FeDCOR: Un Registro Institucional CORDRA

Henry Jerez y Giridhar Manepalli
Corporation for National Research Initiatives
Reston, VA 20191
{ hjerez, gmanepalli }@cnri.reston.va.us

Michael L. Nelson
Old Dominion University
Department of Computer Science
Norfolk, VA 23529
mln@cs.odu.edu

Introducción

El descubrimiento y acceso a colecciones heterogéneas de información es una de las áreas más desafiantes en el estudio de bibliotecas digitales. El crecimiento de la capacidad de almacenamiento y la velocidad de acceso a las redes de computadoras así como la presencia indiscriminada de acceso al Internet nos provee las herramientas necesarias para enfrentar este desafió. El proyecto CORDRA (Arquitectura para el Registro, Resolución y Descubrimiento de Repositorios de Objetos de Contenido) es un esfuerzo colaborativo entre la Corporación Nacional para Iniciativas de Investigación (CNRI), el Laboratorio de Arquitecturas para Sistemas de Aprendizaje (LSAL) y financiado por la iniciativa de Aprendizaje Avanzado Distribuido (ADL) del Departamento de Defensa de los Estados Unidos de América.

Este proyecto tiene el objetivo de crear una infraestructura global para la federación de repositorios de contenidos. Si bien el proyecto se inicio en el ámbito de la educación basada en Internet; este halló inmediatamente el requerimiento de acoger cualquier tipo de contenido necesario para apoyar las actividades de enseñanza. Es por esta razón que el proyecto fue diseñado para incorporar y federar prácticamente cualquier tipo de contenido.

La estructura del proyecto contempla la incorporación de grupos de repositorios organizados en federaciones que buscan integrar su contenido en un Registro central.

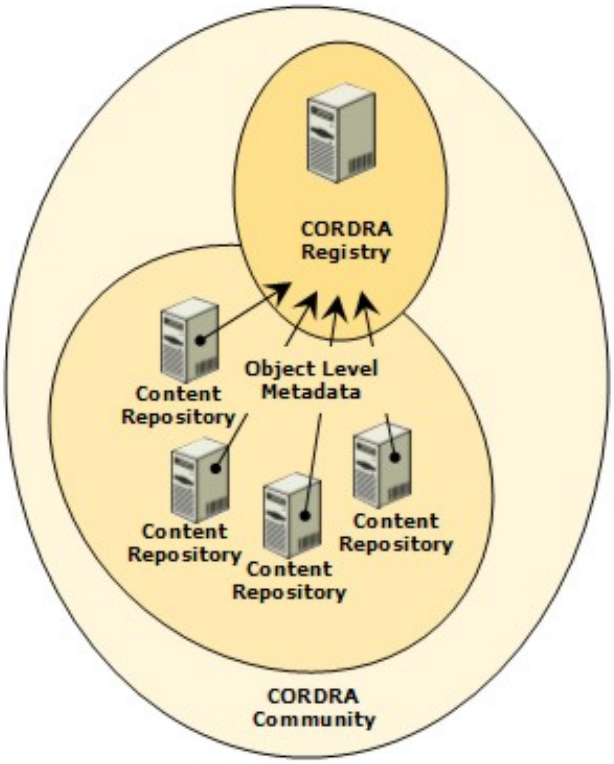

Figura 1. Etapa Inicial en la Federación de Repositorios

Estas federaciones a su vez se registran en Registros de Registros (RoRs) conformando federaciones de federaciones. Estas federaciones pueden a su vez ser integradas en uno o más Registros maestros. El nivel superior no se constituye en un área de control si no mas bien un punto de entrada para acceder a alguna colección o contenido registrado en una de las diferentes federaciones.

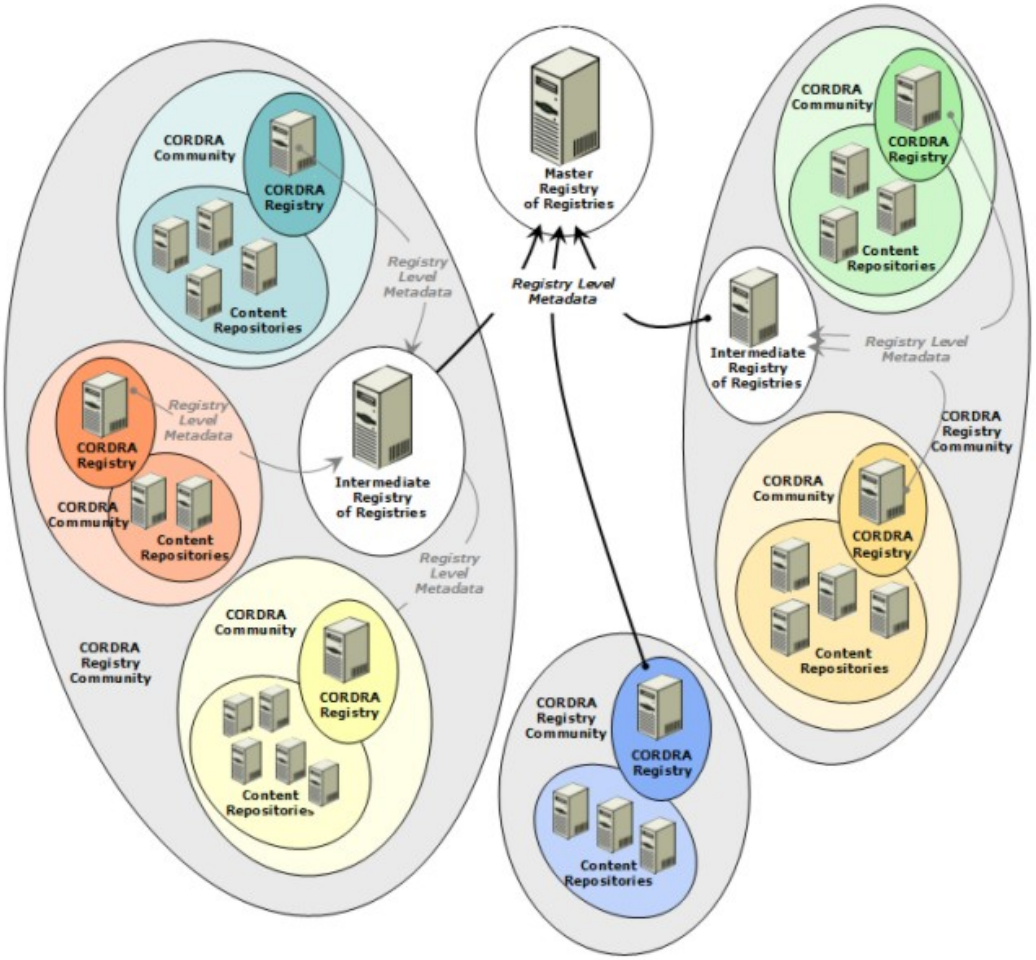

Figura 2. Federación Global CORDRA

Las diferentes federaciones utilizan una variedad de estándares de meta datos, políticas de acceso, principios organizacionales y demás características individuales. Todas ellas sin embargo obedecen el modelo abstracto global de CORDRA y un conjunto de estándares para la federación de federaciones. El primero de estos Registros que se encuentra actualmente en producción recibe el nombre de ADL-R [1] y es utilizado por el departamento de defensa para federar los contenidos de entrenamiento a distancia elaborados por las diferentes fuerzas y sus contratistas en Estados Unidos.
La primera instancia de Registros CORDRA a nivel académico recibe el nombre de FeDCOR y se encuentra en fase experimental al interior del consorcio Ibero Americano de Ciencia y Tecnología ISTEC como resultado de su colaboración con CNRI.

FeDCOR (Federación DSpace utilizando CORDRA).

FeDCOR es un Registro CORDRA responsable por la federación de repositorios DSpace. DSpace [2] es un sistema de repositorios digitales diseñado para capturar, almacenar, indexar, preservar y redistribuir contenido en varios formatos digitales. Muchas instituciones de investigación y académicas que aplican el modelo de bibliotecas digitales han adoptado DSpace como su herramienta de archivo institucional de preferencia.

Aun cuando DSpace ha sido muy exitoso en el ámbito de repositorios institucionales individuales; este todavía adolece de falencias a la hora de consolidarse dentro de federaciones. De hecho; actualmente no existe la posibilidad de efectuar búsquedas distributivas a través de múltiples repositorios DSpace. La federación exitosa de estos repositorios permanece por lo tanto como un desafió abierto en la comunidad de bibliotecas digitales [3].

Es en este espíritu que la elaboración de un Registro de repositorios DSpace utilizando la tecnología CORDRA obedece dos objetivos principales:

1. La creación de una federación de repositorios DSpace efectiva.
2. La creación del primer Registro CORDRA para la comunidad de bibliotecas digitales.

El diseño e implementación de este Registro resulta útil tanto para la comunidad DSpace como para la comunidad CORDRA.

El diseño lógico de FeDCOR se ilustra en la figura 3 a continuación.

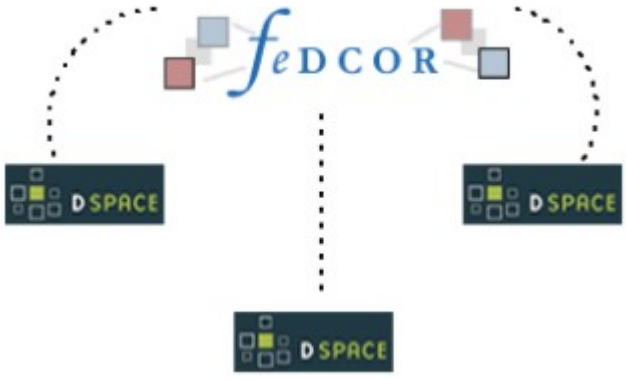

Figura 3: Diseño lógico de FeDCOR

Los repositorios DSpace actúan como repositorios de contenido tradicionales dentro de FeDCOR; y los meta datos provistos por DSpace para cada objeto de contenido son tratados como una instancia más de meta datos acerca de ese objeto. Dspace asocia Handles [4] con la abstracción del objeto de contenido que incorpora instancias de meta datos y secuencias binarias de información que componen el objeto. Esto permite al Registro consolidar las múltiples instancias de los diferentes objetos en múltiples repositorios; siempre y cuando los repositorios respeten la asociación con handles y utilicen el plug-in de DSpace que les permite utilizar servidores Handle independientes de cada repositorio individual.

El diseño de FedCOR obedece los mecanismos de acceso de datos provistos por DSpace y requiere de la definición extensiva de estructuras de datos, reglas de operación y taxonomías que obedezcan la arquitectura CORDRA.

Las siguientes secciones detallan tanto el diseño como la implementación de FeDCOR en función al diseño original de CORDRA .

(1) El Diseño de FeDCOR

Un Registro CORDRA contiene; por referencia, los objetos de contenido registrados; así como las instancias de meta datos correspondientes. El Registro de la comunidad debe por lo tanto reflejar el estándar de meta datos concensuado dentro de la comunidad. Al mismo tiempo; el Registro debe ser capaz de acomodar un nivel superior de meta datos globales a ser propagados dentro de la comunidad CORDRA. Afortunadamente el diseño original de CORDRA provee independencia de meta datos en múltiples capas y niveles. Los tres niveles posibles tal como se muestra en la figura 4 son:

1. Meta datos específicos del objeto de contenido registrado, también conocidos como meta datos al nivel de la Comunidad (Community Level Metadata)
2. Meta datos específicos del Registro (Registry Metadata)
3. Meta datos específicos de CORDRA (CORDRA Federation Metadata)

En FeDCOR los tres diferentes niveles deben ser estrictamente definidos. De acuerdo a nuestra experiencia con ADL-R podemos asumir que los meta datos específicos de CORDRA son un sub-grupo de la unión de meta datos de la comunidad y el Registro.

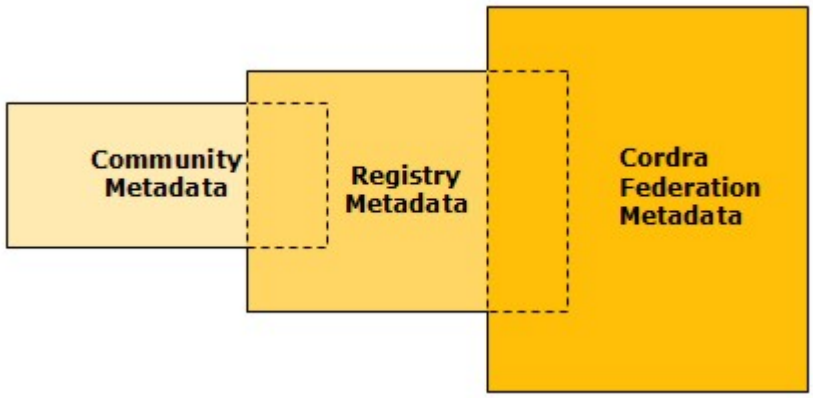

Figure 4: Capas de Meta datos dentro del Registro

Cabe recalcar que los meta datos específicos de CORDRA evolucionaran de acuerdo a las necesidades de múltiples comunidades y el propio modelo interno de la comunidad. Esto se traduce en un modelo variable en el cual las diferentes comunidades deben concentrarse solamente en los meta datos a nivel del Registro y los objetos de contenido.

Definición de los meta datos del objeto de contenido

La instancia de meta datos de los objetos de contenido (COMI) es aquel grupo de meta datos relacionado a cada objeto recuperado de DSpace. Dado que DSpace utiliza el

formato extendido de meta datos Dublin Core y la mayoría de sus instalaciones es compatible con el protocolo de recopilación de meta datos OAI-PMH [5], el cual es un protocolo estándar utilizado por la comunidad de bibliotecas digitales; FeDCOR lo adopta como su estándar de facto para acceder a los repositorios DSpace. El formato Dublín Core [6] utilizado por las comunidades tradicionales de DSpace es compatible con OAI-PMH y por lo tanto con el esquema oai_dc de OAI-PMH.

Los registros de meta datos están directamente relacionados en DSpace tanto con los meta datos como con las secuencias de datos encapsulados en forma de objetos complejos; tal y como se menciono anteriormente. Dado que nosotros consideramos el caso más general en el cual DSpace respeta la existencia de los handles más allá de cada repositorio; podemos utilizar el mismo identificador dentro del Registro. Utilizamos este identificador para identificar una Entidad de Representación del objeto de contenido (CORE) [1] la cual es utilizada como la estructura básica de agrupación para los COMIs dentro del Registro. La figura 5 ilustra esta relación.

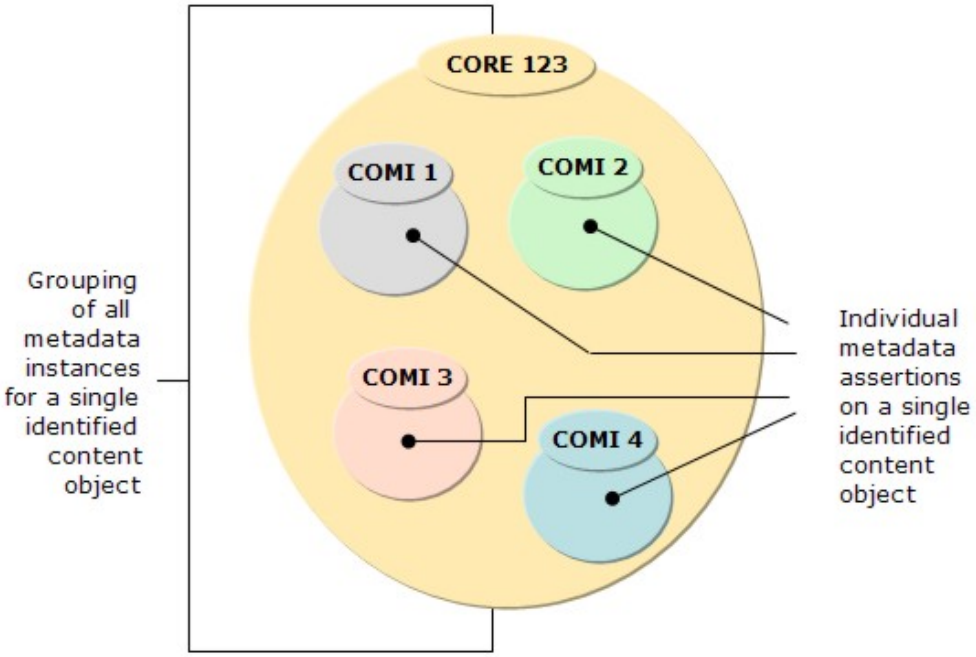

Figure 5: Representación interna del objeto de contenido y sus instancias de meta datos

Además de tener un identificador único para el contenido, CORDRA requiere un identificador para cada instancia de meta datos. FeDCOR sigue la misma hermenéutica de ADL-R por la cual el Registro genera y maneja los handles para cada instancia de meta datos que se encuentra dentro de cada repositorio DSpace en particular. Como

habíamos mencionado antes, el handle generado por DSpace se utiliza como el handle del objeto de contenido.

Meta datos al nivel del Registro

Los meta datos al nivel del Registro reflejan las características particulares de cada ítem en el Registro de cada comunidad. Estos detalles son una realización importante de las características de CORDRA como una plataforma de contenido heterogéneo. Ellos ayudan a asociar cada instancia con el Registro al que pertenecen. En el caso de FeDCOR, el handle de cada objeto de contenido así como el de su instancia de meta datos son almacenados a este nivel. También se incluye a este nivel la fecha y tiempo de la última actualización de cada instancia.

Meta datos específicos de CORDRA

Tal como se mencionó previamente, los meta datos específicos de CORDRA son un sub-grupo de la unión de los meta datos específicos de cada objeto de contenido y de aquellos específicos del Registro. Tal y como se menciona en el artículo de *D-Lib Magazine* [1], las características finales así como los procedimientos asociados con este nivel están aún bajo estudio y desarrollo. Afortunadamente, la colección total de estos datos será construida ya sea por medio de adiciones transparentes o la participación de agentes de recolección que interactúan con las comunidades CORDRA; lo cual garantiza la independencia de implementaciones tempranas del sistema.

Reglas de operación

La operación de CORDRA con múltiples niveles de repositorios depende de los sistemas de identificación utilizados. A fin de garantizar una implementación confiable y persistente se tomo la decisión de utilizar el sistema Handle [4]. Así mismo, la presencia de identificadores persistentes y únicos para los objetos de contenido y sus instancias de meta datos es mandatoria. La fecha de actualización es también necesaria a fin de permitir múltiples vistas integradas del sistema.

FeDCOR hace cumplir las reglas de operación mencionadas previamente en adición a los esquemas de validación; antes de proceder al Registro de cada objeto de contenido presente en los repositorios DSpace.

(2) Implementación de FeDCOR

Las diferentes estructuras de datos, taxonomías y reglas de operación deben ser integradas en un modelo de implementación viable. Cabe recalcar que FeDCOR es una adaptación del modelo general de CORDRA y su primera implementación ADL-R. Dicha implementación será publicada como software de acceso público en el futuro mediato.

Los diferentes componentes de este modelo son ilustrados en la Figura 6 y sus características principales son :

1. El componente principal de coordinación del Registro compuesto por el motor de Registro y la interfaz de comunicación CORDRAWEB. Este módulo es responsable por la coordinación de todos los módulos del Registro.
2. El módulo de validación es responsable por la validación de los documentos XML, así como las reglas de operación tanto a nivel del Registro como de cada comunidad.
3. El repositorio de Registro provee los recursos necesarios para el almacenamiento de las diferentes instancias de meta datos y las representaciones de los objetos de contenido.
4. El motor de indexación basado en la distribución lucene es responsable por la indexación de los meta datos específicos de cada objeto de contenido.
5. El servidor Handle del Registro almacena los handles creados con fines internos, así como los específicos de cada instancia de meta datos.

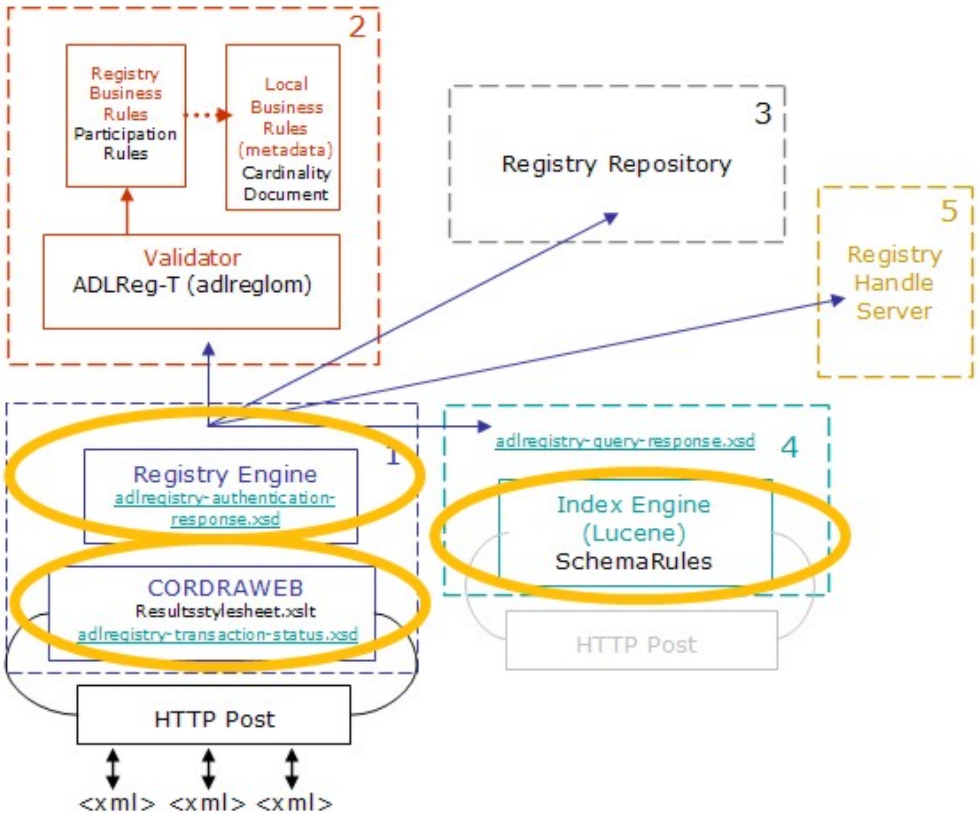

Figura 6: Componentes de la Arquitectura ADL-R que necesitan ser modificados

Los componentes marcados en la figura 6 que necesitan ser adaptados para los fines de FeDCOR son:

- El motor de Registro
- El módulo CORDRAWEB
- El motor de indexación

El Motor de Registro

El motor de Registro a través de su principal componente registrylib es responsable por la coordinación de todos los módulos que componen la arquitectura. Como parte de la estructura de operación la información recopilada de los diferentes repositorios institucionales es validada, indexada y posteriormente almacenada dentro del Registro. Registrylib coordina la aplicación de las diferentes reglas de operación así como las reglas de flujo de estos datos entre los componentes del Registro. Las diferentes operaciones coordinadas por el Registro son: inserción, actualización, retiro y eliminación.

En el caso de una inserción, los meta datos recopilados son primero validados contra las reglas de operación; y posteriormente, indexados y almacenados dentro del Registro. Resulta interesante señalar que cada nueva instancia es adicionada como una nueva secuencia de datos al interior de una entidad de representación del objeto de contenido (CORE). Cada CORE tiene un identificador extraído de la primera instancia de meta datos registrada para cada objeto de contenido. Si el mismo objeto está presente en otro Registro DSpace, las instancias de meta datos adicionales son almacenadas dentro del mismo CORE. Por lo tanto; múltiples instancias o copias del mismo objeto de contenido pueden ser almacenadas por referencia dentro del mismo Registro FeDCOR. Debe notarse sin embargo que esta medida no absuelve el problema adicional de determinar cuando dos objetos en diferentes instancias de DSpace son copias del mismo objeto; especialmente cuando éstas no han sido identificadas como copias utilizando el mismo handle.

En ADL-R el módulo de validación procede a validar primero los documentos XML y posteriormente las reglas de operación de la comunidad LOM que no pueden ser reflejados en esquemas XML. En el caso de FeDCOR este paso adicional no es necesario dado que las reglas de operación de la comunidad son validadas al nivel de cada repositorio. FeDCOR sólo necesita validar las reglas de operación del Registro antes de proceder a la indexación de los datos.

Adicionalmente, dado que los repositorios DSpace manejan los handles de cada objeto de contenido, FeDCOR no se ocupa de la administración de estos handles; concentrándose tan sólo en los correspondientes a las instancias de meta datos.

CORDRAWeb (Interfaz de aplicación)

CORDRAWeb, o la interfaz de aplicación provee una interfaz de acceso a los diferentes servicios provistos por el Registro. En el caso de FeDCOR, la federación de repositorios DSpace es posible gracias a la adición de un agente de recolección. Este agente se comunica a través de OAI-PMH con los diferentes repositorios DSpace. (Obsérvese Poblando FeDCOR en la siguiente sección)

El contenido y los meta datos registrados en FeDCOR son expuestos a través de CORDRAWEB; y las búsquedas realizadas a través de esta interfaz retornan tanto los

meta datos indexados como los handles asociados con las diferentes instancias de DSpace. Estos handles permiten recuperar los diferentes objetos al interior de los repositorios institucionales viabilizando la federación de contenido.

A diferencia de ADL-R, el cual tiene una actitud más pasiva; FeDCOR interactúa directamente con los repositorios DSpace para recolectar su información. Esto se logra gracias al agente de recolección mencionado previamente; el cual monitorea los repositorios DSpace y registra automáticamente los cambios detectados. Los detalles de este procedimiento son expresados en la sección Poblando FeDCOR

El motor de indexación

El motor de indexación; al cual FeDCOR accede a través de una interfaz estándar http, provee los recursos de catalogación al Registro. Este motor provee un sistema de catalogación extensible y dinámica que permite a la arquitectura adaptarse a los múltiples tipos de meta datos.

Las características de los campos de indexación son controladas a través de reglas de esquemas reflejadas en un archivo de configuración. Este archivo indica las diferentes sendas de XML hacia los elementos a ser traducidos a campos de indexación. Dado que FeDCOR utiliza un estándar de meta datos diferente al de ADL-R; cambios en este archivo fueron necesarios para reflejar los nuevos campos de indexación

(3) Poblando FeDCOR

La población de FeDCOR depende de la existencia y la actividad de repositorios DSpace que participan de la federación. Es por esta razón que a fin de monitorear a estos participantes y obedecer las características de CORDRA; el cual requiere un Registro de repositorios [7], se introduce el concepto de un Registro de Repositorios Institucionales IRR.

Los datos contenidos dentro de IRR ayudan a identificar los diferentes repositorios así como sus características de autentificación y de acceso. Esta información es utilizada de acuerdo a los dos esquemas de acceso que componen FeDCOR: el modelo PUSH o de empuje y el PULL o de jalado.

Los diferentes repositorios pueden registrarse, excluirse o consultarse a través de la interfaz de IRR que es muy similar a la del propio Registro principal.

PULL

En el modelo PULL los repositorios se registran en el IRR que almacena los detalles de autentificación y la dirección final de cada proveedor de datos DSpace.

Los repositorios DSpace registrados en el IRR son monitoreados periódicamente a fin de detectar cambios en los objetos de contenido almacenados en ellos. El agente realiza este

monitoreo utilizando el protocolo OAI-PMH y registra cualquier cambio detectado propagándolo hacia FeDCOR.

Los datos recopilados (harvested) son validados y posteriormente almacenados e indexados. Adicionalmente, handles para las diferentes instancias de meta datos son creados.

PUSH

En el modelo PUSH los repositorios proceden a instalar un plug-in para DSpace que es distribuido por FeDCOR. El propósito de este plug–in es el de comunicarse con el IRR cuando encuentra cambios en el repositorio DSpace a fin de iniciar una nueva recopilación por parte del agente OAI-PMH de FeDCOR. Dado que este plug-in reside al interior del repositorio institucional, este tiene mayores derechos de acceso para monitorear y controlar las diferentes colecciones contenidas en el.

El plug-in está diseñado para monitorear repositorios DSpace que periódicamente se encuentran desconectados o que bloquean el acceso remoto regular de recopiladores automáticos. Este plug-in provee a los repositorios la capacidad de controlar la interacción con el Registro y el agente de recolección.

El diagrama de flujo del agente de recolección se muestra en la Figura 7.

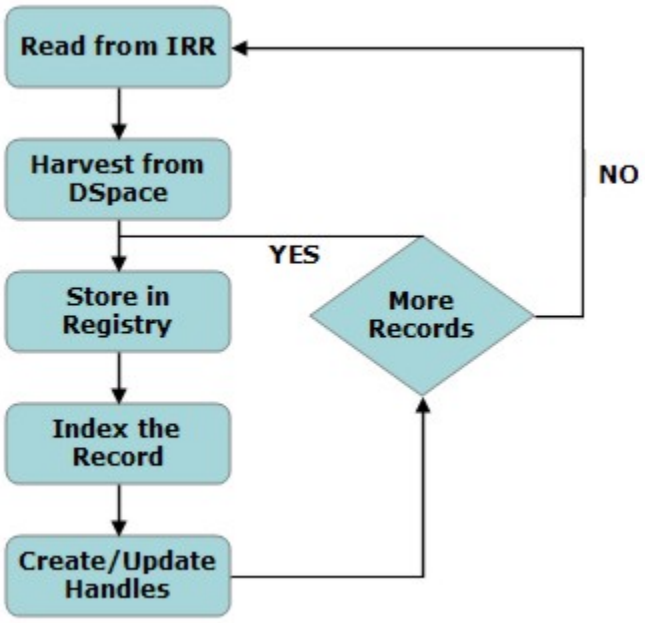

Figura 7: Diagrama de flujo de los agentes de recolección.

Implementación FeDCOR de la Arquitectura CORDRA

Como se ilustra en la figura 8; la arquitectura de FeDCOR preserva la mayoría de los componentes de ADL-R. Se adiciona sin embargo la primera implementación de un Registro de repositorios y un agente de recopilación inteligente; así como el plug-in de Registro y actualización automático. El resultado es una federación CORDRA que reutiliza la mayoría del código fuente de ADL-R y provee un Registro para una comunidad diferente y provee una serie de servicios CORDRA básicos.

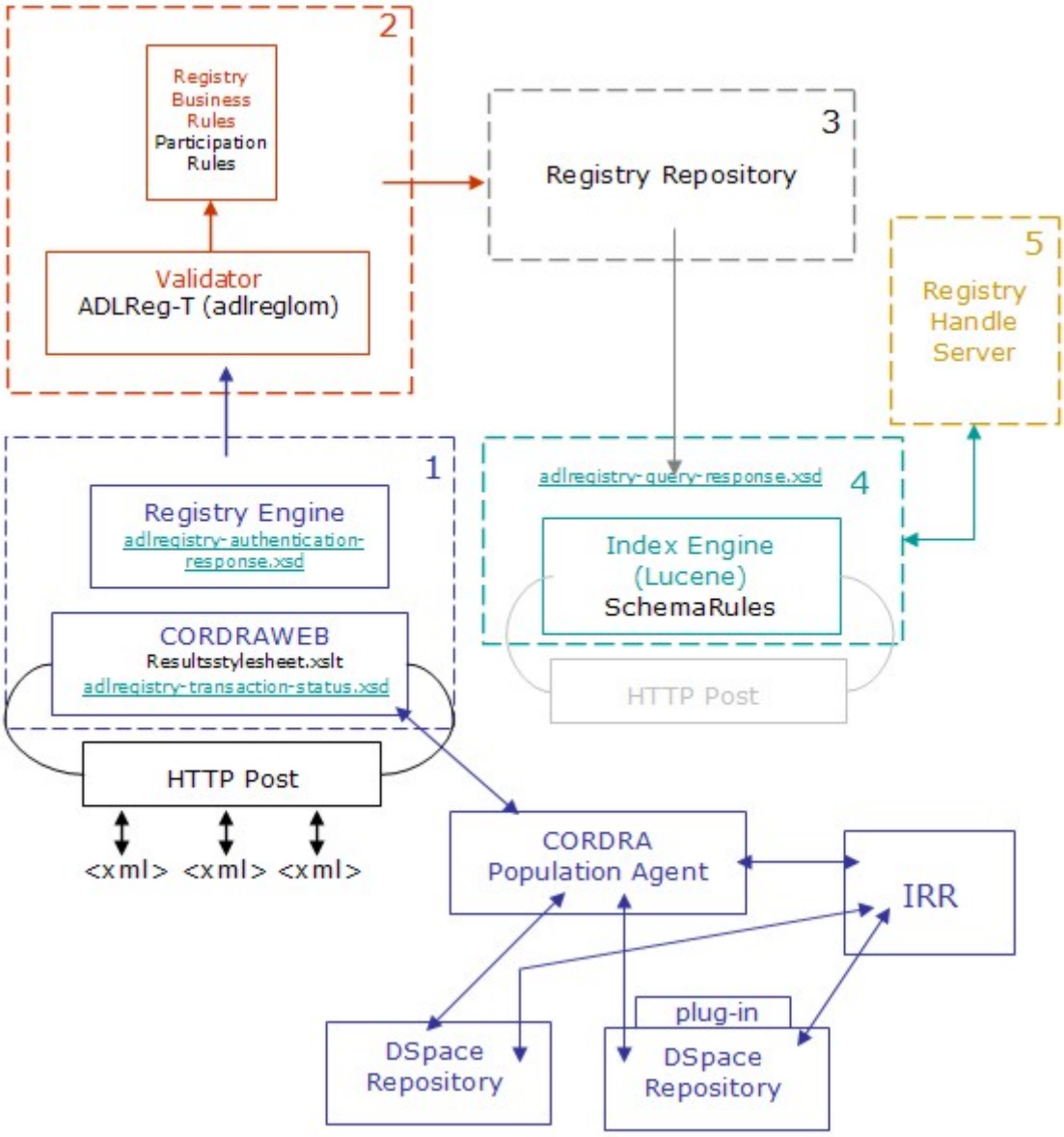

Figura 8: Componentes de FeDCOR

Investigaciones Futuras

Integración de FeDCOR en CORDRA

FeDCOR debe ser integrado dentro de la arquitectura final de CORDRA a fin de explotar todo el potencial de la tecnología y el modelo de federación discutido en este documento. La figura 9 ilustra el modelo CORDRA con la integración de FeDCOR. Esta arquitectura permitirá a los usuarios acceder a una multitud heterogénea de comunidades y contenido a través de un sistema combinado. Esta arquitectura que mantiene independencia a diferentes niveles permitirá a los usuarios efectuar búsquelas tanto a nivel global como local al nivel de sus propias comunidades. Al mismo tiempo proveerá la posibilidad de propagar las búsquedas de contenido automáticamente siguiendo el concepto original de búsqueda y descubrimiento a cualquier nivel con un grupo genérico agregado de meta datos disponible al nivel superior.

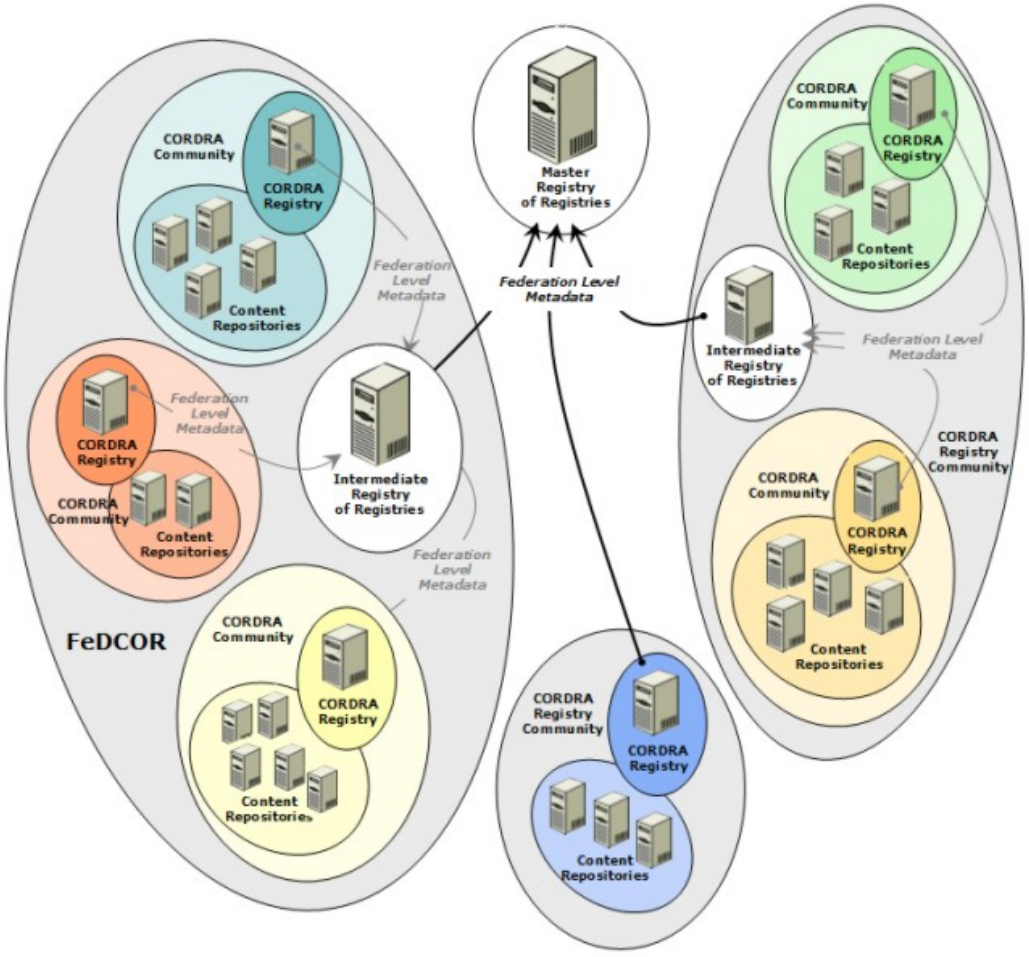

Figure 9: FeDCOR integrado con CORDRA

Actualmente nuestros esfuerzos se encuentran centrados en la construcción del primer Registro de Registros RoR CORDRA que permitirá integrar meta datos de múltiples

comunidades. Este Registro nos permitirá consolidar la extensibilidad del modelo CORDRA y proceder con el aprovisionamiento de nuevos servicios basados en la arquitectura.

Futuros resultados de nuestro trabajo con FeDCOR, así como del desarrollo de CORDRA serán comunicados a la comunidad oportunamente; y tenemos la intención de hacer públicas las métricas resultantes de pruebas realizadas con FeDCOR a través de artículos a publicarse probablemente en D-Lib Magazine.

Actualmente, FeDCOR se encuentra en etapa de pruebas en colaboración con el Consorcio Ibero Americano para la Educación en Ciencia y Tecnología- STEC, el cual se encuentra evaluando una versión de FeDCOR desde Abril del 2006.

Conclusiones

FeDCOR permite la federación de comunidades DSpace utilizando la arquitectura CORDRA. Este Registro no es tan sólo el primer Registro CORDRA de la comunidad de Bibliotecas Digitales sino también la primera federación DSpace públicamente disponible de la que estamos al tanto.

Además de servir como una prueba de campo de la aplicabilidad de CORDRA a otras comunidades; FeDCOR beneficia a la comunidad DSpace al proveer un federador genérico para sus repositorios.

Agradecimientos

Los autores desean expresar su gratitud a los Co-Labs de la iniciativa de Aprendizaje Avanzado Distribuido que financió la mayor parte del trabajo de ADL-R así como los foros de discusión de los usuarios y desarrolladores de DSpace por contribuir con su experiencia y conocimiento a este proyecto. Los autores desean además reconocer el trabajo inicial del equipo de prototipos del Laboratorio Nacional de Los Alamos en la federación de repositorios utilizando OAI-PMH dentro de la arquitectura ADORE [8] que sirvió de inspiración a este proyecto.

Este trabajo fue presentado como parte del trabajo de maestría de Giridhar Manepalli en la Universidad Old Dominion bajo la supervisión del Dr. Michael Nelson.

Referencias

[1] Jerez, Henry, Giridhar Manepalli, Christophe Blanchi, and Laurence W. Lannom. "ADL-R: The First CORDRA Registry". *D-Lib Magazine*, Volume 12, Number 2, February 2006. <doi:10.1045/february2006-jerez>.

[2] DSpace Federation. <http://www.dspace.org/>.

[3] Kraan, Wilber, and Jon Mason. "Issues in Federating Repositories". *D-Lib Magazine*, Volume 11, Number 2, March 2005. <doi:10.1045/march2005-kraan>.

[4] The Handle System. <http://www.handle.net/>.

[5] Open Archives Initiative Protocol for Metadata Harvesting. Document Version 2004/10/12T15:31:00Z. Open Archives Initiative, October 19, 2005. <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

[6] *Dublin Core Metadata Initiative*. November 7, 2005. OCLC Research. November 13, 2005. <http://dublincore.org/>.

[7] Rehak, Dan; Philip Dodds and Larry Lannom. "A Model and Infrastructure for Federated Learning Content Repositories", *Interoperability of Web-Based Educational Systems Workshop*, Volume 143 or *CEUR Workshop Proceedings*, May 10, 2005. <http://cordra.net/cordra/information/publications/2005/www2005/cordrawww2005.pdf>

[8] Jerez, Henry; X. Liu; P. Hochtenbach; and H. Van de Sompel, "The multi-faceted use of the OAI-PMH in the LANL Repository". *Proceedings of the fourth ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL 2004.