

MODELADO DE DATOS DE UN SISTEMA DE INFORMACIÓN PARA LA PRODUCCIÓN CIENTÍFICA BASADO EN EL ESTÁNDAR CERIF USANDO MICROSOFT EXCEL

Sinaí López Castillo

Universidad de Guanajuato

s.lopez.castillo@ugto.mx

María Susana Ávila García

Universidad de Guanajuato

susana.avila@ugto.mx

Isnardo Reducindo

Universidad Autónoma de San Luis Potosí

isnardo.reducindo@uaslp.mx

Resumen

El acelerado progreso de la ciencia en la actualidad se basa en la velocidad con la que el conocimiento científico se distribuye gracias a las herramientas tecnológicas de esta era digital. En este trabajo se presenta un prototipo de sistema de información para contribuir con la rápida y masiva propagación del conocimiento científico, modelado sobre hojas de cálculo y que permite gestionar la producción científica generada en el Departamento de Estudios Multidisciplinarios de la Universidad de Guanajuato. El desarrollo se basa en el estándar CERIF (Common European Research Information Format), que permite interoperabilidad con otros sistemas a nivel global. Como resultado, el sistema para gestionar la producción derivada de la investigación dentro del DEM cuenta ahora con datos normalizados y elimina redundancias, lo que permitirá escalar la estructura a gestores basados en SQL y desarrollar un sistema más potente, que además contará con interoperabilidad a nivel global. Así, la información gestionada

podrá estar a disposición de usuarios externos al DEM y abonará una difusión óptima del conocimiento.

Palabra(s) Clave: Estándar CERIF, Interoperabilidad, Sistemas de información.

Abstract

Now days the fast progress of the science is based on the velocity with the scientific knowledge is distributed around the world, thanks to the technological tools to this digital era. This work presents an information system prototype in order to contribute with the fast and massive scientific knowledge propagation, modeled using spreadsheets and that allows manage the scientific production generated into Departamento de Estudios Multidisciplinarios de la Universidad de Guanajuato. The development is based on the CERIF (Common European Research Information Format) standard, which allows interoperability with other systems at a global level. As a result, the system which will manage the information generated of the production derived from the research in DEM now has standardized data and eliminates redundancies, which will allow to scale the structure of SQL-based managers and develop a more powerful system which will also have interoperability at a global level. Thus, the information managed may be available to users outside DEM and will allow an optimal dissemination of knowledge.

Keywords: *CERIF standard, Information systems, Interoperability.*

1. Introducción

Los Sistemas de Información para la Investigación se han convertido en una herramienta indispensable principalmente para los investigadores y estudiantes en épocas recientes. Estos sistemas les permiten desarrollar sus actividades académicas y de producción científica de manera más ágil, al contar con fuentes actuales de información de manera rápida. Dichas herramientas han sido adoptadas a nivel global de manera independiente por un gran número de investigadores, como lo menciona Joint [2008], se les denomina CRIS (Current Research Information Systems) por sus siglas en inglés.

Al igual que el resto de los sistemas de información, el gestionar la información referente a la producción científica ha llevado a la necesidad de la creación de estándares que establecen características principales que deben contener estos diseños, tal es el caso del estándar CERIF (Common European Research Information Format). Tal y como lo mencionan Nabavi & Jeffery [2016], este estándar se trata de un modelo central de arquitectura de datos, que representa al contexto de investigación mediante entidades y relaciones que permiten el intercambio de esta información; este modelo central ha sido utilizado para el diseño de muchos sistemas de información europeos. Alrededor del mundo diversas instituciones han implementado el estándar CERIF dentro de sus sistemas.

Un ejemplo de ello es el German Science System (GSS), en donde los creadores pasaron por una serie de etapas para poder adaptar su sistema al estándar. Durante estas etapas, mencionan los autores Ivanovic, Surla, Trajanovic, Misic & Konjovic [2017] que fue posible notar que el CERIF no se adapta exactamente a las necesidades de todos los sistemas. En su caso, fue necesario apoyarse en otro estándar que permitiera hacer una extensión de CERIF. De esta manera, en el GSS el estándar CERIF es una extensión sobre el grupo de datos que anteriormente se tenía para desarrollar el sistema.

Otro caso es del Ministry of Education, Science and Technological Development (MESTD) de la República de Serbia, donde Jeffery, Houssos, Jörg, and Asserson. [2014] describen como se utilizó este estándar para el modelado de datos, en este caso con una compatibilidad mayor. El hecho de que el modelado de datos de CERIF contenga información relacionada a las actividades de investigación como: investigadores, proyectos científicos, instituciones, publicaciones, patentes, programas de financiamiento, entre otras; permitió que se adaptara perfectamente a las características del CRIS del MESTD. Para lograr la compatibilidad con CERIF, fue necesario identificar las entidades relacionadas en común y establecer los atributos necesarios para completar el modelado de acuerdo a CERIF.

Por otro lado, lamentablemente en México no existe una política nacional impulsada por las instancias correspondientes como el CONACyT (Consejo

Nacional de Ciencia y Tecnología), que promueva de manera enérgica la adopción de estas herramientas tecnológicas dentro de todas las instituciones públicas cuyo propósito es la generación y propagación del conocimiento, como las universidades y los centros de investigación. Bajo este contexto, si existiera un CRIS a nivel Nacional donde todas las instituciones académicas, por ejemplo, las que conforman el RENIECYT (Registro Nacional de Instituciones y Empresas Científicas y Tecnológicas), permitiría compartir los trabajos científicos que se realizan en todas estas instituciones auspiciadas por recursos Federales. De esta manera, todos los investigadores en el país podrían estar al tanto de lo que sus pares dentro del territorio nacional están realizando. Esto abriría las puertas a un mayor número de colaboraciones en beneficio de fomentar las redes temáticas e incentivar el desarrollo nacional con la obtención de mejores resultados en las investigaciones.

El estándar CERIF busca regular el desarrollo de los CRIS agregando a estos una mayor interoperabilidad entre ellos. Es decir, con la implementación de dicho estándar es posible acrecentar el intercambio de información de la producción científica entre diversas instituciones. Para hacer posible esto, este estándar propone un modelado de datos compuesto por veinticinco entidades, clasificadas en cuatro grupos y las cuales se mencionan a continuación:

- Entidades Base: Persona, Institución Proyecto.
- Entidades Resultantes: Publicación Resultante, Patente Resultante, Producto Resultante.
- Entidades del Contexto de Investigación: se determinan mediante el enlace de las entidades base y las entidades resultantes.
- Entidades Enlace: se relacionan dos entidades independientemente del nivel en el que se encuentren, es decir, puede relacionar entidades base, resultado o de contexto de investigación.

Básicamente, el estándar CERIF establece estos cuatro grupos de entidades de manera global para facilitar la interoperabilidad. En otras palabras, puede ser utilizado dentro de instituciones muy diversas, pero el intercambio de información

estará generalizado en los cuatro grupos de entidades como bloques. Por ejemplo, el modelo de datos toma en cuenta como entidad a "Persona", esta entidad no especifica qué tipo de personas pueden ser registradas en el sistema y que roles tiene dentro de este, es decir, puede ser investigador, autor, colaborador, etc.

Un estudio comparativo realizado a nivel internacional por Pinto, Simões & Amaral. [2014], donde se localizaron 43 CRIS de todo el mundo con la finalidad de conocer las diferencias y similitudes, tanto entre los sistemas con estructura CERIF y como en aquellos que su estructura no era completamente adaptada a CERIF. Dicho estudio valida el mayor nivel de interoperabilidad de los CRIS basados en CERIF, por lo que puede ser considerado como un estándar que beneficia el desarrollo de este tipo de sistemas de información. La tabla 1 muestra las ventajas y desventajas de CERIF para los CRIS obtenidas a partir de los resultados de un estudio realizado por Jeffery & Asserson [2006]. Cabe aclarar que como los autores del estudio mencionan, en todo sistema y modelo una mala aplicación o adaptación puede desviar a cualquier sistema de su objetivo principal, por lo que se debe ser cuidadosos en su implementación.

Tabla 1 Ventajas y desventajas de CERIF.

VENTAJAS	DESVENTAJAS
<ul style="list-style-type: none">• Gran cobertura en gestión de información para la investigación, es un diseño especializado para los CRIS.• Interoperabilidad, sus entidades de enlace permiten relacionarse con cualquier otro sistema.• El modelado de datos puede extenderse, adaptándose a las características del sistema• Evita la redundancia y mejora la integridad de los datos mediante la normalización de sus tablas de clasificación.	<ul style="list-style-type: none">• Se establece que es posible ser implementado en una arquitectura relacional u orientada a objetos, pero realmente solo se adapta a uno relacional.• Todas las entidades propuestas por CERIF tienen la misma estructura y algunas de ellas no están bien definidas.

Fuente: [Jeffery & Asserson, 2006]

Hasta este punto, la información mencionada sobre CERIF vislumbra una perspectiva demasiado alta de las cosas que se pueden alcanzar con éste. Por

otro lado, aunque su diseño se adapta en gran parte a las necesidades de los CRIS, existen algunas inconsistencias detectadas por diversos autores, lo cual deja entre ver que son necesarios ajustes y mejoras para este estándar.

Partiendo de esta idea se identificó en la Universidad de Guanajuato, específicamente en el Departamento de Estudios Multidisciplinarios (DEM), la necesidad de la creación de un CRIS para los productos derivados de la investigación dentro de dicho departamento. Como la mayoría de las instituciones que producen conocimiento en México, en el DEM no se cuenta con un sistema que permita que tanto usuarios internos como externos puedan conocer los trabajos de investigación que se generan, tanto por investigadores como por los alumnos. Para dar solución a dicha necesidad, se propone un prototipo de modelado de datos sobre hojas de cálculo con la finalidad de gestionar toda la información relacionada a la producción científica del DEM, atendiendo las necesidades particulares de la institución y adaptando el modelo al estándar CERIF. El prototipo de CRIS desarrollado, puede ser una primera aproximación para buscar un sistema que pueda servir como referente para un sistema a nivel Nacional.

2. Metodología

CRIS para el DEM

Como parte de este trabajo se diseñó y desarrolló un prototipo de sistema para un CRIS que permita gestionar la producción científica del DEM de la Universidad de Guanajuato. Los CRIS involucran una serie de aspectos para ser útiles para el usuario [Zelepukhina, Danilova, Burmistrov & Tarasevich, 2014], por lo que este trabajo se enfoca en el prototipo de la base de datos que permite almacenar la información de los productos científicos que generan los investigadores. La estructura de datos fue diseñada de acuerdo a la teoría del modelo relacional [Kumar, 2011]. A partir del modelo de datos, el prototipo de sistema se desarrolló sobre hojas de cálculo, específicamente Microsoft Excel, con la intención de obtener provecho de la rápida implementación que se puede realizar sobre este software, y así poder realizar pruebas rápidamente y mejorar el diseño.

Para desarrollar el prototipo en principio se identificó la información relacionada a las actividades de producción científica dentro del DEM. Esta información fue recabada por un grupo de estudiantes del departamento quienes utilizaron dos métodos de recolección de datos: a) fuentes de información disponibles en línea, como sitios web (Universidad de Guanajuato, LinkedIn, Google Académico, etc.), y b) entrevistas semi-estructuradas realizadas a todos los investigadores adscritos al DEM. Una vez recabada la información necesaria, se procedió a realizar el modelado de datos relacional, en donde se establecieron las entidades y atributos correspondientes con los que sería posible gestionar la información dentro del sistema, como se muestra en la figura 1.

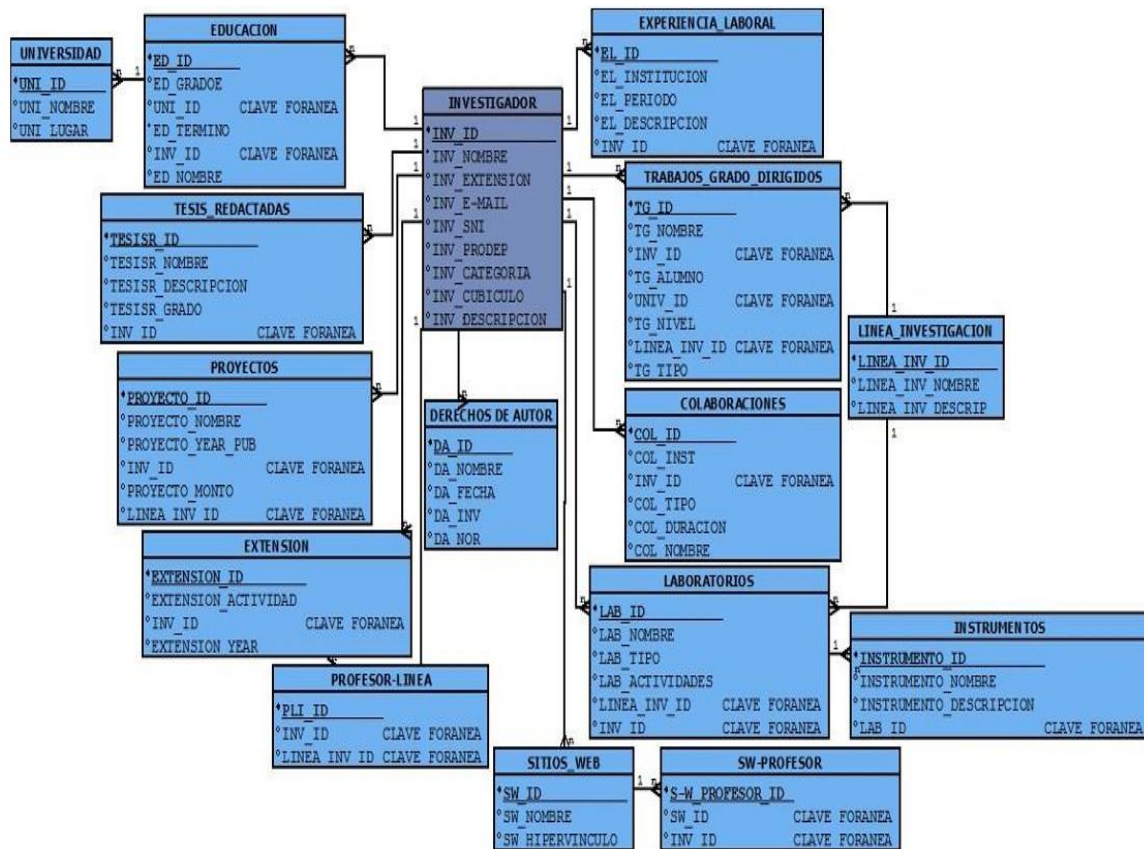
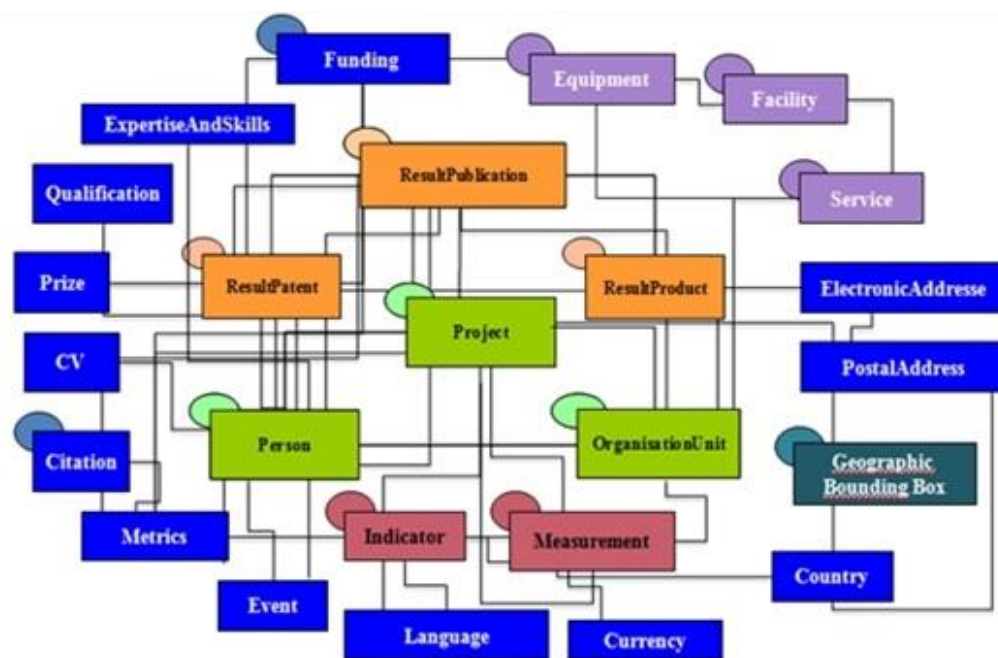


Figura 1 Modelo relacional de la Base de Datos del CRIS-DEM.

Después, se llevó el modelo de datos a Excel, donde las entidades se establecieron una por hoja y los atributos se definieron como columnas dentro de cada hoja correspondiente a su respectiva entidad. En la tabla 2 se muestra la

de modelado de datos para el CRIS del DEM aplicando el estándar CERIF. Esto con el objetivo de normalizar, dar un mayor soporte teórico y permitir la interoperabilidad del prototipo de sistema en futuras implementaciones. Para lograr dicho modelado, se analizó el diagrama de modelado de datos de CERIF, el cual se puede apreciar en la figura 3.



Fuente: [Jörg y cols., 2012]

Figura 3 Diagrama de modelado de datos para el estándar CERIF.

Una vez que analizado el diagrama de las entidades propuestas por CERIF, se procedió a adaptar las características del CRIS-DEM a un modelado de datos sobre dicho estándar. En primera instancia, se adaptaron aquellas entidades que coincidían con el estándar. En la tabla 3 se pueden apreciar las entidades que tienen en común ambos modelos.

Después, se extendieron los atributos del CRIS-DEM para que la estructura que tiene la producción científica pudiera adecuarse totalmente a CERIF. Una vez establecido como se adaptaría al estándar, se elaboró un primer diseño para el modelado de las entidades y la forma en la que éstas deberán estar relacionadas, obteniendo como resultado un diagrama Entidad-Relación, elaborado con el software de diagramación DIA, diagrama que se muestra en la figura 4.

figura 5. Esto con la intención de contar con un diseño más detallado del modelo de datos del CRIS-DEM-CERIF.

PERSON			PROJECT			ORGANIZATION		
KEY	ATTRIBUTE	DESCRIPTION	KEY	ATTRIBUTE	DESCRIPTION	KEY	ATTRIBUTE	DESCRIPTION
PK	P_ID	ID		PJ_ID	ID		O_ID	ID
	P_NAME	NAME OF PERSON		PJ_NAME	NAME OF PROJECT		O_NAME	ORGANIZATION NAME
	P_TEL	TEL NUMBER		PJ_YEAR	YEAR OF EJECTION	FK	ID_COUNTRY	COUNTRY ID
	P_EMAIL	EMAL	FK	P_ID	PERSON ID			
	P_SNI	LEVEL OF NATIONAL RESEARCH SYSTEM	FK	FD_ID	FUNDING ID			
	P_PRODEP	LEVEL OF PRODEP	FK	E&S_ID	EXPERTISE AND SKILLS ID			
	P_CATEGORY	PROFFESOR CATEGORY						
	P_DESCRIPTION	SHORT DESCRIPTION						
RESULT PATENT			RESULT PRODUCT			RESULT PUBLICATION		
KEY	ATTRIBUTE	DESCRIPTION	KEY	ATTRIBUTE	DESCRIPTION	KEY	ATTRIBUTE	DESCRIPTION
	RP_ID	ID		RPR_ID	ID		RPB_ID	ID
	RP_NAME	PATENT NAME		RPR_NAME	NAME		RPB_NAME	NAME
	RP_ASSIGNATION	PATENT ASIGNATION	FK	P_ID	PERSON ID	FK	JA_ID	JOURNAL ARTICLE ID
FK	P_ID	PERSON ID		PRP_POLICY	PRODUCT POLICY	FK	LT_ID	LEADED THESIS ID
						FK	RT_ID	REDACTED THESIS ID
						FK	BC_ID	BOOK CHAPTER ID
						FK	B_ID	BOOK ID

Figura 5 Modelo de la estructura de datos del CRIS-DEM-CERIF en formato UML.

Una vez que se cuenta con el diseño de las entidades, atributos y sus relaciones, se procedió a realizar las pruebas de concepto del prototipo del CRIS-DEM-CERIF, mediante hojas de cálculo de Microsoft Excel. Para esto, una vez más se estableció una hoja para cada entidad, y dentro de cada entidad se definieron como columnas los atributos correspondientes. Cada entidad fue representada con un color, con la finalidad de representar las relaciones, resaltando con el color correspondiente de la entidad de origen a los atributos que aparecen como clave foránea en otra hoja. A continuación, este prototipo se alimentó con la información de la producción científica recabada de una muestra de 16 profesores adscritos al DEM. Un ejemplo de una entidad con sus respectivos atributos y datos de la muestra se puede apreciar en la entidad de "Proyectos" que se presenta en la tabla 4.

Esta adaptación al estándar CERIF permite que el CRIS del DEM sea interoperable con otros sistemas para compartir la información de los productos científicos, el hacerlo en Excel es un primer paso para comprobar su funcionalidad y facilitar su uso en una primera instancia por los investigadores al recolectar información. El propósito es ampliar este diseño y llevar el desarrollo a un sistema

más robusto con gestores de BD más potentes basados en SQL, lo que será de interés para otras instituciones de educación superior y centros de investigación en el país.

Tabla 4 Ejemplo de la entidad Proyectos con atributos y datos almacenados (Excel).

PJ_ID	PJ_NAME	PJ_YEAR	P_ID	FD_ID	E&S_ID
PJ_1	DETERMINACION DEL VALOR ECONOM	2013	P10	FD_3	E&S_4
PJ_2	PROGRAMA DE RESCATE DE ESPACIOS	2011	P10	FD_4	E&S_8
PJ_3	MERCADOS DE TIERRA Y RENTA DEL SU	2011	P10	N/A	E&S_1
PJ_4	PROPUESTA DE INVESTIGACION PARA	2011	P10	PRODEP	E&S_7
PJ_5	FORTALECIMIENTO AL LABORATORIO I	2012	P6	N/A	E&S_10
PJ_6	APOYO A LA INCORPORACION DE NUE	2008	P6	N/A	E&S_11
PJ_7	DISEÑO E INTEGRACION DE UN SISTEM	2009	P6	FD_3	E&S_27
PJ_8	SENSOR DE PRESION Y DE DIRECCION I	2008	P6	N/A	E&S_34
PJ_9	DISEÑO DE FILTROS OPTICOS PARA SE	2008	P6	N/A	E&S_35
PJ_10	DETERMINACION DE GASES DE EFECTO	2007	P6	N/A	E&S_1
PJ_11	GAS SENSING USING OPTICAL CORREL	2007	P6	N/A	E&S_2

3. Resultados

La adaptación del modelo CERIF al CRIS para gestionar la producción derivada de la investigación dentro del DEM, permite que los datos estén normalizados. Además, al adaptar el sistema del DEM a este estándar, se eliminaron redundancias de datos. Es decir, existían atributos y entidades en el modelo anterior que generaban redundancias, por ejemplo para la entidad "EXPERIENCIA LABORAL" uno de sus atributos es "INSTITUCIÓN" es decir el nombre de ésta. Para algunos casos había nombres de empresas, laboratorios, entre otros, pero para la mayoría de los casos el nombre era de instituciones de educación específicamente universidades. Por lo tanto, muchas veces se capturaron datos redundantes de las universidades que ya se encontraban en la entidad "UNIVERSIDADES", y al adaptarse el estándar CERIF fue posible que todas estas instituciones, universidades y centros de investigación, quedaran juntas dentro de la entidad "ORGANISATION". Esto permitió eliminar la entidad "UNIVERSIDAD" y hacer la relación con las entidades que requieran de esta información. El modelo basado en CERIF, gracias a su generalidad estructural de entidades, permite que la información de una entidad pueda adaptarse a varias relaciones con otras entidades, por ejemplo:

- PERSONA-PROYECTO: En donde persona trabaja como rol de investigador.
- PERSONA-TESIS_REDACTADA: En donde persona trabaja como asesor.
- PERSONA-LIBRO: En donde persona se considera como autor.

En estos casos cada entidad puede tener varias relaciones, lo que evita que se creen entidades que no son necesarias y que pueden afectar el modelado de los datos. Algunas entidades pudieron adaptarse adecuadamente a lo que CERIF propone, pero fue necesario crear algunas otras para que este modelo pudiera cubrir todas las características que requiere el DEM. Por ejemplo, en la entidad de publicaciones fue necesario relacionarla con otras que tuvieron que ser creadas como: TESIS_REDACTADAS, TESIS_DIRIGIDAS, CAPITULOS DE LIBRO, LIBROS. Estas publicaciones tienen características particulares, por lo tanto, cada una debe tener su propia entidad que describa sus propias características.

Las diferencias que se encontraron con respecto a las necesidades del DEM y al estándar CERIF, posiblemente son a causa de las variantes culturales y sociales que se tienen con Europa. Debido a esto, probablemente existan ciertas inconsistencias con aquellos sistemas europeos diseñados bajo este mismo estándar, pero en teoría deberían poder interoperar con el prototipo aquí desarrollado gracias a los cuatro grupos de entidades generales del estándar CERIF.

4. Discusión

El estándar CERIF propone una serie de entidades y relaciones que permiten a los diseñadores de sistemas de información para la investigación, adaptar sus sistemas a esta norma. Lo que se busca con CERIF, es que a cada sistema diseñado bajo su modelo le sea posible conectarse y sea interoperable con todos aquellos sistemas que estén estructurados bajo este mismo estándar. Esto permite incrementar la cantidad de información que puede ser compartida y consultada, y poder potenciar mejores resultados dentro de la investigación sobre temas particulares.

Al igual que todas las instituciones que generan productos científicos, la Universidad de Guanajuato, y más en específico el DEM, pretende adoptar un sistema para la investigación con la finalidad de compartir y difundir la información de carácter científico que se genera dentro de sus aulas y laboratorios. La adaptación de CERIF-DEM del modelado de datos a CERIF, permitió darle una estructura normalizada al diseño, además que con la prueba de concepto es posible corroborar la factibilidad de una futura implementación en sistemas gestores de bases de datos relaciones bajo el diseño aquí presentado.

El desarrollo de este sistema puede traer resultados positivos para el DEM, ya que con la información almacenada bajo este modelo, se puede acelerar la construcción del conocimiento que pueden impulsar la generación de tecnologías dentro de la sociedad, al poner a la disposición de usuarios y externos al DEM, la información científica que ahí se genera. En un futuro, esta propuesta de CRIS bajo CERIF podría ser considerada para impulsar un sistema a nivel nacional, que permita que todas las universidades y centros de investigación puedan intercambiar con facilidad su información científica.

5. Bibliografía y Referencias

- [1] Ivanovic, D., Surla, D., Trajanovic, M., Misic, D. & Konjovic, Z. (2017). Towards the Information System for Research Programmes of the Ministry of Education, Science and Technological Development of the Republic of Serbia. *Procedia Computer Science*, 106(June 2016), 122-129. doi: 10.1016/j.procs.2017/03.44
- [2] Jeffery, K. & Asserson, A. (2006). CRIS: Central Relating Information System. *Enabling Interaction and Quality: Beyond the Hanseatic League*. 109-118: <https://goo.gl/c7xsCS>.
- [3] Jeffery, K., Houssos, N., Jörg, B. & Asserson, A. (2014). Research information management: the CERIF approach. *International Journal of Metadata, Semantics and Ontologies*, 9(1), 5-14: <https://goo.gl/kQ2cPF>.
- [4] Jörg, B., Gartner, R., Clements, A., Baker, D. & Zielinski, M. (2012). CERIF 1. 3 Full Data Model (FDM).

- [5] Joint, N. (2008). Current research information systems, open access repositories and libraries: ANTAEUS. *Library Review*, 57(8), 570-575. doi: 10.1108/00242530810899559
- [6] Kumar, S. (2011). Introduction to Data Base models. En *Database Systems: Concepts, Design and Applications*.
- [7] Nabavi, M., Jeffery, K. & Jamali, H. R. (2016). Added value in the context of research information systems. *Program*, 50(3), 325-339. doi: 10.1108/PROG-10-2015-0067
- [8] Pinto, C. S., Simões, C. & Amaral, L. (2014). CERIF - Is the standard helping to improve CRIS? *Procedia Computer Science*, 33, 80-85. doi: 10.1016/h.procs.2014.06.013
- [9] Quix, C. & Riechert, M. (2017). Modelling National Research Information Contexts Based on CERIF. *Procedia Computer Science*, 106 (June 2016), 253-259. doi: 10.1016/h.procs.2017.03.023
- [10] Zelepukhina, V. A., Danilova, T. S., Burmistrov, A. S. & Tarasevich, Y. Y. (2014). Particular experience in design and implementation of a Current Research Information System in Russia: national specificity. *Procedia - Procedia Computer Science*, 33, 168-173: <https://goo.gl/dTAjbZ>.