# **Diccionario de Datos del Proyecto**

**Versión:** 1.1

**Fecha:** 6 de Marzo de 2025

**Autores:** Leonardo Valdés Palafox, Andrea Marlene Ortega Almendares

**Estándar seguido:** **ISO/IEC 11179 (Metadata Registry Standard)**

**Clasificación:** **Documento Entregable**

## **Índice de Contenido**

1. **Introducción**
   1. Propósito del Diccionario de Datos
   2. Alcance del documento
   3. Normativas aplicables
2. **Modelo de Datos del Proyecto**
   1. Definición de entidades principales
   2. Relaciones entre entidades
   3. Diagramas conceptuales
3. **Especificación de Elementos de Datos**
   1. Identificadores únicos y claves primarias
   2. Atributos, restricciones y validaciones
   3. Uso de datos y reglas de calidad
4. **Reglas de Integridad y Seguridad**
   1. Consistencia referencial de datos
   2. Estrategias de control de acceso
   3. Validaciones automáticas
5. **Escenarios de Uso**
   1. Flujo de procesamiento de datos
   2. Casos de uso detallados
6. **Buenas Prácticas en Gestión de Datos**
   1. Estrategias de almacenamiento
   2. Indexación y optimización
   3. Auditoría y trazabilidad
7. **Referencias**

## **1. Introducción**

#### **1.1 Propósito del Diccionario de Datos**

Este documento define la **estructura de datos** del sistema de extracción y análisis de información de **X**, proporcionando una referencia clara y estandarizada para garantizar **coherencia, integración y seguridad** en el manejo de datos.

#### **1.2 Alcance del Documento**

Incluye:

**Modelo de datos y relaciones estructuradas**.

**Atributos clave y validaciones técnicas**.

**Normas de integridad, seguridad y auditoría de datos**.

**Casos de uso detallados para ilustrar procesamiento y consulta**.

## **2. Modelo de Datos del Proyecto**

#### **2.1 Definición de Entidades Principales**

Las entidades principales que organizan la información recopilada son:

|  |  |
| --- | --- |
| **Entidad** | **Descripción** |
| **Tweet** | Representa las publicaciones extraídas de X. |
| **Usuario** | Almacena información sobre perfiles de usuarios. |
| **Regla de Extracción** | Define parámetros y criterios para la recopilación. |
| **Historial de Extracción** | Registra cada ejecución del scraping para auditoría. |

#### **2.2 Relaciones Entre Entidades**

**Un Tweet pertenece a un Usuario** (relación uno a uno).

**Las Reglas de Extracción aplican sobre múltiples Tweets** (relación uno a muchos).

**El Historial de Extracción almacena múltiples Tweets por ejecución** (relación uno a muchos).

## **3. Especificación de Elementos de Datos**

#### **3.1 Identificadores Únicos y Claves Primarias**

Cada entidad posee un identificador único que permite mantener **consistencia y trazabilidad** en el sistema.

|  |  |  |
| --- | --- | --- |
| **Entidad** | **Identificador Único** | **Ejemplo** |
| **Tweet** | tweet\_id (cadena) | "785623412" |
| **Usuario** | usuario\_id (cadena) | "u1587" |
| **Regla de Extracción** | regla\_id (entero) | 102 |
| **Historial de Extracción** | historial\_id (entero) | 2001 |

#### **3.2 Atributos Clave, Restricciones y Validaciones**

Los atributos contienen validaciones que **garantizan calidad y seguridad** en la administración de datos.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Entidad** | **Atributo** | **Tipo de Dato** | **Restricción** | **Validación** |
| **Tweet** | contenido | Texto | Máx. 280 caracteres | No puede contener URLs |
| **Usuario** | username | Texto | Único | No puede incluir caracteres especiales |
| **Regla de Extracción** | palabras\_clave | Texto | Mín. 3 palabras | Debe cumplir con expresiones regulares |
| **Historial de Extracción** | fecha\_ejecucion | Fecha | No puede ser futura | Formato YYYY-MM-DD |

#### **3.3 Uso de Datos y Reglas de Calidad**

**Datos estructurados y clasificados** para evitar pérdidas de información.

**Control de calidad** mediante validaciones antes de almacenamiento.

**Depuración automática** para eliminar datos inconsistentes.

## **4. Reglas de Integridad y Seguridad**

#### **4.1 Consistencia Referencial de Datos**

**Relaciones claras entre entidades** para evitar datos sin referencia.

**Restricciones de claves foráneas** para garantizar coherencia.

#### **4.2 Estrategias de Control de Acceso**

**Roles de usuario con permisos diferenciados**.

**Autenticación obligatoria antes de acceso a datos sensibles**.

**Protección contra modificaciones no autorizadas**.

#### **4.3 Validaciones Automáticas**

**Filtrado de contenido antes de almacenamiento**.

**Verificación de existencia de registros previos** para evitar duplicaciones.

**Cifrado de datos personales para protección de privacidad**.

## **5. Escenarios de Uso**

#### **5.1 Flujo de Procesamiento de Datos**

1️⃣ **El sistema ejecuta scraping de X** según las reglas configuradas.

2️⃣ **Los tweets recolectados pasan por un proceso de validación y filtrado**.

3️⃣ **Los datos se almacenan y se registran en el historial de extracción**.

4️⃣ **Los analistas pueden consultar y analizar los datos extraídos**.

#### **5.2 Casos de Uso Detallados**

✔ **Consulta de tweets** por palabras clave con filtros personalizados.

✔ **Generación de reportes** basados en patrones de publicación.

✔ **Alertas automáticas** ante detección de tendencias en X.

## **6. Buenas Prácticas en Gestión de Datos**

#### **6.1 Estrategias de Almacenamiento**

**Registro optimizado** para grandes volúmenes de datos.

**Estructuración adecuada** para evitar redundancias.

#### **6.2 Indexación y Optimización**

**Búsqueda rápida de registros** mediante estrategias de indexación.

**Optimización de consultas** para mejorar rendimiento.

#### **6.3 Auditoría y Trazabilidad**

**Registro de todas las operaciones de extracción**.

**Control de cambios y modificaciones en reglas de extracción**.

## **7. Referencias**

* **ISO/IEC 11179**: Metadata Registry Standard.
* **Ley Federal de Protección de Datos Personales** (Congreso de la Unión, 2010).
* **Política de Privacidad de X**: <https://x.com/privacy>.

### **Clasificación del Documento**

✔ **Entregable**: Referencia técnica para analistas y administradores de datos. ✔ **Basado en ISO/IEC 11179** para garantizar trazabilidad y consistencia.