Propuesta de un sistema para la evaluación de calidad de datos a través de rutinas automatizadas y aportes de expertos incorporados en la arquitectura informática del Programa de Evaluación y Monitoreo de la Biodiversidad

Rey, Juan
Castro, Carolina
González, Iván
Vélez, Danny
Bastidas, Ricardo
Velasquez-Tibata, Jorge
López, Daniel
Grajales, Valentina

Programa Evaluación y Monitoreo de la Biodiversidad
Instituto de Investigación de Recursos Biológicos Alexander von Humboldt
Bogotá, D.C.
2017

Instituto de Investigación de Recursos Biológicos Alexander von Humboldt
Sede Principal: Calle 28A # 15-09 Bogotá, D.C., Colombia I PBX: (57)(1) 3202767 I NIT 820000142-2

Resumen

En cumplimiento de su misión, el Instituto Humboldt genera una gran cantidad de datos e información sobre biodiversidad en varios formatos y de diversa naturaleza. Dentro de los datos e información más común y relevante están los registros biológicos, que son la evidencia de la presencia de un individuo de una especie en un lugar y tiempo determinado. En el Programa de Evaluación y Monitoreo del Instituto Humboldt (PEM) los registros biológicos son un insumo fundamental en los productos que se generan en los diferentes equipos de trabajo, por lo cual en muchos casos se duplican esfuerzos en la obtención y uso de los mismos. Conscientes de esta problemática y en aras de hacer más eficientes los procesos y flujos de datos e información dentro del PEM se propone un sistema para el almacenamiento y evaluación de calidad de datos a través de rutinas automatizadas, cuyo objetivo principal es el de integrar procesos de gestión de datos e información que permitan a su vez fortalecerlos y hacerlos más eficientes, con el fin de contribuir a la visión de un Instituto articulado en sus procesos y equipos de trabajo.

Palabras clave: Calidad de datos, registros biológicos, automatización, gestión, articulación

Abstract

In fulfillment of its mission, the Humboldt Institute generates a large amount of biodiversity data and information in various formats and diverse nature. Species occurrences are one of the most common and relevant data, which evidence the presence of an organism of any species in a specific place and time. At the Evaluation and Monitoring Program of the Humboldt Institute (PEM), species occurrences are a fundamental input for the products generated in different workteams, reason why in some cases there are duplication of efforts to obtain and use them. Conscious of this problem and in order to make all the processes and data and information flows more efficient within the PEM, a system is proposed for storage and evaluate data quality through automated routines, whose main purpose is to integrate data and information management processes while strengthening them and making them more efficient in order to support the vision of an Institute articulated in its processes and workteams.

Keywords: data quality, occurrences, automation, management, articulation

CONTENIDO

Resumen 2

Abstract 2

CONTENIDO 3

LISTA DE TABLAS 5

LISTA DE FIGURAS 6

INTRODUCCIÓN 7

OBJETIVOS 8

OBJETIVO GENERAL 8

OBJETIVOS ESPECÍFICOS 8

METODOLOGÍA 9

1. DIAGNÓSTICO 9

1.1. Descripción de fuentes de información 9

1.2. Diagnóstico de las herramientas actuales 9

1.3. Diferencias entre las rutinas del procedimiento entre LBA e I2D 10

1.4. Requisitos de funcionamiento integrado 10

2. METODOLOGÍA DE ANÁLISIS Y DISEÑO DE LA SOLUCIÓN 10

2.1. Metodología de desarrollo 11

2.2 Flujo general de la solución 12

2.3 Flujo de validación de datos 12

3. ARQUITECTURA DE REFERENCIA 12

3.1 Arquitectura y requerimientos del sistema 12

RESULTADOS 13

1. DIAGNÓSTICO 13

1.1. Descripción de fuentes de información 13

1.2. Diagnóstico de las herramientas actuales 15

1.2.1. Estructuración 15

1.2.2. Taxonómico 16

1.2.3. Geográfico 16

1.2.4. Otros 17

1.2.5. Evaluación de las herramientas 20

1.3. Diferencias entre las rutinas del procedimiento entre LBA e I2D 21
1.3.1. Calidad de datos en la I2D 21
1.3.2. Calidad de datos en el LBA 23
1.4. Requisitos de funcionamiento integrado 25
2. ANÁLISIS Y DISEÑO DE LA SOLUCIÓN 25
2.1. Metodología de desarrollo 26
2.1.1 Principios de Scrum 26
2.1.2 Valores propios de Scrum 27
2.1.3 Roles de Scrum 27
2.2. Modelo de dominio 28
2.3. Arquitectura general de la solución 29
2.3.1. Flujo general de la solución 30
El flujo general de la solución tiene dos opciones de acuerdo a las necesidades del grupo que lo utilice como se muestra en la figura 10. 30
2.3.3. Flujo de validación de datos 31
2.4. Arquitectura de referencia 32
2.5. Diagrama de componentes 33
2.6. Especificación del sistema 33
2.6.1. Requerimientos Funcionales 34
2.6.2. Requerimientos no funcionales 34
2.7. Diagrama de casos de uso 35
2.8. Historias de usuario 35
3. JUSTIFICACIÓN Y ANÁLISIS DE REQUERIMIENTOS DE LA BASE DE DATOS 36
3.1 Requerimientos generales 37
3.2 Modelamiento de datos (Relacional, NOSQL) 38
3.3 Requisitos orientados al usuario 39
3.4 Requerimientos no funcionales del sistema de base de datos 39
BIBLIOGRAFÍA 41

LISTA DE TABLAS

Tabla 1. Descripción de las metodologías disponibles, puntos clave, características especiales y fallas identificadas. 11
Tabla 2. Lista de fuentes de información de las cuales provienen los registros biológicos incorporados en la I2D. 13
Tabla 3. Lista de fuentes de información de las cuales provienen los registros biológicos utilizados en el LBA. 15
Tabla 4. Referencias bibliográficas de los documentos utilizados por la I2D para la obtención del endemismo de las especies. 18
Tabla 5. Fuentes no dinámicas utilizadas por el LBA para la obtención del endemismo de las especies. 18
Tabla 6. Referencias bibliográficas de los documentos utilizados por la I2D para la construcción de una lista de especies de Colombia a 2016. La lista consta de un total de 33.085 especies distribuidas por grupo taxonómico. 19
Tabla 7. Fuentes utilizadas por el LBA para la construcción de una lista de organismos. 19
Tabla 8. Resultados de la comparación en la validación de nombres científicos entre las herramientas CoL y Taxize. 20
Tabla 9. Resultados de la comparación en la validación de coordenadas entre las herramientas R 0.2 y Java. 20
Tabla 10. Atributos para mejorar de las herramientas evaluadas. 21
Tabla 11. Matriz de requerimientos por cada uno de los tipos de consulta. 37
Tabla 12. Requisitos orientados al usuario. 39

LISTA DE FIGURAS

Figura 1. Consulta de registros seleccionados. El polígono central en rojo representa los registros seleccionados para realizar el análisis..... 9
Figura 2. Ventana de descarga de datos para el LBA..... 14
Figura 3. Validaciones realizadas por la I2D en los conjuntos de datos a incorporar como parte de su proceso de calidad de datos. 22
Figura 4. Estructura del proceso de calidad de datos llevado a cabo por el LBA. 23
Figura 5. Descripción del inicio del proceso de calidad del LBA que corresponde a la estructuración de información. 23
Figura 6. Cálculo de flags como parte del proceso de calidad del LBA. 24
Figura 7. Esquema general flujo de trabajo propuesto por la I2D y el LBA. 25
Figura 8. Modelo de dominio del sistema de calidad de datos. 29
Figura 9. Modelo de dominio del sistema de calidad de datos. 30
Figura 10. Flujo general de la solución propuesta. 31
Figura 11. Flujo de validación de datos..... 32
Figura 12. Estructura de la arquitectura de referencia..... 33
Figura 13. Diagrama de componentes del sistema de calidad de datos. 33
Figura 14. Diagrama de casos de uso del sistema de calidad de datos. 35

INTRODUCCIÓN

En cumplimiento de su misión, el Instituto Humboldt genera una gran cantidad de datos e información sobre biodiversidad en varios formatos y de diversa naturaleza. Dentro de los datos e información más común y relevante están los registros biológicos, que se refiere a la evidencia de la presencia de especies en un lugar y tiempo determinado. Dichos registros son utilizados en una gran variedad de procesos de generación y síntesis de información y conocimiento dentro del Instituto, sin embargo, también contribuyen en la generación de líneas base (o estado del arte) de información y a dar respuesta a una amplia gama de solicitudes de información internas y externas al Instituto.

Existen varias fuentes de donde provienen dichos registros, están las colecciones biológicas, los proyectos que generan información de observaciones y por último las fuentes secundarias que incluyen: literatura, bases de datos sobre biodiversidad con información ya disponible (ej. *Global Biodiversity Information Facility* - GBIF) y conjuntos de datos no públicos que se encuentran bajo custodia directa de investigadores de diversas instituciones.

Dentro de los procesos que se adelantan en el Programa de Evaluación y Monitoreo del Instituto Humboldt (PEM) son frecuentemente utilizados los registros biológicos, duplicando en muchos casos los esfuerzos en la obtención y uso de los mismos, procesos durante los cuales es fundamental la aplicación de validaciones específicas que garanticen unos mínimos de calidad que aseguren la apropiada generación de los productos finales, por lo que muchas rutinas de calidad y limpieza de los datos también se duplican dentro del programa dependiendo de la unidad o línea donde se utilizan los registros biológicos.

Conscientes de esta problemática y en aras de hacer más eficientes todos los procesos y flujos de datos e información dentro del PEM, en el Plan Operativo Anual 2017 del Instituto se planteó realizar en conjunto, desde diferentes enfoques del programa, la propuesta de un sistema para la evaluación de calidad de datos a través de rutinas automatizadas y aportes de expertos incorporados en la arquitectura informática del programa. Este documento tiene como fin último generar una propuesta que permita integrar procesos de gestión de datos e información a la vez que fortalecer y hacer más eficientes dichos procesos para apoyar la visión de un Instituto articulado en sus procesos y equipos de trabajo relacionados con el manejo de datos.

OBJETIVOS

OBJETIVO GENERAL

Proponer la arquitectura de un sistema de calidad de datos a través de rutinas automatizadas y de aportes de expertos incorporados en la arquitectura informática del programa.

OBJETIVOS ESPECÍFICOS

- Realizar un diagnóstico de los flujos de trabajo utilizados por los equipos del PEM.
- Identificar procesos de transformación y procesamiento comunes asociados a los datos.
- Evaluar el desempeño de las rutinas de calidad.
- Determinar requerimientos funcionales y no funcionales del sistema de calidad.
- Realizar el levantamiento de requerimientos y la especificación del sistema de persistencia.
- Proponer una arquitectura de referencia que satisfaga los requerimientos de los equipos.

METODOLOGÍA

1. DIAGNÓSTICO

1.1. Descripción de fuentes de información

Se describieron las fuentes de información que nutren los procesos de validación y almacenamiento tanto para el Laboratorio de Biogeografía (LBA) como para la Infraestructura Institucional de Datos e Información (I2D). Se determinó el publicador de los datos, el volumen de los datos (número de registros), el flujo de información (número de registros por unidad de tiempo) y los estándares en los cuales vienen estructurados, para cada fuente.

1.2. Diagnóstico de las herramientas actuales

Considerando que para cada una de las validaciones realizadas por ambos equipos de trabajo se utilizan diferentes herramientas, se realizó una comparación entre estas. El ejercicio pretendió evaluar aspectos de eficacia y eficiencia sobre un conjunto de datos común y con esto evaluar las fortalezas de cada herramienta de manera cuantitativa. La comparación se realizó para las validaciones geográficas y taxonómicas. Los datos usados fueron extraídos de una sección latitudinal a través del portal de datos del SiB Colombia (data.sibcolombia.net) como se muestra en la figura 1.

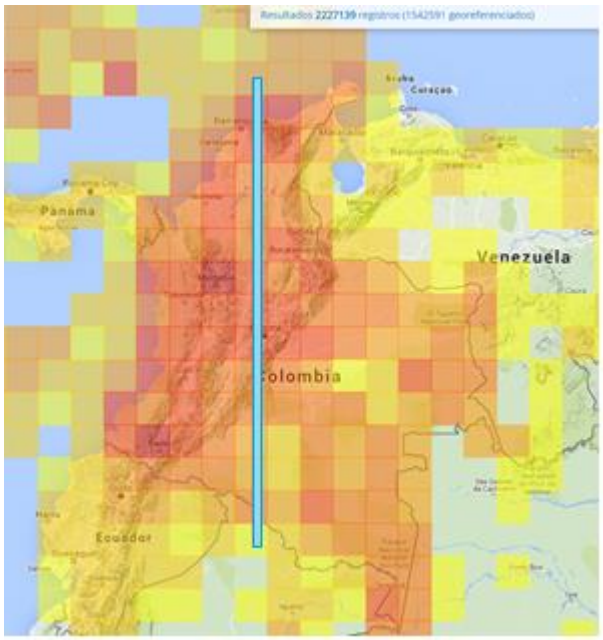

Figura 1. *Consulta de registros seleccionados. El polígono central en rojo representa los registros seleccionados para realizar el análisis.*

Dado que la I2D y el LBA tienen requerimientos diferentes hay ciertas especificaciones para considerar el rendimiento de las validaciones. Para la evaluación se tuvieron en cuenta los siguientes criterios:

- Validación taxonómica: Para la I2D se necesita que el nombre científico sometido sea reconocido como existente y si existe un error sugiera cuál es el nombre científico correcto. Para el LBA que el nombre científico sea válido científicamente y que se pueda extraer su taxonomía superior. Los aciertos fueron definidos como la capacidad de reconocer el nombre científico sometido en las diferentes bases de datos.
- Validación geográfica: Para ambos equipos se evalúa la consistencia de las divisiones administrativas de niveles 0, 1 y 2 (país, departamento y municipio, respectivamente) extraídas a partir de las coordenadas usando shapefiles oficiales con las divisiones administrativas asociadas a cada registro en la base de datos (campos DwC: country, stateProvince y county). Los aciertos fueron definidos como la coincidencia entre la ubicación reportada y la extraída.

Las herramientas utilizadas fueron:

- Validación taxonómica: Para la I2D la librería taxize del software R. Para el LBA la base de datos MySQL de Catalogue of Life 2015.
- Validación geográfica: Para la I2D la herramienta desarrollada por el SiB Colombia en el software Java. Para el LBA el validador geográfico Paynter, disponible en https://github.com/LBAB-Humboldt/GEOGRAPHICAL_VERIFICATIONS.

Una vez finalizados los análisis en cada verificador, se calculó el número de aciertos científicos identificados como y se relacionó con el número total de datos.

1.3. Diferencias entre las rutinas del procedimiento entre LBA e I2D

Durante la descripción de las herramientas y procesos implementados en las rutinas de calidad, también se describió el orden en que estos son ejecutados. Dada la naturaleza de la información, los propósitos de las limpiezas y el uso final de estos, se requiere un esquema y orden de ejecución diferente. Por esta razón se generó un diagrama de flujo general y detallado de cada paso y proceso a los que son sometidos los conjuntos de datos.

1.4. Requisitos de funcionamiento integrado

Según los hallazgos en el desempeño de los métodos, las fases más demoradas en el flujo de trabajo y las dificultades más frecuentes en el proceso de estructuración y validación se realizará un listado de aspectos que deben ser considerados en el momento de nuevos desarrollos tecnológicos. Estas recomendaciones son las nuevas características que harían del sistema integrado ideal al día de hoy para los procesos del programa.

2. METODOLOGÍA DE ANÁLISIS Y DISEÑO DE LA SOLUCIÓN

Levantamiento de requerimientos y definición de especificación mediante entrevistas con los stakeholders del Instituto Alexander von Humboldt. Se construyen modelos aplicando patrones de desarrollo y lenguaje unificado UML.

2.1. Metodología de desarrollo[1]

Para el desarrollo de la solución propuesta, es necesario hacer una revisión de algunas metodologías de desarrollo de software existentes, para seleccionar una que se ajuste a las necesidades tanto de la solución, como a las del equipo de desarrollo involucrado.

A continuación, en la tabla 1, se presenta una breve descripción de las diferentes metodologías disponibles actualmente, junto con sus puntos clave, características especiales y las fallas identificadas.

Tabla 1. *Descripción de las metodologías disponibles, puntos clave, características especiales y fallas identificadas.*

METODOLOGÍA	PUNTOS CLAVE	CARACTERÍSTICAS ESPECIALES	FALLAS IDENTIFICADAS
Adaptative Software Development	Cultura adaptativa, Colaboración, impulsado por la misión, desarrollo iterativo basado en la misión.	Las organizaciones son vistas como sistemas adaptativos. Creando un orden emergente a partir de una red de individuos interconectados.	ASD se refiere más a conceptos y cultura que a la práctica de software.
Agile Methods	Aplicación de principios ágiles para modelar: cultura ágil, organización del trabajo para apoyar la comunicación, simplicidad.	El pensamiento ágil se aplica también al modelado.	Esta es una buena filosofía complementaria para los profesionales de modelado. Sin embargo, sólo funciona dentro de otros métodos.
Crystal family of methodologies	Familia de métodos, donde cada método tiene los mismos valores y principios fundamentales subyacentes. Técnicas, roles, herramientas y estándares varían.	Principios de diseño de método. Capacidad para seleccionar el método más adecuado basado en el tamaño del proyecto y la criticidad	Sólo dos de los cuatro métodos sugeridos existen.
Dynamic Systems Development Method	Aplicación de controles a Rapid Application Development, equipos DSDM empoderados, activos para dirigir el método de desarrollo.	Primer método de desarrollo de software realmente ágil, usa de prototipos y varios roles de usuario: "embajador", "visionario" y "asesor".	Aunque el método está disponible, sólo los miembros del consorcio tienen acceso a los documentos técnicos que tratan con el uso actual del método.
Extreme Programming	Desarrollo orientado al cliente, pequeños equipos, creaciones diarias.	Refactorización - el rediseño contínuo del sistema para mejorar su rendimiento y su capacidad de respuesta al cambio.	Mientras que las prácticas individuales son adecuadas para muchas situaciones, a las prácticas administrativas se le presta menos atención.
Rational Unified Process	Completa el modelo de desarrollo SW incluyendo soporte de herramientas. Asignación de roles impulsada por la actividad.	Modelamiento del negocio, familia de herramientas de soporte.	RUP no tiene limitaciones en su alcance. Hace falta una descripción de cómo adaptar cambios específicos.
Scrum	Equipos de desarrollo pequeños auto organizados. Ciclos de lanzamiento de productos de 30 días.	Impone un cambio de paradigma desde el "definido y repetible" a la "vista de desarrollo de nuevos productos de Scrum".	Aunque Scrum detalla en cómo administrar el ciclo de lanzamiento de 30 días, las pruebas de integración y aceptación no se detallan.

[1] Adaptado de: Abrahamsson, P., Salo, O., Ronkainen, J., & Warsta, J. (2002). Agile software development methods.

2.2 Flujo general de la solución

Para el entendimiento del flujo general que tendrán los datos en la solución a desarrollar, se realizaron entrevistas con cada uno de los integrantes de cada equipo (I2D y LBA) con el objetivo de identificar los principales procesos que rigen la dinámica de cada uno en términos de calidad de los datos. Dichos procesos involucran: las fuentes de datos (entradas), transformaciones y validaciones (procesos), y un sistema de persistencia en el cual la información será almacenada (salidas).

2.3 Flujo de validación de datos

Teniendo en cuenta las necesidades y formas de trabajo de cada uno de los equipos involucrados, el sistema de calidad de datos requiere una solución que permita la creación de flujos de trabajo por parte de cada equipo con un alto grado de automatización. Estos flujos de trabajo estarían compuestos por procesos configurables por parte de los investigadores.
En esta etapa, se evalúan los tipos de datos, transformaciones, reglas y restricciones que permiten tener los datos en un formato y estructura adecuada que facilitarán su posterior análisis. Esto incluye la revisión de los estándares y esquemas definidos anteriormente por cada uno de los equipos.

3. ARQUITECTURA DE REFERENCIA

En la sección de resultados se presenta la arquitectura recomendada de acuerdo al análisis de los requerimientos expuestos. La arquitectura de referencia presenta libertad en el uso de tecnologías de comunicación, almacenamiento y lenguajes de programación.

3.1 Arquitectura y requerimientos del sistema

Durante las reuniones con los equipos de la I2D y el LBA en las que se discutieron los flujos de trabajo, necesidades y procesos en común en cuanto a la calidad de datos se llegaron a unas decisiones para la unificación de los mismos. Estas decisiones permiten tener un panorama general de los requerimientos funcionales del sistema, pero es necesario realizar un levantamiento de requerimientos formal una vez se dé inicio al proyecto. En cuanto a la arquitectura, se realizan los siguientes diagramas:

- Modelo de dominio para identificar las entidades más importantes del sistema de calidad de datos.
- Arquitectura general de la solución donde se plantean los componentes técnicos que permiten la ejecución de tareas y procesos del sistema.
- Flujo general de la solución donde se muestra la integración de los flujos de trabajo tanto de la I2D como del LBA en términos de calidad de datos.
- Flujo de la validación de datos el cual refleja el flujo detallado en cada uno de los niveles (fuentes, validación/transformación, persistencia y análisis).

Estos diagramas permiten tener una visión global hacia donde se deben dirigir los esfuerzos en el desarrollo del sistema.

RESULTADOS

1. DIAGNÓSTICO

1.1. Descripción de fuentes de información

Las fuentes de las cuales provienen los registros biológicos que se incorporan en la I2D se pueden agrupar en tres grandes grupos, (1) Convenios y contratos, (2) Investigaciones internas y (3) Recuperación de información histórica (Tabla 2). Los conjuntos de datos se reciben acompañados de un metadato y a esto es lo que se le denomina un "recurso". Actualmente hay un total de 404 recursos de los cuales 257 son registros biológicos. El promedio de recursos incorporados anualmente en la I2D es de 81 y el volumen de registros por recurso varía ampliamente dependiendo del tipo de proyecto, pero puede estar entre los 4 y los 200.000 registros.

Tabla 2. *Lista de fuentes de información de las cuales provienen los registros biológicos incorporados en la I2D.*

FUENTE	VOLUMEN DE REGISTROS (número de registros a 2017)	VOLUMEN DE RECURSOS (número de recursos a 2017)	FLUJO DE REGISTROS (promedio de número de registros por año)	FLUJO DE RECURSOS (promedio de número de recursos por año)	ESTÁNDAR
Contratos y convenios (Información incorporada por entidades o investigadores externos a través de un convenio de cooperación o un contrato de prestación de servicios)	415.409	210	83.082	42	Darwin Core
Investigaciones internas (Información incorporada por investigadores del Instituto Humboldt generada en algún proyecto interno)	275.194	30	55.039	6	Darwin Core
Recuperación de información histórica (Información proveniente de conjuntos de datos generados antes de consolidación de la I2D y que por lo tanto no se encontraban ni centralizados ni estructurados)	401.251	16	80.250	3.2	Indefinido

Por otro lado, el LBA al usar una ventana que excede a Colombia, según la figura 2, requiere acceder a diferentes y variadas fuentes de información.

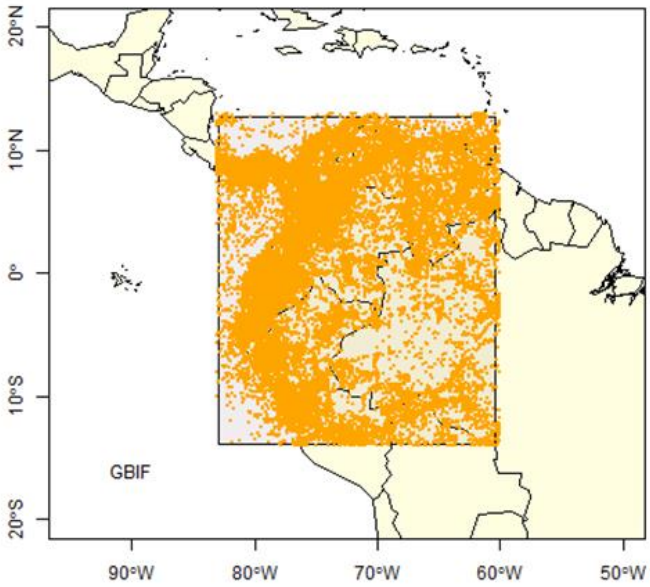

Figura 2. *Ventana de descarga de datos para el LBA*

Es por esta razón que el LBA consulta las siguientes fuentes de datos que actualiza de manera periódica:

- GBIF: Consulta espacial para la región (http://www.gbif.org/occurrence/search?GEOMETRY=-83.00+1.00%2C-83.00+-14.00%2C-60.00+-14.00%2C-60.00+13.00%2C-83.00+13.00#)
- VertNet: Consulta para los países de Colombia, Perú, Venezuela, Ecuador, Panamá y Brasil.
- speciesLink: Consulta para los países de Colombia, Perú, Venezuela, Ecuador, Panamá y Brasil.
- eBird: Consulta para los países de Colombia, Perú, Venezuela, Ecuador, Panamá y Brasil.

Adicionalmente se cuenta con el aporte de datos de otras fuentes que fueron generados una sola vez y no son actualizados. Estos conjuntos son privados y son entregados con restricción de visualización en BioModelos.

En la tabla 3 se muestran las fuentes consultadas, el número de registros que aporta, el porcentaje que representa para el total del conjunto de datos, porcentaje de datos repetidos en relación a las demás fuentes, el estándar que utiliza para documentar, la periodicidad de su consulta y su privacidad.

Tabla 3. *Lista de fuentes de información de las cuales provienen los registros biológicos utilizados en el LBA.*

Fuente	Número	%Aporte	% Repetidos	Estándar de campos	¿Consulta periódica?	Privado
Biomodelos	342	0.002	0.00	Irregular	No	No
Cardelina	1683	0.009	7.13	Irregular	No	Sí
DATAcuaticas	11992	0.064	0.00	Irregular	No	Sí
eBird	4841759	25.735	0.00	Darwin Core	Sí	No
Ecopetrol	8075	0.043	39.74	Irregular	No	No
Fishnet	127104	0.676	0.90	Irregular	No	No
GBIF	10404436	55.301	12.11	Darwin Core	Sí	No
I2D	2236332	11.886	80.80	Darwin Core	Sí	No
I2D_2017	86612	0.460	7.62	Darwin Core	No	No
I2D_priv2016	101347	0.539	0.00	Darwin Core	No	Sí
Invasoras	12653	0.067	28.76	Irregular	No	No
Magnoliaceae	311	0.002	0.00	Irregular	No	Sí
Odonata	262	0.001	0.00	Irregular	No	No
Peces-BC	292	0.002	0.68	Irregular	No	Sí
PecesIAvH	8668	0.046	0.00	Irregular	No	Sí
Politica	73	0.000	5.48	Irregular	No	Sí
Sheffield and Norwegian University	11640	0.062	0.00	Irregular	No	No
SiB-Brasil	100538	0.534	3.33	Darwin Core	No	No
SINCHI	9545	0.051	0.00	Irregular	No	Sí
SpeciesLink	402733	2.141	12.85	Irregular	Sí	No
VertNet	437403	2.325	20.31	Darwin Core	Sí	No
Xenocanto	10065	0.053	22.17	Irregular	No	No
Zamias2016	324	0.002	17.59	Irregular	No	Sí

1.2. Diagnóstico de las herramientas actuales

1.2.1. Estructuración

Los procesos desarrollados por los enfoques varían en campos específicos debido a la naturaleza y requerimiento de los datos, sin embargo se comparten estándares de calidad en campos como:

fechas, separadores, etc. En la I2D se reciben en su mayoría, para incorporación, archivos en excel estructurados en estándar Darwin Core. Por otro lado, el LBA debe hacer un mapeo de los archivos originales que se hace de manera manual en Excel. Para compilar los datos y ajustarlos a estructuras de trabajo se utilizan rutinas en R.

1.2.2. Taxonómico

Los procesos de validación taxonómicos se realizan sobre los campos de "Nombre científico" (scientificName) y toda la taxonomía superior (kingdom, phylum, class, order, family, genus y specificEpithet). Para la evaluación de la calidad de los nombres se usan diferentes herramientas. La I2D utiliza exclusivamente la librería "Taxize" de R, mientras que el LBA utiliza como fuente la última versión disponible del "Catalogue of life".

Taxize:

La I2D desarrolló un script basado en el paquete "taxize: Taxonomic Information from Around the Web" (Chamberlain SA & Szöcs E 2013), el cual interactúa con diferentes "APIs" de la red para tareas taxonómicas, tales como verificación de nombres de especies, resolución de "jerarquías" taxonómicas.

El uso de herramientas de validación de nombres taxonómicos se debe principalmente a que los nombres taxonómicos a menudo varían debido a las revisiones de las especies a niveles genéricos o específicos, a la unión o división de taxa inferiores (géneros, especies) o entre los taxa superiores (familias) y los cambios de nombre de ortografía (Chamberlain SA & Szöcs E 2013).

Este paquete toma recursos de diferentes "APIs" que permite a los usuarios buscar en muchos sitios de la red nombres de especies (nombres científicos y comunes) y descargar la taxonomía superior o inferior. El script puede consultarse y descargarse a través del siguiente enlace: https://github.com/I2DHumboldt/Script-validaci-n-taxon-mica-en-R

Catalogue of Life (CoL):

Es una base de datos SQL que contiene tablas con información de nombres de especies, taxonomía superior y sinónimos. Esta base de datos requiere el nombre científico para extraer de él su nombre validado o sinónimo y su correspondiente taxonomía superior. Para la reconciliación es necesario someter el género y epíteto específico por separado. Para separar el nombre científico se utiliza una función que adicionalmente remueve calificadores del nombre científico, autorías, familias y demás acompañamientos que se documentan en el campo original. Las rutinas son ejecutadas en R usando la librería RMySQL y pueden consultarse en: https://github.com/LBAB-Humboldt/dataDownload

1.2.3. Geográfico

Los procesos de validación geográficos se realizan sobre los campos "decimalLatitude", "decimalLongitude", "country", "stateProvince" y "county". En estos se verifica que las coordenadas sean números naturales decimales con referencia geográfica (positiva o negativa) y que coincidan con país, departamento y municipio documentado. La gran diferencia entre la validación de los dos equipos de trabajo es el rango geográfico y temporal de la validación, ya que la I2D utiliza un script desarrollado en Java por el SiB Colombia, el cual usa la última información básica IGAC a escala

1:100.00 año 2014; mientras que el LBA utiliza un script en R en el que usa capas desde el año 1964 hasta el 2014 y adicionalmente la validación abarca varios países.

Script Java - SiB Colombia:
Es un script basado en lenguaje de programación Java desarrollado por el SiB Colombia, en el que se tiene un archivo en formato texto que deben incluir las columnas: ID, Latitud, Longitud, Departamento y Municipio (sin encabezados). Se ejecuta el .bat y se obtiene una salida con los mismos elementos mencionados anteriormente y cuatro columnas adicionales (municipio interpretado, departamento interpretado y columnas lógicas de coincidencia). Este script permite la validación a nivel de municipio y departamento únicamente en territorio continental Colombiano.

Paynter:
Esta herramienta basada en lenguaje de programación R fue creada por el LBA con el objetivo de realizar verificación de coordenadas y vacíos geográficos de bases de datos de registros biológicos. A partir de esta herramienta se puede realizar la verificación de las coordenadas geográficas de los registros en bases biológicas para 29 países de Suramérica y el caribe entre los que se destacan Colombia, Ecuador, Perú, Panamá, Venezuela y Brasil. La verificación se hace en cinco pasos donde se prueba si la ubicación de la coordenada dentro del país, el departamento y el municipio son correctos. Como resultado se obtiene una base de datos con etiquetas adicionales que denotan la consistencia geográfica de los registros. El conjunto de datos no es modificado en sus contenidos originales sino que se agregan columnas indicativas de la calidad. Para cada nivel geográfico evaluado (país, departamento, municipio) se indica si hubo coincidencia con los valores extraídos con la coordenada y la sugerencia de la ubicación por la coordenada según el shapefile más reciente.

1.2.4. Otros

Esta información hace referencia a los cruces para responder consultas, solicitudes, obtención de cifras o desarrollar análisis internos en los enfoques. Los cruces se realizan una vez la información taxonómica se valide. Este proceso se realiza en la I2D únicamente para dar respuesta a solicitudes, mientras que el LBA lo hace para su base de datos en general.

Endémicas:
El endemismo de las especies se obtuvo a partir de documentación científica específica para cada grupo taxonómico tal y como se muestra en la tabla 4.

Tabla 4. Referencias bibliográficas de los documentos utilizados por la I2D para la obtención del endemismo de las especies.

GRUPO TAXONÓMICO	REFERENCIA BIBLIOGRÁFICA
Mamíferos	Wilson D.E, Reeder, D.M. (eds.). 2005. Mammals species of the World. A taxonomic and geographic reference. Third edition. The Johns Hopkins University Press, Baltimore.
Aves	Dickinson, E.C. (Ed.)(2003) The Howard and Moore Complete Checklist of the Birds of the World. Revised and enlarged third edition. Princeton University Press, Princeton.
Reptiles	Uetz, P. & Jirí Hošek (eds.), The Reptile Database, http://www.reptile-database.org, accessed Dec 8, 2013
Anfibios	AmphibiaWeb. 2016. <http://amphibiaweb.org> University of California, Berkeley, CA, USA. Accessed 18 Oct 2016
Peces	Froese, R. and D. Pauly. Editors. 2016. FishBase. World Wide Web electronic publication. www.fishbase.org, version (06/2016). http://Intreasures.com/colombiam.html
Plantas	Bernal, R., S.R. Gradstein & M. Celis (eds.). 2015. Catálogo de plantas y líquenes de Colombia. Instituto de Ciencias Naturales, Universidad Nacional de Colombia, Bogotá. http://catalogoplantasdecolombia.unal.edu.co

El LBA, por su parte tiene dos fuentes no dinámicas de las que derivó un listado de especies endémicas (tabla 5).

Tabla 5. Fuentes no dinámicas utilizadas por el LBA para la obtención del endemismo de las especies.

FUENTE	NÚMERO DE ESPECIES
Biota	577
Contratistas IAvH	712

Amenazadas:

Para ambos equipos se utilizan las categorías de amenaza internacional y nacional. Las fuentes de las cuales se obtuvo dicha información son:

- Categoría de amenaza nacional: Lista de especies definida en la resolución 0192 de 2014 (MADS) (Está pendiente la actualización de esta categoría de amenaza según la resolución 1912 del 15 de septiembre de 2017).
- Categoría de amenaza global: http://www.iucnredlist.org/

Medios de establecimiento:

Esta información se obtuvo a partir de las listas de especies para Colombia de diferentes fuentes según el grupo taxonómico, en el que se obtuvo categorías de distribución y sus medios de establecimiento como: especies nativas, naturalizadas, cultivadas, residentes, migratorias o introducidas (Tabla 6).

Tabla 6. *Referencias bibliográficas de los documentos utilizados por la I2D para la construcción de una lista de especies de Colombia a 2016. La lista consta de un total de 33.085 especies distribuidas por grupo taxonómico.*

GRUPO TAXONÓMICO	FUENTE	NÚMERO DE ESPECIES
Plantas	Bernal, R., S.R. Gradstein & M. Celis (eds.). 2015. Catálogo de plantas y líquenes de Colombia. Instituto de Ciencias Naturales, Universidad Nacional de Colombia, Bogotá. http://catalogoplantasdecolombia.unal.edu.co	25885
Hongos	Bernal, R., S.R. Gradstein & M. Celis (eds.). 2015. Catálogo de plantas y líquenes de Colombia. Instituto de Ciencias Naturales, Universidad Nacional de Colombia, Bogotá. http://catalogoplantasdecolombia.unal.edu.co	1666
Reptiles	Uetz, P., Freed, P. & Jirí Hošek (eds.). 2016. The Reptile Database, http://www.reptile-database.org, accessed [19-10-2016]	612
Anfibios	AmphibiaWeb. 2016. <http://amphibiaweb.org> University of California, Berkeley, CA, USA. Accessed 18 Oct 2016. Acosta Galvis, A. R. & D. Cuentas 2016. Lista de los Anfibios de Colombia: Referencia en línea V.05.2015.0 (18-10-2016). Página web accesible en http://www.batrachia.com; Batrachia, Villa de Leyva, Boyacá, Colombia.	835
Mamíferos	Solari, S., Muñoz-Saba, Y., Rodríguez-Mahecha, J. V., Defler, T. R., Ramírez-Chaves, H. E., & Trujillo, F. 2013. Riqueza, endemismo y conservación de los mamíferos de Colombia. *Mastozoología neotropical*, *20*(2), 301-365. Disponible en: <http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S0327-93832013000200008&lng=es&nrm=iso>. ISSN 1666-0536. Ramírez-Chaves, H. & Suárez-Castro, A.F. 2014. Adiciones y cambios a la lista de mamíferos de Colombia: 500 especies registradas para el territorio nacional. Mammalogy Notes|Notas Mastozoológicas 1: 31-34.	545
Aves	Stiles, G., Cuervo, A., Rosselli, L., Bohórquez, C., Estela,F., Arzuzas, D. 2011. Species lists of birds for South American countries and territories: Colombia. Versión 16/10/2016. http://www.museum.lsu.edu/~Remsen/SACCCountryLists.htm Salaman, P., Donegan, T. & Caro, D. Listado de Aves de Colombia 2009. Conservación Colombiana 8: 1-89.	1873
Peces	Maldonado-Ocampo, J. A., Vari, R. P. & Usma, J. S. 2008. Checklist of the Freshwater Fishes of Colombia. Biota Colombiana, 9 (2): 143-237. Disponible en: <http://www.redalyc.org/articulo.oa?id=49120960001> ISSN 0124-5376	1451

El laboratorio no cuenta con listas a nivel de grupos como la I2D sino de todos los organismos (Tabla 7).

Tabla 7. *Fuentes utilizadas por el LBA para la construcción de una lista de organismos.*

FUENTE	NÚMERO DE ESPECIES
Programa de Ciencias de la Biodiversidad IAvH	60
Revista Biota	111
Invasive Species Compendium - Cabi	224
Inter-American Biodiversity Information Network (IABIN)	95
IUCN/SSC Invasive Species Specialist Group (ISSG)	50
Resoluciones Ministerio Ambiente	277

CITES:

Ambos equipos se refieren a la autoridad CITES para Colombia consultando y descargando el listado en: UNEP-WCMC (Comps.) 2015. The Checklist of CITES Species Website. CITES Secretariat, Geneva, Switzerland. Compiled by UNEP-WCMC, Cambridge, UK. Available at: http://checklist.cites.org. [Accessed (19-10-2016)].

1.2.5. Evaluación de las herramientas

Una vez realizada la evaluación de la eficiencia y la eficacia de las herramientas utilizadas para la validación taxonómica y geográfica, expuestas anteriormente, se obtuvieron los siguientes resultados:

Para el caso de las herramientas taxonómicas, los aciertos usando taxize fueron mayores gracias a que la función consulta varias bases de datos y permite hacer una búsqueda no exacta del nombre del taxón (Tabla 8). Por otro lado los tiempos de consulta son más reducidos usando CoL por ser una base SQL local y porque taxize hace consultas web para cada nombre buscado.

Tabla 8. *Resultados de la comparación en la validación de nombres científicos entre las herramientas CoL y Taxize.*

HERRAMIENTA	% VALIDACIÓN
CoL	84.52%
Taxize	99.68%

Por otro lado, los resultados obtenidos en la comparación de las herramientas utilizadas para la validación geográfica por la I2D y el LBA se muestran en la tabla 9.

Tabla 9. *Resultados de la comparación en la validación de coordenadas entre las herramientas R 0.2 y Java.*

HERRAMIENTA	% VALIDACIÓN
R 0.2	43.71%
Java	3.10%

Cada uno de los métodos tiene fortalezas que justifican su uso. A pesar de esto presentan limitaciones que para el trabajo de los equipos se convierten en atributos a mejorar. En la tabla 10 se muestra una relación de tales atributos.

Tabla 10. *Atributos para mejorar de las herramientas evaluadas.*

TEMA	HERRAMIENTA	FORTALEZAS	DEBILIDADES
Geográfico	SiB	Fácil uso Tiempo de respuesta	Exactitud Limitación de capas
Geográfico	verifGeo R	Exactitud alta Capacidad de actualización	Dificultad de usar Complejidad de interpretación
Taxonomía	taxize	Exactitud alta Varias fuentes de información Coincidencia difusa	Altos tiempos en consultas web
Taxonomía	CoL	Bajos tiempos en consulta Capacidad de grandes volúmenes	Menores aciertos Coincidencia exacta

1.3. Diferencias entre las rutinas del procedimiento entre LBA e I2D

1.3.1. Calidad de datos en la I2D

El proceso de validación de datos en la I2D se realiza cada vez que un investigador o contratista solicita la incorporación de datos y consiste en la revisión de los mismos para asegurar unos mínimos de calidad. Los conjuntos de datos deben estar estandarizados en Darwin Core y las validaciones que se realizan abarcan la estructura, los campos geográficos y taxonómicos como se muestra en la Figura 3. Al finalizar cada validación se envía un reporte al responsable para que realice las revisiones y modificaciones sugeridas, cualquier modificación realizada durante el proceso se documenta y se envía dentro del reporte enviado. Este proceso se repite cuantas veces sea necesario hasta que los datos cumplan con los lineamientos definidos por la I2D.

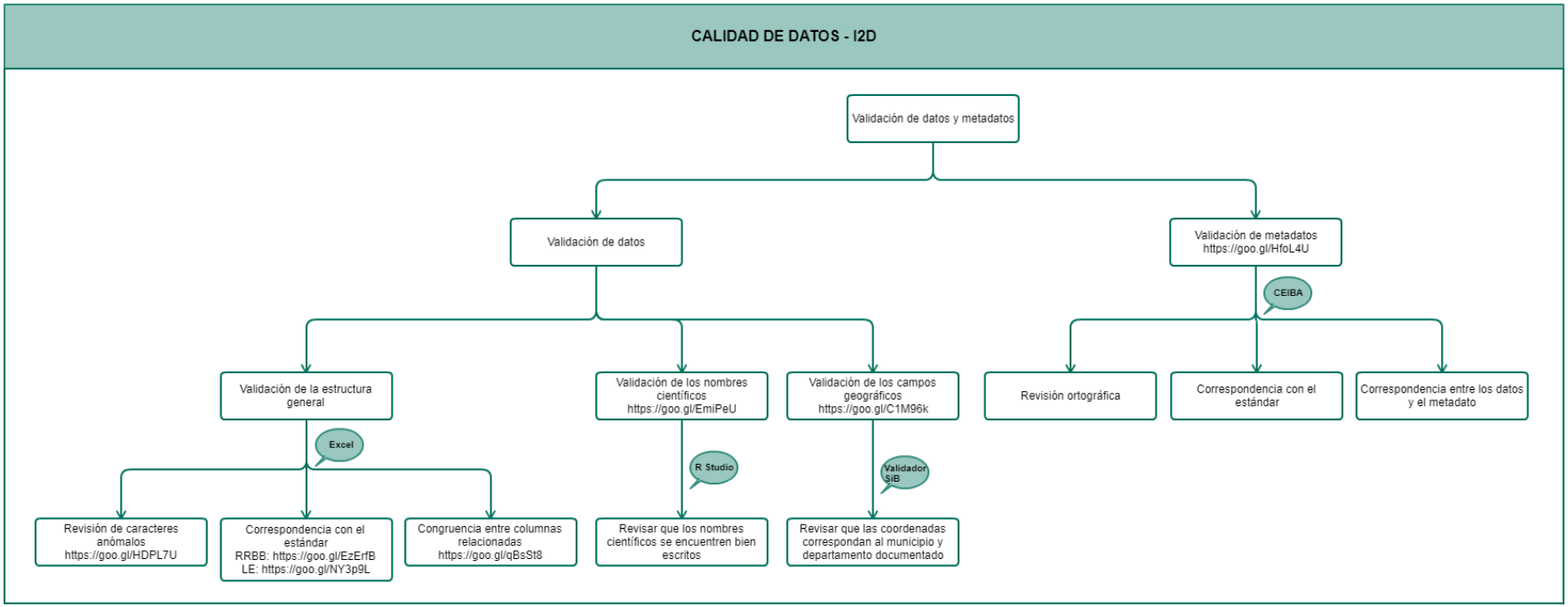

Figura 3. *Validaciones realizadas por la I2D en los conjuntos de datos a incorporar como parte de su proceso de calidad de datos.*

1.3.2. Calidad de datos en el LBA

El laboratorio hace un cálculo de etiquetas o flags (nuevas columnas) que evalúan diferentes aspectos. El procedimiento empieza con los datos originales que se estandarizan en sus encabezados, se consolidan en una sola tabla y termina con el conjunto final. Cada aspecto se ejecuta secuencialmente hay retroalimentación a las fuentes (Figura 4).

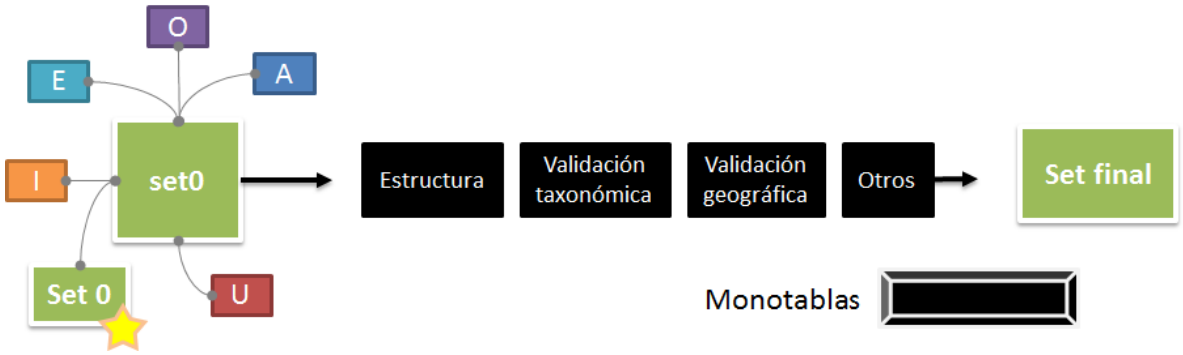

Figura 4. *Estructura del proceso de calidad de datos llevado a cabo por el LBA.*

De manera detallada se presentan las etiquetas implementadas en cada una de las cajas negras de la figura anterior.

El primer paso es estructurar la información como se muestra en la figura 5.

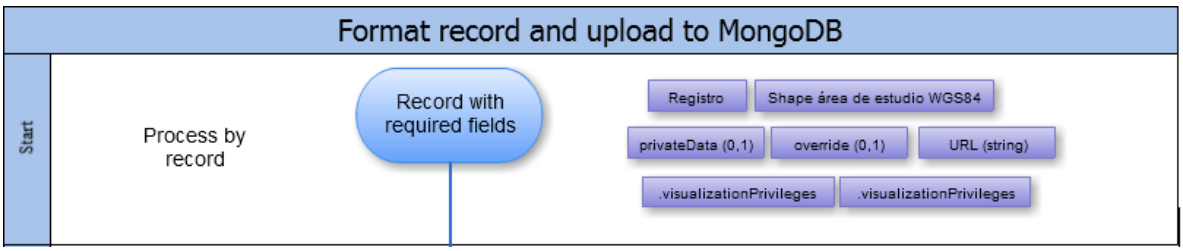

Figura 5. *Descripción del inicio del proceso de calidad del LBA que corresponde a la estructuración de información.*

Una vez estructurada la información se procede al cálculo de los flags según se muestra en la figura 6.

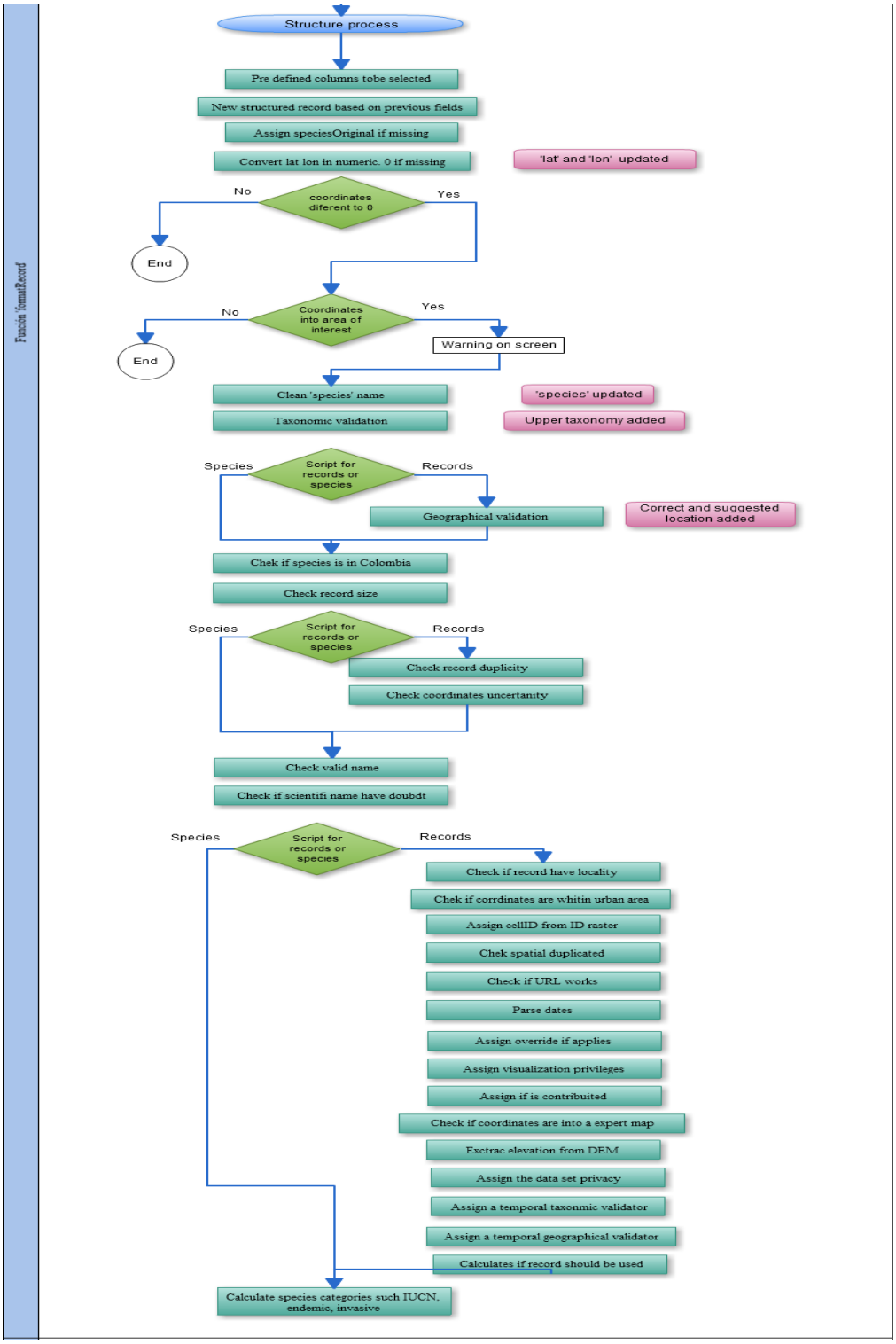

Figura 6. *Cálculo de flags como parte del proceso de calidad del LBA.*

1.4. Requisitos de funcionamiento integrado

La propuesta del nuevo flujo de trabajo comprende una mejora entre las técnicas de validación así como la reducción de pasos manuales. Lo anterior con el fin de:

- Mejorar la exactitud de las validaciones.
  - Estructuración
    - Tener en cuenta las consideraciones del anexo, en donde se describe todos los elementos del estándar para los diferentes tipos de datos.
  - Taxonómicas
    - Usar ambos con posibilidad de seleccionar cuál fuente usar
    - En caso de seleccionar ambas fuentes se propone usar primero Catalogue of Life y después taxize
  - Geográficas
    - Usar Paynter e incorporar con la capa más actual
    - Implementar Paynter en Shiny para facilitar su uso por una base amplia de usuarios
    - Facilitar la personalización de argumentos como las capas geográficas a incluir y el parámetro de similitud de texto.
  - Duplicados
    - Identificar registros potencialmente duplicados que ya se registren en la base de datos.
- Reducir tiempos de trabajo.

El flujo de trabajo pretende omitir pasos manuales y que en el uso de una función procese las validaciones (Figura 7). El único paso que se sale de este procedimiento es el mapeo sobre los elementos de las tablas que debe hacerse de manera manual

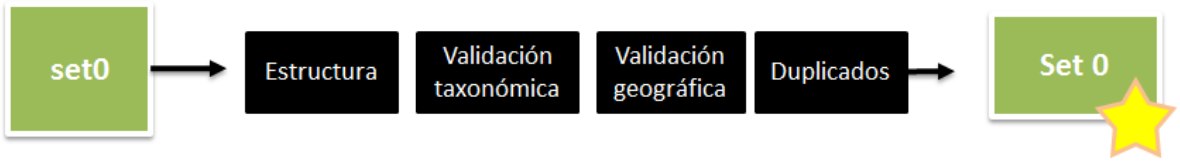

Figura 7. *Esquema general flujo de trabajo propuesto por la I2D y el LBA.*

2. ANÁLISIS Y DISEÑO DE LA SOLUCIÓN

El objetivo de esta sección es presentar una propuesta de solución de calidad de datos a partir del levantamiento de requerimientos realizado con el equipo de I2D y LBA. Para el análisis de requerimientos se realizaron reuniones de trabajo y entrevistas para extraer las necesidades mediante observación, mesa de trabajo, lluvia de ideas y finalmente documentación. La documentación resultado de estas sesiones de trabajo corresponde a la sección de diagnóstico de este documento. Para tener una visión general del problema, y de las necesidades, se aplica

metodología de levantamiento de requerimientos conducido por dominio del problema. Posteriormente se presenta un modelo de dominio para conocer las interacciones generales entre los interesados (*stakeholders)*, finalmente se presentan modelos de flujos de datos y una arquitectura de referencia.

El modelo de componentes de la solución no busca ser exhaustivo, se recomienda realizar sesiones de levantamiento de requerimientos mediante historias de usuario que permitan puntualizar los requerimientos funcionales de la solución.

Las sesiones de levantamiento de requerimientos incluyeron análisis de las necesidades del equipo y de la solución respecto a la base de datos y motores de persistencia. Se presentan los requerimientos no funcionales de la base de datos, arquitectura de la base de datos y lineamientos generales a considerar para la elección y puesta en marcha de los sistemas de almacenamiento.

Antes de presentar los modelos resultado de este análisis, se expone la metodología de desarrollo de software ágil Scrum, sugerida para su aplicación durante la etapa de implementación de la solución.

2.1. Metodología de desarrollo[2]

Debido a su amplio uso en el campo de desarrollo de software se escoge Scrum como metodología a seguir para el diseño e implementación del sistema de calidad de datos. Adicionalmente, el enfoque de Ingeniería de Datos y Desarrollo ha venido desarrollando sus actividades diarias basadas en este marco conceptual de trabajo.

Scrum tiene como finalidad crear un marco de trabajo para abordar proyectos de innovación en ambientes complejos donde de una manera creativa e iterativa se crean soluciones y productos concertados y priorizados, que tendrán maximizada su utilidad y funcionalidad. A continuación se presenta una breve introducción a los principales conceptos del marco de trabajo de Scrum. En su mayoría estas definiciones han sido adaptadas de *Proyectos Ágiles con Scrum*[3] (Alaimo & Salías, 2015).

2.1.1 Principios de Scrum

1. Individuo e interacciones sobre procesos y herramientas:
   Los integrantes del equipo Scrum deciden y toma responsabilidad en el desarrollo del proyecto e interactúa con otras partes de la organización cuando un tema se sale de su conocimiento.
2. Software funcionando sobre documentación:
   La documentación en Scrum pasa a ser de un segundo plano, lo importante es un producto funcional, que cuente con una documentación que el equipo Scrum crea que es necesaria para su implementación y replicación.

[2] El Programa de Evaluación y Monitoreo de Instituto Humboldt recibió capacitación en la metodología Scrum para el desarrollo ágil de proyectos, por lo que se inclina a seguir haciendo uso de esta metodología.

3. Colaboración con el cliente sobre la negociación contractual:
   El Product Owner será el encargado de garantizar que la organización obtenga el mayor beneficio de los productos desarrollados por el equipo Scrum, sirviendo como puente entre las dos partes.
4. Respuesta al cambio sobre seguir un plan:
   Debido a que todos los integrantes del marco de Scrum están al tanto de todas las actividades desarrolladas como también de los pendientes, en cada Sprint (ciclo iterativo de desarrollo) se tiene la posibilidad de agregar tareas que al inicio de las actividades no se tenían contempladas, permitiendo esta agregar el mayor valor posible al producto final.

2.1.2 Valores propios de Scrum

1. **Foco**: Concentración en las actividades específicas concertadas para el desarrollo del producto por Sprint.
2. **Coraje**: El equipo Scrum estará en la capacidad para asumir retos que reúnan a los integrantes de Scrum como equipo.
3. **Apertura**: La información para la toma de decisiones estará siempre abierta para consulta por todos los miembros del equipo Scrum.
4. **Compromiso**: Autonomía en el desarrollo de actividades, por lo que se espera un compromiso total con las actividades asignadas.
5. **Respeto**: Se fomenta el respeto entre cada uno de los integrantes del equipo Scrum.

Como pilares de esta metodología de trabajo se tienen:

- **Transparencia**: Los aspectos importantes de cada proyecto deben ser visibles para todos los involucrados en el mismo.
- **Inspección**: Los responsables de los resultados tendrán asignada la tarea de hacer revisión y control de los productos convenidos con el equipo de desarrollo.
- **Adaptación**: Si en los procesos de revisión e inspección se identifican características a desarrollar que aumentan el valor final del producto, se implementará una estrategia de viabilidad y ajuste.

2.1.3 Roles de Scrum

A continuación se describen brevemente los principales roles en Scrum:

Product Owner

El Product Owner es la única persona responsable de gestionar la Lista del Producto (Product Backlog), con el propósito de lograr maximizar el mayor valor del producto resultado. Sus principales actividades incluyen:

- Claramente identificar y describir los ítems del backlog para construir y compartir el entendimiento del problema y la solución a cualquier nivel con el equipo de desarrollo.
- Tomar decisiones sobre la priorización de los ítems del backlog.
- Determinar la conformidad con el resultado del desarrollo de algún ítem del backlog.
- Mostrar transparencia en cuanto a las futuras actividades para el equipo de desarrollo.

Equipo de desarrollo

Está formado por profesionales que realizan el trabajo de entregar un incremento del producto hecho al finalizar un Sprint que podría ponerse en producción. Sólo miembros del Equipo de Desarrollo crean el Incremento. Sus principales actividades incluyen:

- Son autoorganizados
- El equipo es multifuncional
- No existen subgrupos, independientemente que existan especialidades específicas.
- Las responsabilidades siempre recaen sobre el equipo en conjunto.
- Todos los integrantes de este equipo se reconocen como desarrolladores al mismo nivel.

Scrum Master

Es el rol responsable de que el equipo Scrum siga los valores, procesos y prácticas que el equipo acordó usar. El Scrum master define qué interacciones con personal externo al equipo son benéficas y cuáles no. Siempre estará al servicio del equipo. Sus principales actividades incluyen:

- Aclaración de dudas y preguntas, es un líder facilitador
- Fomentar un ambiente de trabajo donde se pueda maximizar el rendimiento y el valor de los productos.
- Asegurar la buena relación entre los miembros del equipo y también con los de la organización.
- Bloquear interrupciones externas innecesarias.

Historias de usuario

Es la división del trabajo de desarrollo que necesita ser hecho en incrementos funcionales. Se espera que cada historia una vez implementada conlleve a un incremento en la funcionalidad y el valor del producto. El desarrollo de historia de usuario conlleva a los siguientes beneficios:

- Se minimiza el riesgo de tardanzas en la retroalimentación para desarrolladores por parte del Product Owner o el cliente.
- Promueve una separación clara de responsabilidades en la definición del qué y el cómo

Las historias de usuario se componen de los siguientes atributos:

- Independiente
- Negociable
- Agrega valor a los usuarios o clientes
- Estimable
- Pequeña
- Comprobable

2.2. Modelo de dominio

El modelo de dominio presenta las entidades o artefactos (abstractos o concretos) que interactúan en el problema a resolver. Presenta las relaciones más importantes (representadas como verbos o acciones) y la cardinalidad detectada durante las sesiones de levantamiento de requerimientos.

En la figura 8, se puede observar que la entidad más importante es el validador, este debe ser entendido como la representación abstracta de cualquier validador que se quiera implementar. Actualmente se presentan dos tipos de validadores para los registros biológicos (independientemente del formato de entrada): validador taxonómico y validador geográfico. Nótese la importancia de las salidas posibles del validador, reportes o base de datos. La salida por la entidad reportes no implica -necesariamente- el almacenamiento de los resultados en una base de datos y viceversa. Por otro lado los datos almacenados en la base de datos, pueden incorporar información de reportes, algunos de estos reportes son etiquetas o *flags* anexos a los registros biológicos para indicar su estado de calidad. Esta última es la estrategia aplicada por el equipo de LBA.

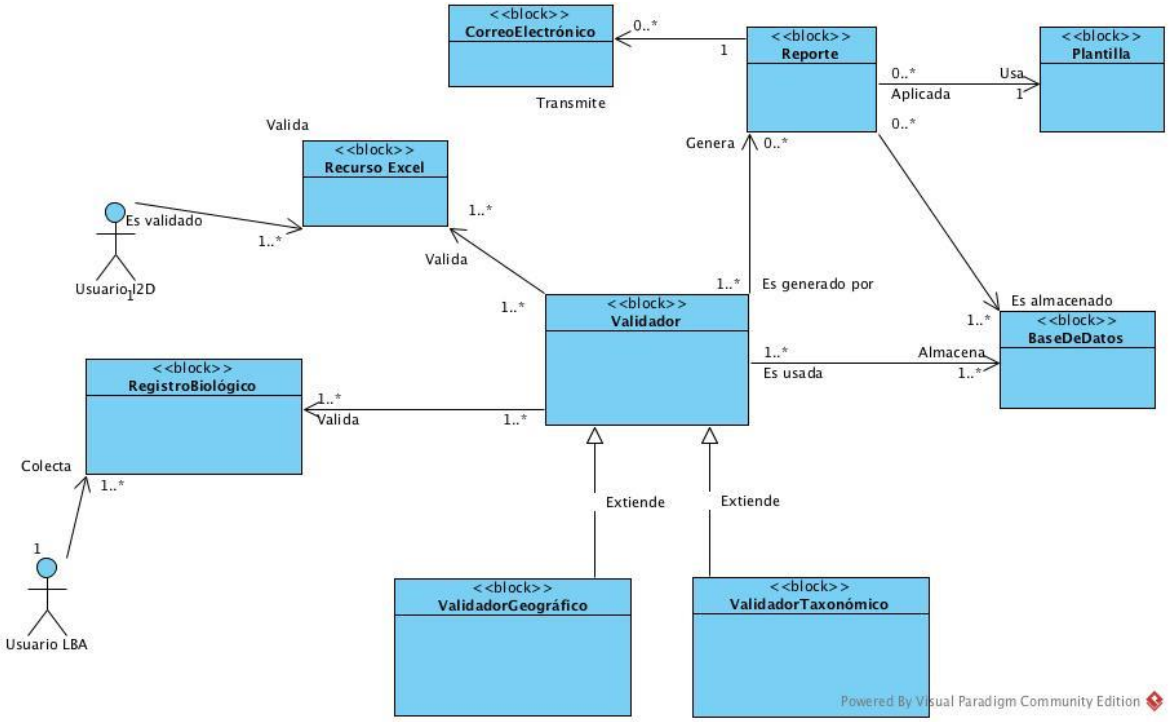

Figura 8. *Modelo de dominio del sistema de calidad de datos.*

Este modelo de dominio constituye el insumo principal para la construcción de los modelos que se presentan a continuación.

2.3. Arquitectura general de la solución

La arquitectura general de la solución de calidad de datos constituye una propuesta que indica, de forma general, las "fichas" o partes clave de la solución. La estructura presentada no implica que la implementación deba realizarse en N máquinas; en la práctica, todo el proceso se puede ejecutar en una sola máquina siempre y cuando persistan las partes que la conforman.

La figura 9 expone la arquitectura general, el usuario interactúa a través de una consola, o servicio de gestión, las tareas y procesos que puede ejecutar para aplicar calidad a un conjunto de datos. El

usuario envía el conjunto de datos al sistema usando el formato de archivo (previamente normalizado), este archivo es almacenado temporalmente para ser ejecutado por los "workers" de calidad, estos son los servicios principales que ejecutan las tareas de validación expuestas en el modelo de dominio, cuyo flujo de datos se explica más adelante. Un servicio (denominado coordinador) se encarga de recibir los registros o conjuntos de datos a los que se debe aplicar calidad, adiciona estas tareas pendientes en una cola de entrada. Los "workers" de validación trabajan como cajeros de un banco, atienden aquellas tareas que están primero en la cola e inician los algoritmos de validación taxonómico y geográfico. Se proponen múltiples instancias de estos servicios de validación, de modo que puedan atender grandes cantidades de datos en la medida que los requerimientos futuros se incrementen.

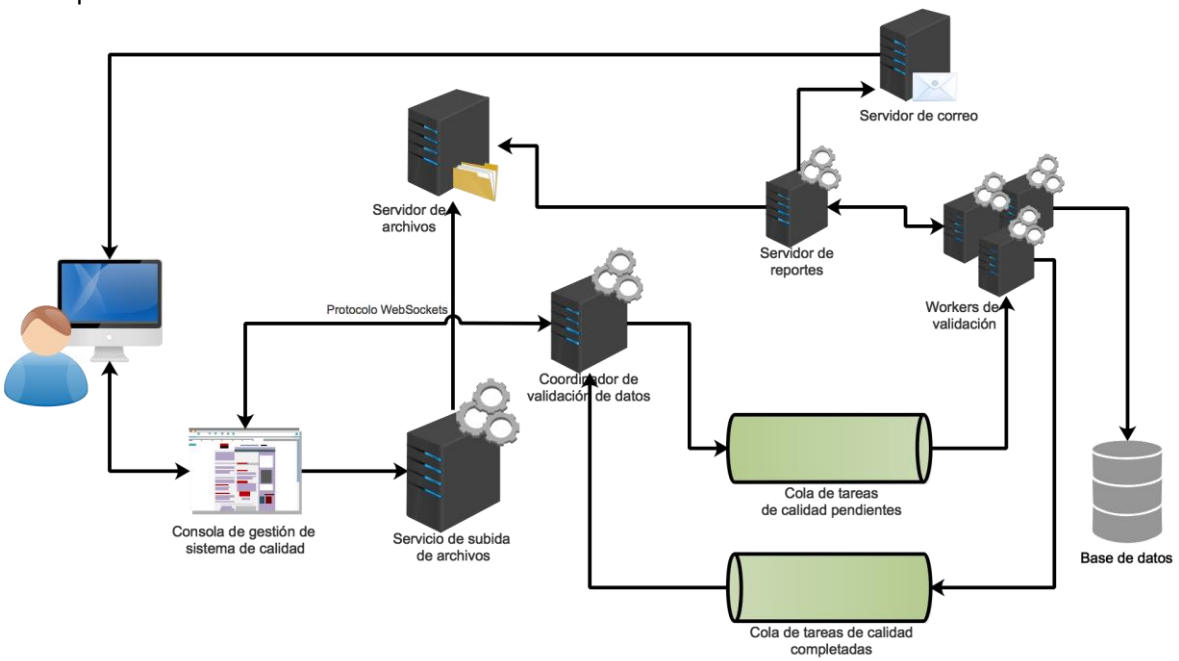

Figura 9. *Modelo de dominio del sistema de calidad de datos.*

Los resultados del proceso de validación de calidad son almacenados en base de datos, o anexados en reportes de calidad que se almacenan en un sistema de archivo. Finalmente se notifica al coordinador de las tareas completadas mediante una cola de salida, éste a su vez informa al usuario que el proceso de calidad ha completado. El sistema envía un correo electrónico al usuario notificando la disponibilidad del informe de calidad para el conjunto de datos procesado; esto ocurre para aquellos datos que nos son almacenados en base de datos.

2.3.1. Flujo general de la solución

El flujo general de la solución tiene dos opciones de acuerdo a las necesidades del grupo que lo utilice como se muestra en la figura 10.

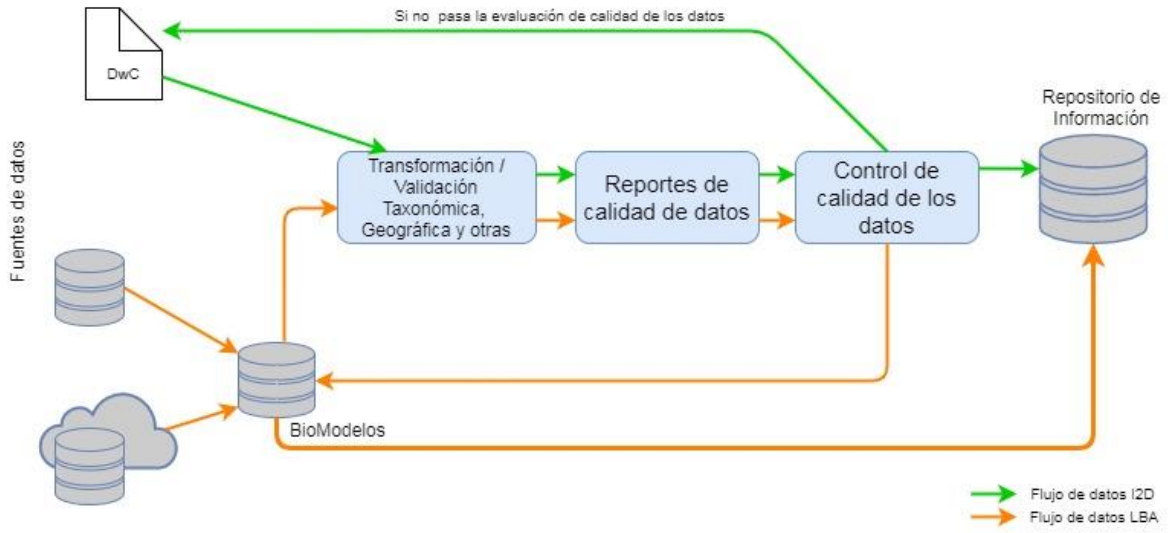

Figura 10. *Flujo general de la solución propuesta.*

I2D

El conjunto de datos ingresa por medio de un archivo bajo un formato previamente establecido por el equipo de la I2D, siguiendo el estándar Darwin Core. A continuación los datos de este archivo son transformado y validados de acuerdo a los algoritmos especificados y se genera un reporte. El siguiente paso es el control de calidad de los resultados de la validación y transformación junto con otras verificaciones no automatizadas donde si se aprueba este control, los datos pasarán directamente al sistema de persistencia o de lo contrario se regresarán al autor para que haga las respectivas correcciones.

LBA

Los datos pueden llegar de múltiples fuentes como otras bases de datos y archivos de texto plano que son normalizadas utilizando el estándar Darwin Core. Estos conjuntos de datos pasan por el proceso de transformación/validación donde se crean nuevos campos de acuerdo a los diferentes procesos. Posteriormente se genera un reporte de la calidad de los datos. Con este reporte, se pueden realizar ajustes de control de calidad por parte de los investigadores antes de ingresar al sistema de persistencia.

2.3.3. Flujo de validación de datos

En la figura 11 podemos ver las diferentes capas que componen la visión del sistema de calidad de datos. En la primera capa, se contemplan las posibles fuentes de información que alimentarán el sistema. Seguido a esto, podemos ver el proceso de extracción, transformación y carga el cual corresponde a los diferentes procesos de: revisión de estructura, validaciones (taxonómica y geográfica) y almacenamiento temporal del resultado de estos procesos. Una vez los resultados han sido almacenados temporalmente, pasarán a ser parte de un sistema integrado de persistencia. Finalmente, estos datos estarán a disposición de todos los investigadores para la elaboración de análisis y reportes.

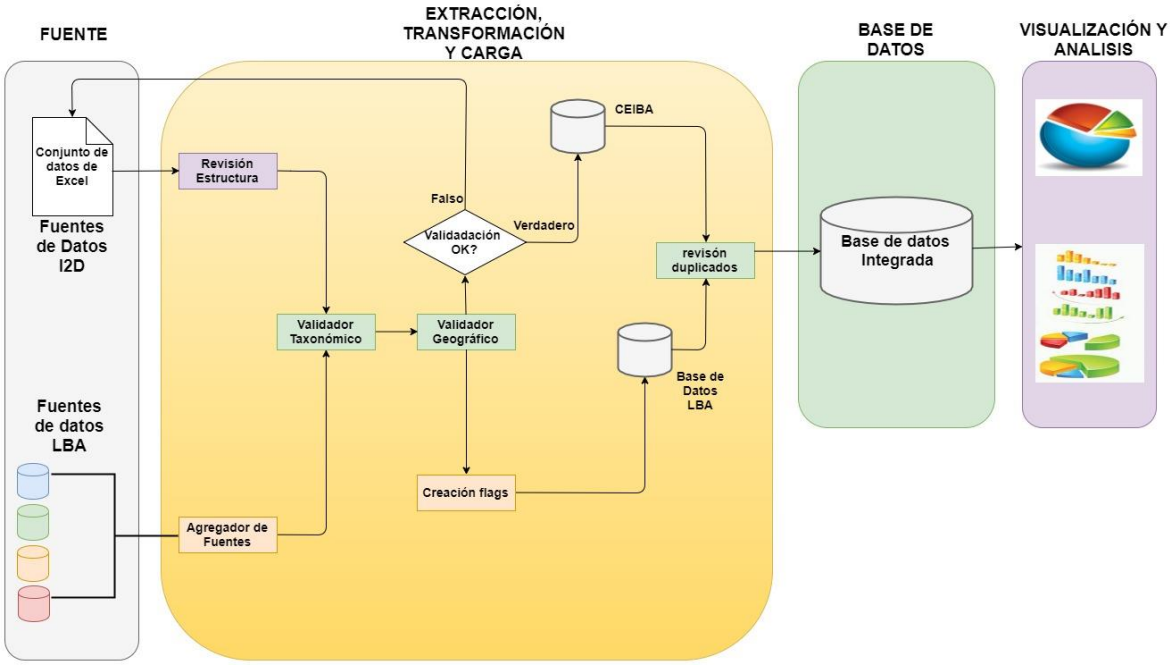

Figura 11. *Flujo de validación de datos.*

2.4. Arquitectura de referencia

La arquitectura de referencia propuesta (Figura 12) para la solución del sistema de calidad de datos está compuesta principalmente por 3 componentes:

Componente de transformación y validación: se encarga de recibir los datos que siguen el estándar Darwin Core ya sea desde un archivo plano o de otras fuentes previa normalización con intervención de los investigadores. Una vez se han recibido los datos, este componente se encarga de validarlos y transformarlos en los casos previamente discutidos con los dos grupos, incluyendo la validación taxonómica y geográfica.

Componente de almacenamiento y reporte: una vez los datos han sido validados, este componente se encarga de generar y enviar un reporte vía correo electrónico que contiene los errores encontrados y las acciones llevadas a cabo por el componente anterior.

Componente de almacenamiento: finalmente, este componente se encarga del almacenamiento de los datos validados y transformados previamente. Adicionalmente, se ejecutaran rutinas que permitan identificar registros que ya hayan sido anteriormente insertados y los actualiza de ser necesario, para evitar la duplicidad de información en la base de datos consolidada.

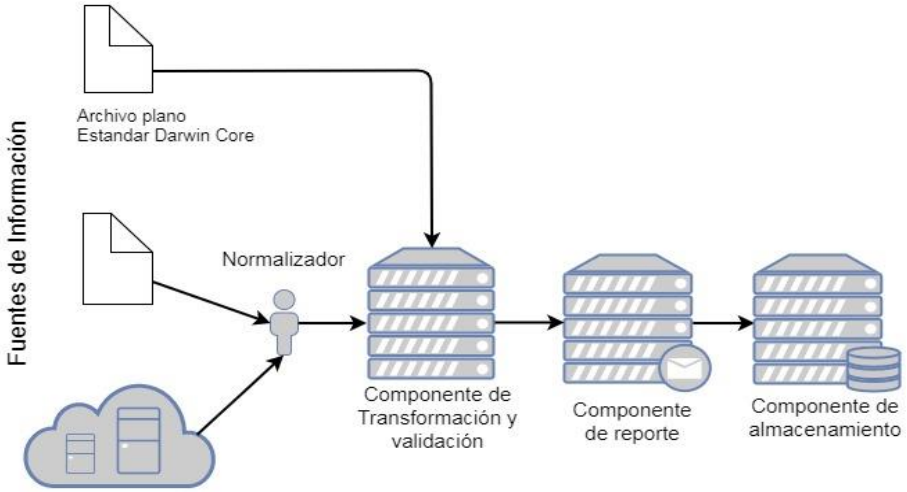

Figura 12. *Estructura de la arquitectura de referencia.*

2.5. Diagrama de componentes

El diagrama de componentes representa a las partes integrantes (componentes) del sistema de calidad de datos. Estos componentes corresponden a unidades de código o servicios de ejecución y muestra las dependencias entre estos componentes. Los componentes físicos corresponden a archivos, bibliotecas y librerías de código compartidas, módulos y paquetes.

El diagrama de componentes del sistema de calidad de datos, ver figura 13, modela la vista estática y dinámica del sistema, muestra la organización y las dependencias entre los conjuntos de componentes.

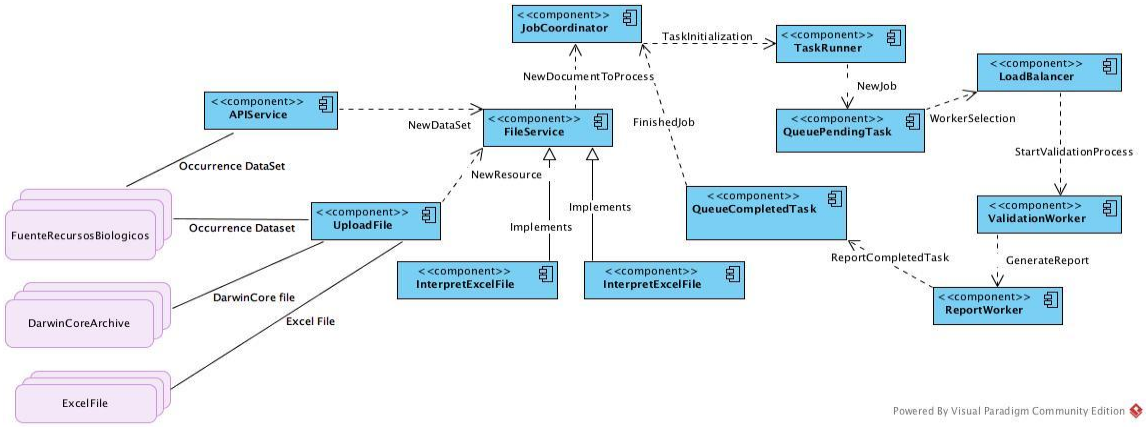

Figura 13. *Diagrama de componentes del sistema de calidad de datos.*

2.6. Especificación del sistema

A continuación se presentan aquellos elementos del sistema que agregan especificidad al mismo. Estos son los requerimientos funcionales, no funcionales, casos de uso e historias de usuario que deberán ser validados durante la fase de desarrollo.

2.6.1. Requerimientos Funcionales

- Soportar subir archivos siguiendo el formato utilizado por la I2D para la publicación de datos.
- Validar la estructura general de los datos que ingresan de forma automática cuando sea posible de acuerdo al documento presentado por la I2D.
- Seleccionar que tipo de fuente utilizar para la verificación taxonómica entre taxize y Catalogue of Life.
- Seleccionar las capas históricas a utilizar en la verificación geográfica.
- Establecer el parámetro de similitud a utilizar en la verificación geográfica.
- Generar un reporte como resultado del proceso de validación/transformación que especifique errores y cambios realizados.
- Enviar el reporte generado por medio de correo electrónico a los investigadores interesados.
- Almacenar los datos que han pasado por el flujo de datos de la I2D de forma exitosa.
- Almacenar los datos que han pasado por el flujo de datos del LBA, con sus respectivas anotaciones de acuerdo a los resultados del proceso.

2.6.2. Requerimientos no funcionales

Eficiencia

- El sistema de calidad de datos podrá ser accedido por máximo 5 usuarios simultáneos sin que se presente una reducción en el tiempo de respuesta. Garantizando la seguridad y confiabilidad de los procesos.
- Los tiempos de respuestas deben estar acorde con la complejidad de la transformación o validación en curso y el volumen de datos. Como referencia, se consideran aceptables tiempos menores a 15 segundos para el procesamiento de un solo registro.

Disponibilidad

- El sistema estará disponible las 24 horas del día los 7 días de la semana, de no ser posible lo anterior, se debe garantizar el cubrimiento de la jornada laboral de los investigadores de la I2D y LBA.

Escalabilidad

- El sistema deberá estar en la capacidad de aumentar el nivel de procesamiento y almacenamiento conforme el volumen de datos se incremente.
- En el sistema se podrán implementar desarrollos adicionales a mediano plazo, por lo tanto el sistema se desarrollará con un enfoque incremental.

Facilidad de uso

- El sistema contará con manuales de usuario que describan detalladamente su uso.
- La interfaz de usuario permitirá una fácil interacción con el sistema, permitiendo una inserción de “datasets” para su posterior proceso.
- En el caso de presentarse algún tipo de error en el sistema, este avisará al usuario por medio de un mensaje en pantalla y se cancelará la operación

Flexibilidad

- El sistema de calidad de datos estará en la capacidad de evaluar conjuntos de datos en formato Darwin Core y formatos provenientes de las bases de datos mencionadas en el apartado “ Descripción de fuentes de información”.

Mantenibilidad

- El sistema de calidad de datos contará con documentación de cada uno de sus componentes incluyendo código fuente y arquitectura utilizadas.

Operatibilidad

- El acceso al sistema de calidad de datos por parte del soporte técnico podrá realizarse remotamente.
- El nivel de acceso al sistema dependerá del rol determinado para cada usuario.

Seguridad

- El manejo y gobernabilidad de los datos dependerá de la política de datos implementada por la I2D.
- Únicamente se permitirá acceso a usuarios designados por el equipo de la I2D y LBA a los cuales se les designaran credenciales de acceso.
- El sistema contará como mínimo con roles de administrador y de investigador.

2.7. Diagrama de casos de uso

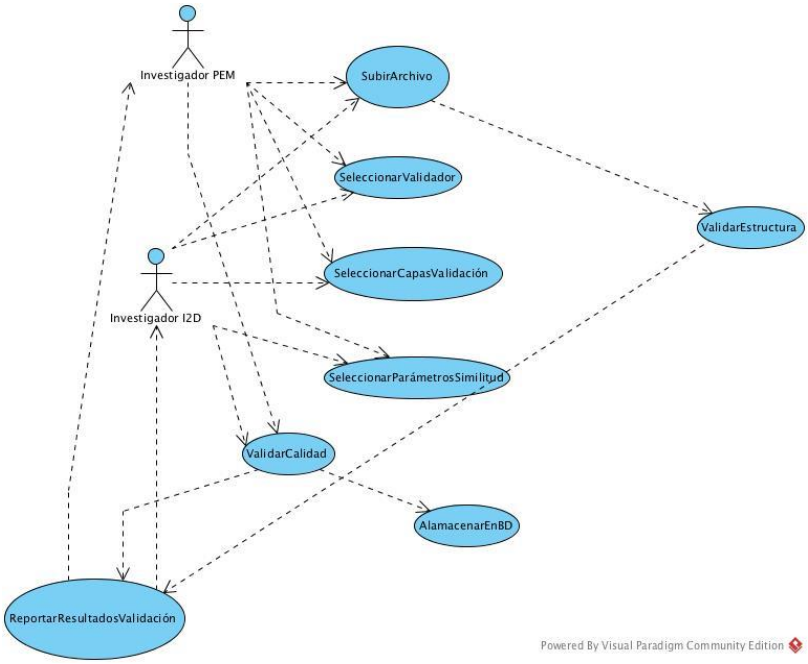

Figura 14. *Diagrama de casos de uso del sistema de calidad de datos.*

2.8. Historias de usuario

Las historias de usuario generadas a partir de las reuniones y de la propuesta general de la solución son las siguientes:

- Como investigador de la I2D, quiero poder subir un archivo con el conjunto de datos siguiendo el formato establecido por el equipo para realizar la validación y transformación de su contenido.
- Como investigador de la I2D, quiero que se valide automáticamente la estructura general de los datos que se ingresan al sistema, en los campos acordados previamente.

- Como investigador del PEM, quiero poder seleccionar que tipo de fuente utilizar para la verificación taxonómica entre taxsize y Catalogue of Life.
- Como investigador del PEM, quiero poder seleccionar las capas históricas a utilizar en la verificación geográfica.
- Como investigador del PEM, quiero poder establecer el parámetro de similitud a utilizar en la verificación geográfica.
- Como investigador del PEM, quiero que se genere un reporte como resultado del proceso de validación/transformación que especifique errores y cambios realizados.
- Como investigador del PEM, quiero recibir el reporte generado por medio de correo electrónico en cuanto se termine el proceso de validación/transformación.
- Como investigador de la I2D, quiero que los datos que sean validados con éxito sean almacenados en un sistema de persistencia.
- Como investigador del LBA, quiero que los datos que pasen por el proceso de validación/transformación sean almacenados en un sistema de persistencia, adicionando las anotaciones respectivas.

Caracterización de usuarios

1. Investigador: dentro del sistema los investigadores serán los encargados de diseñar los flujos de trabajo, ingresar los datos, ejecutar los procesos y recibir los resultados.
2. Administrador: en el sistema, este rol velará por el funcionamiento adecuado de la plataforma, la asignación de permisos, la inclusión de cambios y otros ajustes.

Supuestos y restricciones

- Software Libre - Restricciones de uso inherentes al Software
- Pre-existencia de algunos scripts para el desarrollo de varios de los procesos desarrollados por los dos equipos.
- Definición de los diferentes tipos de acceso a los datos
- Se trabajará sobre la infraestructura tecnológica del Instituto Humboldt
- Infraestructura física necesaria (Servidores...etc)
- Capacidad técnica necesaria. conocimiento en las herramientas tecnológicas para el desarrollo

3. JUSTIFICACIÓN Y ANÁLISIS DE REQUERIMIENTOS DE LA BASE DE DATOS

Los requerimientos del sistema están basados en las actividades diarias que desarrollan los equipos de la I2D y el LBA. Estas actividades se relacionan con la generación de estadísticas y de consultas específicas sobre de los datos que se van almacenando en cada uno de sus repositorios. Adicionalmente, en muchos casos reciben solicitudes por parte de la dirección del Instituto Humboldt y por parte de la ciudadanía en general que tienen que resolverse con la mayor brevedad posible. Es por esto que es deseable tener un sistema donde se puedan hacer estas consultas directamente sin tener que estar consultando repetidamente en diferentes repositorios de datos. Por lo tanto, a continuación se describen los requerimientos generales a nivel de consultas más frecuentes, con el fin de guiar el desarrollo del esquema de la base de datos integrada.

3.1 Requerimientos generales

Después de sostener una reunión con la I2D y el LBA se recolectan los siguientes requerimientos:

1. Búsquedas alfanuméricas.
2. Registros biológicos dentro de un departamento, municipio o polígono en general.
3. Atributos taxonómicos cruzados con listas de Colombia.
4. Cuantos registros hay por categoría taxonómica en cualquier jerarquía.
5. Nombre de especie por categoría taxonómica.
6. Cuales especies están en categoría de amenaza.
7. Agregación por taxonomía por cualquier campo.
8. Consultas por registros biológicos por atributo de calidad.
9. Cuales registros tienen medidas.
10. Mantener históricos de las actualizaciones y cambios en la base de datos (definir campos que guardan historia).
11. Registros biológicos por polígono (formato “shapefile”).
12. Qué modelos se encuentran en determinado espacio geográfico (Lat- Long).
13. Almacenamiento del metadato EML.
14. Almacenar listas de especies amenazadas.
15. Consultas descargables.

Los requerimientos en cuanto a consultas a la base de datos integrada se pueden clasificar en tres grandes categorías que son:

- Consultas alfanuméricas.
- Consultas espaciales.
- Agregaciones.

En la tabla 11 se mencionan los requerimientos por cada uno de los tipos de consulta.

Tabla 11. *Matriz de requerimientos por cada uno de los tipos de consulta.*

REQUERIMIENTO	CATEGORIAS		
	ALFANUMÉRICA	ESPACIAL	AGREGACIÓN
1	X		
2		X	
3	X		
4	X		
5	X		
6	X		
7			X
8	X		X

REQUERIMIENTO	CATEGORIAS		
	ALFANUMÉRICA	ESPACIAL	AGREGACIÓN
9	X		
10	X	X	X
11		X	
12		X	
13	X		
14	X		
15	X	X	X

Los equipos de la I2D y LBA manifiestan que las consultas alfanuméricas y las especiales son de mucha importancia debido a que son actividades que realizan diariamente y es deseable que tengan una priorización en el desarrollo del proyecto.

Como primera aproximación al esquema de la base de datos, se tomará el estándar Darwin Core. Esto debido a que es el estándar utilizado por los dos equipos.

3.2 Modelamiento de datos (Relacional, NOSQL)

Lenguaje de consulta

El usuario debe tener la capacidad de hacer consultas a la base de datos de una manera simple y sencilla, no debería ser un lenguaje complejo.

Consistencia disponibilidad y particionamiento (CAP)

De acuerdo al teorema CAP, en una base de datos solo se podrán tener simultáneamente 2 de las siguiente propiedades: Consistencia (C), disponibilidad (A) y tolerancia particiones (P). Por lo tanto se tiene que:

AP: disponibilidad y tolerancia a particiones, pero no la consistencia.
CP: consistencia y tolerancia a particiones, pero no la disponibilidad
CA: consistencia y disponibilidad, pero no tolerancia a particiones

Teniendo en cuenta los requerimientos del equipo de la I2D la condición deseable es CP.
Sistemas manejadores de datos que vienen por defecto con esta configuración son:

- MongoDB
- Hbase
- Redis

Sin embargo, las propiedades mencionadas en el teorema CAP en algunos motores de bases de datos son configurables.

3.3 Requisitos orientados al usuario

En la tabla 12 se describen los requisitos orientados al usuario según su categoría y su prioridad.

Tabla 12. *Requisitos orientados al usuario.*

CATEGORÍAS	DESCRIPCIÓN	PRIORIDAD
Disponibilidad	Veinticuatro horas del día los siete días de la semana (24/7)	Alta
Eficiencia	15 minutos	Medio
Escalabilidad	50 millones de registros en 5 años	Alto
Facilidad de uso	Consultas directas a base de datos API	Medio
Fiabilidad	Consistencia y Particionamiento (CP)	Alta
Mantenibilidad	updates, bulk loads, inserts, mantenimiento físico, documentación	Baja

3.4 Requerimientos no funcionales del sistema de base de datos

Disponibilidad

La alta disponibilidad de la base de datos es una característica de mucho interés en el proyecto, debido a que usuarios concurrentes estarán accediendo a la base de datos para realizar consultas y reportes. Técnicamente es muy difícil que un sistema se encuentre el 100% del tiempo disponible, debido a que cada sistema tiene características específicas, diversos tipos de procesamientos, usuarios y eventos fortuitos, se buscará que la base de datos esté en la medida de lo posible disponible la mayor parte de tiempo para el usuario final (> 90%).

Eficiencia

El tiempo de respuesta de una consulta simple no deberá tomar más de 5 segundos. Sin embargo en el caso de consultas complejas que impliquen un alto procesamiento un tiempo aceptable de respuesta aceptable será 15 minutos, teniendo en cuenta que los tiempos se podrán reducir drásticamente al manejar buenas prácticas en la realización de las consultas y por supuesto teniendo un modelado de datos óptimo.

Escalabilidad

La base de datos debe ser capaz de crecer dinámicamente independiente de la infraestructura física que se utilice ya sea un datacenter o en la nube. Se espera que los datos crezcan alrededor de 50 millones de registros en 5 años.

Facilidad de uso

La facilidad de uso se encuentra en un nivel medio de prioridad debido a que inicialmente los usuarios serán investigadores de la I2D y LBA que están dispuestos a profundizar en el aprendizaje

de lenguajes de consulta. No obstante, a mediano plazo se debe contar con una interfaz que facilite la consulta de datos y construcción de reporte por medio de API's.

Modelo de datos flexible

Esta es una característica deseable debido a que la integración se debe realizar de una manera dinámica que permita combinar información de diferentes fuentes. Además, en el caso que sea necesario modificar el esquema, no se cree un traumatismo en las aplicaciones que apunten a la base de datos.

Fiabilidad

La base de datos estará disponible en lo posible ante la ocurrencia de algún fallo, para ello se ha definido que a base de datos debe estar desarrollada teniendo en cuenta el teorema CAP donde se desea tener Consistencia (C) y Particionamiento (P) simultáneamente.

Mantenibilidad

Aunque la mantenibilidad de la aplicación se clasifica como baja, es importante construir una documentación que permita entender el funcionamiento detallado de la aplicación, permitiendo así, que cualquier desarrollador pueda construir, hacer modificaciones o actualizaciones sobre lo ya establecido.

BIBLIOGRAFÍA

Literatura citada

Abrahamsson, P., Salo, O., Ronkainen, J., & Warsta, J. (2002). Agile software development methods.

Alaimo M., & Salías M. (2015). Proyectos Ágiles con Scrum: Flexibilidad, aprendizaje, innovación y colaboración en contextos complejos. Buenos aires: Kleer.

Chamberlain, S. & Szocs, E. (2013). Taxize - taxonomic search and retrieval in R. F1000Research, 2:191. URL: http://f1000research.com/articles/2-191/v2.

Otra literatura de referencia

Fowler, M. (2003). *Patterns of enterprise application architecture*. Boston: Addison-Wesley.

Fowler, S. J. (2017). *Production-ready microservices: building standardized systems across an engineering organization*. In (pp. 1 online resource (1 volume)). Retrieved from http://proquest.safaribooksonline.com/9781491965962

Kleppmann, M. (2017). *Designing data-intensive applications : the big ideas behind reliable, scalable, and maintainable systems*. In (pp. 1 online resource). Retrieved from http://proquest.safaribooksonline.com/9781491903063

Morris, K. (2016). *Infrastructure as Code*. In (pp. 1 online resource). Retrieved from http://proquest.safaribooksonline.com/9781491924334

Newman, S. (2015). *Building microservices : designing fine-grained systems*. In (pp. 1 online resource (1 volume)). Retrieved from http://proquest.safaribooksonline.com/9781491950340

Richard, B. B. C. J. J. P. N. M. (2016). *Site Reliability Engineering*. In (pp. 1 online resource). Retrieved from http://proquest.safaribooksonline.com/9781491929117

Tanenbaum, A. S., & Steen, M. v. (2007). *Distributed systems: principles and paradigms* (2nd ed.). Upper Saddle RIver, NJ: Pearson Prentice Hall.

Bass, L., Clements, P., & Kazman, R. (2003). *Software architecture in practice* (2nd ed.). Boston: Addison-Wesley.

Cervantes, H., & Kazman, R. (2016). *Designing software architectures: a practical approach*. Boston: Addison-Wesley.

Duggan, D. (2012). *Enterprise software architecture and design entities, services, and resources*. Hoboken: Wiley.

Evans, E. (2004). *Domain-driven design: tackling complexity in the heart of software*. Boston: Addison-Wesley.

Ford, N., Parsons, R., & Kua, P. (2017). *Building evolutionary architectures support constant change*. S.l.: O'REILLY MEDIA, INC, USA,.

Lankhorst, M. (2017). *Enterprise architecture at work: modelling, communication and analysis* (Fourth edition. ed.). Berlin, Germany: Springer.

Martin, R. C. (2017). *Clean architecture: a craftsman's guide to software structure and design*. Boston, MA: Addison-Wesley.

Mitra, T. (2016). *Practical software architecture: moving from system context to deployment*. New York: IBM Press, Pearson plc,.

Pastor, Ó., Molina, J. C., & SpringerLink (Online service). (2007). *Model-driven architecture in practice a software production environment based on conceptual modeling*. Berlin; New York: Springer.

Richards, M. (2015). *Software Architecture Patterns*. S.l.: O'Reilly Media, Inc.

Sangwan, R. S. (2015). *Software and systems architecture in action*. Boca Raton, FL: CRC Press is an imprint of the Taylor & Francis Group, an informa business.