Informe de Avance (Fase 1): Grafo de Conocimiento sobre Agricultura de Precisión

Matería: Interoperabilidad y explotación de datos en ecosistemas heterogéneos

Integrantes:

- Jean Villavicencio
- Samuel Reyes
- Sebastian Mendieta

1) Introducción y objetivo

Este trabajo tiene como meta construir un grafo de conocimiento a partir de publicaciones científicas sobre Agricultura de Precisión. En esta primera fase definimos el alcance, la fuente de datos a usar será: Semantic Scholar. El dominio y la viabilidad, tenemos un volumen objetivo de 500–1000 artículos, y dejamos trazada una metodología clara para la fase de extracción, transformación, carga y análisis.

2) Fuente de datos: Semantic Scholar (API)

Usaremos exclusivamente Semantic Scholar por tres razones: primero por el acceso programático robusto vía Academic Graph API, segundo los metadatos ya estructurados como papers, autores, citas, venues que facilitan el modelado como grafo, y por último la documentación y ejemplos suficientes para un pipeline reproducible.

2.1 APIs relevantes a consumir

- **Paper Search:** /graph/v1/paper/search para la búsqueda por palabras clave para obtener el corpus inicial (IDs de artículos).
- **Paper Details:** /graph/v1/paper/{paperId} para los metadatos completos del paper (título, abstract, autores, citas, referencias, venue, campos de estudio, etc.).
- **Author Details (opcional, recomendado):** /graph/v1/author/{authorId} para enriquecer nodos de Autor (afiliaciones, desambiguación).

La API permite solicitar solo los campos necesarios con el parámetro fields y manejar paginación, con API key individual se alivian límites de tasa.

3) Dominio de trabajo: Agricultura de Precisión

Lo que buscamos es un corpus equilibrado de 500–1000 publicaciones que hable de la intersección entre Agricultura de Precisión y tecnologías digitales como

sensores, teledetección, IoT, ML, SIG para manejo sitio-específico ya sea riego, fertilización, fitosanidad “en la cantidad correcta, lugar correcto y momento correcto”.

Consulta sugerida (ejemplos, para la API/web): "precision agriculture" AND (sensor OR "remote sensing" OR drone OR UAV OR "variable rate" OR IoT OR GIS OR "machine learning")

Filtros recomendados: year:2015-, fieldsOfStudy:Computer Science,Agricultural Science, publicationTypes:JournalArticle,Conference

Estas consultas suelen devolver un volumen en el rango objetivo y con buena densidad de citaciones cruzadas (útil para el grafo). En caso de exceder 1000, ajustamos con filtros de año/venue/citas; si quedara corto, ampliamos sinónimos o relajamos los filtros.

4) Descripción del dominio

La Agricultura de Precisión integra datos geoespaciales y temporales como mapas de rendimiento, humedad, índices de vegetación, sensores en campo y maquinaria, imágenes de drones/satélites y modelos de ML/IA para decidir con precisión dónde, cuánto y cuándo intervenir.

El objetivo es producir más y mejor con menos insumos, reduciendo impacto ambiental y costos. En el plano científico, el tema es ideal para un grafo de conocimiento porque los papers conectan tecnologías como sensores, UAV, IoT, prácticas como riego/fertilización variable, cultivos, regiones y resultados medibles como el rendimiento, eficiencia hídrica, huella ambiental, generando relaciones ricas y consultables.

5) Esquema del grafo y mapeo desde la API

Entidades (nodos):

- Paper (artículo)
- Author (autor/a)
- Venue (revista/conferencia)
- FieldOfStudy (área/tema)
- (Opcional, por enriquecimiento textual) Concept, Method, Dataset, Crop, Sensor cuando aparezcan de forma consistente en metadatos/resúmenes.

Relaciones (aristas):

- AUTHORED_BY (Paper → Author)

- CITES (Paper → Paper)
- PUBLISHED_IN (Paper → Venue)
- HAS_TOPIC (Paper → FieldOfStudy)
- (Opcional) USES_METHOD, MEASURES_ON, APPLIES_TO_CROP según extracción ligera de conceptos.

Tabla de mapeo (resumen):

- paperId, title, abstract, year, citationCount, fieldsOfStudy, publicationVenue/journal → propiedades del Paper y vínculos a Venue/Field.
- authors[*].authorId/name → Nodos Author y relación AUTHORED_BY.
- citations[*].paperId / references[*].paperId → Relación CITES.

Este mapeo replica el "blueprint" bibliométrico probado y encaja perfectamente con la Academic Graph API.

6) Metodología (ETL) y plan de trabajo

Extracción (E):

1. Ejecutar la query en /paper/search para reunir ~800–900 paperIds.
2. Iterar /paper/{id} solicitando fields mínimos (título, abstract, autores, año, venue, citas, referencias, fieldsOfStudy).
3. Guardar JSON crudo y, si conviene, usar batch endpoints para eficiencia.

Transformación (T):

4. Normalizar IDs (DOI/S2), deduplicar, desambiguar autores, limpiar venues.
5. Proyectar a CSV de nodos y CSV de relaciones como paper_cites_paper.csv, paper_authoredby_author.csv, paper_publishedin_venue.csv, paper_has_topic.csv.
6. (Opcional) Extraer conceptos frecuentes del abstract para relaciones como USES_METHOD o APPLIES_TO_CROP con parsers simples de términos clave.

Carga (L) y explotación:

7. Importar en Neo4j/Memgraph con LOAD CSV.
8. Validar con consultas de control: top-cited, autores prolíficos, clústeres por tema/venue, evolución temporal.

9. Visualizar subgrafos (Gephi/pyvis) y preparar preguntas de análisis (ej.: “¿Qué métodos se asocian más a mejoras en rendimiento en cultivos de maíz 2018–2025?”).

Riesgos y mitigación:

- **Corpus >1000:** afinar filtros (año, campos de estudio, venues) o añadir términos más específicos (p.ej., “variable rate application”, “UAV NDVI”).
- **Corpus <500:** ampliar sinónimos/tecnologías (p.ej., “site-specific management”, “precision viticulture”, “smart farming”).
- **Límites de tasa:** usar API key individual y planificar descargas en lotes.

7) Viabilidad (500–1000 documentos)

La propia API y la experiencia de diseño de consultas indican que “precision agriculture” combinada con vocabulario tecnológico como los sensores, UAV, IoT, ML, GIS, VRA alcanza un rango estable entre 500 y 1000 resultados recientes y relevantes. Si la búsqueda base excede el umbral, el filtrado temporal y por campos de estudio devuelve un subconjunto manejable sin perder representatividad para el grafo.