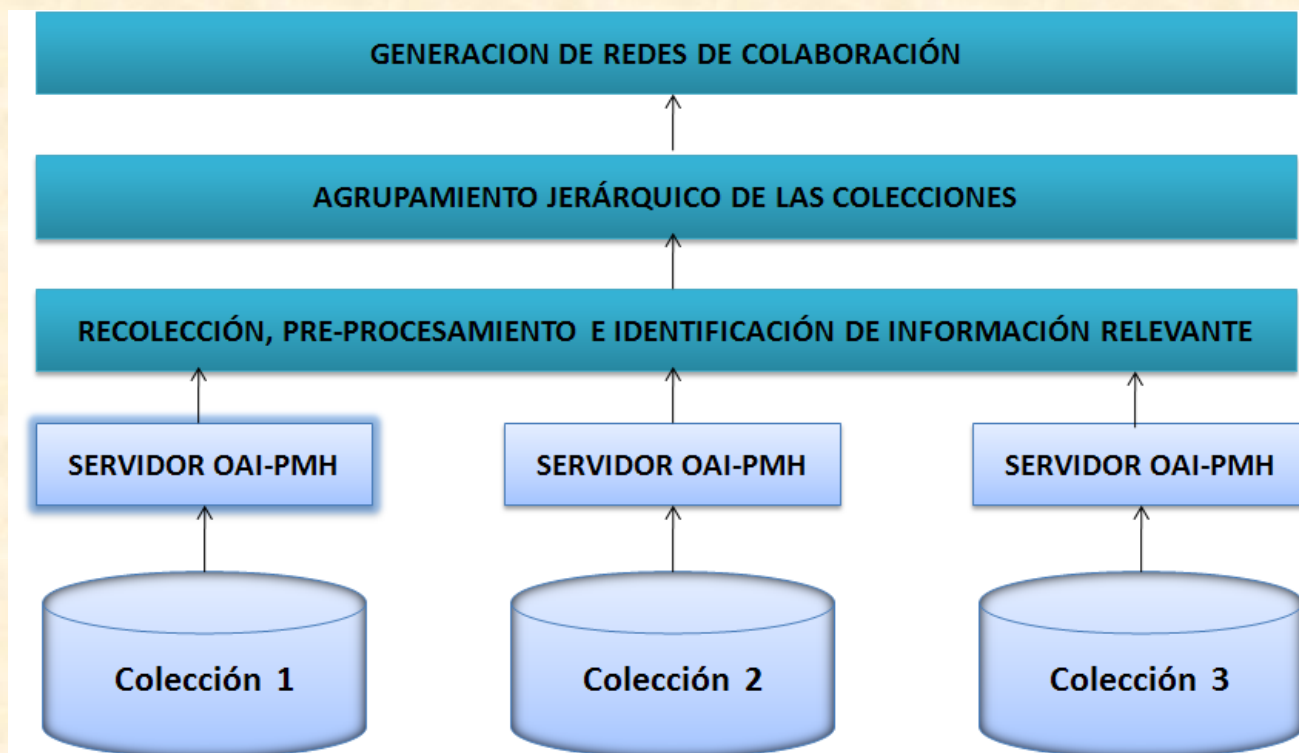


Figura 2. Metodología

- ✘ ETAPA 1. Recolección, pre-procesamiento e identificación de información relevante de las colecciones de documentos (CIRIA, Phronesis y Redalyc) . Esta etapa nos permite generar la entrada para el algoritmo de clasificación temática y almacenar información relevante de los documentos, podemos observar lo anterior en la figura 3.

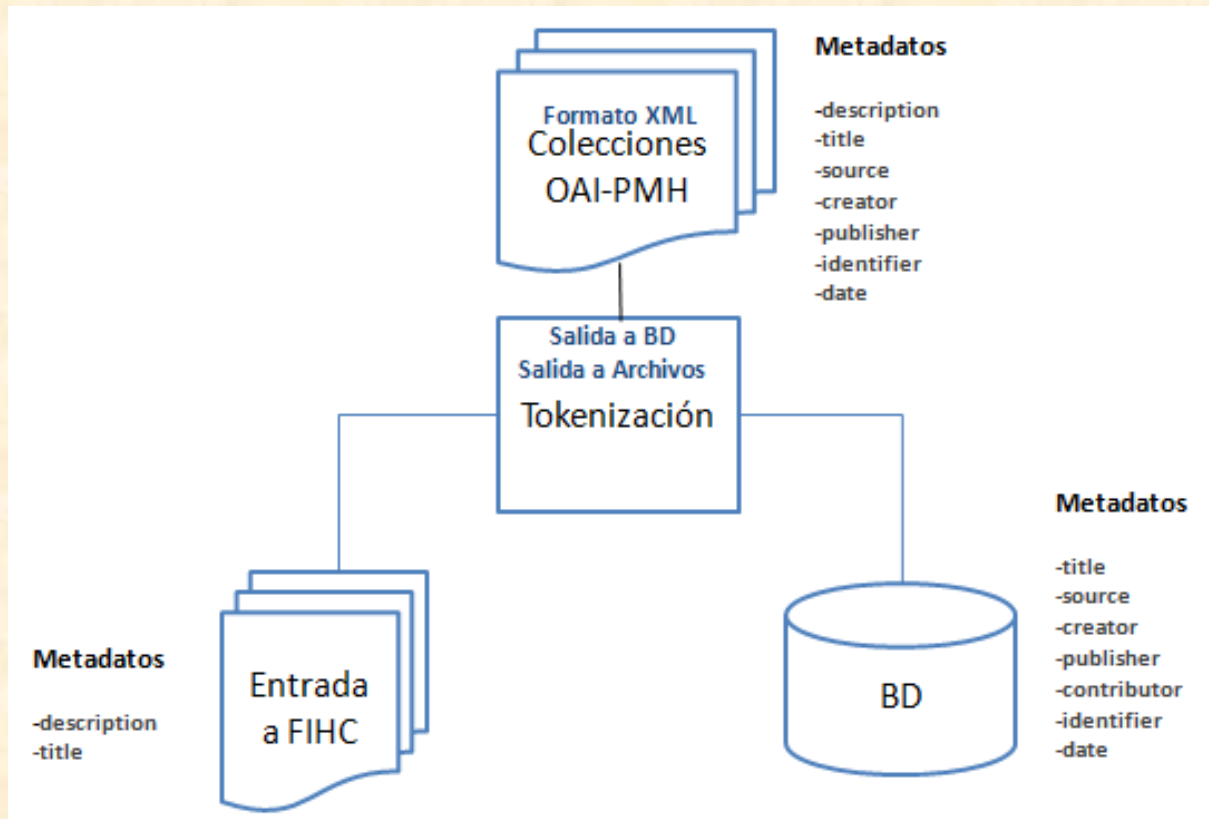


Figura 3. Etapa 1

- ✘ ETAPA 2: Agrupamiento jerárquico de las colecciones a través de la implementación del algoritmo de agrupamiento jerárquico basado en la frecuencia de un conjunto de elementos (FIHC), con el objetivo de agrupar los documentos de acuerdo a su similitud de contenido (figura 4):

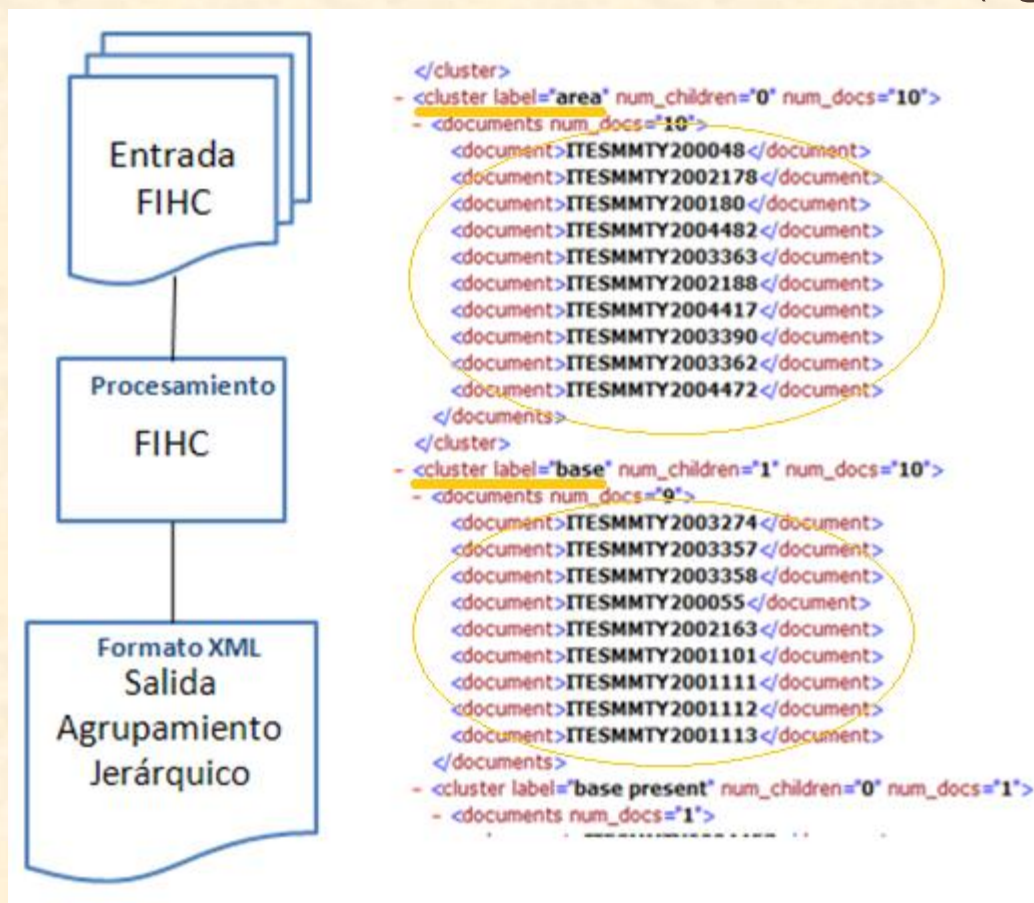


Figura 4. Agrupamiento jerárquico

El agrupamiento jerárquico que se obtiene a partir del uso del FIHC corresponde a un árbol de agrupamiento de documentos (figura 5) que ofrece como ventajas la facilidad de búsqueda y etiquetado de grupos:

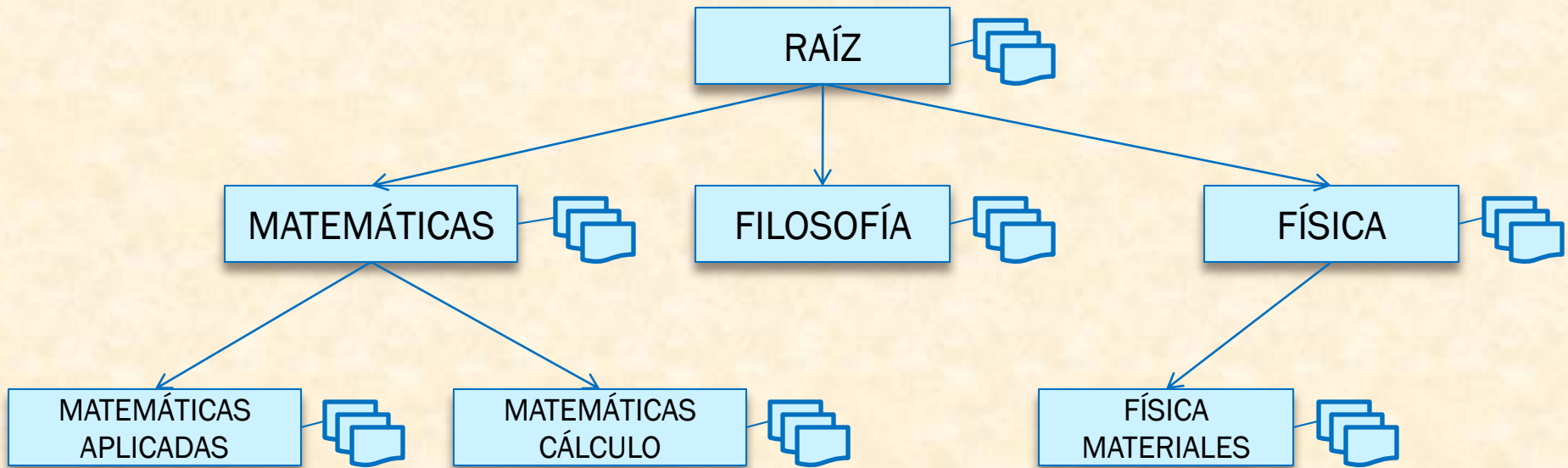


Figura 5. Árbol de agrupamiento de documentos

- ✘ ETAPA 3: Generación de redes de colaboración mediante la adecuación del uso del algoritmo de Xiaoming Liu utilizado originalmente para obtener una red de coautoría, obtenemos una red de colaboración que expresa en términos cuantitativos la similitud de entre documentos obtenida en la etapa 2, lo anterior la podemos observar en figura 6:

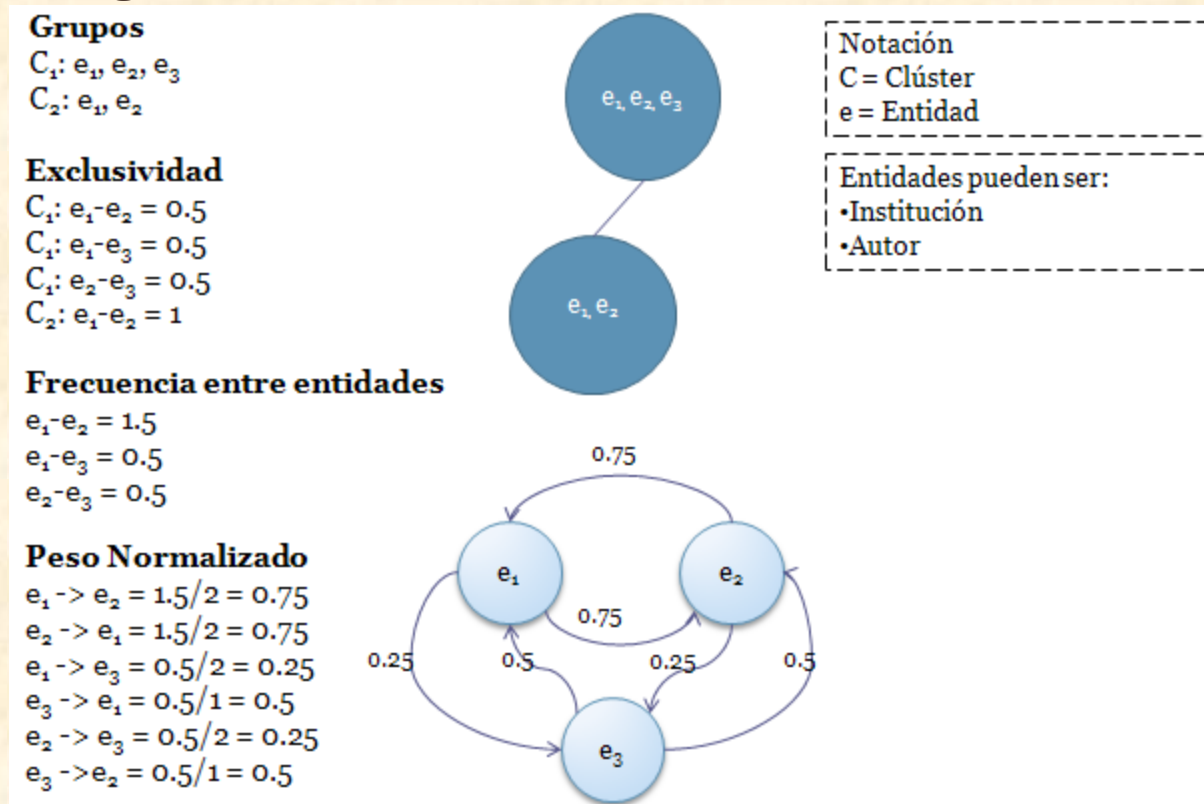


Figura 6. Red de colaboración

- **Frecuencia de relación entre entidades**

- Entidades que con frecuencia tengan relación podrían tener un mayor peso en su relación.

- **Número total de entidades en los grupos**

- Si un grupo cuenta con muchas entidades, la posible relación entre entidad tendría un peso menor (debido a que hay varias opciones para establecer una relación).
- La notación de que usaremos para las entidades será $E = (e_1, e_2, e_3, \dots, e_n)$ y para los clusters $C = (c_1, c_2, c_3, \dots, c_k, \dots, c_m)$ y $f(c_k)$ será el número de entidades en un grupo c_k .

- **Exclusividad**

- Representa el grado de relación existente entre una entidad e_i y una entidad e_j en un grupo en particular c_k . Dando más peso de relación en grupos con pocas entidades .y menor peso en el caso contrario.

$$g_{i,j,k} = 1/(f(c_k) - 1)$$

- **Frecuencia de relación**

- Consiste en la suma de todos los valores de la relación $g_{i,j,k}$ para todos los grupos donde estén relacionadas las entidades e_i y e_j .

$$S_{i,j} = \sum_{k=1}^m g_{i,j,k}$$

- **Normalización del peso**

- Se asegura que los pesos de las relaciones de una entidad sumen uno.

$$W_{i,j} = S_{i,j} / \sum_{k=1}^n S_{i,k}$$

En base resultados obtenidos en el paso anterior y la información almacenada en la BD en la etapa 1, podemos generar la red de temática de colaboración deseada con toda la información relevante asociada a la entidad (Figura 7).

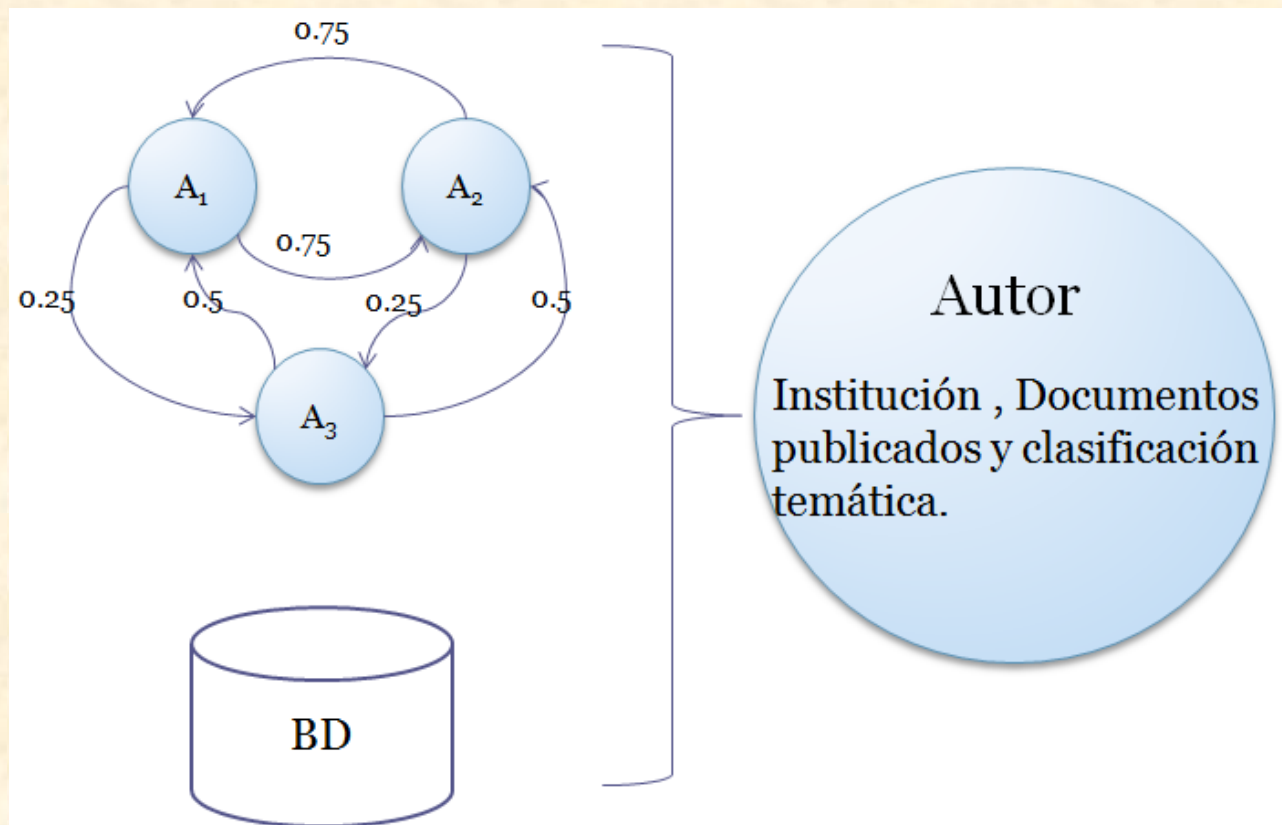


Figura 7. Información relevante para la entidad Autor

RESULTADOS

- ✘ El desarrollo de una herramienta que permita a los miembros de RABiD obtener a partir de sus colecciones digitales obtener una red de colaboración a nivel de instituciones y a nivel de autores.
- ✘ Las figuras 8 y 9 muestran la estructura que representa la red de colaboración a nivel de instituciones y a nivel autores respectivamente.

```

<?xml version="1.0" encoding="utf-8"?>
<overview xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="SN.xsd" >
<institutionQuery>
  <dataset>
    <institution id="01" acronym="UDLAP" name="Universidad De Las America Puebla" country="Mexico" >
      <author name="Jose Maria Morelos de la O">
        <subject>Tema 2</subject>
      </author>
      <author name="Antonio Presidio">
        <subject>Tema 8</subject>
      </author>
      <author name="Daniela Romo">
        <subject>Tema 3</subject>
      </author>
    </institution>
    <institution id="02" acronym="UAEM" name="Universidad Autonoma del Estado de Mexico" country="Mexico" >
      <author name="Victor Marcelo Ebrat">
        <subject>Tema 3</subject>
      </author>
      <author name="Marcelino Cornelio">
        <subject>Tema 10</subject>
      </author>
    </institution>
  </dataset>
  <relationSet>
    <relation from="01" to="02" strength="34" normalizedStrength="0.15" />
    <relation from="02" to="01" strength="98" normalizedStrength="0.293" />
  </relationSet>
</institutionQuery>

```

Figura 8. Red de colaboración a nivel de instituciones

```

<?xml version="1.0" encoding="utf-8"?>
<overview xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="TN.xsd" >
<AuthorQuery>
  <dataset>
    <author id="mx01" name="Jose Maria Morelos de la O" Institution="Universidad De Las Americas Puebla" acronym="UDLAP" country="Mexico">
      <document title="Tema 2-30 mx" url="www.repositoriy.org/files/fileNN" date="No date">
        <subject>Tema 2</subject>
      </document>
      <document title="Tema 46 mx" url="www.repositoriy.org/files/fileNN" date="No date">
        <subject>Tema 46</subject>
      </document>
      <document title="Tema 30-50 mx" url="www.repositoriy.org/files/fileNN" date="No date">
        <subject>Tema 30</subject>
      </document>
    </author>
    <author id="mx02" name="Marcelino Cornelio" Institution="UAEM" acronym="Universidad Autonoma del Estado de Mexico" country="Mexico">
      <document title="Tema 10-20-28 mx" url="www.repositoriy.org/files/fileNN" date="No date">
        <subject>Tema 10</subject>
      </document>
      <document title="Tema 40-41 mx" url="www.repositoriy.org/files/fileNN" date="No date">
        <subject>Tema 40</subject>
      </document>
    </author>
  </dataset>
  <relationSet>
    <relation from="mx01" to="mx02" strength="34" normalizedStrength="0.15" />
    <relation from="mx02" to="mx01" strength="98" normalizedStrength="0.293" />
  </relationSet>
</AuthorQuery>

```

Figura 9. Red de colaboración a nivel de autores

REFERENCIAS

- ✘ [1] C. L. Borgman. Challenges in building digital libraries for the 21(st) century. In E. P. Lim, S. Foo, C. Khoo, S. Urs, T. Costantino, E. Fox, and H. Chen, editors, *5th International Conference on Asian Digital Libraries (ICADL 2002)*, volume 2555, pages 1{13, Singapore, Singapore, 2002. Springer-Verlag Berlin. 29 BERLIN BW29S.
- ✘ [2] W.B. Frakes and R. Baeza-Yates. *Information retrieval: data structures and algorithms*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1992.
- ✘ [3] B. C. M. Fung, K. Wang, and M. Ester. Hierarchical document clustering using frequent itemsets. In D. Barbara and C. Kamath, editors, *3rd SIAM International Conference on Data Mining*, pages 59{70, San Francisco, Ca, 2003. Siam. 19 PHILADELPHIA BX26M.
- ✘ [4] L. Iverson. Collaboration in digital libraries: A conceptual framework. In H. Chen, M. Christel, and E. P. Lim, editors, *4th Joint Conference on Digital Libraries*, pages 380-380, Tucson, AZ, 2004. Assoc Computing Machinery. 2 NEW YORK BAM92.
- ✘ [5] A. K. Jain, M.Ñ. Murty, and P. J. Flynn. Data clustering: A review. *Acm Computing Surveys*, 31(3):264-323, 1999. 204 ASSOC COMPUTING MACHINERY NEW YORK 302KZ.
- ✘ [6] X. M. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel. Co-authorship networks in the digital library research community. *Information Processing and Management*, 41(6):1462-1480, 2005. 36 PERGAMON- ELSEVIER SCIENCE LTD OXFORD 956XE.
- [7] B.Y. Ricardo, R.N. Berthier, et al. Modern information retrieval. *England: Pearson Education Limited*, 1999.