Arquitectura para un repositorio semántico de documentos de investigación

Architecture for a semantic repository of research documents

Phd. Gustavo Andrés Uribe Gómez[1]
Corporación Universitaria Comfacauca, Colombia
guribe@unicomfacauca.edu.co

Msc. Pedro Harvey Álvarez Sánchez[2]
Corporación Universitaria Comfacauca, Colombia
palvarez@unicomfacauca.edu.co

Fecha Recepción: 21/12/17 - Fecha Aprobación: 23/12/17

Resumen: Dada la necesidad de facilitar la recuperación de información en el ámbito de la investigación se requiere mejorar cada vez más la precisión y la exhaustividad de los repositorios de documentos de investigación. Para ello se requiere especificar el significado de los datos, lo cual puede lograrse mediante los lenguajes definidos por la Web Semántica, tales como el lenguaje para la ontología (OWL). El Modelo Genérico de Componentes (GCM) permite crear sistemas de información que se desarrollan desde ontologías de diversos dominios de conocimiento. En este artículo se presenta la arquitectura de un sistema mediante el GCM y su validación.

Palabras clave: Repositorio semántico, vista de arquitectura gcm, documentos finales de investigación, vistas de repositorio semántico.

Abstract: Due the need to facilitate the information recovery in the research context, the improvement of the precision and recall in repositories of research documents is needed. The Semantic Web is a promising set of technologies for that, including languages to describe taxonomies and ontologies. Thus, through of this article A Methodology Associated Bliss New Technology Presented : Generic Component Model, which is allows you to create information systems that are developed from ontologies much richer in : semantic elements, a situation that allows better search Associated Information Research. This model is characterized by describing the system allowing graininess.

Keywords: Semantic repository, architecture views gcm, final documents research, semantic repository views.

1. Introducción

En los últimos años las comunicaciones han evolucionado enormemente, facilitando la transmisión inmediata de la información; situación que ha impactado el ámbito académico, especialmente en todo lo relacionado con la investigación. Para esta área fundamental de toda institución académica, el uso de la tecnología ha sido una herramienta muy beneficiosa que ha impulsado el conocimiento científico, permitiendo difundir dicho conocimiento en y hacia diferentes espacios, incluida la web.

Sin embargo, es tanta la información que hoy en día se produce que se hace necesario contar con elementos tecnológicos que permitan gestionar la misma, pues las dificultades en este campo saltan a la vista y no permiten un filtrado amigable, preciso, idóneo y efectivo de la información que se requiere.

Esta situación es una preocupación latente a nivel mundial y mucho más en países como Colombia, donde la tecnología no alcanza aún los desarrollos existentes en el primer mundo. Tal es el caso de la ciudad de Popayán, ubicada en el departamento del Cauca, donde es notoria la existencia de limitaciones en el uso de repositorios de información, especialmente asociados a la investigación, a pesar de los esfuerzos realizados por algunas instituciones.

Es por esto que el presente artículo muestra una propuesta tecnológica que podría ayudar a solucionar el inconveniente o apoyar dichos procesos de recuperación de información, con el

1. Ingeniero en Electrónica y Telecomunicaciones, Doctor en Ingeniería Telemática. Docente Investigador de Corporación Universitaria Comfacauca.
2. Ingeniero de Sistemas, Magíster en Gestión de la Tecnología Educativa. Docente Investigador de Corporación Universitaria Comfacauca.

ánimo de agilizar dichas búsquedas y facilitar la selección de información, para de esta manera contribuir al crecimiento de la labor investigativa de las Instituciones de Educación Superior - IES de la ciudad.

Esta propuesta surge del proyecto "Mejoramiento de las métricas de recuperación de información en repositorios de proyectos de investigación institucionales por medio de la web semántica", a través del cual se sugiere un mecanismo para facilitar la obtención precisa y exhaustiva de información, con el fin de mejorar la precisión y exhaustividad de las consultas de documentos asociados a la investigación en las Instituciones de Educación Superior – IES de la ciudad de Popayán.

2. Revisión de la Literatura

En la actualidad existen numerosas herramientas tecnológicas que ayudan en el proceso de búsqueda y selección de información; dentro de las más conocidas aparecen: Google Scholar, SpringerLink, ACM Digital Library, IEEE Explorer, Web of Science, Scopus y ScienceDirect. Estas herramientas presentan dos grandes desventajas relacionadas con su posibilidad de acceso para todo público (debido a que son muy costosas) y su falta de semántica (carencia de mecanismos que permitan definir el significado exacto de las búsquedas y de los metadatos). Estas falencias provocan resultados inexactos y poco exhaustivos, lo que finalmente conlleva a que no se obtenga realmente la información que se necesita y que los nuevos documentos carezcan de un mayor impacto.

Debido a lo anterior, existen algunos esfuerzos por construir repositorios de información, especialmente asociados a la investigación, que abaratan costos y aplican algunas tecnologías semánticas, con el fin de mejorar la recuperación de información. Estas herramientas se utilizan cada vez más en diferentes áreas del conocimiento: En educación virtual [1]–[3]; en medicina [4]; en búsqueda semántica de noticias [4], y en documentación en ingeniería informática [5].

Dichas herramientas utilizan tecnologías como *Resource Description Framework* (RDF), *Web Ontology Language* (OWL)[6], SPARQL y SKOS, las cuales son estándares propuestos aceptados por la *World Wide Web Consortium* (W3C) y que han demostrado su efectividad en la búsqueda y selección de información. Sin embargo, sigue siendo necesario mejorar en ellas sus aspectos semánticos, pues según García-Peñalvo *et al.* [7] es a través de mejoras en estos elementos que se puede llegar a obtener herramientas mucho más robustas, que permitan ubicar eficazmente documentos de manera automática y semejante a como lo haría un humano gracias a su etiquetado semántico previo.

Complementando lo anterior, se puede decir que en estas tecnologías, a pesar de usar SKOS, RDF y/o OWL para la descripción de los metadatos de información; SPARQL como lenguaje de consulta, y Procesamiento de Lenguaje Natural (PLN) para generar automáticamente metadatos acerca de los documentos en el repositorio [2], [5], aún no se identifica una asertiva descripción semántica de los metadatos para el dominio de los aspectos asociados a la investigación, como es el fin de la presente propuesta.

Así mismo, se identificaron otras herramientas un poco más cercanas a los procesos de investigación, tales como: OpenDOAR [8] y P*ublic Knowledge Project (PKP)[9]*; sin embargo, éstas tampoco presentan muchos adelantos en el campo semántico.

Es por lo anterior que se hace evidente la necesidad de mejorar la semántica de las búsquedas y de los repositorios de documentos de investigación pertenecientes a las IES, para, entre otras cosas, fortalecer sus grupos de investigación, mejorar la difusión de sus proyectos e integrar investigaciones desarrolladas en red o cooperación por diferentes universidades.

3. Metodología

El GCM Según (Blobel, 2002) es un marco de referencia que permite la construcción de sistemas informáticos, usando estándares como las ontologías y el Modelo de Referencia para Procesamiento Distribuido y Abierto (RM-ODP). Para [10] permite el desarrollo de sistemas modulares, distribuidos, coherentes con la realidad, abiertos y reusables.

En las siguientes secciones se explican los componentes metodológicos más importantes de este trabajo. Esta metodología ha arrojado exitosos resultados en otros contextos, como por ejemplo en [11], [12].

Este documento propone el mejoramiento de las métricas de recuperación de información en repositorios de proyectos de investigación en

instituciones de la ciudad de Popayán, departamento del Cauca, por medio de la web semántica, utilizando para ello estándares definidos como son: El Modelo de Referencia para Procesamiento Distribuido y Abierto (RM-ODP) y el Modelo Genérico de Componentes (GCM), los cuales se describen a continuación.

3.1. Modelo de Referencia para Procesamiento Distribuido y Abierto (RM-ODP).

Es un estándar para el desarrollo de aplicaciones abiertas y distribuidas, establecido por normas internacionales como: ISO (*International Organization for Standardization*) e ITU-T (Unión Internacional de Telecomunicaciones), las cuales definen estándares para el desarrollo de aplicaciones abiertas y distribuidas. Cabe señalar que el RM-ODP tiene como objetivo definir un modelo de referencia que permita integrar toda una serie de estándares sobre el desarrollo de aplicaciones abiertas y distribuidas, manteniendo consistencia entre ellos.

De esta manera, el RM-ODP contribuye con el desarrollo de aplicaciones proporcionando un marco de trabajo conceptual y una arquitectura que integra aspectos relacionados con la distribución, interoperabilidad y portabilidad de sistemas software, logrando la transparencia en los usuarios. También contribuye con un marco de coordinación para la normalización del desarrollo de estas aplicaciones y permite definir de forma clara y precisa aquellos conceptos que aparecen en el desarrollo de componentes distribuidos, proporcionando un vocabulario y un marco semántico común a todos los participantes o usuarios de las aplicaciones.

En este orden de ideas, el RM-ODP define cinco puntos de vista [13]:

a) El punto de vista de la empresa: Centrado en que el software se enfoque en los objetivos, el alcance y las políticas definidas por la empresa.

b) El punto de vista de la información: Enfocado en la semántica de la información y en el procesamiento de la información realizada en la aplicación.

c) El punto de vista computacional: Permite la distribución funcional del desarrollo, por medio de módulos que se comunican mediante interfaces.

d) El punto de vista de ingeniería: Describe el sistema y su entorno, centrándose en los procesos necesarios para soportar la interacción distribuida entre los objetos del software.

e) El punto de vista de tecnología: Orientado en la selección de tecnología en el sistema.

Se espera que al describir el sistema, mediante cada una de estas vistas y los lenguajes que describe el estándar, se obtenga un sistema con una arquitectura de procesamiento distribuido y abierto.

3.2. Modelo Genérico de Componentes (GCM)

La principal metodología usada en la descripción del sistema es el Modelo Genérico de Componentes (GCM) [14] (Figura 1). Esta metodología, derivada de la Teoría General de Sistemas permite describir sistemas mediante tres dimensiones: De composición de componentes, de dominios del sistema y de puntos de vista del sistema.

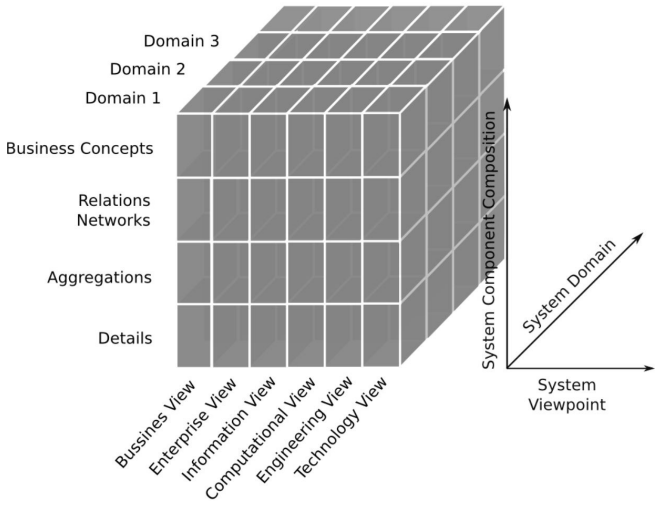

Figura 1. Modelo Genérico de Componentes (GCM). Fuente: [14].

En primera instancia, la dimensión de composición describe el sistema acorde con sus componentes y las relaciones entre ellos. Todo esto considerando cuatro niveles de granularidad o detalle. Esta vista permite describir las propiedades estructurales de los sistemas mediante las relaciones "es parte de" o "está conectado con". Por su parte, la dimensión de los dominios del sistema describe los diferentes aspectos del sistema, por ejemplo, un dominio describiría los aspectos médicos del sistema, mientras otro dominio describiría los aspectos contables o de gestión de recursos.

Cada dominio corresponde al interés de un grupo de personas y está descrito acorde a una ontología,

terminología u otro mecanismo de representación de conocimiento. Finalmente, la dimensión de los puntos de vista del sistema corresponde con las vistas para el desarrollo de software usadas por el RM-ODP, adicionando la vista del negocio, que es una descripción independiente de la computación [15]. Basados en esta vista se completan los modelos del RM-ODP, que representan la propuesta de modelo ejecutable que se implementará finalmente mediante el proceso de desarrollo de software.

Este modelo genérico de componentes, basado en estas dimensiones, permite la descripción de la estructura y del comportamiento del sistema, el cual se puede complementar mediante lenguajes de modelado de procesos de negocio, lo que permite representar más explícitamente el comportamiento del sistema en su conjunto.

Con el fin de construir una arquitectura comprensible con el GCM es necesario tener en cuenta los siguientes principios de diseño (Buenas Prácticas de Modelado): Ortogonalidad (no vincular aspectos independientes - no incorporar entidades de diferentes niveles de granularidad), generalidad (no introducir múltiples entidades similares), parsimonia (no introducir aspectos irrelevantes) y completitud (que no se restrinja aspectos inherentes) [16].

4. Resultados

En las siguientes secciones se presentarán las tres primeras vistas de la arquitectura del sistema según la metodología provista por el GCM, en procura de su implementación para las IES de la ciudad de Popayán.

4.1. Vista de negocio

Como se mencionó, esta vista describe el sistema a intervenir mediante una aplicación informática. Aquí cada uno de los dominios del sistema es representado mediante el uso de una ontología de nivel superior, donde los dominios aparecen ínter-conectados. Para seleccionar la ontología de nivel superior más apropiada para esta descripción se usó la aplicación ONSET [17], la cual mediante una serie de preguntas determinó cuál era la más adecuada. Para este caso concreto, la ontología de nivel superior seleccionada fue: *"Descriptive Ontology for Linguistic and Cognitive Engineering* (DOLCE)" [18].

En ella el sistema de búsqueda de documentos comprende tantos dominios como los que se puedan definir en el conocimiento científico. Sin embargo, para este caso, se definió un dominio relevante asociado a elementos de investigación. A continuación la Figura 2 muestra parcialmente la descripción de dicho dominio, y la Figura 3 presenta el dominio del desarrollo de software propuesto.

Como se puede apreciar, en la Figura 2 se observan algunos conceptos (clases) correspondientes a la ontología de nivel superior (DOLCE) y las partes principales de un documento asociado a la investigación. Las líneas punteadas corresponden a relaciones "PART-OF" (parte-de), mientras que las

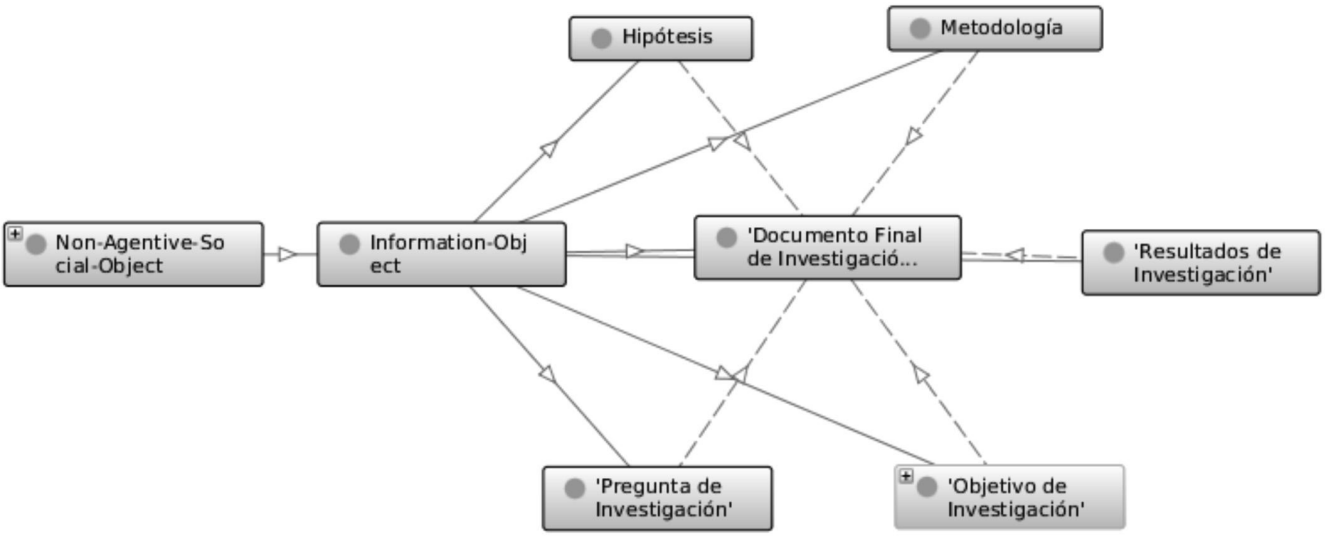

Figura 2. Fragmento de la ontología para el dominio de los procesos de investigación. Fuente: Elaboración propia

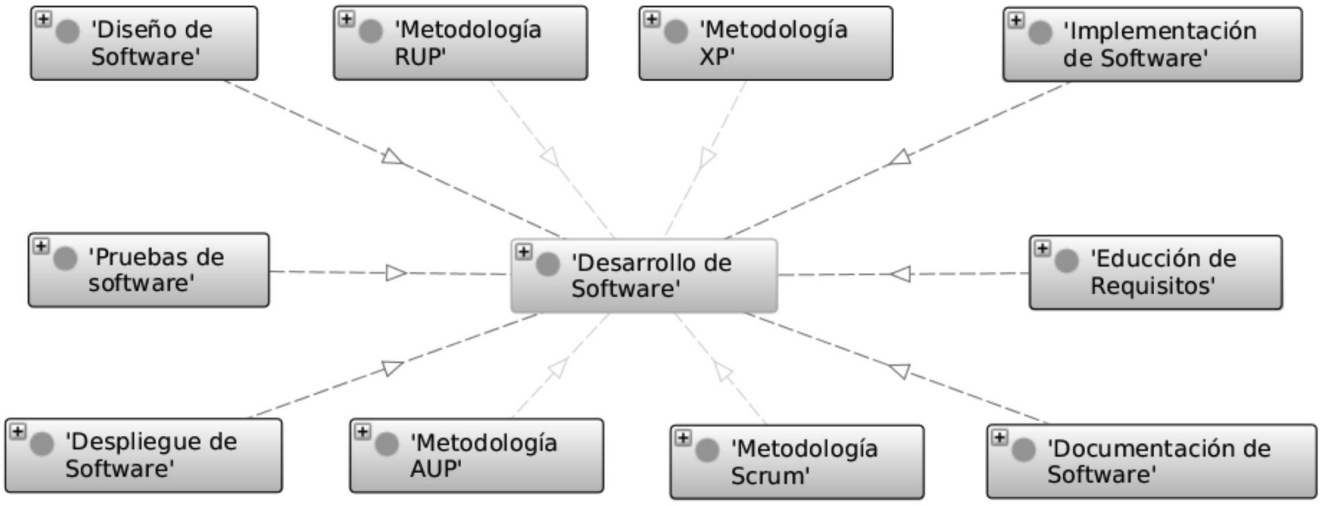

Figura 3: Fragmento de la ontología del desarrollo de software propuesto. Fuente: Elaboración propia.

líneas punteadas corresponden con relaciones "ISA" (es-un). Por su parte, en la Figura 3 se muestran las metodologías de desarrollo de software más utilizadas y las partes más relevantes dentro del desarrollo de software. Las metodologías se relacionan con el desarrollo del software mediante una relación de "PARTICIPANT-IN" y las partes mediante la relación de "PART-OF" (parte-de).

Como se mencionó, el GCM también obliga a ubicar los diferentes elementos en diferentes niveles de granularidad. En el nivel superior de granularidad (*Business Concepts*) se encuentran el desarrollo de software y el documento de investigación, uno para cada dominio. Las diversas partes de estos elementos son ubicados en el nivel de granularidad denominado: "*Relation Networks*".

4.2 Vista empresarial

En esta vista se determinan los objetivos y las políticas del sistema de información propuesto. Esta información es equivalente a describir lo que espera cada usuario del sistema, de qué manera el sistema va a responder y cuál sería la retribución del usuario hacia el sistema. Para expresar algunas de estas características es pertinente hacer uso de diagramas UML de casos de uso y complementarlo, posteriormente, con descripciones detalladas de casos de uso.

En la Figura 4 se pueden observar los casos de uso de más alto nivel. La palabra gestión en estos casos de uso hace referencia a las operaciones fundamentales de los objetos de información, es decir, creación, lectura (consulta), actualización y borrado. Detalles más precisos de esta vista están por fuera del alcance del presente artículo.

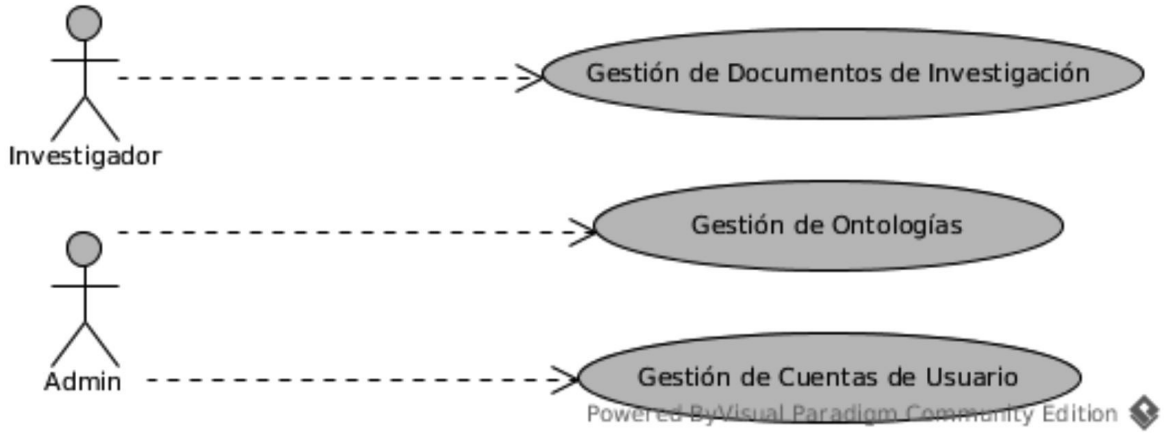

Figura 4. Casos de uso de alto nivel. Fuente: Elaboración propia

6. Conclusiones

Mediante la metodología GCM y algunos elementos del UML fue posible la descripción de dos vistas correspondientes a un repositorio semántico de documentos finales de investigación. Este artículo demuestra parcialmente que el uso de esta metodología permite una clara separación en los diferentes aspectos (dominios), vistas y granularidades de un sistema de información para investigaciones finales.

Adicionalmente, se muestra una metodología apropiada para el desarrollo de ontologías inter-conectadas de diversos dominios. Esto ayuda a que el sistema sea lógicamente coherente, rico semánticamente y facilitador de la reutilización de módulos, debido al uso de estas ontologías con ancestros comunes.

Finalmente, se debe decir que la completitud de la arquitectura es meramente dependiente del futuro esfuerzo por complementar los dominios y conceptos faltantes. Se espera entonces, en un próximo artículo, completar las vistas faltantes del GCM y, de esta manera, mostrar el éxito de las metodologías usadas para el desarrollo de un repositorio de documentos finales de investigación que mejore su exhaustividad y precisión.

7. Referencias Bibliográficas

[1] A. O. Agüero, A. S. Mansolo, Ms. R. Jorge, and P. Lauzán, "Repositorios De Objetos De Aprendizaje De Acceso Abierto Para La Educación De Postgrado. República Bolivariana De Venezuela.," 2010.

[2] M. C. Chiarani, I. G. Pianucci, and G. Leguizamón, "Repositorio de objetos de aprendizaje para carreras informáticas," in *VIII Workshop de Investigadores en Ciencias de la Computación*, 2006.

[3] M. L. Bonilla, "Semántica para repositorios de objetos de aprendizaje," *Scientia et Technica*, vol. 19, no. 4, pp. 425–432, 2014.

[4] P. Castells *et al.*, "Neptuno: tecnologías de la web semántica para una hemeroteca digital," *España: Ministerio de Ciencia y Tecnología*, vol. 1, 2004.

[5] R. J. Espinoza Florez, "Diseño de una herramienta para la anotación semántica automática de documentos basados en ontologías en el dominio de la Ingeniería Informática," 2015.

[6] D. L. McGuinness, "Ontologies for information fusion," in *Proceedings of the Sixth International Conference on Information Fusion*, 2003, pp. 650–656.

[7] F. J. García-Peñalvo, R. Colomo-Palacios, P. Soto-Acosta, I. Martínez-Conesa, and E. Serradell-López, "SemSEDoc: Utilización de tecnologías semánticas en el aprovechamiento de los repositorios documentales de los proyectos de desarrollo de software," 2011.

[8] P. Millington, "OpenDOAR - Home Page - Directory of Open Access Repositories," 06-Sep-2006. [Online]. Available: http://www.opendoar.org/. [Accessed: 03-Jun-2016].

[9] Simon Fraser University Library, "Open Monograph Press | Public Knowledge Project," 2014.

[10] V. Luna, R. Quintero, M. Torres, M. Moreno-Ibarra, G. Guzmán, and I. Escamilla, "An ontology-based approach for representing the interaction process between user profile and its context for collaborative learning environments," *Computers in Human Behavior*, vol. 51, pp. 1387–1394, 2015.

[11] G. A. Uribe, B. Blobel, D. M. López, and S. Schulz, "A generic architecture for an adaptive, interoperable and intelligent type 2 diabetes mellitus care system," *Stud Health Technol Inform*, vol. 211, pp. 121–131, 2015.

[12] G. A. Uribe, B. Blobel, D. M. López, and A. A. Ruiz, "Specializing architectures for the type 2 diabetes mellitus care use cases with a focus on process management," *Stud Health Technol Inform*, vol. 211, pp. 132–142, 2015.

[13] ISO, "Information technology — Open Distributed Processing — Reference model: Overview," Nov-2008. [Online]. Available: http://www.itu.int/rec/T-REC-X.901-199708-I/en. [Accessed: 11-Jun-2014].

[14] M. Brochhausen *et al.*, "Discussion of' biomedical ontologies: toward scientific debate'.," *Methods of information in medicine*, vol. 50, no. 3, p. 217, 2011.

[15] B. Blobel and P. Pharow, "Architectural Approaches to Health Information Systems for Empowering the Subject of Care," *Medical and Care Compunetics 5*, vol. 1, p. 355, 2008.

[16] M. Lankhorst and others, *Enterprise Architecture at Work: Modelling, Communication and Analysis (The Enterprise Engineering Series)*, 2nd ed. Heidelberg: Springer, 2009.

[17] M. Keet, "ONSET: foundational ONtology Selection and Explanation Tool," 2012. [Online]. Available: http://www.meteck.org/files/onset/. [Accessed: 09-Jun-2016].

[18] Laboratory for Applied Ontology, "Laboratory for Applied Ontology - DOLCE," 2006. [Online]. Available: http://www.loa.istc.cnr.it/old/DOLCE.html. [Accessed: 09-Jun-2016].