

Una Representación en UML del Metamodelo Estándar ISAD(G) e ISAAR(CPF) para la Descripción de Archivos Digitales

J. Sáenz, C. Costilla
Grupo de Bases de Datos
Universidad Politécnica de Madrid
jsaenz@eui.upm.es, costilla@dit.upm.es

E. Marcos, JM. Cavero
Grupo KYBELE
Universidad Rey Juan Carlos
{e.marcos, j.m.cavero}@escet.urjc.es

Resumen

Este trabajo se encuadra dentro del proyecto DAWIS (Digital Archive Web Information Systems) que se está desarrollando actualmente en las universidades Politécnica y Rey Juan Carlos de Madrid. El objetivo de DAWIS es la construcción de un entorno web que posibilite la generación sistemática y (semi-) automática de archivos digitales, y su integración virtual para ser consultados a través de la Web. En este trabajo se propone una descripción en UML de un metamodelo de archivo digital, basado en los estándares ISAD(G) e ISAAR(CPF). Esta descripción en UML jugará un papel principal en la especificación de una ontología para archivos digitales. Tal ontología será un modelo de referencia para la creación de archivos digitales, a la vez que un metamodelo que, posteriormente, permita la integración de diferentes archivos digitales ya existentes.

Palabras clave: archivo digital, ISAD(G), ISAAR, modelado conceptual, ontología, web, generación automática, integración, UML.

1 Introducción

Un archivo digital (AD) es una colección muy grande de documentos en formato digital, junto con información descriptiva acerca de dichos documentos (metadatos), almacenados en un repositorio o base de datos, con el fin de poder gestionarlos de la forma más adecuada y de proporcionar una recuperación sistemática de la información contenida en el AD. Dada la naturaleza de la información, el repositorio contendrá datos semi-estructurados, estructurados y multimedia.

La digitalización y el procesado automático de archivos pueden mejorar considerablemente la gestión de ese tipo de información. Pero, quizás, la principal ventaja es que cualquier persona, en cualquier lugar y momento, pueda acceder a diferentes tipos de datos de forma transparente. La comunidad de Bases de Datos y de Inteligencia Artificial tienen intereses comunes en las actividades de investigación dirigidas a la integración de fuentes de datos, que cada vez son más numerosos [1, 14, 23]. Además, la integración de fuentes de datos recobra especial relevancia en el entorno web. La integración de distintas clases de datos, proporcionados por diferentes fuentes, homogéneas y heterogéneas, ha sido bastante investigada durante

los últimos años y se han propuesto diferentes arquitecturas de integración y federaciones de datos [2, 3, 5, 10, 11, 13, 20].

A pesar de los importantes progresos que estamos viviendo, todavía se siguen construyendo sistemas de integración de datos de forma bastante artesanal, lo que requiere tediosos procesos de intensivos trabajos y donde se corren serios peligros de caer en diversos errores y carencias funcionales. La llegada de lenguajes y medios para crear e intercambiar datos semi-estructurados, como XML, y la web semántica con DAML+OIL y OWL está acelerando cada vez más la necesidad de contar con diversos sistemas de integración de datos, y todo esto agudiza y otorga más relevancia al problema ya descrito.

DAWIS es un proyecto cuyo objetivo es la construcción de un entorno web de forma sistemática y (semi-)automática que facilite y dé soporte para ofrecer un acceso integrado a múltiples ADs. En dicho entorno, la idea de DAWIS es proponer una solución genérica para posibilitar la construcción de ADs que quieran ser publicados en la web, así como la integración consultiva de los ya existentes. Ésta es la razón por la que resulta necesario definir técnicas específicas, herramientas, modelos, lenguajes, etc, como medio para conseguir la gestión sistemática de ADs y la construcción de una arquitectura para su integración genérica, así como la construcción de portales web que faciliten al usuario el acceso integrado a los ADs.

Para los objetivos de DAWIS resulta necesario, en primer lugar, definir qué es un AD y qué elementos debe contener; es decir, definir una ontología, de acuerdo con Gruber [15]. Esta ontología se puede describir mediante distintos lenguajes [23], en función de cómo vaya a ser utilizada. En DAWIS, servirá como: a) un modelo de referencia al que cada AD deberá ajustarse; b) un metamodelo para la integración de distintos tipos de AD a través de la Web.

En este trabajo describimos una ontología estática, utilizando UML como lenguaje de modelado, ya que no está pensada para poder realizar inferencias de conocimiento mediante ella, por eso la denominamos estática. Lo que buscamos ahora es organizar conceptualmente los elementos descriptivos de un archivo cualesquiera. Para lo cual, esta ontología estática debe estar basada en reglas nacionales e internacionales proporcionadas por los expertos en materia archivística. Por ello, este trabajo considera dos estándares: el *General International Standard Archival Description* (ISAD(G)) [18] y el *International Standard Archival Authority Record for Corporate Bodies, Persons and Families* (ISAAR(CPF)) [7], ambos promovidos y adoptados por el *International Council on Archives* (ICA).

El resto del trabajo está organizado como sigue: las secciones 2 y 3 describen - en UML- un metamodelo de AD que es conforme a los estándares mencionados y que constituirá una importante taxonomía básica o fundamental; a partir de la cual, se puedan codificar, en sucesivos trabajos, las correspondientes ontologías dinámicas en DAML+OIL o en OWL (aquí omitido). Finalmente, en la sección 4 se presentan algunas conclusiones y futuros trabajos.

2 Descripción de las normativas ISAD(G) e ISAAR(CPF)

Los archivos del mundo real tienen, en sí mismos, una naturaleza distribuida (oficinas, edificios, ciudades, etc.) y están pensados para almacenar una inmensa cantidad de información. Adicionalmente, el servicio que puede ofrecer un archivo (su objetivo funcional) es de lo más variopinto que podamos imaginar (desde los destinados a custodiar y ofrecer determinada herencia cultural o históricos, hasta los de la policía, los de las administraciones públicas –ayuntamientos, consejerías, parlamentos, etc.-). De esta manera, algunos archivos pueden contar con ciertas normas reguladoras sobre su funcionamiento (reglamentos de las administraciones públicas, por ejemplo) mientras que otros no. A todo esto hay que añadir que, en el caso de ADs, necesitamos contar con los más diversos formatos y diferentes medios (vídeo, audio, fotos, mapas, planos, gráficos, textos, etc.).

Esto dicho dificulta el proceso de obtención de información válida de entre toda la información contenida en los ADs. Además, hoy existen muy pocas herramientas para crear y gestionar (de forma sistemática) ADs de acuerdo a estándares internacionales (no en vano se considera hoy a la rama de la archivística digitalizada como la ‘hermana pobre’ de las bibliotecas digitales).

Hasta hace pocos años, el principal problema ha sido la adopción de un estándar internacional para la descripción de ADs que regulara las formas de recuperar y compartir información entre archivos dispersos por todo el mundo. En algunos países existen estándares adoptados a nivel nacional como *Archives, Personal Papers, and Manuscripts* (APPM) en USA, *Rules of Archival Description* (RAD) [6] en Canadá, y *Manual for Archival Description* (MAD) [8] en Gran Bretaña. Basándose en ellos, se han adoptado algunos estándares internacionales, como *Dublin Core Metadata Initiative* [12], la *Open Archive Initiative* (OAI) [21, 22], ISAD(G) e ISAAR(CPF), por citar los más importantes. Todos estos estándares definen un conjunto de metadatos para la descripción de cualquier recurso archivístico de forma clara y precisa.

Para el proyecto DAWIS, se han adoptado ISAD(G) e ISAAR(CPF) porque son estándares internacionales impulsados por el Committee on Descriptive Standards, son unos modelos claros, sencillos y completos, y actualmente no cuentan con herramientas software que los soporten. Sin embargo, otros modelos, como por ejemplo, DUBLIN CORE que sí poseen este tipo de herramientas.

Este trabajo propone una taxonomía para los ADs basada en los estándares ISAD(G) e ISAAR(CPF), con el fin de que sirva de base para establecer una referencia o fundamento para la construcción del entorno web, así como de herramientas informáticas para la generación (semi-)automática de ADs, que es el objetivo general de DAWIS.

ISAD(G) está inspirado en APPM, MAD y RAD, y su objetivo es servir de guía para la descripción de ADs mediante un conjunto de elementos y reglas generales que pueden ser aplicadas independientemente de la forma o el medio del material a

archivar. Además, este conjunto de reglas generales está pensado para poder satisfacer los siguientes requisitos y características:

- asegurar la creación de descripciones consistentes, apropiadas y autoexplicativas
- facilitar la recuperación y el intercambio de información sobre material archivado
- posibilitar la compartición de datos de autoridades
- hacer posible la integración de descripciones desde diferentes localizaciones en un sistema de información unificado

Para poder concretar cuál será la Unidad Archivística que –en cada caso- se desea describir, la norma ISAD(G) establece una organización en jerarquía (de agregación o composición) [19]. ISAD(G) entiende por Unidad de Descripción (UD) el objeto a describir, y dicho objeto puede estar localizado a cualquier nivel de los que se proponen en dicha jerarquía. Para ello, ISAD(G) define un conjunto de 26 elementos descriptivos (metadatos), independientemente de su formato y presentación, organizados en siete áreas de información: *de identidad, de contexto, de contenido y estructura, de condiciones de acceso y uso, de material asociado, de notas y de control de descripción.*

La jerarquía contiene cuatro tipos de niveles de recursos: *fondo, serie, expediente y elemento o unidad documental.* La descripción de cualquier UD requiere, como esenciales, sólo seis de los veintiséis metadatos, los demás son opcionales. Se consideran como metadatos esenciales los 6 siguientes: *código de referencia, título, creador, fecha(s), extensión de la unidad de descripción y nivel de descripción.*

Acercas de la información sobre autoridades archivísticas, el Committee on Descriptive Standards adoptó en 1995 un conjunto de metadatos estándar, ISAAR(CPF), que describe la información que representa a una autoridad en cada una de las tres formas en que puede ser descrita: *corporación, persona o familia;* y está organizado en tres áreas: *de control de autoridad, de información y de notas.*

En ISAD(G) e ISAAR(CPF) no se ha estandarizado cada formato de los elementos de descripción, sino que se deja en manos, y depende, de cada país e idioma en el que son descritos. Por tanto, la estructura del metamodelo, como se propone en la siguiente sección será independiente de esos formalismos.

Los elementos mediante los cuales ISAD(G) e ISAAR(CPF) representan un AD van a ser la referencia para el establecimiento posterior de una ontología que permita integrar estos elementos con los propios de otros sistemas de descripción de ADs.

3 Metamodelo ISAD(G) e ISAAR(CPF)

En esta sección se propone una descripción en UML de las normas ISAD(G) e ISAAR(CPF), basada en los 26 elementos descriptivos y en las reglas para su organización. Esta descripción se puede considerar como un punto de partida para establecer una ontología estática de AD, que será usada como un modelo de referencia para la generación (semi-)automática de ADs, y como un metamodelo para la integración de consultas a través de la Web.

De los seis elementos esenciales que define ISAD(G) para describir un AD, cinco de ellos constituyen el área de identidad (*código de referencia o signatura, título, fecha(s), nivel de descripción y soporte*), y el sexto (*nombre del productor o creador de la UD*) pertenece al área de contexto de la UD. Estos elementos esenciales se representan en el modelo como miembros de la clase UNID_DESCRIP (véase fig. 1) excepto el nivel de descripción, que se ha modelado como una generalización, según se muestra en la fig. 2. Adicionalmente, hemos modelado al elemento fecha(s) y a la referencia a la autoridad archivística que hace la descripción, como relaciones de composición y agregación, con las clases correspondientes.

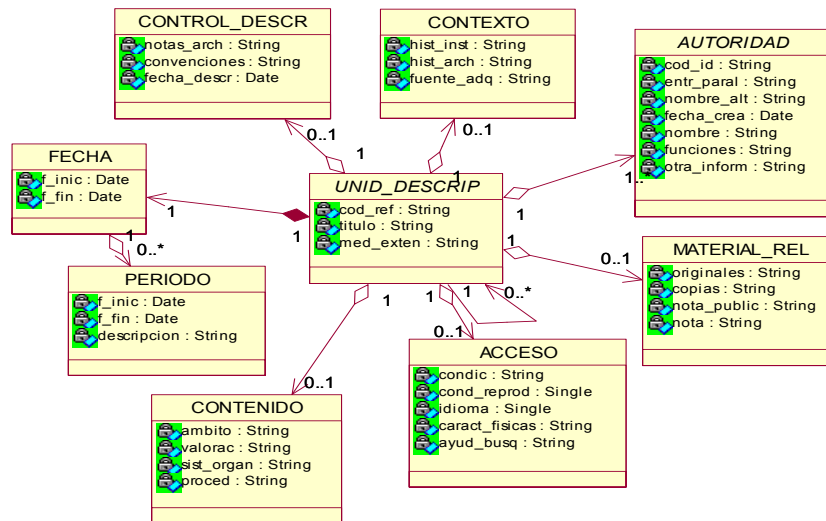


Figure 1. Elementos descriptivos de un Archivo Digital

El resto de los elementos (20), llamados elementos accesorios, se agrupan en seis áreas. En el modelo propuesto, cada una de estas áreas se ha representado mediante una clase, excepto el área de notas, que se ha agrupado con la de material asociado, en la clase MATERIAL_REL y la relación de agregación de la clase

UNID_DESCRIP consigo misma. Como esos elementos son opcionales, todas las asociaciones con la clase UNID_DESCRIP han sido modeladas como relaciones de agregación con cardinalidad 0..1.

Un AD puede contener fondos, series, expedientes y elementos documentales. Un fondo puede contener otros fondos (subfondos), series y/o expedientes. Una serie puede contener otras series (subseries) y/o expedientes. Un elemento documental sólo puede estar contenido en un expediente. Las clases y asociaciones propuestas para representar esa información se muestran en la fig. 2. Se han modelado como una generalización al considerar que todas ellas son diferentes formas de unidades de descripción, aunque reguladas por una jerarquía bien definida, en la que el fondo está al más alto nivel y el elemento documental al nivel más bajo.

La información sobre las autoridades archivísticas se ha descrito en base al estándar ISAAR(CPF), propuesto y adoptado por el International Council of Archives. La Fig. 1 ha mostrado que cada descripción debe estar respaldada por una autoridad archivística, representada mediante una relación de agregación entre las clases UNID_DESCRIP y AUTORIDAD, con cardinalidad 1..1.

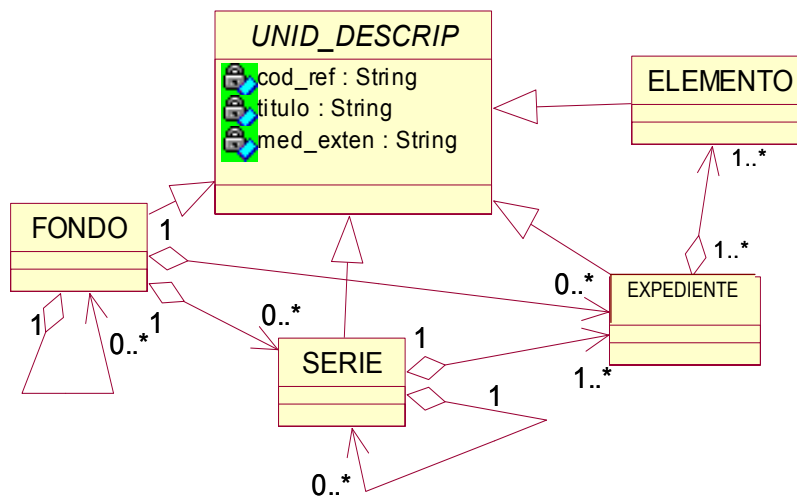


Figure 2. Niveles de descripción de un AD

Las distintas clases de autoridades y sus características se muestran en la fig. 3, mediante una relación de generalización.

Como puede apreciarse en la figura 3, la clase AUTORIDAD es la clase base genérica que describe los elementos comunes a las clases CORPORACION, PERSONA y FAMILIA. Estas clases especializadas identifican cada uno de los

correspondientes tipos, y describen los correspondientes atributos específicos (metadatos).

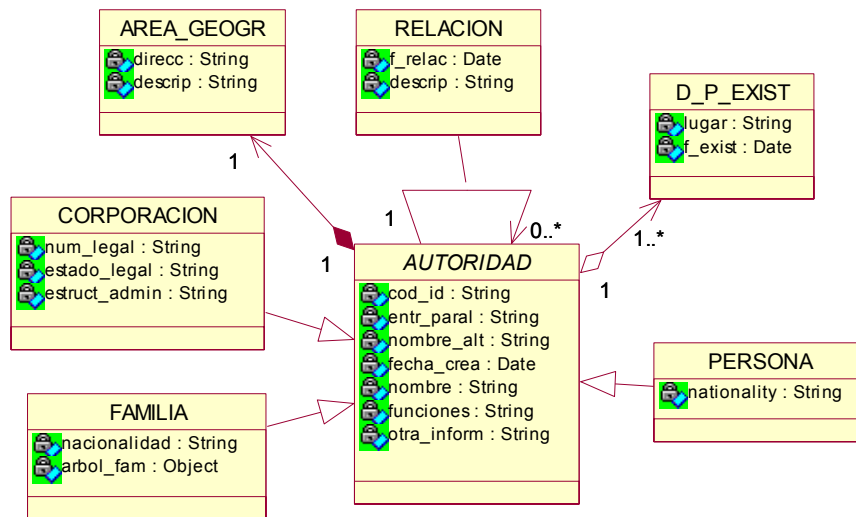


Figure 3. Descripción de autoridad

4 CONCLUSIONES

El objetivo del proyecto DAWIS es proponer un entorno web que sirva como solución genérica para la creación (semi-) automática y la integración de ADs, de forma unificada, y que facilite las consultas sobre ellos, independientemente de su localización, a través de la Web. Una parte de este objetivo general deberá basarse en la definición de ciertas ontologías específicas para los ADs.

En este trabajo, proponemos una descripción en UML de un AD, basada en los estándares ISAD(G) e ISAAR(CPF), como una primera aproximación estática a una ontología. Actualmente, estamos trabajando en una extensión de esta descripción conceptual, integrando Dublin Core, un modelo de archivo parlamentario y el conocimiento de expertos en archivística.

Los siguientes pasos, entre otros, serán la codificación de ontologías dirigidas a propiciar la integración de ADs para su acceso consultivo en la web, la implementación de la correspondiente arquitectura de integración [13], así como el desarrollo de un prototipo de herramienta que permita la generación (semi-) automática de ADs. Dicha herramienta será configurable, tanto en cuanto a los metadatos a usar en cada AD en particular, como en la terminología usada para describir esos metadatos.

RECONOCIMIENTOS

Este trabajo está parcialmente financiado por el Ministerio de Ciencia y Tecnología (MCYT- TIC2002-04050-C02, proyecto DAWIS).

REFERENCIAS

- [1] AnHai Doan and Robert McCann, *Building Data Integration Systems: A Mass Collaboration Approach*, 18th Int. Joint Conference on Artificial Intelligence (IJCAI), 2003. También disponible en <http://www.isi.edu/info-agents/workshops/ijcai03/papers>, August, 2003.
- [2] Ashish N and Knoblock C, *Wrapper Generation for Semistructured Data*, ACM Sigmod Record, V. 26, N. 4, December, 1997.
- [3] Bermúdez de Andrés J, *Una lógica de descripciones en un nivel meta-ontológico para la gestión de sistemas de información globales*, Tesis Doctoral, Dpto. Lenguajes y Sistemas Informáticos, Universidad del País Vasco, Noviembre, 2001 http://sun3.lib.uci.edu/~blandis/faf2000_content.pdf. Downloaded 2003-03-05.
- [4] Bill Landis, *What is ISAD(G)?*, Description Section, Society of American Archivists, 2000, http://sun3.lib.uci.edu/~blandis/faf2000_content.pdf. Downloaded 2003-03-05.
- [5] Brisaboa N, Parama J, Penabad M, Places A, Rodriguez F. *Solving Languages Problems in a Multilingual Digital Library Federation*. EURASIA-ICT'2002. LNCS, Springer Verlag, pp. 503-510, Teheran, Irán, October, 2002.
- [6] Canadian Committee on Archival Description and Canadian Council of Archives. *Rules of Archival Description (RAD) (Revised Version – December 2002)*. Ottawa, Canada, 1990. http://www.cdncouncilarchives.ca/rad_part1.pdf. Downloaded 2003-03-05.
- [7] Commission on Descriptive Standards. International Council on Archives. *ISAAR(CPF): International Standard Archival Authority Record for Corporate Bodies, Persons, and Families, 15-20 November 1995*. http://www.ica.org/biblio/cds/isaar_eng.pdf. Downloaded 2003-03-05.
- [8] Cook M., *MAD2: Reassessing the Experience*. Archivaria 35 (Spring 1993): 15-23.
- [9] Cranefield S. *UML and the Semantic Web*. Semantic Web Working Symposium, July 30 - August 1, 2001 Stanford University, California, USA <http://www.semanticweb.org/SWWS/program/full/paper1.pdf>. Downloaded 2003-03-07
- [10] Da Rocha G, Palma S, Moreira de Souza J, *Spatial Data Integration in a Collaborative Design Framework*. *Communications of the ACM*, Vol. 46, No. 3, pp. 86-90, March 2003.
- [11] Domenig R. y Dittrich K.R. *An Overview and Classification of Mediated Query Systems*. ACM SIGMOD RECORD, Vol. 28, N. 3, pp. 63-72, September, 1999.
- [12] Dublin Core Metadata Initiative. *Dublin Core Metadata Element Set, Version 1.1: Reference Description*. 2003-02-04. <http://dublincore.org/documents/2003/02/04/dces/>. Downloaded 2003-03-05.

- [13] Eibe S, Costilla C, Menasalvas E y Acuña CJ, *DAWIS: Una Arquitectura de Integración Web para el Acceso Integrado a Archivos Digitales*. Presentado a JISBD'03, pendiente de evaluación, Alicante, Nov. 2003.
- [14] García-Molina H, Papaconstantinuo Y, Quass D, Rajaraman A, Sagiv Y, Ullman J and Widom J., *The TSIMMIS project: Integration of heterogeneous information sources*. Journal of Intelligent Information Systems, Vol. 8, N. 2, 1997.
- [15] Gruber T. *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. International Workshop on Formal Ontology, March 1993, Padova, Italy
- [16] Hensen S. *Archives, Personal Papers and Manuscripts: A Cataloging Manual for Archival Repositories Historical Societies and Manuscript Libraries*. Society of American Archivists; 2nd edition (December 1989)
- [17] Hensen S. *RAD, MAD, and APPM: The Search for Anglo-American Standards for Archival Description*. Archives and Museum Informatics 5, Summer 1991: 2-5
- [18] International Council on Archives. *ISAD(G): General International Standard Archival Description, 2nd Edition, 1999*. http://www.ica.org/biblio/cds/isad_g_2e.pdf. Downloaded 2003-03-05.
- [19] *ISAD(G) General International Standardization Archive Description*, 2nd edition, ISBN 0-9696035-5-X, International Council on Archives, Ottawa 2000
- [20] Mena E, *Observer: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies*. Tesis Doctoral, Dpto. Informática e Ingeniería de Sistemas, Universidad de Zaragoza, 1998.
- [21] Open Archives Initiative. *Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting Protocol Version 2.0 of 2002-06-14 Document Version 2002/06/13T19:43:00Z*. <http://www.openarchives.org/OAI/guidelines.htm>. Downloaded 2003-03-05.
- [22] Open Archives Initiative. *The Open Archives Initiative Protocol for Metadata Harvesting Protocol Version 2.0 of 2002-06-14 Document Version 2003/02/21T00:00:00Z*. Downloaded 2003-03-05. <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>
- [23] Wache H, Vögele T, Visser U, Stuckenschmidt H, Schuster G, Neumann H, Hübner S. *Ontology-Based Integration of Information –A survey of Existing Approaches*. Proceedings of IJCAI-01 Workshop: Ontologies and Information Sharing. Seattle, WA, 2001, pp 108-117. También disponible en <http://www.isi.edu/info-agents/workshops/ijcai01/papers>