

# CoVaMaT: Modelo conceptual de una herramienta para el soporte a la gestión de variedad en Sistemas Big Data

Agustina Buccella, Alejandra Cechich, LÍam Osycka,  
Juan Luzuriaga, Carolina Villegas, Marcos Cruz, Franco Corgatelli,  
Rodolfo Martínez, Rafaela Mazalu  
GIISCO Research Group  
Departamento de Ingeniería de Sistemas  
Universidad Nacional del Comahue  
Neuquen, Argentina  
agustina.buccella@fi.uncoma.edu.ar

## 1. Resumen

Un cambio importante con respecto a depósitos de datos tradicionales, es que en los Sistemas Big Data (SBDs) la naturaleza no estructurada de algunos datos puede provenir de diversas fuentes, entre ellas sensores, redes sociales, entorno y la misma empresa. La diversidad de esos datos puede analizarse abordando distintas características. Precisamente, la propiedad de los SBDs con respecto a diversidad de los datos se denomina *variedad*. Nuestro proyecto propone modelar *variedad* mediante casos documentados a través de las variaciones que diferentes variables pueden tomar en un contexto. Sin embargo, para la aplicación de la propuesta, es indispensable una herramienta de soporte que construya incrementalmente repositorios de variaciones a ser reusadas.

Palabras Clave : Reusabilidad - Líneas de Producto de Software - Big Data

## 2. Contexto

La línea presentada se inserta en el contexto del Proyecto UNComa: Modelado de Variedad en Sistemas Big Data Directora: Dra. Agustina Buccella, Co-directora; Dra. Alejandra Cechich (2022-2025).

## 3. Introducción

En [3] hemos definido una primera aproximación de los elementos que componen una arquitectura de referencia para SBDs basada en reuso. Uno de los componentes principales de esta arquitectura agrupa los denominados activos de dominio constituidos por artefactos de software que son creados para el dominio en el que se está trabajando. Es importante resaltar que estos activos deben generarse a partir de activos basados en conocimiento. De esta forma, se deben crear artefactos enfocados en que puedan ser reusados en el mismo dominio e incluso en otros dominios relacionados (artefactos para reuso), y/o que puedan desarrollarse en base a otros artefactos ya creados (artefactos con reuso).

Para identificar *variedad* en activos de dominio, previamente propusimos dos enfoques: (1) identificación desde los datos (bottom-up) y (2) identificación desde los requerimientos (top-down). El enfoque bottom-up de la propuesta parte de la definición de un problema dependiente del dominio e intenta detectar características variantes a partir de los datos existentes en uno o varios repositorios. Por otro lado, en el enfoque top-down se parte del análisis de los requerimientos y del dominio para contrastar suposiciones a través de los datos. Detalles de esta propuesta pueden verse

en [4, 6, 7]. La documentación de activos de dominio se describe mediante un enfoque basado en funcionalidades, donde cada funcionalidad se documenta a través de una hoja de datos funcional (datasheet), representando el conjunto de variaciones que permitan reusar activos de dominio y detectar variedades de contexto.

### 3.1. CoVaMaT: Context-based Variety Management Tool

La propuesta planteada cobra sentido cuando se tiene una forma de poder almacenar esas variedades formalizadas en forma de datasheets y cuando, además, las variaciones pueden asociarse a casos puntuales. Por ello, se debe plantear una solución que permita almacenar, consultar, modificar y visualizar la variedad de un dominio y todos los casos trabajados, siendo éstos recuperados y reutilizados cuando la variedad de contexto lo permite. La solución se basa entonces en el diseño de una herramienta software, nombrada Context-based Variety Management Tool (CoVaMaT), que tiene la intención de exponer el conocimiento adquirido dentro de un dominio de una forma declarativa y que pueda ser de soporte tanto para el proceso bottom-up como top-down en la identificación de variedad.

Previo a la presentación del modelo conceptual de la herramienta, es necesario introducir las siguientes definiciones:

- **Instanciación de Caso**: consiste en la descripción de un caso de dominio a través de la selección de variaciones que lo describen. Por ejemplo, en el dominio de recursos hídricos, los cuerpos de agua pueden ser ríos, lagos, etc. El caso en cuestión puede corresponder a uno de estos valores (ej. ríos). Además, para ese río en particular, pueden resultar relevantes valores específicos de calidad de agua, como el oxígeno disuelto, pH, etc., y de los cuales se hará también una selección. El conjunto de valores seleccionados para todas las variables que sean relevantes en el dominio, constituirá el caso instanciado.
- **Caso de Reuso**: se utiliza para componer una consulta que recupere aquellos casos

ya instanciados en el repositorio que posean características similares. Esta consulta se compone de un contexto determinado por variaciones seleccionadas (ej. río con variables de calidad de agua con foco en la turbidez). El objetivo es analizar trabajo anterior realizado bajo condiciones especificadas en el caso de reuso.

La arquitectura lógica de CoVaMaT presenta un producto que empaqueta tres servicios, con procesos que los atienden y eventos que los invocan. A continuación se explicarán estos procesos, centrando en el nivel conceptual de negocios de la herramienta, de acuerdo a la notación propuesta por el lenguaje de modelado de procesos ArchiMaté.

Documentación de Variedad: S-1, P-1 y E-1

- **S-1 Servicio de documentación de variedad**: le permite al usuario documentar los diferentes tipos de variedad que se encuentran en un dominio cuando se trabaja con el ciclo de actividades de Big Data (adquisición, transformación, análisis y presentación de datos).
- **P-1 Documentar variedad en dominio**: es el proceso que especifica cómo debe trabajarse al momento de identificar un tipo de variedad, y de construir los objetos de negocios que documenten esta variedad.
- **E-1 Comienzo de búsqueda de variedad**: es el evento que se dispara cuando se comienza a una nueva iteración del modelo dentro de un dominio.

Luego de finalizar todas las ejecuciones de (P-1) se tiene un conjunto de datasheets almacenadas representando la variedad en el dominio. Es así, que a partir de estas, puede representarse un caso de dominio seleccionando las variaciones de las diferentes datasheets.

Instanciación de caso: S-2, P-2 y E-2

- **S-2 Servicio de instanciación de caso**: le permite al usuario documentar un caso

<sup>1</sup><https://www.opengroup.org/archimate-forum/archimate-overview>

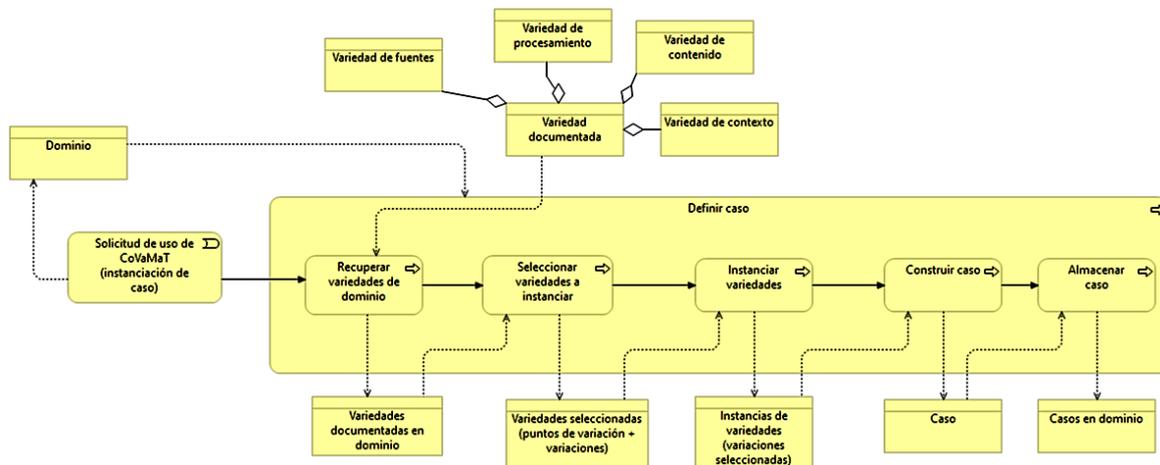


Figura 1. P-2: Definir Caso

de dominio compuesto por instancias de las diferentes variedades encontradas. El caso será identificado particularmente por su variedad de contexto.

- P-2 Definir caso : es el proceso que especifica cómo debe confeccionarse el caso de dominio que será almacenado.
- E-2 Solicitud de uso de CoVaMaT (instanciación de caso): es el evento que se dispara cuando se quiere almacenar un caso en el dominio para su futuro reuso y consulta.

En la Figura 1 puede verse el proceso modelado. Primero entonces, se recuperan todas las variedades almacenadas del dominio, luego se deben seleccionar cuáles se utilizarán para definir el caso y construir al objeto de negocio Variedades seleccionadas (puntos de variación + variaciones). Con este objeto, puede procederse a crear eCaso. Finalmente, el objeto es almacenado dentro de todos los casos del dominio.

Consulta de activos: S-3, P-3, P-4 y E-3

- S-3 Servicio de consulta de activos de dominio: con este servicio se pueden consultar activos de dominio (que serían los casos y las variedades asociadas a ellos) almacenados en la base de conocimiento de CoVaMaT. La consulta será en base a un contexto definido por el usuario que solicita usar el servicio, recuperando entonces los casos que responden al mis-

mo.

- P-3 Definir caso de reuso : en este proceso se construye un caso a partir de la variedad de contexto almacenada en el dominio.
- P-4 Consultar activos de dominio en casos similares : en este proceso se toma el caso de reuso definido en (P-3) para poder recuperar otros con contexto similar y poder observar qué variedades instanciaron.
- E-3 Solicitud de reuso : es el evento que se dispara cuando se quiere recuperar conocimiento del dominio en función de casos trabajados sobre un contexto particular.

En la Figura 2 se observan los pasos a seguir para recuperar el conocimiento de experiencias previas. Primero, se deben recuperar todas las diferentes datasheets que conforman la variedad de contexto del dominio, almacenadas en el objeto Variedad de contexto creado en (P-1). Luego, similar a (P-2), se deben seleccionar qué datasheets se usarán para definir el contexto del caso de reuso, siendo almacenadas en el objeto de negocio Variedades de contexto seleccionadas (punto de variación + variaciones).

Una vez definidas qué datasheets de contexto se usarán, se procede a instanciarlas; y se almacena en el objeto de negocio Instancias de variedades de contexto (variedades seleccionadas). A partir de las datasheets elegidas

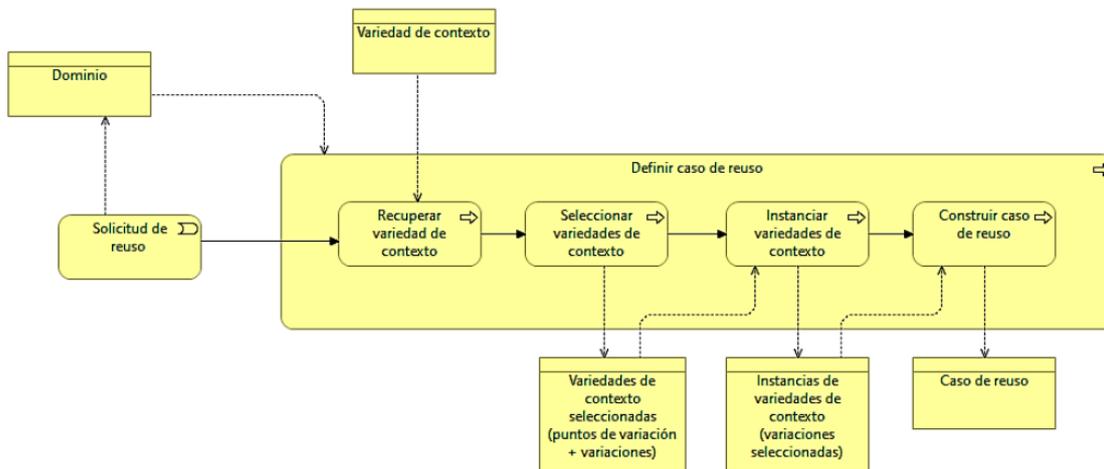


Figura 2. P-3: Definir caso de reuso

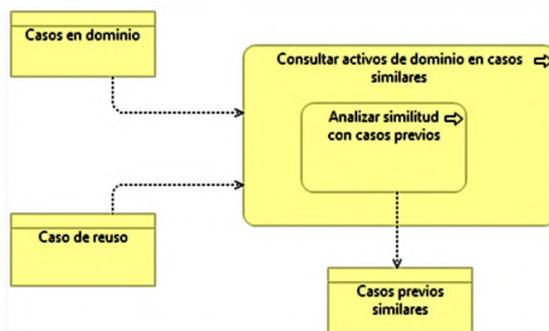


Figura 3. P-4: Consultar activos de dominio en casos similares

y las instanciaciones de cada una, se construye el objeto Caso de reuso. Con esto finaliza la ejecución de (P-3) y comienza la ejecución de (P-4).

En el proceso siguiente, Figura 3, se utilizan los objetos de negocio Casos en dominio (generado en P-2) y Caso de reuso en (P-3) como entradas para realizar una consulta de similitud, con el fin de filtrar dentro de todos los casos de dominio a aquellos que posean la misma variedad de contexto por la que se está consultando. De esta forma puede verse qué se ha hecho previamente en el dominio cuando se presentaron situaciones con un contexto particular.

#### 4. Líneas de Investigación y Desarrollo

En proyectos previos, hemos realizado amplios avances en lo que respecta al área de LPSs definiendo y refinando una metodología de desarrollo a nivel de subdominios [1] que tiene la particularidad de favorecer el reuso basado en una taxonomía de servicios [2, 5].

Actualmente, nuestra línea de investigación migra hacia un cambio con respecto a la construcción de sistemas Big Data con reuso. La variedad en SBDs ha sido relacionada con diversas propiedades como interoperabilidad, seguridad, reusabilidad, etc. En este contexto, y respondiendo a la pregunta de investigación: ¿Cómo puede modelarse la variedad de la información de dominio de manera de incorporar reusabilidad en el desarrollo de SBDs? , nuestra línea de trabajo ha avanzado en los resultados que se listan a continuación.

#### 5. Resultados Obtenidos / Esperados

Al momento, hemos planteado una arquitectura de referencia [3] y un modelo de procesos [6, 7, 4] para la gestión de la variedad en SBDs. Como primeras aproximaciones experimentales, en [8, 9], presentamos caracterizaciones de contexto de variables que influyen la turbidez en los cuerpos de agua y el índice de vegetación normalizado (NDVI). Esto per-

mite combinar datos relevantes del contexto de aplicación, lo que favorecería la identificación de situaciones recurrentes. La propuesta se ejemplifica con un caso de estudio llevado a cabo en el área del Alto Valle del Río Negro.

Continuamos trabajando en la implementación de CoVaMaT y también en colaboración con el Instituto de Tecnología Agropecuaria (INTA)-Alto Valle para la aplicación del proceso de modelado en el análisis de la napa freática, en función de la variedad de fuentes acuíferas de diversas zonas geográficas (variedad contextual)<sup>2</sup>.

## 6. Formación de Recursos Humanos

El proyecto reúne aproximadamente a 13 investigadores, entre los que se cuentan docentes y alumnos de UNComa, y colaboradoras expertas del dominio de aplicación, específicamente pertenecientes al INTA. A su vez, el proyecto cuenta actualmente con tres docentes y un magister. Varios de los docentes-investigadores de GIISCo-UNComa han terminado o se encuentran próximos a terminar carreras de grado o postgrado.

## Referencias

- [1] A. Buccella, A. Cechich, M. Arias, M. Pol'la, S. Doldan, and E. Morsan. Towards systematic software reuse of gis: Insights from a case study. *Computers & Geosciences* 54(0):9 – 20, 2013.
- [2] A. Buccella, A. Cechich, M. Pol'la, M. Arias, S. Doldan, and E. Morsan. Marine ecology service reuse through taxonomy-oriented SPL development. *Computers & Geosciences* 73(0):108 – 121, 2014.
- [3] A. Buccella, J. Luzuriaga, A. Cechich, L. Osycka, F. Paterno, M. Pol'la, M. Cruz, R. Martinez, R. Mazalu, and M. Moyano. Reusabilidad en el contexto de desarrollo de sistemas para big data. In *Actas del XXIII Workshop de Investigadores en Ciencias de la Computación*, Chilecito, La Rioja, pages 525–529, 2021.
- [4] A. Buccella, J. Luzuriaga, A. Cechich, L. Osycka, C. Villegas, M. Cruz, F. Corgatelli, R. Martinez, R. Mazalu, and M. Moyano. Modelado de variedad de activos de dominio en sistemas big data. In *Actas del XXIV Workshop de Investigadores en Ciencias de la Computación*, Mendoza pages 113–117, 2022.
- [5] Agustina Buccella, Alejandra Cechich, Juan Porfiri, and Domenica Diniz Dos Santos. Taxonomy-oriented domain analysis of gis: A case study for paleontological software systems. *ISPRS International Journal of Geo-Information*, 8(6), 2019.
- [6] L. Osycka, A. Buccella, and N. A. Cechich. Identificación de variedad contextual en modelado de sistemas big data. In *Memorias del XXVII Congreso Argentino de Ciencias de la Computación (CACIC)*, pages 367–376. Red de Universidades con Carreras en Informática, 2021.
- [7] L. Osycka, A. Buccella, and N. A. Cechich. Data variety modeling: A case of contextual diversity identification from a bottom-up perspective. In *27th Argentine Congress, CACIC 2021*, Salta, Argentina, October 4-8, 2021, Revised Selected Papers. *Communications in Computer and Information Science* 1584, pages 124–138. Springer, 2022.
- [8] Gastón Vidart, Alejandra Cechich, Agustina Buccella, and Ayelén Montenegro. Análisis de turbidez basado en caracterización de contextos. *Memorias de las JAIIO*, 8(4):170–183, dic. 2022.
- [9] C. Villegas, A. Buccella, and A. Cechich. Caracterización de variables para el análisis del Índice de vegetación. In *Memorias del XXVIII Congreso Argentino de Ciencias de la Computación (CACIC) - Workshop Alumnos*, pages 978–982, 2022.

<sup>2</sup>Convenio de Cooperación Técnica FaIF-INTA Resolución CD FaIF N°109/21