

ANEXO I

PROPUESTA DE CONTENIDO DEL RETO DE LA CPM DEL PROYECTO AIDEA

“Plataforma Innovadora de Análisis Integrados de Datos Espaciales de Andalucía”

Contenido

1. INTRODUCCIÓN	2
2. NECESIDAD NO CUBIERTA	3
3. ANTECEDENTES	6
4. ESTADO DEL ARTE.....	8
5. OBJETIVOS GENERALES	9
6. OBJETIVOS ESPECÍFICOS	10
7. RESULTADOS ESPERADOS: REQUISITOS FUNCIONALES.....	11
7.1 Objetivo 1: Garantizar una amplia disponibilidad de datos temáticos multidimensionales integrados espacialmente.	11
7.1.1 Estado del arte	11
7.1.2 Resultados esperados. Requisitos funcionales.	13
7.1.3 Grado de innovación	14
7.2 Objetivo 2: Desarrollo de una plataforma con amplias capacidades para el almacenamiento, explotación y difusión de datos normalizados. Requisitos funcionales.....	16
Bloque I: Tratamiento de datos espaciales (TDE)	17
7.2.1. TDE01. Ciclo de vida del dato	17
7.2.2. TDE02. Importación y normalización de datos espaciales	18
7.2.3. TDE03 Adscripción de variables a unidades espaciales de referencia	20
7.2.4. TDE04. Servicio de análisis en la parte servidora (CPU/GPU).....	22
7.2.5. TDE05. Servicio de datos y de difusión.	25
7.2.6. TDE06. Servicio de Clientes web de acceso, análisis y explotación orientados a soluciones.....	28
Bloque II: Gobierno de la plataforma (GP).	30
7.2.7 GP01 Modelo flexible de computación en la nube	30
7.2.8 GP02 Modelo flexible de gestión de la arquitectura de microservicios	31
7.2.9 GP03 Gestión integral del nivel de servicio de la plataforma.....	33
7.2.10 GP04 Modelo de interacción y construcción extensible de aplicaciones de análisis.....	34
7.3 Objetivo 3: Desarrollo de aplicaciones (productos) de valor añadido para el soporte a la gestión en ámbitos específicos de los servicios públicos.	35
7.3.1 Ámbito sanitario.....	35
7.3.2 Agricultura.	36
7.3.3 Ámbito forestal.	37

7.3.4	Ámbito de reto demográfico.....	37
8.	PRESUPUESTO ESTIMATIVO.....	38
9.	INDICADORES DE IMPACTO A CONSIDERAR.....	39
9.1	Multiplicar el número de gestores que realizan análisis integrados de datos espaciales para la mejora de los servicios públicos.....	39
9.2	Ampliar las capacidades de análisis y explotación de los datos espaciales.....	39
9.3	Multiplicar el uso y explotación de grandes conjuntos de datos.....	40
9.4	Favorecer el uso de la Inteligencia Artificial en la explotación de datos espaciales.....	40
9.5	Mejorar la toma de decisiones sobre aspectos de los servicios públicos vinculados a los tiempos de accesibilidad a cualquier tipo de elemento.....	41
9.6	Favorecer el mercado de generación de soluciones basadas en datos espaciales para empresas del sector TIG.....	41
9.7	Impulsar la I+D+i en el sector TIG.....	41
	BIBLIOGRAFÍA.....	43

1. INTRODUCCIÓN

Los sistemas de información basados en datos espaciales, desde principios de los años noventa, se han convertido en herramientas indispensables como soporte a la planificación y gestión de múltiples actividades públicas: medio ambiente, agricultura, infraestructuras, ordenación del territorio, minería, protección civil, etc.

Son muchos los ejemplos de sistemas de información que se han desarrollado, por parte de la Junta de Andalucía, y de otras administraciones: Rediam, SIG-PAC, SIGMA, Catastro, PNOA, SIOSE..., los cuales siguen siendo pilares fundamentales para asistir la toma de decisiones en multitud de temáticas que afectan a los servicios públicos. Además, en los últimos años, el uso de este tipo de sistemas se ha venido incrementando en otros ámbitos de los servicios públicos, como son los directamente dirigidos al ciudadano: salud, vivienda, educación, transportes, protección civil, etc.

Aunque muchos de éstos nacieron, a modo de grandes inventarios, para la evaluación y seguimiento de elementos muy específicos, de un determinado ámbito de la gestión (SIGPAC, Catastro, SIPNA, SIOSE, etc.), a lo largo del tiempo ha quedado demostrado que el máximo potencial de los datos se obtiene de la posibilidad de interrelacionar unas informaciones, con muchas otras, y no del análisis aislado de un discreto conjunto de variables.

Es, precisamente, la posibilidad de realizar análisis que comprendan la integración de datos muy diferentes, la que permite modelizar fenómenos de interés (erosión, riesgo de incendio, riesgo de inundación, previsión meteorológica, potencialidad forestal y agrícola, idoneidad de hábitats para especies, etc.), la extracción de estadísticas e indicadores (kilómetros de red viaria por espacio natural protegido, población residente a 100 metros de una red fluvial, etc), y, en definitiva, la que ha llevado a que, en los últimos años, se hayan desarrollado soluciones que hacen uso de algoritmos de inteligencia artificial (machine learning, deep learning) que implican una enorme cantidad y diversidad de parámetros, para obtener modelos predictivos, de tendencias, de reconocimiento automático de patrones y de clasificación, entre muchas otras aplicaciones.

Como consecuencia, estas herramientas están ampliamente consolidadas en el ámbito de muchas actividades públicas, y privadas, siendo muy numerosos los ejemplos de aplicaciones efectivas y exitosas.

Sin embargo, a día de hoy, y con carácter general, se constata una gran deficiencia en el mundo de los sistemas de datos espaciales de las administraciones públicas. Ésta consiste en el escaso uso,

respecto al potencial, de explotaciones que impliquen la integración de informaciones diversas, y más aún, cuando se trata de fuentes de información de gran tamaño, y de procesos de análisis muy complejos, que impliquen una gran capacidad de computación.

Así, es indudable que, estos análisis integrados, sea para realizar una modelización, calcular un indicador, o para extraer unas sencillas estadísticas derivadas del cruce de datos espaciales, que podrían ser de gran utilidad para muchos procesos de toma de decisiones, no se llevan a cabo, a pesar de disponer de datos y herramientas para ello.

Como consecuencia, actualmente, se percibe una importante infrutilización del potencial que tiene el análisis integrado de los enormes conjuntos de datos residentes en estos sistemas, y se constata que el impacto que tienen estas herramientas sobre la mejora de la gestión de múltiples servicios públicos es muy reducido respecto a su potencial.

Entre las causas que explican estas limitaciones de los sistemas de datos geográficos de las administraciones destacan dos:

- La explotación integrada de datos mediante herramientas SIG es un proceso complejo, y mayoritariamente manual, que requiere de la implicación de personal altamente capacitado en el uso de TIG (Tecnologías de la Información Geográfica). Este hecho restringe el potencial de usuarios a un número muy reducido dentro de las organizaciones, que, además, tendrán que realizar procesamientos complejos y costosos en tiempo.

- Las limitaciones de la capacidad de almacenamiento y procesamiento de estos sistemas, fruto de una serie de problemas técnicos inherentes al dato geográfico, y muy especialmente a la topología del dato vectorial, han bloqueado su incorporación a los nuevos contextos computacionales (Big data, paralelización, procesamiento en la nube, etc). Como consecuencia, se adolece de la capacidad de digerir y explotar la ingente, e incesantemente creciente, cantidad de datos disponibles (plataformas espaciales y aerotransportadas, drones, redes de medición automáticas, redes sociales, etc), se dificulta grandemente la aplicación de las tecnologías de análisis vinculadas a la Inteligencia Artificial, y se impiden muchos procesos de automatización, los cuales requieren de datos previamente normalizados espacialmente.

En los últimos años, de forma fraccionaria, se vienen produciendo una serie de avances en las tecnologías de la información, que, adaptados y aplicados al campo de la geoinformación, pueden dar solución a los problemas antes referidos. Se trata de tecnologías tales como ETL, Spatial Big data, modelos Discrete Global Grid Systems, arquitecturas de microservicios, Cloud computing, geovisores y clientes ligeros de visualización y explotación de datos estadísticos y geográficos, etc.

Todo este conjunto de tecnologías, en avanzado estado de desarrollo, y ya, algunas, de uso habitual en muchos campos de actividad, aún no se han unificado en una herramienta integral, que supere las barreras antes mencionadas, y permita multiplicar exponencialmente las capacidades de realizar análisis, y hacerlas accesibles a un gran número de gestores.

La potenciación de las capacidades de explotación de los datos, y la generalización de aplicaciones para hacerlas accesibles a un gran espectro de usuarios, y derivarlas hacia una mejora de los servicios públicos, es la finalidad del proyecto “**Plataforma Innovadora de Análisis Integrado de Datos Espaciales de Andalucía**”, cuyo acrónimo es AIDEA.

2. NECESIDAD NO CUBIERTA

En la sociedad actual, la información se ha convertido en uno de los pilares fundamentales para multitud de actividades. La gestión de las responsabilidades públicas no está ajena a esta realidad, y, como se señaló anteriormente, son muchos los esfuerzos realizados para desarrollar herramientas de

gestión de información que les ayuden a mejorar los procesos de toma de decisiones en múltiples campos.

Sin embargo, en lo que afecta a sistemas basados en información espacial, se aprecian unas importantes limitaciones que minimizan extraordinariamente su uso potencial para la mejora de los servicios públicos. La búsqueda de soluciones que permitan superar estas deficiencias constituye la meta a perseguir, de forma que se dé satisfacción a que las necesidades de informaciones procedentes de análisis espacial estén al alcance de un amplio espectro de gestores, y que su uso sea rutinario, y no puntual, como en el momento presente.

Actualmente, los diferentes sistemas de información de las administraciones funcionan como grandes repositorios de conjuntos de datos de naturaleza, composición y formatos muy diversos (lineales, poligonales, puntuales, ráster, con diferentes escalas y resoluciones, etc), sobre los que se aplican múltiples procesos de análisis espacial (álgebra de mapas, análisis espacial, geoprocementos, filtros, etc) para proceder a su análisis y explotación, ya sea para obtener estadísticas de determinados parámetros, modelizaciones que reflejan el comportamiento de determinados fenómenos, mapas en forma de geovisores de consulta, etc.

Como se señaló con anterioridad, entre las explotaciones que tienen mayor interés para la gestión, se hayan las que se producen a partir del análisis integrado de diferentes variables. Para la realización de análisis integrados y para modelizaciones, se requiere, como paso previo, un costoso proceso de normalización espacial de las variables, siendo éste una tipología de proceso que, a día de hoy, requiere de una alta capacitación en SIG, y, cuando se trata de fuentes de datos muy grandes, de una altísima capacidad de computación. Los problemas inherentes a estos requisitos de capacitación, y a las limitaciones técnicas para poder solventar los problemas de alta capacidad de almacenamiento y de procesamiento, y de maximizar los procesos automáticos de carga y explotación de la información, son los principales cuellos de botella para que los datos espaciales se usen de forma extensiva, quedando múltiples aplicaciones potenciales sin llevarse a cabo.

Solventar estos problemas constituye la necesidad fundamental a ser afrontada por este proyecto.

Para satisfacer esa necesidad genérica, se deberán solucionar una serie de retos tecnológicos de los sistemas actuales, que se describen a continuación:

- Establecimiento de **procesos metodológicos que faciliten la integración espacial normalizada de los conjuntos de datos** de interés transversal para muchas políticas públicas. Desde 2007 se ha avanzado mucho en la interoperabilidad en el ámbito de la información geográfica, mediante la normalización de los sistemas de referencia, y en la de modelización de los conjuntos de datos espaciales (especialmente en el proceso de implantación de la directiva Inspire), pero no existen procesos estandarizados de los procedimientos para su integración espacial. Esta deficiencia deberá ser abordada por el proyecto para garantizar la disponibilidad para los usuarios de datos integrados espacialmente, y preparados para su explotación, frente a los problemas de que, esta costosa tarea, se tenga que realizar individualmente, y usando criterios no armonizados. Cabe señalar que, a día de hoy, son contados los ejemplos de conjuntos de datos que se proveen en sistemas de referencia pensados para posibilitar esta integración, y que este proceso requiere del desarrollo de metodologías ex-profeso para cada conjunto de datos, constituyendo un núcleo fundamental de necesidad de innovación de este proyecto.
- Facilitar al máximo los **procesos de automatización de la captura, gestión, almacenamiento, normalización, adscripción a sistemas de referencia normalizados y explotación** de los datos espaciales. A día de hoy los procesos siguen siendo manuales, y

requieren de una alta capacitación, estando, por tanto, disponibles solo con carácter puntual, y no extensivo. Resulta una necesidad simplificar estos procesos para los usuarios, incluso para los más avanzados.

- Se requieren de capacidades de almacenamiento y gestión de **ingentes cantidades de datos**, actualizados en alta frecuencia, procedentes de múltiples fuentes (internet de las cosas, redes sociales, imágenes satelitales y aeroportadas, etc), y de proceder a su normalización, depuración y almacenamiento. La incorporación de **procesos automatizados** de carga y normalización, y el uso de tecnologías de Spatial Big Data que hagan más eficiente el almacenamiento y la gestión de los datos, suponen una necesidad a afrontar para dar respuesta a este problema.
- Igualmente, otra necesidad será superar la **insuficiente capacidad computacional** de los sistemas, para dar respuesta a análisis complejos y muy exigentes en términos de computación. Para ello se exige poder desplegar procesos en la nube, y poder realizar el necesario procesamiento en paralelo. Dado que una parte fundamental de la información original residente en los sistemas procede de datos poligonales y lineales, donde la topología impone una importante restricción para la computación distribuida, se hace necesario superar este problema mediante unidades de integración espacial que conlleven la resolución de estos problemas (estructuras de rejillas multiescalares y multidimensionales, y DGGs, unidades administrativas de gestión, etc).
- Para extender el uso de la información espacial se requiere **hacer accesible su explotación a personal no experto en TIG**. Éste es, sin duda, el más importante de los cuellos de botellas que tiene que solucionar el proyecto, para asegurarse un impacto positivo en la mejora de los servicios públicos. Los sistemas actuales, más allá de los clientes y geovisores que permiten unas amplísimas capacidades de localización, consulta y visualización de los datos, no poseen clientes que faciliten los análisis integrados de forma amigable. Se requiere generalmente de herramientas SIG, cuyo uso solo está al alcance de personal especialista. AIDEA deberá desarrollar soluciones que faciliten la explotación integrada de los datos, de forma que la realización de geoprosesos, la aplicación de algoritmos de análisis y clasificaciones y la obtención de mapas y estadísticas resultantes de modelos estén al alcance de un amplísimo elenco de usuarios, no especialistas. Estos procesos podrán implicar análisis de grandes conjuntos de datos, que impliquen, o no, procesos paralelizables y distribuidos en la nube.
- Una necesidad adicional será la generación de una e-infraestructura dotada de una alta flexibilidad, con amplias capacidades para la gestión, almacenamiento y explotación de los datos, abierta a una amplia tipología de usuarios, y dotada de una estructura de gobierno que automatice los procesos de consulta y explotación, de forma que sea transparente para el usuario dónde y cómo se realiza el despliegue. Así se superarán, igualmente, las restricciones de uso de los despliegues en cloud, por motivo de conocimientos técnicos de estas tecnologías, facilitando su empleo por todo tipo de usuario.
- Entre las deficiencias de los sistemas vigentes se encuentra que, **algunas dimensiones fundamentales de los estudios geoespaciales aplicados a la gestión de responsabilidades públicas, como es la accesibilidad, no estén disponibles**. Para la provisión de servicios públicos según accesibilidad de la población a éstos, la evaluación de la movilidad, el cálculo de emisiones por transporte terrestre, los costes de transporte de biomasa, caracterización de las de olivar por distancia a almazaras, y para muchos otros estudios, se necesita, y no está disponible, una solución que permita conocer, y mantener esta variable, y que, ésta pueda ser integrable con otras para derivar indicadores, mapas y muchas otros parámetros y productos de interés para la gestión.

- **Facilitar el desarrollo de aplicaciones específicas a medida**, que demanden el análisis integrado de un gran volumen de datos espaciales de naturaleza diversa, y en continua actualización, para la gestión de temáticas estratégicas de diferentes servicios públicos (salud, emergencias, transporte, medio ambiente, agricultura, etc.). En el momento actual, cualquier aplicación especializada debe partir de cero, debiendo normalizar y personalizar las herramientas. Se requiere desarrollar una plataforma dotada de alta flexibilidad para nuevos desarrollos, y donde se concentre el núcleo fundamental de datos de referencia actualizados permanentemente, y las herramientas para su explotación, se podrán realizar aplicaciones muy especializadas y personalizadas, que podrán requerir nuevos datos de la propia entidad demandante, con características muy específicas (confidencialidad, formato, resolución, etc.) y nuevos algoritmos de procesamiento, para dar respuesta a ámbitos muy concretos. Con ello se persigue potenciar que el uso de las TIGs redunde, no solo en aplicaciones de consultas de amplio espectro, destinada a muchos gestores y público en general, sino también facilitar la puesta en marcha de herramientas muy especializadas que partirán de un conjunto de datos normalizados previamente, y de herramientas de análisis y explotación avanzadas, junto a capacidades de alta computación, facilitando, así, grandemente algo que, a día de hoy, es sumamente complejo. Esta dimensión tiene un especial interés para las empresas TIGs que ofrecen aplicaciones a medida, en tanto que línea de actividad que, sin duda, se verá potenciada.

3. ANTECEDENTES

El proyecto AIDEA tiene como elemento fundamental proporcionar una gran base de datos normalizados espacialmente, y mantenida en el tiempo, que contenga información sobre las variables de referencia para la gestión de buena parte de las políticas públicas que son responsabilidad de la Junta de Andalucía.

La característica fundamental de esta base de datos geográficos normalizados espacialmente, consiste en proveer informaciones que ya están preparadas para que cualquier proceso de análisis integrado, pueda llevarse a cabo de forma sencilla, salvando las dificultades que para ello impone la dimensión geométrica y la topología del dato geográfico, la naturaleza temática de éstos, o su formato, abriendo así la posibilidad de su análisis y gestión en contextos computacionales de alto rendimiento (spatial big data, cloud computing, machine learning, etc).

A ese respecto, cabe ser destacado, que, desde el mundo de las TIGs, la principal propuesta, aunque no única (ver apartado 7.1 del presente documento), para abordar el reto de esta necesaria normalización espacial, consiste en la integración de todas las variables y parámetros, esencialmente en formatos ráster y vectoriales, en los sistemas de referencia conocidos como “sistema de rejillas multiescalares y multidimensionales”, en español, o “Discrete Global Grid Systems” en inglés. Éstos permitirán superar los problemas de paralelización de los datos para los procesos de procesamiento distribuido (cloud computing). Además, estos modelos permiten obtener enormes vectores de datos sobre cada unidad discreta (celdilla). Estos vectores constituyen la entrada natural de datos a los algoritmos de los campos del machine learning y el deep learning, permitiendo la exploración de la información integrada mediante estas potentes técnicas analíticas.

Esta novedosa propuesta de estructuración de los datos espaciales es la respuesta adoptada como estándar desde 2017 por el Open Geospatial Consortium (DGGS) para facilitar el análisis espacial eficiente de un ingente número de datos espaciales heterogéneos, especialmente para escalas globales. Igualmente ha sido adoptado, en este caso como sistemas proyectados, y para escala continental o local, como una de las bases de referencia en el marco de la gran iniciativa de normalización y compartición de datos geográficos en la UE, la directiva Inspire (Geographical Grid Systems).

La exploración de las potencialidades de estos sistemas de referencia multiescalares y multidimensionales, como posibles soluciones a las limitaciones para análisis integrados de grandes conjuntos de datos de las tecnologías de información geográfica, tiene un hito importante en 2021 con la tesis doctoral de la Universidad de Sevilla de Juan Pedro Pérez Alcántara titulada “Plataforma cloud para la integración espacial de geoinformación en estructuras de teselas multiescalares asimétricas, su análisis y su visualización”. En este trabajo de investigación se desarrollan, y testean, a nivel de demostración, métodos para facilitar la creación de un prototipo de plataforma cloud, que permita el análisis temático multicriterio con técnicas de procesamiento distribuido Big Data y Machine Learning, superando las dificultades que estas metodologías presentan sobre datos geográficos, así como que proporcione un medio para el almacenamiento eficiente de información y su tránsito entre servidores y clientes.

Otro importante referente para la puesta en marcha de la iniciativa AIDEA es el PROYECTO CLOUDRISK_A (P18-FR-2574), desarrollado por la Universidad de Sevilla, y finalizado en 2022. El objetivo general de este proyecto se centra en ofrecer un conjunto de metodologías que solventen el problema esencial de la integración espacial de datos de distinta tipología (vécotor/raster), temática (sociodemográficos/físico naturales) y escala (globales/de detalle), utilizando para ello diferentes entidades espaciales, como son las rejillas multiescalares/multidimensionales o los edificios procedentes del Catastro (Atom). Esta integración espacial se muestra crítica en los procesos de evaluación de los riesgos costeros, y en la generación de indicadores de vulnerabilidad, donde los datos multidimensionales constituyen un requisito ineludible. Adicionalmente, el proyecto desarrolló un novedoso entorno científico- tecnológico, tanto para la referida integración espacial de los datos, como para el tratamiento y análisis de los mismos (cloud computing y open data), así como un conjunto de herramientas web (dashboard) que facilitan la gestión de estos riesgos a la sociedad a través de la adopción de las oportunas medidas de adaptación.

El gran antecedente de referencia para la presente iniciativa es, sin duda, el proyecto de Compra Precomercial Cloud_IA. Éste tiene como objeto el desarrollo de las soluciones que permitan la implementación de un prototipo de e-infraestructura de análisis integrado de geoinformación ambiental, con un amplio alcance temático: generar una gran base de datos geográficos normalizados espacialmente, en estructuras de rejillas multiescalares/multidimensionales, desarrollar una plataforma basada en arquitectura de microservicios dotada de unas soluciones de gobierno que permita una amplia flexibilidad a la hora de garantizar despliegues cloud y de gestión de la arquitectura de microservicios, así como garantizar una amplísimo sistema de construcción extensible de aplicaciones de análisis, que, además de incluir capacidades de explotación de datos basadas en machine learning e IA, ayuden a facilitar la explotación de los datos a usuarios no expertos.

Este proyecto se inició en julio de 2022, y tiene prevista su finalización en 2025. Aunque esté a mitad de desarrollo, su fase 1, de exploración de soluciones, que terminó en marzo de 2023, constituye un importante referente. Así se ha podido comprobar, en esta fase inicial, la viabilidad de las soluciones innovadoras propuestas por los tres adjudicatarios para solucionar los 10 grandes retos que plantea la implantación del prototipo de e-infraestructura (ver el apartado 5 del Documento de Prescripciones Técnicas de la CPP Cloud_IA, descargable en el siguiente [enlace](#)).

Igualmente resaltar que, la fase 2 de la CPP de Cloud_IA tiene como cometido la implementación del prototipo de e-infraestructura, el cual estará completado a principios de 2025, momento en el que se iniciará la fase 3 de validación en entorno real, de la que resultará una solución que se espera llegue a un TRL7. Quiere decirse con ello que, en el momento de comenzar a redactar los pliegos para la potencial contratación de AIDEA, se contará con un importante elenco de tecnologías testadas en un “entorno relevante” (demostradores de la fase I de Cloud_IA). En la medida de lo posible, los

resultados del proyecto Cloud_IA se podrán a disposición de AIDEA, para ser tenidos en cuenta en la redacción de los documentos para su potencial contratación.

Con relación a otro objetivo de AIDEA, no contemplado en la CPP Cloud_IA, como es el de la generación de un módulo específico para la incorporación de la variable de accesibilidad por carretera, entre todos los puntos de la Comunidad Autónoma, existe una metodología desarrollada en el marco del proyecto Indalo, coordinado por la Agencia de Medio Ambiente y Agua de Andalucía. Esta iniciativa, enmarcada en el programa de la CE Lifewatch-ERIC, tiene, en una de sus líneas de desarrollo, el establecimiento de una [“Red de observatorios y plataforma web para la integración espacial de las variables demográficas, espacios urbanos e infraestructuras y cálculo de indicadores de impacto y vulnerabilidad en los ecosistemas costeros: costa de Andalucía”](#), la cual ha sido desarrollada por el departamento de Geografía Física y Análisis Geográfico Regional de la Universidad de Sevilla. En el marco de estos trabajos, finalizados en 2023, se ha desarrollado una experiencia para el cálculo de la accesibilidad a 3000 puntos de la costa, desde cualquier punto, con datos, del territorio andaluz. Esta solución constituye un referente para extender su implementación y explotación para cualquier punto de la geografía autonómica, aunque habrá que abordar indudables retos tecnológicos derivados de la enormidad de datos a manejar.

Por último, existe un nuevo elemento innovador a introducir en AIDEA, no abordado por la CPP Cloud_IA, como es la introducción del lenguaje natural como herramienta de interacción con la plataforma. Por un lado, existen multitud de frameworks para el desarrollo de chatbots o asistentes virtuales, como por ejemplo IBM Watson o DialogFlow, que muestran que la tecnología necesaria para el desarrollo de estas herramientas está cada vez más madura, mientras que en el ámbito de la transformación digital se están realizando esfuerzos para la adopción de estas técnicas, como muestra el trabajo que se está llevando a cabo en el proyecto BUBO: *“Mejorando la operación y desarrollo de servicios digitales por medio de la colaboración bot-humano”* (PID2021-126227NB-C21/22) de la Universidad de Sevilla. Por otro lado, el auge de los grandes modelos de lenguaje (LLM) y la IA generativa, como GPT, Bard o Bing Chat, ofrece una oportunidad clara de impulsar la innovación en los modos de interacción con la plataforma que permita un mayor aprovechamiento de los datos espaciales incorporados.

Con este acúmulo de antecedentes, la CSMAEA entiende que el actual contexto de desarrollo tecnológico, con los adecuados impulsos innovadores, permitiría la implementación, y puesta en explotación operativa, de una plataforma integral con las características de AIDEA.

Como paso previo, y con vistas a evaluar la viabilidad de llevar a cabo su desarrollo, se plantea la presente Consulta Preliminar al Mercado, con la finalidad de recoger información del sector empresarial y de los agentes del conocimiento, sobre aspectos fundamentales de una potencial contratación.

4. ESTADO DEL ARTE

En el contexto del proyecto que se presenta a la CPM se articulan muchas tecnologías, en desigual estado de desarrollo, cuyos análisis del estado del arte, se llevan a cabo, por separado, en subepígrafes diferenciados del apartado 7 del presente documento.

No obstante, como proyecto de e-infraestructura novedosa, con las características descritas, cabe resaltar que, durante la realización de la fase 1 del proyecto CPP Cloud_IA, la CSMAEA contrató, en 2022, la elaboración de un Informe de Vigilancia Tecnológica a la empresa LINKNOVATE, del cual se obtuvieron conclusiones muy relevantes sobre la inexistencia de soluciones operativas de estas características.

Este estudio, bajo el título de “*Plataformas DGGs para representación de datos territoriales*”, y realizado con buscadores basados en IA en la red a nivel global (Google Patents e Innoscout), tenía como finalidad la identificación de los documentos y productos más relacionados con la “búsqueda de soluciones integradas e innovadoras que sirvan para la creación de una e-infraestructura de información ambiental, con gran capacidad de almacenamiento y procesamiento en la nube, que permita la generación de soluciones a medida para gestores, investigadores y ciudadanía no expertos en TIGs, en ámbitos relacionados con la gestión ambiental y territorial”.

A partir de estos requerimientos se realizó una búsqueda exhaustiva de organizaciones provenientes tanto del sector público como privado que pudieran cumplir con la totalidad de requisitos exigidos.

Si bien no se encontró ninguna organización que lo haga, ni ningún documento relevante que refleje la existencia de una herramienta cuyo nivel de integración y funcionalidad cumpla con todas las exigencias del proyecto, sí se detectaron algunas que cumplieran de forma parcial algunos de los requerimientos, así como abundantes iniciativas en el ámbito académico.

En su apartado de Conclusiones, el informe establece que *"en la evaluación del grado de innovación, y teniendo en cuenta que tanto el TRL de partida del proyecto Cloud-IA como el de esta vertical concreta parten de un tramo bajo comprendido entre los TRL 1 y 5, se presentan ampliamente en las anteriores secciones los productos en mercado provenientes mayoritariamente de iniciativas privadas (con diferentes prestaciones y características), así como los desarrollos innovadores más cercanos a fase piloto de empresas y organismos de I+D, que no llegan al nivel de los requisitos descriptores y grado de ambición del proyecto Cloud-IA.*

Por estos motivos el grado de innovación del reto de interés se considera, que con los requerimientos de coste, eficacia y características adecuadas puede ser elevado y elegible para un proceso de Compra Pública Precomercial (CPP)."

Como se expuso en el epígrafe de “Antecedentes” del presente documento, durante la realización de la fase 1 de la CPP Cloud_IA, finalizada en marzo de 2023, se ha podido comprobar la viabilidad de las soluciones innovadoras propuestas por los tres adjudicatarios para solucionar los 10 grandes retos que plantea la implantación del prototipo de e-infraestructura. No obstante, aún queda camino por recorrer, pues estas tecnologías solo han sido validadas en un entorno de pruebas, y de forma aislada, y cada una por separado, y no funcionando en conjunto. Quedarían pendiente, el diseño del prototipo de plataforma integrada, y su validación en entorno real, lo que conllevará llegar a una potencial solución, en 2025, con un nivel de madurez tecnológica cercano a TRL7.

5. OBJETIVOS GENERALES

Con carácter general, AIDEA tiene como meta multiplicar exponencialmente el impacto de los datos espaciales residentes en los sistemas de información de las administraciones, para la mejora de la gestión y planificación de múltiples políticas públicas (medio ambiente, salud, agricultura, vivienda, transporte, etc.), e, igualmente, favorecer la actividad en el sector privado de las TIGs, y el de la investigación, al poner a disposición de la sociedad un gran repositorio de datos espaciales preparados para su integración y análisis con las herramientas más novedosas (machine learning, IA, deep learning, etc).

Para alcanzar este objetivo, se persigue una solución integral, aún no disponible en el mercado, que facilite:

- La generación y mantenimiento semiautomatizado de un gran repositorio de datos normalizados espacialmente, obtenidos a partir de los conjuntos de datos geográficos de referencia, así como de la generación de nuevas variables, como la referente a la accesibilidad.
- Multiplicar las capacidades de almacenamiento y computación, superando las limitaciones de los sistemas actuales, y, en especial, los inconvenientes técnicos que impiden la paralelización, y subsiguientes despliegues cloud. Solo así se podrán digerir, y explotar la creciente disponibilidad de datos, de naturaleza muy diversa, que se ponen a disposición de los usuarios (redes de medición, drones, imágenes de satélite, redes sociales, etc.).
- Extender la capacidad de exploración y explotación integrada de datos geográficos, incluyendo análisis de machine learning e IA, a un amplísimo número de usuarios, no necesariamente expertos en TIGs, y, por tanto, haciendo que la alta capacitación en estas tecnologías no sea el requisito indispensable para análisis avanzado y extenso de la información geográfica.
- Desarrollo de una herramienta para la captura, actualización y explotación de una variable fundamental para muchos servicios públicos directamente vinculados al ciudadano (salud, educación, transporte, acceso a equipamientos públicos, etc.), como es la accesibilidad a cualquier punto del territorio.
- Abrir la posibilidad, o facilitar grandemente, del desarrollo de aplicaciones específicas de alcance complejo, para el análisis de parámetros de interés en ámbitos concretos de la gestión pública.
- Estimular el mercado la generación de soluciones a medida para empresas del sector TIG, que verán favorecidos los procesos de generación de clientes y geovisores avanzados centrados en la gestión de variables de interés para un servicio público concreto, para el propio sector empresarial y para la ciudadanía.
- Potenciar todo tipo de procesos de investigación de contenido geospacial, al poner a disposición de los agentes del conocimiento un gran repositorio de datos, previamente preparados para su análisis mediante técnicas avanzadas, incluyendo machine learning e IA, entre otros.

6. OBJETIVOS ESPECÍFICOS

Objetivo 1- La puesta a disposición para la Junta de Andalucía, y para la sociedad en general, de un amplísimo elenco de variables integradas y normalizadas espacialmente, que pueda ser mantenida y actualizada constantemente, y que facilite su explotación en contextos computacionales de alto rendimiento (Spatial Big data, procesamiento en paralelo, etc.), supone un pilar fundamental del proyecto. Este enorme repositorio de datos deberá abarcar todo el conjunto de variables geoespaciales que son de interés general y transversal a muchas políticas públicas, descritas en el apartado 7.1, e, igualmente, a aquellas propias de las aplicaciones específicas que se mencionan en el apartado 7.3.

Cabe resaltar, por su carácter novedoso y por su importancia para la gestión de muchas políticas públicas, la variable de accesibilidad entre elementos de todo el territorio de la Comunidad Autónoma, que se detalla en el apartado 7.1.

Objetivo 2- La creación de una plataforma que permita el mantenimiento automático, la consulta, la difusión y la explotación integrada de las variables normalizadas necesarias para las principales políticas públicas de Andalucía, mediante el desarrollo de clientes personalizados que faciliten su análisis de forma amigable para un amplio elenco de usuarios no especializados, incluyendo el lenguaje natural como fórmula de interacción. Esta plataforma, además de recursos locales, deberá disponer de una alta capacidad de despliegue de procesos en cloud, y de mecanismos de gobierno flexibles que hagan transparente al usuario la tipología de nube y de modelo de computación a usar. Igualmente incorporará una amplia variedad de soluciones de análisis integrados de datos, incluyendo clientes personalizables de fácil uso, y con capacidades de procesos de análisis basados en tecnologías de machine learning e IA.

Objetivo 3- La generación de un número limitado de aplicaciones específicas que, haciendo uso de la plataforma, respondan a necesidades de gestión de amplio impacto en diferentes ámbitos de los servicios públicos: sanitario, agrícola, planificación y gestión hidrológica, forestal y de obtención de indicadores socioeconómicos. Estas aplicaciones específicas, desarrolladas a medida, representan productos en los que la cadena de valor de todo el proyecto se materializa, y ejemplifica todo el potencial del mismo. Igualmente constituirán una referencia de gran interés para potenciar las posibilidades de AIDEA como infraestructura que favorece y facilita, por su datos y por sus herramientas de explotación, y por la flexibilidad de su configuración, el desarrollo de aplicaciones personalizadas para múltiples aspectos de la gestión, que constituirán, sin duda, un impulso a la actividad comercial en el sector de las empresas TIG que ofrecen este tipo de servicios de valor añadido, así como para muchas actividades de I+D que requieran análisis espaciales.

7. RESULTADOS ESPERADOS: REQUISITOS FUNCIONALES

Es evidente que la actualidad existe una clara demanda de datos integrados espacialmente, tanto por la sociedad, las empresas o la administración, es decir, de datos multidimensionales asociados a diferentes entidades espaciales. En los últimos años, se han testado para ello, tanto estructuras de rejillas o teselas multiescalares (cuadradas, triangulares o hexagonales), como su integración en entidades administrativas (secciones censales, por ejemplo) y/o entidades espaciales al máximo nivel de detalle necesarias para de gestión directa sobre el territorio (parcelas, unidades de gestión del SICPAC, edificios catastrales, etc.). Por ello, el primer resultado esperado de este reto se centrará en garantizar el **acceso, el análisis, la exploración y geovisualización a un amplio conjunto de datos multidimensionales normalizados espacialmente para toda la Comunidad Autónoma de Andalucía**, que se concretará físicamente en el desarrollo de una **plataforma web** a desarrollar para su integración, gestión, acceso y explotación, así como en **aplicaciones temáticas para casos pilotos**, cuyos requerimientos funcionales temáticos y técnicos se desarrollaran en los siguientes puntos de resultados esperados de este apartado

En el presente apartado se aborda, con mayor nivel de detalle, lo que se pretende funcionalmente obtener con el proyecto, haciendo mención del estado del arte y al grado de innovación que, se estima, requieren cada uno de sus diferentes apartados u objetivos.

La definición del detalle de la solución se abordará de forma desglosada en tres bloques, correspondientes a cada uno de los objetivos específicos que se quieren conseguir.

7.1 Objetivo 1: Garantizar una amplia disponibilidad de datos temáticos multidimensionales integrados espacialmente.

7.1.1 Estado del arte

Como se expuso con anterioridad, la creciente demanda de datos multidimensionales, integrados espacialmente, contrasta con la actual escasez de disponibilidad de éstos. Al margen de los claros avances en la normalización espacial de los datos temáticos individuales, que ha supuesto la directiva Inspire, sigue pendiente una eficiente normalización para los datos multitemáticos integrados espacialmente. Aun así, en los últimos años se han producido avances en la accesibilidad de estos datos, si bien de forma parcial, y muy dependiente de la naturaleza topológica y/o la continuidad espacial de las variables integradas. Básicamente podemos distinguir tres casos:

Cuando se trata de **variables físicas, biofísicas o territoriales** que tienen la propiedad de ser **continuas espacialmente** (su valor cambia cuando nos desplazamos en x o y), se ha avanzado mucho en la propuesta de nuevas estructuras de datos basados generalmente en estructuras matriciales multidimensionales (modelo ráster). Así, hoy en día, para enormes volúmenes de datos físicos o biofísicos (temperatura, contaminación, imágenes de satélite, ortofotos, MDE, modelización climática, etc) se han desarrollado estructuras muy eficientes para su almacenamiento y gestión (HDF, NetCDF, GRID, CUBES, etc) e incluso para su difusión web, con estándares OGC (Cloud Optimized GeoTiff –COG-).

Cuando se trata de **variables socioeconómicas, estadísticas o territoriales**, generalmente **discontinuas espacialmente**, se suelen utilizar entidades administrativas (desde datos asociados a nivel nacional o a secciones censales en España o a las clásicas NUTS a nivel europeo). Aunque es una forma eficiente para su difusión y tratamiento estadístico, plantean graves problemas cuando en su análisis participe la componente de localización, ya que hay una clara inconsistencia espacial entre continente (la unidad administrativa) y el contenido (la variable temática –población, renta, etc.-). Por otra parte, la heterogeneidad de tamaño de la mismas hace aflorar la conocida problemática de MAUP (Modifiable Areal Unit Problem).

En parte, para solventar algunas de estas debilidades, se han propuesto como unidades de integración espacial diferentes estructuras de datos espaciales que segmentan el territorio de forma homogénea: las **estructuras de teselas regulares** (cuadrados, triángulos o hexágonos). Desde la perspectiva espacial se han diferenciado:

Estructuras de teselas regulares generadas sobre superficies proyectadas, adecuadas para áreas de escalas medias y de detalle. Algunos ejemplos:

GRID de la European Forum for Geostatistics –EFGS-, que ha propuesto tamaños de teselas cuadradas diferentes: 1 km, 10 km y 100 km, proyectadas sobre la proyección ETRS89 LAEA
Rejilla cuadrada de 1 Km utilizada por el INE para la difusión del Censo de Población y Vivienda de 2011 (proyección ETRS89 LAEA)

Los cientos de variables sociodemográficas y territoriales publicadas por el IECA asociadas a una rejilla cuadrada de 250m de lado (proyección ETRS89 LAEA)

La Universidad de Sevilla (grupo de investigación RNM-177) ha incorporado el uso de rejillas multiescalares en diferentes proyectos como INDALO , CLOUDRISK_A (https://kerdes.cica.es/cloudrisk_a/indice/index.html) y tesis doctorales (<https://idus.us.es/handle/11441/116623>)

El proyecto de compra Precomercial **Cloud_IA**, incorpora las estructuras de teselas multiescalares para la integración espacial de un número aún mayor de variables multidimensionales.

Diversas publicaciones recogen el uso de estructuras de teselas, tanto a nivel nacional (Goerlich Gisbert, F.J., & Cantarino Martí, I. -2012-) como comarcal o local (<http://rua.ua.es/dspace/handle/10045/46646>)

Por otra parte, la mayor parte del software SIG (comercial u open source) ofrecen la posibilidad de teselar regularmente el espacio sobre superficies proyectadas.

Estructuras de teselas regulares generadas directamente sobre el elipsoide o la esfera. En estos casos se aborda la teselación (cuadrados, triángulos o hexágonos) a escala global y, por lo tanto, las teselas multiescalares se proyectan sobre el elipsoide o la esfera: se trata de **Discrete Global Grid Systems (DGGS)**. En consecuencia, los DGGS se consideran cada vez más como sistema de referencia de datos espaciales aplicables a nivel mundial, como alternativa al sistema de coordenadas geográficas (Goodchild, 2018; Purss et al., 2019). Igualmente, en muchas aplicaciones, software y plataformas, se utilizan como unidades espaciales para la integración de datos puntuales en el contexto del “big data” (data binning). Ejemplos de desarrollos para la generación y gestión de estos sistemas de teselas globales son, entre otros:

Uber H3 (Uber Technologies Inc, 2018)

rHEALPix (Gibb et al., 2016)
DGGRID (Sahr, 2019)
Google S2 (Veach et al., 2017)
OpenEaggr DGGS (Bush, 2017)

7.1.2 Resultados esperados. Requisitos funcionales.

El primer resultado esperado del proyecto es, por lo tanto, generar y garantizar el acceso, el análisis, la exploración y geovisualización de un amplio conjunto de datos multidimensionales normalizados espacialmente para toda la Comunidad Autónoma de Andalucía.

Fuentes de información y variables a integrar espacialmente:

Las diferentes fuentes de datos y variables a integrar serán las procedentes de los principales organismos públicos o registros oficiales que proporcionan información territorial, ambiental o socioeconómica georreferenciada sobre Andalucía. Al menos se deberán integrar:

Variables derivadas de las fuentes de información más demandadas, partiendo de las que proporcionará el proyecto CPP CLOUD_IA para finales de 2024 (ver anexo 1 del Documento de Prescripciones Técnicas de la CPP Cloud_IA, descargable en el siguiente [enlace](#)).

Variables temáticas adicionales ligadas a las fuentes de información más utilizadas en la gestión de **algunas actividades y políticas temáticas** que finalmente ofrecerán soluciones operativas para diferentes ámbitos de gestión a través de aplicaciones web de carácter específico. Aunque el número de aplicaciones/soluciones se definirá a través del propio proceso de tramitación y ejecución de la CPI, al menos habría que valorar la incorporación de variables ligadas a la gestión de las siguientes temáticas en Andalucía:

AGRICULTURA: las derivadas de la gestión de la política agraria en Andalucía (SIGPAC)

MEDIOAMBIENTE: variables generadas y publicadas sobre riesgos climáticos en diferentes escenarios.

GESTIÓN FORESTAL: variables para evaluar la biomasa de los montes públicos de Andalucía, incluyendo explotación de datos de Lidar, y la incorporación de datos de parcelas de inventarios, junto a otras variables incorporadas en Cloud_IA (relieve, clima, suelo, etc.).

GESTIÓN DEL AGUA: Unidades de gestión, demarcaciones, masas de agua, unidades de explotación, etc.

GESTIÓN DE SERVICIOS DE SALUD. Ámbitos de gestión sanitaria. La peculiaridad de estos datos es que gran parte de los necesitados para la gestión pueden ser datos sensibles y sometidos a un riguroso procedimiento de protección.

Incorporar, de forma innovadora, **variables ligadas a la accesibilidad por carretera y callejero**, ya calculada para varios escenarios (día laborable, fines de semana, verano) e incorporadas a la estructura de rejillas/teselas de la totalidad de Andalucía. Solo esto supondría la generación de matrices de tiempo de accesibilidad que generarían 3.240.000 millones de registros (tesela origen, tesela destino, tiempo, orden) para las teselas de 250 m y para un solo escenario.

Unidades espaciales para de integración de datos

Los resultados esperados de la integración espacial de variables multidimensionales se generarán a partir de dos tipos de entidades espaciales:

Estructuras de teselas regulares generadas sobre superficies proyectadas. En este proyecto, se proponen el uso de una estructura multiescalar de **teselas cuadradas basadas en la proyección ETRS89 LAEA.** Esta selección, entre las muchas opciones posibles, se justifica por:

Garantizar la compatibilidad con gran parte de las experiencias previas realizadas por entidades públicas suministradoras de datos básicos (UE, INE e IECA), proyectos (Cloudrisk_A o CLOUD_IA) y experiencias de investigación.

Una adecuada selección de los niveles escalares garantizaría la propiedad perseguida de que todas las teselas de los niveles inferiores estén autocontenidas en los superiores (topología vertical). Una primera propuesta inicial de los **niveles escalares a utilizar**, con las propiedades anteriores, serían las propuestas en CLOUD_IA.

Aunque, desde la perspectiva semiológica, es más débil que el uso de teselas hexagonales, éstas no pueden garantizar la propiedad anterior y, además, en el proyecto, se propone el cálculo de la accesibilidad por carretera incorporada a las teselas cuadradas para solventar la debilidad de su topología radial (diferente distancia de las teselas vecinas en los lados y en los vértices).

El uso prioritario de estructuras de teselas proyectadas se completará con el diseño de un procedimiento para su integración en estructuras de teselas globales: **Discrete Global Grid Systems (DGGS).**

Entidades espaciales al máximo nivel de detalle necesarias para de gestión directa de las **aplicaciones temáticas (soluciones).** Estas entidades se han revelado críticas en el análisis y la gestión a escala de detalle. En ejemplo del uso de los edificios de catastro (atom) desarrollado por la universidad de Sevilla (proyecto CLOUDRISK_A) es revelador en este sentido (<https://universidad-sevilla.carto.com/u/univ-sevilla-admin/builder/734bde62-470f-4ade-a9a2-424c1b0558fb/embed>). Al menos, en una primera propuesta, se deberán contemplar:

Edificios de Catastro (formato ATOM) para las políticas urbanísticas, de vivienda, etc.

Subparcelas del SIGPAC por ejemplo para la gestión de PAC.

Montes públicos

Unidades de gestión ambiental para la planificación de espacios naturales protegidos, de riesgos o de gestión de agua, etc.

unidades de gestión sanitaria.

Otras

7.1.3 Grado de innovación

Con relación al carácter innovador que conlleva la generación de un amplio conjunto de datos multidimensionales normalizados espacialmente para toda la Comunidad Autónoma de Andalucía, en este reto, se podrían citar los siguientes:

En primer lugar, **el número y volumen de variables integradas espacialmente** (tanto en estructuras de teselas como en unidades espaciales de gestión) y el **acceso** a las mismas, de

forma operativa y garantizada a futuro, a través de una plataforma web. Aunque en los últimos años ha aumentado el número de experiencias en el uso de estructuras de teselas/rejillas, en ninguna de ellas se alcanza el número y la diversidad temática de las incluidas en este reto para implementar el proyecto CPI. Además, se avanza al incorporar, de forma complementaria, el uso de otras entidades espaciales de más detalle (edificios de catastro, subparcelas SIGPAC, etc.) mejor adaptadas a las necesidades de gestión. Finalmente, no existe en la actualidad una vía de acceso a esta multitud y diversidad de datos integrados espacialmente a través de una infraestructura web que permita adicionalmente su análisis, exploración y geovisualización y el desarrollo de aplicaciones web temáticas (para las soluciones temáticas seleccionadas).

En segundo lugar, **los procesos de integración espacial**, formalizados a través de diferentes algoritmos (tanto en estructuras de teselas como en unidades espaciales de gestión), **conlleven siempre un alto grado de innovación**, e incluso en algunos casos de I+D. En los casos en que la información original es de carácter puntual (plazas turísticas, estaciones de aforo, etc..) los procesos de integración son relativamente simples (solo existe la relación topológica de inclusión), al contrario de lo que ocurre en los casos de fuentes originales poligonales (datos estadísticos, catastro, de renta, etc.) donde los algoritmos de integración son bastante complejos y exigen el uso de técnicas dasimétricas específicas para cada variable. El número algoritmos de adscripción será elevado en este reto y, además, se le exigirá exhaustivos controles de validación y control de calidad. Este hecho, el uso de controles calidad y validación oficiales le proporcionan un valor adicional como métodos innovadores, a diferencia de otras experiencias previas.

Sin duda, la incorporación explícita de los **tiempos de accesibilidad por carretera (ya calculados)** a la estructura de teselas multiescalares y multidimensionales, **constituye uno de los principales rasgos innovadores** de este reto. La accesibilidad es un criterio clave en la toma de decisiones en políticas públicas, tanto desde una perspectiva económica (potencial de acceso a los mercados) como desde un punto de vista social (equidad en el acceso a los servicios) y ambiental (presión humana sobre espacios de alto valor ambiental). Aunque en una primera etapa, se utilizaron redes simplificadas, con orígenes y destinos que representaban unidades espaciales como secciones censales o zonas de transporte (en estudios urbanos) o municipios y provincias (en estudios regionales), en la actualidad es posible la utilización de redes exhaustivas, con topología compleja y velocidades de circulación medidas a través de dispositivos móviles, como Google Maps o TomTom Speed Profiles. Esos estudios, con multitud de relaciones origen-destino, redes de transporte complejas y cálculos precisos sobre tiempos de acceso, han sido llevados a cabo generalmente en ámbitos urbanos. Sin embargo, cuando se analizan áreas extensas con redes exhaustivas y multiplicidad de orígenes y destinos los tiempos de computación son mucho mayores. Esto es especialmente crítico en estudios de **accesibilidad a diferentes tipos de servicios**, que en principio exigen un cálculo diferente para cada tipo de servicios. La única experiencia previa de la incorporación de la accesibilidad a una estructura de teselas es la realizada por el Grupo de Investigación RNM-177 de la Universidad de Sevilla (en colaboración con otro grupo – T_GIS- de la UCM), que han trabajado en el proyecto INDALO en dar una solución al cálculo detallado de accesibilidad en áreas extensas, tomando como ejemplo la accesibilidad (presión antrópica) sobre las playas de Andalucía. Para la exploración masiva de los datos de accesibilidad se utilizaron geovisores y dashboards en la plataforma CARTO (ver <https://universidad-sevilla.carto.com/u/univ-sevilla-admin/builder/05418126-6c67-4b6d-ba38-9e36f3ac5366/embed>).

Sin duda, el reto asociado esta CPI es dar un paso más en este rasgo innovador **al ampliar el ámbito de análisis y extenderlo a toda Andalucía de forma masiva**. Es decir, lo que se

propone es que todas las teselas multiescalares de 250 m. o 125m. tengan calculadas los tiempos de accesibilidad por carretera (con redes exhaustivas, topología compleja y velocidades de circulación medidas a través de dispositivos móviles) contra todas las teselas del mismo nivel escalar, lo cual generaría matrices orígenes/destino muy voluminosas (1.700.000 x 1.700.000 para teselas de 250 m.), es decir 3.240.000 millones de registros (tesela origen, tesela destino, tiempo, orden) para las teselas de 250 m. y para un solo escenario. Teniendo en cuenta que se contemplan varios escenarios, y que la accesibilidad es esencial para muchas actividades de gestión, y crítica en alguna de ellas (servicios de salud, por ejemplo), su generación y posterior gestión en la plataforma (aplicación para el filtrado de los orígenes, destinos y escenarios) es un rasgo que contempla un altísimo grado de innovación.

7.2 Objetivo 2: Desarrollo de una plataforma con amplias capacidades para el almacenamiento, explotación y difusión de datos normalizados. Requisitos funcionales.

En este apartado se presenta un conjunto mínimo de requisitos que deben ser abordados y que están contextualizados en base al foco principal de AIDEA: el desarrollo e implantación de un sistema funcional que permita integrar espacialmente (teselas/rejillas multiescalares y multidimensionales e integración en recintos de gestión) y analizar de manera flexible los datos geoespaciales más demandados por y para la gestión medioambiental, y otras políticas públicas.

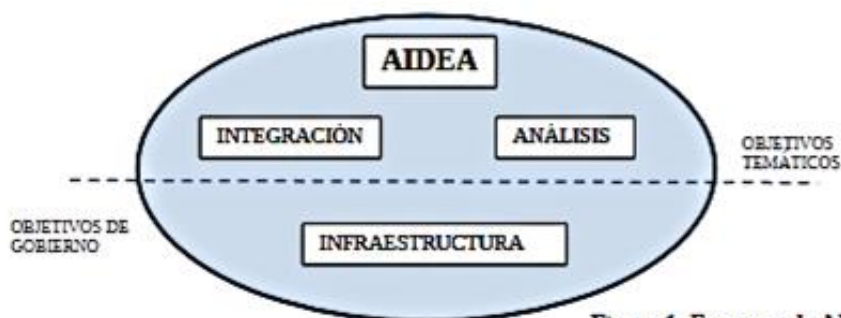


Figura 1. Esquema de AIDEA

Tal como se observa en la Figura 1, AIDEA representa la conjunción de tres dimensiones: Integración, Análisis e Infraestructura, que deben ser abordadas de manera integral en base a soluciones innovadoras para conseguir satisfacer un grado alto de flexibilidad en términos de extensibilidad, despliegue y costes de operación. En este contexto, se establecen los siguientes focos principales para cada una de las dimensiones:

El reto principal de la Integración es conseguir un modelo extensible de importación y almacenamiento e integración espacial unificado de distintos tipos de datos que permita una posterior explotación eficiente de la información, garantizando una evolución y manipulación controlada de los datos.

Se debe desarrollar un modelo de análisis de los datos integrados eficiente y escalable, que pueda ser compatible con distintos tipos de despliegue y modelos de computación de manera sensible al coste, así como permitir su acceso desde dentro y fuera de la organización.

Se debe conseguir una Infraestructura que dé soporte a la integración y análisis permitiendo la evolución controlada de los distintos artefactos desarrollados, estableciendo un modelo de gobierno

de la arquitectura y operación que gestione los distintos niveles de servicio y coste de manera automatizada, en base a políticas configurables.

Para ello, se proponen una serie de requisitos funcionales mínimos que debe abarcar la plataforma para garantizar los puntos anteriormente explicitados y organizados en bloques funcionales. Estos requisitos se asocian a dos grandes conjuntos de módulos funcionales, que se refieren, por un lado, a la propia gestión y mantenimiento de la plataforma (gobierno de la plataforma –GP-) y, por otro lado, a la gestión y tratamiento específico de la información espacial (tratamiento de datos espaciales –TDE-). Cada uno de los módulos funcionales de cada bloque serán nombrados como GP01, GP02...GPnn, o TDE01, TDE02...TDEnn.

Todo este conjunto de funcionalidades debe generar una serie de módulos que podrán agruparse y estructurarse de forma flexible de muy diferentes maneras, siendo deseable la opción de una **colección de microservicios** que gocen de cierta independencia, pero que, a su vez, se encuentren suficientemente coordinados a través de diferentes mecanismos de orquestación. A continuación, se exponen los requisitos mínimos señalados en diferentes bloques de objetivos.

Bloque I: Tratamiento de datos espaciales (TDE)

7.2.1. TDE01. Ciclo de vida del dato

El dato debe estar en el centro de gravedad de la plataforma a desarrollar. Por lo tanto, es esencial ofrecer una funcionalidad completa en cuanto a la gestión del ciclo de vida del dato. En este sentido, serán necesarios servicios que permitan gestionar, tanto la persistencia, como el registro de acceso a los datos almacenados, de manera que se pueda obtener un modelo de auditoría sobre los procesos y transformaciones de datos realizados en la plataforma, así como los actores responsables. Toda la operativa a este respecto deberá regirse de acuerdo con los estándares relevantes y lo definido en los Planes de Gestión de Datos (DMP) que necesariamente se definirán en la plataforma.

7.2.1.1 Estado del arte

Para la gestión del ciclo de vida del dato se han de tener en cuenta el extensivo trabajo existente en cuanto a buenas prácticas y guías publicadas por organizaciones internacionales (Comisión Europea, ODI, Data Curation Centre, etc.), como es el caso de la propuesta realizada para INDIGO-DataCloud en el contexto de LifeWatch (Aguilar, 2017) que define las actividades a contemplar en el ciclo de vida del dato.

Igualmente resulta de interés referenciar las iniciativas más ligadas a los aspectos temáticos de AIDEA, como por ejemplo el trabajo de Ivanova et al. (<https://doi.org/10.5194/isprs-archives-XLII-4-W20-33-2019>) en el que los autores reflejan la importancia de los principios FAIR aplicados a datos geoespaciales, o la necesidad de seguir las distintas iniciativas al respecto del metadatado de datos geoespaciales por parte de OGC o de INSPIRE.

7.2.1.2 Requisitos funcionales

Por un lado, el artefacto principal que debe regir el ciclo de vida del dato, esto es el Plan de Gestión de Datos de la plataforma, deberá ser pieza fundamental dentro de los servicios que se ofrezcan en la plataforma para la gestión del ciclo de vida del dato. A este respecto, se deben definir con precisión, para cada tipo y fuente de datos a utilizar, el propósito de su recopilación o generación, los formatos

que se usarán para su representación, si se reutilizarán datos existentes y de qué forma, el origen de dichos datos, su tamaño y la utilidad de los mismos en el contexto de los usuarios de potenciales de AIDEA, entre otras cuestiones.

Siguiendo con la referencia discutida anteriormente para INDIGO-DataCloud, se deberá dar soporte a todas las actividades del ciclo de vida a definir, entre las que necesariamente se deberán contemplar la planificación, la recopilación y adquisición de datos, la conservación y curación de manera que se asegure la correcta depuración, integración y normalización de los datos sin importar su fuente, el análisis o procesamiento, su consumo y publicación para facilitar la reutilización, y por último la preservación a largo plazo. El resto de los servicios de la plataforma tendrán necesariamente que integrarse con los servicios de gestión del ciclo de vida del dato de una manera transparente y eficiente.

Finalmente, se requerirá que la solución desarrollada se atenga a los principios FAIR, de manera que se asegure la capacidad de búsqueda, acceso, integración y reutilización de todos y cada uno de los datos disponibles en la plataforma tanto por el resto de los servicios de la misma como para usuarios externos. Para ello además de los procesos asociados se tendrá que definir el catálogo de metadatos a utilizar.

7.2.1.3 Grado de innovación

Conseguir un grado suficiente de automatización e integración de los procesos de gestión del ciclo de vida del dato, frente al resto de la plataforma, implicará un esfuerzo de innovación importante, en especial en cuanto a propuestas para incorporar los principios FAIR en la e-infraestructura. Existen distintas soluciones posibles que tendrán que adecuarse a los objetivos del sistema. Por otro lado, es de esperar que, para la auditoría del dato se puedan incorporar aspectos innovadores que promuevan un uso responsable de la plataforma, a la vez que incentiven el aprovechamiento y reutilización de los datos y procesos asociados.

7.2.2. TDE02. Importación y normalización de datos espaciales

Teniendo en cuenta que el éxito de este tipo de proyectos se apoya en la integración de una ingente variedad y tipología de datos espaciales que, a su vez, van a suponer enormes volúmenes de información, el diseño de herramientas **ETL (Extract, Transform, Load)** resulta un aspecto absolutamente crítico. Los procesos básicos que se contemplan en este tipo de herramientas son los de **extraer datos** de fuentes muy diversas, **transformar los múltiples formatos** que presentan todos estos datos para, finalmente, **almacenarnos de forma estructurada** en un repositorio común (**DATA LAKE**) que constituye la segunda pieza básica de este apartado.

7.2.2.1 Estado del arte

En la actualidad, la adopción de paradigmas de **BIG DATA** en proyectos vinculados a la información espacial ha requerido de esfuerzos intensivos e innovadores en el campo de las herramientas **ETL** (<https://doi.org/10.3390/ijgi8110475>), quedando claro, por el momento, que no existen soluciones comerciales que satisfagan las especificidades de proyectos como el que aquí se contempla y que, por tanto, se hacen necesarios diferentes tipos de desarrollos propios (https://univ-pau.hal.science/hal-03778251/file/Article_ETL_Accept%C3%A9.pdf) para generar módulos específicos (ya sean *on-premise* o tipo *cloud*), en los que el usuario, no necesariamente especialista, seleccione entre diferentes opciones en función de los tipos de datos a los que pretende acceder.

Por lo que se refiere a grandes bases de datos, o repositorios centrales de datos espaciales (**DATA LAKE**), el debate tecnológico se ubica en la elección de sistemas de gestión de bases de datos (SGBD) **SQL** o **NoSQL** (<https://ieeexplore.ieee.org/document/8258324>; <https://www.mdpi.com/2504-2289/7/2/97>). En el primer caso, se trata de tecnologías cuyo uso se ha generalizado en el campo de los datos espaciales, con proyectos como **POSTGIS**, **MYSQL**, **SQL-SERVER** que, en su momento, supusieron un salto cualitativo en la potencialidad de los SIG, pero que pueden ver limitadas sus capacidades cuando son requeridos procesos de despliegue *cloud* o de paralelización. Por tanto, es el segundo caso, esto es, los SGBD NoSQL (**CASSANDRA**, **MONGODB**, **NEO4J**), 1

[https://www.researchgate.net/publication/342472214 Comparative study of Spatial Hadoop and Geospark for Geospatial Big Data Analysis](https://www.researchgate.net/publication/342472214_Comparative_study_of_Spatial_Hadoop_and_Geospark_for_Geospatial_Big_Data_Analysis)

Al margen del debate central al que se ha hecho referencia, otros aspectos de gran relevancia en este contexto es el que se refiere a la indexación espacial (**SPATIAL INDEX**) que debe generarse en estas nuevas estructuras de datos, donde, con especial referencia a la alternativa NoSQL, aún no existen alternativas definitivamente generalizadas y aceptadas. (<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9779567>).

7.2.2.2 Requisitos funcionales

Por lo que se refiere a la generación de un **módulo de herramientas ETL**, tanto la diversidad de tipos de datos espaciales como la que se refiere a las propias fuentes de información que se requieren para nutrir al sistema, obligan a contemplar múltiples tipos de servicios de descarga, que se adapten a esa variada especificidad. Así, no solamente se trata de la **dualidad clásica de los datos espaciales (vector/raster)**, sino que debe contemplarse la enorme **variabilidad de formatos** en los que los mismos se encuentran (shp, geodb, gml, geotiff, json, csv, etc.), así como las **diferentes formas de acceso/descarga** que las fuentes de información ofrecen (carga en local, open data, servicios OGC, Apis de acceso, requerimientos de certificados, etc.).

Los datos que se relacionan en el epígrafe “Fuentes de información y variables a integrar espacialmente” del apartado 7.1.2 de este documento, pretenden únicamente servir de muestra de la diversidad de tipología, formatos y fuentes que el proyectos requiere, apelándose en cualquier caso a una lógica y esperable multiplicación de esta diversidad, conforme la plataforma vaya siendo requerida para atender a otras parcelas de gestión, tal como se ha comentado en apartados anteriores (medioambiente, salud, agricultura, transporte, etc). A partir de este amplio catálogo, por tanto, **el reto se centra en el diseño y desarrollo de un módulo ETL**, que asuma esta **diversidad** y que, al mismo tiempo, presente un importante grado de **flexibilidad y capacidad de ampliación**, considerando las necesidades futuras de información que puede presentar la plataforma.

Ahora bien, si en una primera instancia, debe diseñarse este módulo de descarga/importación con la suficiente diversidad de servicios para atender las múltiples tipologías de datos y fuentes de información, como un segundo paso deberán contemplarse diferentes mecanismos y herramientas que se dirijan a la **depuración y normalización** de los diferentes conjuntos de datos incorporados al sistema como datos originales. Efectivamente, la finalidad de ese primer módulo de descarga/importación es conseguir almacenar los datos originales tal como se obtienen de sus respectivas fuentes. No obstante, considerando que el objetivo último del proyecto se apoya en la integración de múltiples tipos de datos, se hace estrictamente necesaria una serie de pasos encaminados a una primera homogenización/normalización de estos datos. Este proceso debe contemplar tanto la **corrección de errores de naturaleza geométrica/topológica** o de **naturaleza alfanumérica**, como la interpretación de las diferentes colecciones de datos y su **normalización propiamente dicha, a través de su transformación en variables almacenadas en un formato**

homogéneo (por ejemplo, JSON o GEOJSON) correspondiente al repositorio común de la plataforma (**DATA LAKE**).

Este último aspecto mencionado constituye nuevamente un elemento crucial de la plataforma. Efectivamente, un gran repositorio central o **DATA LAKE** se convierte en una pieza de la máxima trascendencia en un proyecto de estas características, si se tienen en cuenta los dos requisitos, ya ampliamente descritos, de soportar una gran diversidad de tipos de datos espaciales y de mostrar capacidad para gestionar unos extraordinarios volúmenes de información, que sitúan el proyecto en el ámbito del **BIG DATA**. Otro de los grandes retos de este repositorio se relaciona con su **capacidad de comunicarse y asistir eficientemente al conjunto de servicios** que se apoyan en el consumo de estos datos (visualización, análisis, difusión, etc.) y que constituyen la parte del sistema más vinculada a la explotación de la información. Para ello, las opciones quedan muy abiertas, en un rango que iría desde esquemas que utilicen un tipo de vinculación más directa (repositorio-servicios), a otros que interpongan mecanismos de comunicación/relación intermedios que procuren mejores rendimientos (*dbases as services*).

7.2.2.3 Grado de Innovación

En cuanto al **módulo de herramientas ETL**, los esfuerzos en innovación se refieren a la propia generación de un catálogo de procedimientos de descarga de desarrollo propio, a partir de lenguajes tipo Python o Java, y librerías específicas de uso generalizado como pandas, geopandas, GDAL, tensorflow, etc. Especialmente de interés, resultarían los procesos de innovación en el acceso y descarga de datos a partir de fuentes con requerimientos de identificación, certificado, etc., persiguiendo la automatización de estos procedimientos que deberían resultar lo más transparente posible para los usuarios.

En el caso del repositorio central (**DATA LAKE**), dado el actual debate en torno a las opciones tecnológicas disponibles en un contexto de **BIG SPATIAL DATA**, e incluso de inconsistencia y volatilidad de dichas opciones, quizás las apuestas más innovadoras pueden ir de la mano del diseño de sistemas híbridos que no se apoyen en piezas monolíticas que comprometan todo el sistema.

7.2.3. TDE03 Adscripción de variables a unidades espaciales de referencia

Si un aspecto crucial de la plataforma es acceso flexible a los múltiples y numerosos conjuntos de datos espaciales necesarios para el proyecto, así como su almacenamiento estructurado en un repositorio central (**DATA LAKE**), otro de los aspectos clave es el que supone la integración espacial de las variables almacenadas. Para ello, ya se ha comentado que la estrategia pasa por la utilización de determinadas **unidades espaciales de referencia**, a la que se adscriben todas las variables disponibles en el sistema. De esta forma, una vez conseguida la homogenización de formatos a través de los procesos ETL y el almacenamiento en un repositorio común, ahora se trata de una homogenización espacial, por la que todas las variables comparten una misma expresión y representación espacial. Las unidades de referencia que se contemplan en el proyecto son de dos tipos: **unidades geométricas regulares** y **unidades administrativas o de gestión**. Por su parte, el proceso de adscripción supone el proceso de convertir o transformar información temática vinculada a las geometrías de los datos originales a las teselas definidas por la rejilla. Existen muchos posibles mecanismos de adscripción, que dependen fundamentalmente de la naturaleza de las dimensiones temática y geométrica de la información geográfica de origen y su interacción con las teselas determinadas por la rejilla (Pérez Alcántara, 2021)

En el caso de las **unidades geométricas regulares**, tal como se ha descrito en apartados anteriores, su principal virtud consiste en aunar los principios de **multidimensionalidad** y **multiescalaridad**. En el primer caso, el carácter multidimensional hace referencia a la posibilidad de contar en la misma unidad (celdilla), con todas las variables que espacialmente colisionen con ella en el proceso de adscripción. De esta forma, cada unidad cuenta con un conjunto diferente de **variables originales** (por ejemplo, superficie y población), así como con todas aquellas que puedan derivarse (**variables derivadas**), a partir de cálculos de distinto tipo realizados entre éstas (por ejemplo, población/superficie = densidad). En el segundo caso, la multiescalaridad se vincula a la posibilidad de realizar cambios de escala (resolución) en los conjuntos de celdillas (GRID), gracias a la autocontención (herencia) de este tipo de rejillas geométricas. Al margen de estas propiedades, la otra gran ventaja de este tipo de estructuras regulares es la de su idoneidad para la realización de operaciones de análisis complejos, tipo **machine learning**, constituyendo un escenario espacialmente adecuado para procesos de computación distribuida.

Por lo que se refiere a las **unidades administrativas o de gestión**, su principal virtud es la de hacer posible la integración multidimensional de información, en contextos con un significado y relevancia real desde el punto de vista de la gestión y las políticas públicas. Así, el poder contar con la distribución de determinadas variables originales o derivadas en un conjunto de municipios, secciones censales, distritos o parcelas catastrales, implica el conocimiento real del estado de un indicador o conjunto de indicadores en esos ámbitos de gestión y, por tanto, la posibilidad de agilizar la toma de decisiones en esos mismos ámbitos, en los que pueden existir valores de referencia o umbrales determinados según instrumentos normativos o de planificación y gestión.

El proceso por el que se asignan valores de una variable a las unidades de referencia se denomina adscripción, la cual, fundamentalmente, implica procesos de colisión espacial o lógica entre los datos espaciales y las unidades de referencia, y procesos de asignación de valores a partir de algún tipo de algoritmo.

7.2.3.1 Estado del arte

Las aportaciones y referencias esenciales en relación con la disponibilidad de diferentes opciones de unidades de referencia se han presentado en el apartado 7.1.

En relación con la adscripción de variables, como ya se ha comentado, los procesos de adscripción se componen de dos pasos fundamentales: de una parte, la colisión espacial o lógica entre los datos y las unidades espaciales de referencia y, de otra parte, la asignación de valores de una variable extraída de los datos mediante algún tipo de algoritmo predefinido (Pérez Alcántara, 2021 - <https://idus.us.es/handle/11441/116623>-). En este sentido, las opciones son muy variadas y se diferencian de forma apreciable en función de que se trate de datos espaciales tipo vectorial (<https://link.springer.com/article/10.1007/s12145-021-00737-9>;<http://www.efgs.info/wp-content/uploads/geostat/1a/GEOSTAT1A-final-report.pdf>) o tipo ráster (<https://doi.org/10.1080/15230406.2021.1966648>). Al margen de esta diferenciación, también existe una importante diferencia con relación a las unidades de referencia utilizadas, que se relaciona con la diferente naturaleza geométrica de las mismas, irregular, en el caso de las administrativas y regular en el de las estructuras GRID (<https://cran.r-project.org/web/packages/geomerge/readme/README.html>; <https://www.sciencedirect.com/science/article/pii/S1364815221002528>).

7.2.3.2 Requisitos funcionales

Tal como se ha expuesto, los requisitos funcionales de este objetivo se dirigen a las dos operaciones básicas que están implícitas en la adscripción de variables. Obviando el caso de las unidades de referencia administrativas, que en sí mismas constituyen datos espaciales ya existentes (municipios, secciones censales, parcelas catastrales, parcelas SIGPAC, etc.), una de las funcionalidades de la plataforma será la de **generar estructuras GRID** (unidades de referencia geométricas), incorporando una gran flexibilidad y diversidad de opciones. Así, deberán poder crearse **GRID locales**, con indicación de parámetros definitorios básicos (origen, resolución, tipo de celda -cuadrada, hexagonal, etc.-, o sistema de referencia -SRC-), y, de la misma forma, **GRID tipo DGGS** a partir de algunas de las opciones ya comentadas en el apartado 7.1.

Por otro lado, un grupo de funcionalidades básica y que constituye un **reto de primer orden**, se refiere a la **generación de un extenso catálogo de métodos de adscripción**, en el que, para los diferentes conjuntos de datos disponibles, tanto vectoriales (puntos, líneas y polígonos), como ráster, se expongan las distintas opciones en forma de algoritmos de asignación a unidades de referencia de los valores de las variables consideradas. Al margen de esta diferenciación en base a los modelos de datos básicos (vector y ráster), que determina en gran parte las diferentes formas de coincidencia o colisión entre objeto espacial y unidad de referencia (intersecta, contiene, etc.), otros aspectos a considerar por este catálogo de métodos de adscripción se refieren a la propia naturaleza de las variables que pueden encerrar una extensa tipología (discretas o continuas, cualitativas o cuantitativas, etc.).

7.2.3.3 Grado de Innovación

En este apartado los procesos de innovación se vinculan a los dos bloques funcionales de los que consta. Por lo que se refiere a la **generación de estructuras GRID**, los retos se dirigen al diseño de aplicaciones con altos grados de flexibilidad a la hora de generar rejillas a partir de un amplio conjunto de parámetros de entrada en los que pueda intervenir el usuario. De la misma forma, en relación con los procesos de adscripción, deben explorarse las diferentes opciones que minimicen la **persistencia** física de rejillas junto con los resultados de adscripciones, de forma que no se comprometa la capacidad del sistema. De interés también resultan la exploración y esfuerzos innovadores en relación a las estructuras que deben contemplarse para el almacenamiento de los **resultados de las adscripciones**, previsiblemente en forma de **vectores (clave, valor)** que constituyen la expresión del **carácter multidimensional** de las unidades de referencia que se utilizarán para las fases analíticas y de difusión de resultados. Por último, relacionado con los aspectos anteriores, la **gestión de la multiescalaridad en las estructuras GRID** una vez que estén dotadas de información temática (herencia multiresolución), resulta una cuestión absolutamente clave, en la que las actividades de innovación deberán intensificarse.

7.2.4. TDE04. Servicio de análisis en la parte servidora (CPU/GPU)

Estos servicios se orientan al análisis de las variables ambientales, territoriales y socioeconómicas que han sido integradas espacialmente en las rejillas multiescalares/multidimensionales o en recintos no regulares de gestión en el datalake centralizado. La estructura de adscripción (vectores de adscripción) ofrece una estructura que facilitará la utilización de servicios de análisis ligados a la construcción de indicadores, índices o clasificaciones basados en funciones analíticas del ámbito del análisis espacial o la Inteligencia Artificial y Machine Learning (Nikparvar, B.; Thill, J.-C, 2021). En este sentido, con relación a los procesos analíticos y la generación de nuevos datos derivados utilizando la información multidimensional de las teselas/rejillas multiescalares, dos aspectos podrían resaltarse:

Su estructura, al subdividir el espacio en teselas de “cuasi equalarea” facilitará su análisis con el procesado distribuido (parallel computing) ya que el área de estudio es fácilmente divisible y, posteriormente, los resultados del proceso analítico son fácilmente reensamblados espacialmente recuperando la topología implícita en la rejilla.

Por otra parte, el poseer cada rejilla un identificador único en cada nivel escalar, y tener asociados la totalidad de las variables que se seleccionen, proporcionará su información en una estructura parecida a un vector n-dimensional para cada tesela-rejilla y nivel escalar, que es una de las estructuras de datos que alimentan a los procesos analíticos de Machine Learning (ML) e Inteligencia Artificial (AI).

7.2.4.1. Estado de arte

El tratamiento masivo de datos geoespaciales y las necesidades computacionales para el desarrollo de análisis complejo de datos es un reto de primer orden en el diseño de una plataforma como AIDEA. Constituye por tanto un gran reto desde el punto de vista del diseño de la infraestructura y de la integración de soluciones que permitan un ecosistema flexible y eficaz para desarrollar la panoplia de análisis de datos geoespaciales requerida: análisis geoespacial (vectorial y ráster), generación de indicadores espaciales complejos y herramientas de Machine Learning (ML) y Deep Learning (DL). Estas últimas técnicas de análisis complejas han tenido un amplia aplicación en la información geoespacial, la búsqueda de patrones de distribución espaciales y el desarrollo de algoritmos de predicción de variables espacio-temporales (Ekeanyanwu et al, 2022). Son, por tanto, uno de los componentes esenciales que la plataforma AIDEA debe incluir en el catálogo de microservicios de análisis.

Varios son los elementos centrales que deben ser tenidos en cuenta en el desarrollo de las propuestas que deben recibirse del mercado:

En la actualidad muchos de estos servicios de análisis se basan en la utilización de un conjunto de librerías específicas para el desarrollo de los microservicios de análisis. En este sentido, según la vigilancia tecnológica realizada en relación con este objetivo, ha podido constatarse que existen distintos ecosistemas de software libre muy consolidados que están presentes en la mayoría de los casos, y se encuentra en el foco de los requisitos funcionales asociados a los servicios de análisis con varias soluciones óptimas, desde Tensorflow hasta ScikitLearn entre otros y a modo de ejemplo. La oferta de librerías específicas es enorme (Haan et al. (2023)) y cubre casi cualquier caso de uso. Estos ecosistemas de referencia que permiten los desarrollos propios que deben orientarse y adaptarse a las necesidades y requisitos del proyecto AIDEA que cubren, además, las necesidades de desarrollo de plataformas para el ML y la IA. Estos ecosistemas tienen también la virtud de hacer transparente el uso de las GPU en el caso de que estén disponibles para procesamientos aprovechando las capacidades de las tarjetas gráficas.

Un segundo elemento a tener en cuenta en los esquemas de desarrollos actuales es el diseño deseable de estas herramientas de análisis como microservicios para conseguir la máxima flexibilidad y extensibilidad del sistema. En los esquemas de desarrollo más innovadores, la heterogeneidad de sistemas es un valor en alza. La computación distribuida moderna se basa en tener la virtud de encapsular en microservicios con APIs coherentes, funcionalidades proporcionadas por software diverso que tienen su propio ciclo de vida y actualización. Los sistemas monolíticos basados en una única solución o lenguaje son menos eficientes, y restan flexibilidad a las plataformas. Soluciones DevOps como Docker y Kubernetes por ejemplo hacen que esto sea más fácil de mantener e implementar que nunca, y lo hacen una realidad factible.

La ingente cantidad de datos que van a ser manejados y su característica geoespacial van a obligar al sistema al desarrollo de procesos de computación en paralelo (parallel computing). Dado que el uso de funciones analíticas, en este contexto de spatial big data, es muy exigente en tiempo de computación dicho servicio debería proporcionar una alta escalabilidad y una estrategia clara de computación en paralelo (Werner M, 2019). La selección e integración de herramientas para las necesidades de computación en paralelo, asociadas a técnicas analíticas sobre datos masivos espaciales que requieran de dicho esquema de computación, son elementos críticos en los trabajos que han sido verificados. Por ejemplo, tecnologías como Hadoop-Spark (GeoSpark) que ejecutan procesamiento basados en las técnicas de Map-reduce con gestores de recursos como YARN u otros serán elementos a tener en cuenta en el desarrollo de un esquema de integración funcional. Se espera, por tanto, por parte del mercado, un ecosistema completo para un procesamiento analítico de alto rendimiento, realizado en la parte servidora de los datos geoespaciales masivos que albergará la plataforma.

Finalmente, un aspecto novedoso, y de gran calado para procesar analíticamente en paralelo conjuntos masivos de datos, es el proporcionado por las capacidades de las GPU. Las Unidades de Procesamiento Gráfico (GPU, por sus siglas en inglés) son componentes esenciales en la computación moderna, especialmente en tareas que requieren un alto rendimiento. Aunque inicialmente se crearon para renderizar gráficos en juegos y aplicaciones multimedia, su versatilidad ha llevado a su uso en campos como la ciencia de datos y la inteligencia artificial (Velentzas, P et al, 2023). El aprovechamiento de estas capacidades y su integración en un esquema de computación híbrido en la plataforma AIDEA son elementos esenciales en la consulta al mercado, por el exigente grado de innovación y desarrollo que presentan. Las GPU se han convertido en herramientas esenciales en la ciencia de datos debido a su capacidad para acelerar el procesamiento de datos. Algunas librerías especializadas en análisis de datos como TensorFlow y PyTorch, aprovechan las GPU para realizar cálculos de manera más eficiente, acelerando el cálculo de análisis complejos de ML y DL.

7.2.4.2. Requisitos funcionales

Los principales requisitos funcionales que deben ser tenidos en cuenta para garantizar un servicio eficaz de análisis en la parte servidora son los siguientes:

Capacidad de Procesamiento en Paralelo: el sistema debe ser capaz de dividir tareas en subprocesos que se ejecutan simultáneamente en múltiples nodos para mejorar la eficiencia y el rendimiento del procesamiento.

Distribución de Carga: debe ser capaz de distribuir de manera equitativa la carga de trabajo entre los recursos de procesamiento en paralelo para optimizar el rendimiento

Escalabilidad: debe permitir la gestión flexible de los recursos de procesamiento según sea necesario para manejar cargas de trabajo variables.

Manejo de Datos Distribuidos: debe ser capaz de gestionar grandes volúmenes de datos distribuidos de manera eficiente, incluyendo la sincronización y la comunicación entre nodos.

Gestión de la componente espacial: debe garantizar en los procesos de map-reduce la componente espacial de los datos y la topología implícita de las unidades espaciales de análisis. Del mismo modo, muchas de las operaciones de análisis espacial que pueden procesarse en paralelo son esclavas de las relaciones espaciales entre los datos.

Comunicación con el datalake centralizado: el servicio debe garantizar la conexión (lectura/escritura) entre los trabajadores y las capas de negocio o piezas intermedias necesarias y el almacén de datos. Esta conexión debe ser eficaz y permitir la lectura y escritura rápida de los datos hacia los trabajadores (lectura) y hacia el datalake (escritura de los resultados de los análisis).

Tolerancia a Fallos: debe contar con mecanismos para detectar y recuperarse de fallos en el hardware o el software, garantizando la disponibilidad y la confiabilidad del sistema.

Interfaz de Usuario: debe proporcionar una interfaz de usuario que permita a los usuarios definir tareas, monitorear el progreso y acceder a los resultados de manera intuitiva.

7.2.4.3. Grado de innovación

En relación con los aspectos innovadores que pueden asociarse a este objetivo cabe destacar los siguientes:

Diseño de un ecosistema funcional eficaz para la computación en paralelo de datos geoespaciales masivos (spatial big data). En sí mismo, la integración de tecnologías y los desarrollos a nivel de APIs que expongan los distintos microservicios de análisis, suponen un reto que exigirá, del mercado, una respuesta innovadora. Ést irá especialmente ajustada a la naturaleza de los datos integrados en rejilla, y permitirán un amplio espectro de analíticas complejas.

Dentro de ese diseño, los desarrollos de conexión entre el datalake y los trabajadores y soluciones de computación por los que se opte también suponen un reto importante y un grado de innovación e investigación altos.

La integración de la componente espacial en los procesos de map-reduce, para garantizar la integridad de los datos espaciales y de las operaciones de análisis espacial, así como los procesos y el desarrollo de algoritmos de indexado espacial, serán elementos que exigirán, también, un alto grado de innovación.

El diseño y desarrollo de un sistema de computación híbrido en el que se integren herramientas de procesamiento en GPU, así como la optimización de las capacidades de computación en GPU orientadas a la ciencia de los datos, es un elemento esencial de investigación e innovación dentro de este objetivo.

7.2.5. TDE05. Servicio de datos y de difusión.

Los servicios de datos y de difusión tienen como objetivo la publicación web de los diferentes servicios que permitirían a un usuario interno o externo el acceso a los datos. En este sentido, se han establecido dos vías complementarias:

Servicios OGC. Estos servicios serían los encargados de ofrecer cualquier dato de la plataforma AIDEA (original, normalizado, transformado o derivado) siguiendo los diferentes estándares y especificaciones OGC (<https://www.opengeospatial.org/>). Para cada tipología de datos integrados en la plataforma podrá optarse por el tipo de servicio que se considere más adecuado (WMS, WCS, WFS, WTMS, MVT, KML, etc), y podrá utilizarse, para, diferentes aplicaciones “open source”(map server, geoserver,..) o cualquier desarrollo propio que siga los estándares OGC.

Servicio de acceso a datos. Este tipo de servicios, permitirán el consumo directo de los datos nativos (datos integrados y datos derivados), existentes en las estructuras de teselas multiescalares y

multidimensionales, hacia las aplicaciones de la plataforma y, con los convenientes permisos, a aplicaciones externas preferentemente vía API para garantizar la máxima interoperabilidad con la plataforma. Junto a ello, un requisito esencial para la adecuada gestión del ciclo de vida de los datos, especialmente cuando se proceda a su difusión pública o privada (ingest/publish) para ser utilizados por otros usuarios (generalistas o especializados), debería ser un enfoque orientado a recursos, de manera que se asigne un “Identificador Persistente” al dato. Este identificador podría ser una URL (Uniform Resource Locator) que identifica la dirección en Internet usando el protocolo HTTP. De manera adicional, se pueden contemplar otro tipo de identificadores persistentes como el Digital Object Identifier (DOI). En cualquier caso, los identificadores únicos junto con los correspondientes metadatos deben ser integrados en el servicio del Ciclo de Vida del Dato.

7.2.5.1. Estado del arte

La difusión de la información espacial, y sus atributos a partir de estándares OGC, garantiza la interoperabilidad en el acceso a los datos de la plataforma. Para el acceso estandarizado a datos GIS se emplean, desde la aparición de los primeros estándares, los protocolos WMS, WFS, WCS, WMTS, etc.) del Open Geospatial Consortium (OGC). Estos se caracterizan por permitir la obtención de información como resultados de petición HTTP. Esta ha sido la vía interoperable de acceso a este tipo de datos, si bien, con el desarrollo de las modernas técnicas de definición de servicios web basada en OpenAPI y REST, en la actualidad su metodología de desarrollo ha quedado, en cierta medida, obsoleta.

Como una evolución natural de estos protocolos, el propio Consorcio ha comenzado a trabajar en el diseño de alternativas que proporcionen un servicio con los mismos fines, pero atendiendo a los diseños OpenAPI.

Los distintos grupos de trabajo del Consorcio se encuentran en la actualidad en un proceso de definición de estándares OpenAPI. Según se desprende de los desarrollos que van produciéndose, existen un conjunto de APIs para la web (APIs for the web), algunos de ellos con estándares ya aprobados y otros en desarrollo. La aplicación de estos nuevos estándares, basados en API, empiezan a ser aplicados e implementados en desarrollo de aplicaciones (Lehto, L. and Kähkönen, 2021).

Trabajar con estos estándares (algunos ya aprobados), y desarrollar herramientas que puedan implementar estos estándares, es un elemento clave que debe estar presente en las propuestas que procedan del mercado.

En relación con los servicios de datos nativos de la plataforma, basados en APIs, será fundamental el desarrollo de los mismo conforme a enfoques RESTful. Las propuestas deben ir además contextualizadas en el conjunto de APIs y orquestadores de las mismas que forman parte del gobierno de la infraestructura.

<i>Listado de OGC API</i>		
Estándar	Descripción	Estado
OGC API Features	Se trata de un estándar compuesto que ofrece la capacidad de crear, modificar y consultar datos espaciales features en la web, y especifica requisitos y recomendaciones para las API que desean seguir una forma estándar de compartir datos tipo feature. La especificación es un documento compuesto. La parte central de la especificación describe las capacidades obligatorias que cada servicio de implementación debe admitir y se limita al acceso de lectura de datos espaciales. Las capacidades adicionales que abordan necesidades específicas se especificarán en partes adicionales. Las capacidades futuras previstas incluyen, por ejemplo, soporte para crear y modificar datos, modelos de datos más complejos, consultas más avanzadas y sistemas de referencia de coordenadas adicionales	Aprobado

OGC API Common	Especifica las prácticas comunes que son compartidas por la mayoría o todos los estándares de API de OGC para garantizar la consistencia en toda la familia.	Aprobado
OGC API EDR	Proporciona una familia de interfaces ligeras para acceder a recursos de datos ambientales. Cada recurso abordado por una API de EDR se mapea a un patrón de consulta definido.	Aprobado
OGC API Tiles	Proporciona funcionalidad extendida a otros estándares de OGC API para entregar tiles vectoriales, tiles de mapas y otros datos tileados.	Aprobado
OGC API Records	Actualiza los Servicios de Catálogo y metadatos de OGC para la Web al aprovechar el acceso sencillo al contenido en OGC API - Features.	Borrador
OGC API Processes	Permite el empaquetado de tareas computacionales en procesos ejecutables que un servidor puede ofrecer a través de API y una aplicación cliente puede invocar	Aprobado
OGC API Coverages	Permite acceder, visualizar y consultar conjuntos de datos raster complejos y cubos de datos.	Borrador
OGC API Styles	Define una API web que permite a los servidores de mapas, clientes, así como a los editores de estilos visuales, administrar y recuperar estilos.	Borrador
OGC API Maps	Ofrece un enfoque moderno para el estándar de Servicio de Mapas Web (WMS) de OGC para la provisión de contenido de mapas con renderizado dinámico	Borrador
OGC API DGGS	Permite a las aplicaciones organizar y acceder a datos estructurados según un Sistema Global de Cuadrícula Discreta (DGGS).	Borrador
OGC API Routes	API web que permite a las aplicaciones solicitar rutas de manera independiente del conjunto de datos de enrutamiento subyacente, motor de enrutamiento o algoritmo.	Borrador
OGC API Joins	Permite la unión de datos procedentes de múltiples fuentes, con colecciones de entidades o directamente con otros archivos de entrada	Borrador
OGC API Moving Features	Define una API que proporciona acceso a datos que representan características de elementos móviles.	Borrador
OGC API 3D GeoVolumes	Gestión de contenido geoespacial 3D a través de Internet	Borrador
OGC API Connected systems	Actúa como un puente entre los datos estáticos y dinámicos recopilados por sensores.	Borrador
OGC SensorThings API	Interfaz unificada, abierta y habilitada geoespacialmente para interconectar los dispositivos, datos y aplicaciones de IoT a través de la web	Borrador

7.2.5.2. Requisitos funcionales

- La integración en el esquema tecnológico de una capa de negocio, en la parte servidora, que produzca servicios interoperables OGC (con propuesta de soluciones que deben tender a la reproducibilidad, y con poca o nula dependencia de procesos de generación de los servicios alejados de scripts o códigos de programación). Esta pieza tecnológica será la responsable de configurar y servir los estándares OGC geoespaciales: WMS, WFS, WCS, etc... en los que deben servirse los resultados del proceso de teselación e integración.
- Desarrollos siguiendo los estándares de API basados en OpenAPI, que está implementando la OGC (OGC API) para generar una colección de APIs de servicios que sigan dichos estándares, algunos de ellos ya aprobados y con end points definidos.
- Además de la producción de servicios interoperables basados en estándares OGC y OGC API, el sistema debe desarrollar una estrategia para servir los datos (vectores espaciales y atributos alfanuméricos) a los geovisores web, con el fin de aprovechar las ventajas del tratamiento de los mismos mediante librerías de análisis de datos js y herramientas de visualización avanzadas en la parte del cliente. Estas estrategias que deben definir el mercado se orientaran a la producción de servicios de tiles vectoriales (vector tiles), o geojson, por ejemplo. En cualquier caso, será crítico realizar pruebas de rendimiento sobre datos masivos en el territorio para poder adoptar la mejor estrategia de difusión web.
- El desarrollo de APIs de servicios de datos nativos en rejilla, y de accesibilidad para ser consumidos por clientes externos, se debe basar en las buenas prácticas y especificaciones de la

Open Knowledge Foundation y seguir los principios FAIR. Para ello se comprobará el uso de estándares, modelos semánticos aplicables y formatos reutilizables para facilitar a su vez que estas APIs puedan ser integradas por agentes externos

7.2.5.3. Grado de innovación

Con relación a los aspectos innovadores que pueden asociarse a este objetivo cabe destacar los siguientes:

Desarrollo de un modelo de difusión de datos geoespaciales en la web a partir de la implementación de un conjunto de APIs basadas en las normas (aprobadas o en borrador) OGC API. Para ello, desarrollos sobre servidores de mapas (Geoserver), o la utilización de librerías como pygeoapi, presentan un alto grado de innovación a considerar en el conjunto del objetivo.

La necesaria implementación de APIs propias para el acceso a los datos nativos de la plataforma, basadas en enfoques REST, y que supongan un nodo de difusión de los datos que posibilite que puedan ser consumidos por otras plataformas y aplicaciones, supone un reto que debe ser abordado con desarrollos propios y que debe ofrecer un alto grado de rendimiento, debido al carácter masivo de los datos.

La integración de estas APIs, y su gestión, a partir del desarrollo orquestado de APIs, supone también un ámbito de innovación en el que se esperan propuestas novedosas e innovadoras por parte del mercado.

7.2.6. TDE06. Servicio de Clientes web de acceso, análisis y explotación orientados a soluciones.

Este servicio se basa en el desarrollo de clientes ligeros flexibles, amigables e interactivos que, aprovechando los avances de las tecnologías más recientes (HTML5, CANVAS, WebGL..), permitan realizar, en el cliente, una parte de las funcionalidades que realizaban los técnicos SIG en aplicaciones desktop o directamente sobre la parte servidora. Con ello, se pretende reducir el gap tecnológico y las limitaciones que otras herramientas más sofisticadas imponen a los usuarios no especialistas en TIG. La interacción de todos los usuarios potenciales con la información importada, e integrada, en rejillas multiescalares, que el sistema albergará, se realizará necesariamente a partir de clientes web de geovisualización. Es por ello que el óptimo desarrollo de este objetivo es esencial en términos de utilidad y usabilidad del Sistema. Diseñar, por tanto, soluciones bien adaptadas a los distintos perfiles de usuarios, que deben ser identificados, consolidados y testados en las distintas fases de la implementación de la plataforma, es, de todo punto, uno de los retos, y uno de los objetivos del cual se espera un mayor número de aportaciones de la consulta al mercado, en su parte de definición de requisitos, integración de soluciones y aplicaciones desarrolladas para los casos de estudio, que puedan derivarse de los análisis de clientes potenciales que se llevarán a cabo, y que quedan descritos en el apartado 7.3 del presente documento.

7.2.6.1. Estado del arte

En los últimos años, en el contexto de las Tecnologías de Información geográfica, se han desarrollado espectacularmente los clientes web (geovisores y dashboard), los cuales proporcionan una gran interactividad por parte del usuario. Gran número de soluciones de alta calidad, especialmente orientadas a la difusión de la información ambiental y al *big data* espacial (Anthony C. Robinson et al, 2017) han sido desarrolladas por numerosas instituciones nacionales e internacionales. Estos clientes proporcionan una gran variedad de opciones para la visualización cartográfica multiescalar e

interactiva de los datos que, al ser tan numerosos y detallados, son difíciles de manejar con cartografía convencional. De hecho, estos clientes pasan a ser los equivalentes a los voluminosos anexos cartográficos que exigiría cualquier proceso de evaluación y gestión de los recursos naturales o de procesos ambientales.

Adicionalmente, la incorporación de nuevas tecnologías (html5, webGl, etc) en el desarrollo de estos clientes web incrementan la interactividad de los usuarios con los geodatos originales, de tal forma, que estas herramientas se convierten en magníficos instrumentos de exploración de los datos, con un amplio repertorio de posibilidades al incorporar diferentes tipos de widgets para la selección, filtrado y análisis de éstos sobre el propio cliente web. De esta forma, las nuevas posibilidades de geovisualización de la información y el desarrollo de herramientas interactivas (widgets) integradas en el cliente facilitan una nueva forma de interactuar con los datos, enfatizando el análisis exploratorio de datos espaciales (Sánchez, Martín & Rengifo, 2019).

La integración de componentes (frameworks como deck.gl, librerías para la geovisualización de datos espaciales como openlayers, maplibre, leaflet,..., librerías para la visualización de gráficos como d3.js, librerías para el análisis de los datos en cliente como tensorflow o componentes como react y superset, junto a los diferentes motores de renderizado en cliente, como por ejemplo svg, canvas o webgl) ofrecen, a modo de ejemplo, entre otros que puedan ser propuestos en la consulta al mercado, un amplio abanico de herramientas que permiten el desarrollo de clientes dotados de gran versatilidad, y que permiten, también, el diseño de clientes orientados a soluciones, que son los que se ajustan mejor a las necesidades del proyecto AIDEA. Se espera del mercado un ecosistema integrado y funcional de tecnologías que permitan el desarrollo de estas herramientas de geovisualización flexibles, orientadas a soluciones y con un alto grado de interactividad.

7.2.6.2. Requisitos funcionales

Los principales requisitos funcionales que deben ser tenidos en cuenta para garantizar un servicio eficaz de desarrollo de clientes web de difusión son los siguientes:

- Diseño de una solución (cliente web de geovisualización) global y genérica que permita el acceso rápido y fácil (mediante búsquedas orientadas sobre los metadatos (T00)) a las capas de información originales, derivadas e integradas en rejilla. Este cliente web genérico debe estar orientado a un perfil generalista y no altamente especializado. Por lo tanto, debe ser dotado de una estructura de acceso a los datos altamente intuitiva, además de cumplir con los máximos estándares de accesibilidad.
- Debe poder consumir servicios interoperables OGC externos
- Debe poder consumir los vectores de asignación de valores a las teselas de las rejillas multiescalares derivados de la ejecución del microservicio de teselado.
- Debe garantizar la asignación semiológica en el cliente por parte de los usuarios.
- Debe garantizar la utilización de librerías en el cliente especializadas en la generación de gráficos interactivos por parte de los usuarios a partir del consumo de los datos.
- Debe permitir la llamada a las funcionalidades más avanzadas, mediante la carga de los widgets desarrollados específicamente en el marco de este Proyecto (funcionalidad de Machine Learning o calculadora de indicadores, por ejemplo).
- Debe permitir un microservicio de gestión de componentes, y diseño de clientes orientados a soluciones

7.2.6.3. Grado de innovación

En relación con los aspectos innovadores que pueden asociarse a este objetivo cabe destacar los siguientes:

Desarrollos orientados a una eficaz integración de componentes que permitan una geovisualización dinámica de los datos.

Desarrollo de un microservicio para la gestión de componentes y el diseño de clientes orientados a soluciones.

Se espera el diseño de un cliente nativo para la geovisualización de datos integrados en rejillas, o en recintos de gestión, que presente un alto grado de eficacia en el manejo de grandes volúmenes de datos

También serán objeto de innovación los desarrollos que deben realizarse para la integración de herramientas de análisis de datos complejos en el entorno del cliente

Bloque II: Gobierno de la plataforma (GP).

7.2.7 GP01 Modelo flexible de computación en la nube

Para el correcto desarrollo y operación de la plataforma, se hace necesario definir un modelo de computación en la nube que sea suficientemente flexible para dar un soporte adecuado a los objetivos de AIDEA. Por un lado, se debe definir el nivel de hibridación a implantar, en cuanto a la tipología de nubes privadas y/o públicas a integrar en la plataforma, así como acerca de la granularidad del modelo de computación, para los servicios de la plataforma (directamente sobre infraestructuras en la nube, en contenedores, o como funciones). En caso de proponer un modelo compatible con la computación en nube pública se deberían tener en cuenta, a su vez, requisitos de anonimización y de herramientas para la toma de decisiones de manera automatizada, en función del coste económico y energético, de manera que se puedan estimar, predecir y optimizar dichos costes en infraestructuras de nube pública, soportando también la posibilidad de transferir los procesos entre distintos proveedores.

7.2.7.1. Estado del arte

Existe un conjunto de tecnologías asentadas para dar soporte a la computación en la nube y aportar cierto grado de flexibilidad, eficiencia y escalabilidad, como son el uso de contenedores para encapsular los servicios y gestionarlos mediante sistemas, como Kubernetes, que ofrecen facilidades para una orquestación flexible. Los distintos proveedores existentes ofrecen incluso servicios gestionados con estas características, aunque no soportan en general el nivel de hibridación que se pretende ofrecer en la plataforma AIDEA.

Centrándonos en el modelo de hibridación, existen acercamientos en la literatura que tratan la problemática de cómo optimizar la transferencia (offloading) entre proveedores de infraestructuras en la nube, o incluso, entre nube pública y privada ([Chunlin et al. 2017](#); [Wang et al. 2020](#); [Shalini Lakshmi y Vijayalakshmi 2021](#)). Por otro lado, los retos en la gestión de costes automatizada han sido también ampliamente tratados, sirviendo estudios como los de [Wu et al. 2020](#) o [García et al. 2021](#) como base para proponer soluciones en este ámbito, además de trabajos específicos sobre anonimización o eficiencia energética en la nube ([Mastelic et al. 2014](#); [Berl et al. 2010](#)) que son de utilidad para conocer el estado del arte.

7.2.7.2. Requisitos funcionales

De manera general, la plataforma a desarrollar deberá cubrir los siguientes requisitos funcionales para soportar un modelo flexible de computación en la nube:

- Modelo de gestión automatizada de la transferencia y despliegue multi-cloud: La plataforma debe ser capaz de, dinámicamente, gestionar los despliegues de los servicios incluidos en la misma, de manera independiente a la infraestructura donde se vaya a ejecutar (nube privada o cualquier proveedor de nube pública), además de ser capaz, en función de las métricas que se determinen, de transferir entre las distintas modalidades de despliegue referidas. Idealmente se debe, también, poder soportar modelos de computación distintos a los habituales, como es la ejecución de funciones (serverless).
- Modelo de gestión automatizada de costes: Los costes asociados a los servicios de la plataforma deben poder ser gestionados por ésta, de manera que se pueda monitorizar, en todo momento, los costes incurridos por el uso de los recursos de computación asociados. Este sistema deberá incorporar métricas apropiadas, e información proporcionada por los proveedores externos, de forma que sea posible tener un control exhaustivo y monitorización de los costes, así como poder estimar o predecir potenciales costes en función de las necesidades de computación esperadas. Esto último permitirá soportar la toma de decisiones a la hora de gestionar las políticas de transferencia y despliegue en los proveedores más convenientes.
- Modelo de anonimización para nube pública: Las políticas asociadas con las tipologías de datos sensibles son de suma importancia al utilizar infraestructuras de nube pública, por lo que la plataforma debe proporcionar mecanismos para asegurar la anonimización de datos, cuando sea necesario por las aplicaciones finales desarrolladas.
- Modelo de análisis del impacto energético: En conjunción con los costes económicos, la plataforma debe proporcionar servicios para poder analizar el impacto energético de la ejecución de los servicios desarrollados, de forma que puedan utilizarse un conjunto de métricas asociadas a este aspecto, para la toma de decisiones en cuanto a las infraestructuras a adquirir para soportar la ejecución de la plataforma.

7.2.7.3. Grado de innovación

Sin duda existe un margen de innovación claro en cuanto a la automatización de la transferencia y despliegue en entornos híbridos multi-cloud y multi-proveedor, principalmente si se integra con una gestión automatizada de costes. De esta manera, se espera que se desarrollen mecanismos innovadores para realizar esta gestión global de la computación en la nube de manera flexible, y automatizada, en función de las políticas de costes económicos y energéticos que se definan para cada servicio de la plataforma, tomando como base los avances tecnológicos y metodológicos actuales en este ámbito.

7.2.8 GP02 Modelo flexible de gestión de la arquitectura de microservicios

Como se ha dicho anteriormente, las distintas capacidades de la plataforma AIDEA serán ofrecidas siguiendo una arquitectura de microservicios. El conjunto de éstos debe ser gestionado de una manera flexible, y que permita una evolución controlada de la plataforma. En particular, se debe establecer un conjunto de buenas prácticas que, se propone, deben estar principalmente basadas en el uso de especificaciones y estándares. Igualmente se debe dotar de un modelo de gobierno del ciclo de vida de las APIs, haciendo especial hincapié en la gestión de acuerdos de nivel de servicios de las APIs que permitan operar de manera automatizada la plataforma para conseguir optimizar el uso de la infraestructura. Así mismo, se debe dar soporte a un esquema integral de seguridad y a la gestión del catálogo de APIs, de forma que pueda ser evolucionado de manera controlada durante el ciclo de vida de la plataforma.

7.2.8.1. Estado del arte

De cara al seguimiento de buenas prácticas en el diseño de arquitecturas de microservicios resulta esencial referenciar la [Open API Initiative](#) y su conjunto de especificaciones y herramientas asociadas que pueden resultar de interés para el establecimiento de un buen modelo de gestión de la arquitectura pretendida, así como para la definición y mantenimiento catálogo de servicios de la plataforma. Sobre las tecnologías disponibles se encuentran extensiones en la literatura, como por ejemplo SLA4OAI ([Gamez-Díaz et al. 2019](#)), u ontologías que pueden ser de interés ([Karavisileiou et al. 2020](#)).

Igualmente, importante son los aspectos relativos a la elasticidad y escalabilidad de la arquitectura, de manera que, en función de los acuerdos de nivel de servicios que se tengan, se pueda gobernar adecuadamente la plataforma. Algunas referencias reseñables en este ámbito son las de [Copil et al. 2013](#) centrado en la elasticidad de aplicaciones en la nube o las más generalistas de [Cáceres et al. 2010](#) y [Lehrig et al. 2015](#), sin olvidar las capacidades inherentes que el modelo de computación en la nube desarrollado en virtud del objetivo GP01 pueda proporcionar.

Finalmente, los aspectos de seguridad y privacidad son cada vez más relevantes a la hora de desarrollar y poner en marcha una plataforma de computación en la nube compleja, como es AIDEA, por lo que se deben tener en cuenta, tanto los mecanismos punteros para aportar el nivel de seguridad deseado ([Zhou et al. 2010](#)), como las oportunidades y retos existentes ([Mazhar et al. 2015](#)) para aportar soluciones novedosas en esta área.

7.2.8.2. Requisitos funcionales

Los principales requisitos funcionales a contemplar para el diseño e implementación del modelo de gestión de la arquitectura de microservicios de AIDEA son los siguientes:

- Aplicación de buenas prácticas de diseño de microservicios en base a estándares: La plataforma deberá establecer un conjunto de buenas prácticas para el diseño de los microservicios incluidos y asegurar su cumplimiento. Necesariamente se deberán tener en cuenta especificaciones asentadas en la industria, como Open API Specification, posiblemente aportando extensiones y reutilizando o adaptando herramientas que impulsen el seguimiento de dichas buenas prácticas.
- Gestión del catálogo de microservicios: para la gestión del catálogo de microservicios existentes en la plataforma se deberá adoptar una metodología concreta, en función de las herramientas más adecuadas y siguiendo las buenas prácticas adoptadas.
- Gestión de acuerdos de nivel de servicio: Como parte de la funcionalidad de catalogación, se debe incluir la posibilidad de asociar acuerdos de nivel de servicio que puedan definir las características no funcionales esperadas, y garantizadas para los microservicios de la plataforma. Este aspecto está íntimamente relacionado con lo desarrollado en el objetivo GP03.
- Gestión de la escalabilidad de los microservicios: En base al conjunto de métricas que se definan en los acuerdos de nivel de servicio, la plataforma debe ser capaz de autorregularse de forma que se asegure en todo momento el cumplimiento de las garantías definidas. Estas características servirán de puente entre los desarrollos relativos al GP01 y los de este objetivo GP02.
- Gestión de un modelo de seguridad integral: Los mecanismos de seguridad deben ser primordiales para el desarrollo de la plataforma AIDEA, y, por tanto, se deben incorporar como parte de las buenas prácticas a la hora de diseñar los servicios, y también como una funcionalidad horizontal para toda la plataforma.

7.2.8.3. Grado de innovación

Por un lado, la gestión del catálogo de servicios en general debe incorporar aspectos innovadores, integrando las iniciativas en cuanto a las especificaciones y extensiones que se han reseñado en el estado del arte. De esta forma, se espera que la plataforma proporcione un catálogo que vaya más allá de las herramientas existentes, poniendo el foco en aspectos como proporcionar acuerdos de nivel de servicio suficientemente expresivos y relevantes para el dominio en el que se engloba AIDEA.

Por otro lado, respecto a la funcionalidad relacionada con la escalabilidad la plataforma a desarrollar tendrá necesariamente que aportar innovaciones, donde se aproveche la integración de los acuerdos de nivel de servicio en el propio catálogo, extendiendo el modelo de computación del GP01 para dar soporte a una escalabilidad inteligente, en base a las métricas relevantes que se definan en dichos acuerdos.

7.2.9 GP03 Gestión integral del nivel de servicio de la plataforma

El nivel de servicio de la plataforma en general debe ser garantizado, tanto en relación con los acuerdos de nivel de servicio (ANS) de los propios microservicios que se proporcionan, como con los ANS del soporte a los usuarios de la plataforma. En este contexto, se debe poder explicitar los objetivos de nivel de servicio planteados, así como la plataforma de gestión del seguimiento, monitorización y alerta de las garantías explicitadas a fin de detectar violaciones de los ANS de forma automatizada. Dado que la plataforma tendrá que gestionar recursos económicos, humanos y técnicos finitos y fluctuantes, el sistema deberá ser capaz de adaptar dinámicamente el coste total de propiedad (TCO), así como proporcionar un sistema de gestión de servicios, que no sólo siga las buenas prácticas en la industria (ITSM), sino que soporte la flexibilidad necesaria en este escenario.

7.2.9.1. Estado del arte

La gestión de servicios ITSM es un campo en el que existe una amplia bibliografía y buenas prácticas extendidas en la industria, destacando prácticas como ITIL y multitud de herramientas software de soporte. En este ámbito, algunas referencias que extienden algunos aspectos como el cálculo del OpEx o el TCO ([Andreo et al. 2021](#)). En el ámbito de la gestión de ANS de microservicios existen referencias relevantes sobre telemetría ([Li et al. 2019](#)), gestión de la capacidad ([Fadda et al. 2021](#)) y operacionalización de ANS ([Leonildo et al. 2018](#)), además de un conjunto de estándares asociados como [OpenTelemetry](#) o [SLA4OAI](#).

7.2.9.2. Requisitos funcionales

Para la gestión de ANS en la plataforma, tanto aplicados a los microservicios desarrollados como para los servicios de soporte a usuarios, se tendrán que cumplir los siguientes requisitos funcionales:

- Infraestructura de gestión automatizada de ANS: La plataforma deberá ofrecer soporte a la definición de ANS, tanto de los microservicios desarrollados, como de los servicios de soporte a usuarios. La gestión y actualización de los ANS podrán ser realizadas por usuarios no técnicos, de manera que se puedan ajustar y estimar los costes de operación asociados a la plataforma. Este requisito se relaciona con los descritos en el GP02.
- Modelo de definición de ANS de soporte: La gestión de los servicios de soporte deberán poder hacerse en base a ANS personalizables, por lo que se deberá dar soporte a un modelo suficientemente expresivo y extensible.

- Soporte al análisis de cumplimiento y evolución de los ANS: El sistema debe permitir monitorizar la efectividad de los ANS, utilizando cuadros de mando que permitan tomar decisiones para la evolución y mejora continua de los ANS a todos los niveles.
- Armonización de la gestión de costes y ANS asociados: El sistema deberá tener en cuenta el grado de cumplimiento de los ANS, así como el TCO para la toma de decisiones en cuanto a la escalabilidad (GP02) y el uso efectivo de nubes híbridas (GP01).

7.2.9.3. Grado de innovación

Se espera que se incorporen tecnologías y estándares punteros en la definición y operacionalización de los ANS asociados tanto a los microservicios desplegados en la plataforma como para el soporte a usuarios. Para ello será necesario incorporar innovaciones en las herramientas usuales de gestión de servicios, así como proporcionar una solución transversal con el resto de los objetivos de gobierno de la plataforma que tenga en cuenta la tipología de usuarios de la misma.

7.2.10 GP04 Modelo de interacción y construcción extensible de aplicaciones de análisis

Para impulsar el uso de la plataforma por actores interesados en las aplicaciones de análisis como las indicadas en la sección 7.3, se hace necesario ofrecer una modelo de interacción apropiado con la plataforma para estos usuarios potenciales. Esto implica abordar nuevas metodologías de definición y desarrollo de interfaces de usuario multimodales, incluyendo componentes visuales, interfaces conversacionales, y cuadros de mando que permitan desarrollar dichas aplicaciones de análisis y evolucionarlas en el tiempo de vida de la plataforma, sin olvidar aspectos como accesibilidad, adaptabilidad y soporte multidispositivo.

7.2.10.1. Estado del arte

En lo referente a componentes visuales y mashups, existen trabajos centrados en el dominio de aplicación (Zang et al. 2019) o sobre metodologías para el desarrollo de este tipo de visualizaciones complejas (Vesyropoulos et al. 2018). Por otro lado, hay bastantes referencias centradas en el diseño de cuadros de mando y el análisis visual, como, por ejemplo, proponiendo metodologías para el diseño de cuadros de mando multi-modales (Chowdhury et al. 2021), o más centradas en el dominio de aplicación (Young & Kitchin 2020).

Recientemente, las tecnologías relacionadas con el procesamiento de lenguaje natural (NLP) y la inteligencia artificial generativa están en auge, proporcionando formas alternativas de interacción con las aplicaciones gracias a técnicas, como la extracción de información textual, tecnologías semánticas, grandes modelos de lenguaje y transformadores. Así, frameworks como IBM Watson o DialogFlow para el desarrollo de chatbots o IAs generativas, como GPT o Bard, son herramientas de referencia para facilitar la interacción y el desarrollo de aplicaciones de análisis más inteligentes y adaptables a las necesidades de los usuarios finales.

7.2.10.2. Requisitos funcionales

Los principales requisitos funcionales que cubrir en la plataforma en cuanto a la metodología de interacción y desarrollo de aplicaciones concretas de análisis son los siguientes:

- Metodología de definición y análisis de requisitos: Se hace necesario establecer métodos para obtener los requisitos específicos para las aplicaciones de análisis que se desarrollarán. Esto incluye modelar los aspectos más importantes, utilizando herramientas como ontologías o

grafos de conocimiento, para guiar el desarrollo de estas aplicaciones y las formas de interacción esperadas.

- Modelos de interacción multimodal: La plataforma debe contemplar, no sólo representaciones visuales de la información y los procesos de análisis desarrollados, sino, también, ofrecer la capacidad de interactuar conversacionalmente con un asistente virtual que genera las visualizaciones y cuadros de mando necesarios para responder a las necesidades expresadas.
- Arquitectura flexible para la gestión de componentes visuales: El sistema debe ser capaz de mantener un catálogo de componentes visuales que pueda evolucionar con el tiempo, permitiendo el uso de distintas tecnologías para la construcción de frontends, independientes del proveedor, y con soporte multi-dispositivo.
- Sistema de creación de cuadros de mandos adaptables: La plataforma debe proporcionar un sistema que permita la integración de distintos componentes visuales de manera dinámica, dependiendo de la metodología de definición anterior y las entradas proporcionadas por los usuarios en lenguaje natural o estructurado.

7.2.10.3. Grado de innovación

El mayor impulso de innovación en este ámbito se debe centrar en la disponibilidad de componentes visuales y de interacción basado en lenguaje natural que pueda impulsar la generación de nuevas aplicaciones de análisis cercanas a los requisitos de los usuarios finales. La integración de técnicas de NLP, ontologías y grafos de conocimiento, e IA generativa dentro de las aplicaciones supondría un grado de innovación muy importante en este ámbito de aplicación, automatizando la generación de cuadros de mando que den respuesta a las necesidades expresadas en distintas formas.

7.3 Objetivo 3: Desarrollo de aplicaciones (productos) de valor añadido para el soporte a la gestión en ámbitos específicos de los servicios públicos.

En este apartado se exponen algunas ideas en relación a potenciales aplicaciones específicas que pueden derivarse del proyecto, encaminadas a contribuir a la mejora de la prestación de servicios por parte de diferentes departamentos de la administración pública.

En principio, se identifican algunas de estas aplicaciones con carácter demostrativo del potencial que puede tener AIDEA como plataforma que favorece, como gran punto de partida, el desarrollo de soluciones específicas para abordar temáticas muy concretas, pero de gran relevancia.

En ese sentido, se propone un conjunto discreto de aplicaciones a presentar al mercado, con objeto de recibir información de este sobre capacidades de desarrollo, interés y recursos a comprometer. No obstante, durante la CPM, como un bloque diferenciado, se buscará la participación de gestores de distintos servicios públicos, para que aporten información sobre otras aplicaciones específicas, y sobre requerimientos de éstas.

7.3.1 Ámbito sanitario.

El ámbito de la salud constituye uno de los pilares básicos de las políticas públicas que deben atender las administraciones. En el caso de la aplicación de herramientas TIG a este apartado de la gestión pública, cabe distinguir entre dos facetas de gran importancia. De una parte, todo lo que se refiere a la planificación en los servicios y dotaciones que deben atender la demanda asistencial en materia de salud y, de otra parte, una faceta más operativa que se relaciona con la epidemiología y la atención a determinadas crisis y emergencias sanitarias.

7.3.1.1 Planificación de dotaciones y servicios sanitarios.

Desde el punto de vista de la planificación de servicios y dotaciones de salud, un conocimiento exhaustivo de la localización, distribución y características de la población andaluza y su contraste con los servicios y dotaciones de salud puede contribuir a un diagnóstico certero de las demandas no cubiertas. Para ello, es necesario contar con esa caracterización rigurosa de la **distribución espacial de la población**, con la **localización y caracterización** (tipología, espacialidades, capacidad, etc) **de los servicios de salud**, y con una **relación espacial entre ambas** en términos de **accesibilidad**. Los resultados inmediatos deben arrojar información sobre **demandas poblacionales no cubiertas**, **saturación de servicios sanitarios**, **localización óptima de dotaciones generales y específicas**.

Existen precedentes y estudios en el campo de la investigación en este ámbito, que demuestran la utilidad de las herramientas TIG en la planificación sanitaria,

(<https://rio.upo.es/xmlui/handle/10433/1534>;

<https://hum.unne.edu.ar/investigacion/geografia/labtig/publicaciones/public21.pdf>). No obstante las características y resolución de los datos que se contemplan en este proyecto, así como el conjunto de herramientas que pretenden diseñarse, abren unas excelentes oportunidades para una exhaustiva determinación de la oferta y demanda sanitarias, a través de una serie de **análisis de accesibilidad** de gran exactitud. La versatilidad de estos datos y herramientas no solamente posibilita este tipo de estudios a nivel de **población residente**, sino que los **datos de movilidad y estacionalidad poblacional** permiten otros enfoques de gran utilidad al examinar **incrementos de demanda** y, por tanto, **zonas de saturación sanitaria**.

7.3.1.2 Epidemiología y crisis/emergencias sanitarias.

Desde los primeros estudios de salud que hicieron uso de la información espacial para la gestión de crisis sanitarias (<https://ij-healthgeographics.biomedcentral.com/articles/10.1186/s12942-015-0011-y>), hasta las recientes demostraciones de la utilidad de las TIG en el análisis y gestión de la pandemia COVID-19

(<https://www.sciencedirect.com/science/article/pii/S0048969720335531>;
<https://www.mdpi.com/1660-4601/18/5/2336>), son innumerables las evidencias que ponen de manifiesto la idoneidad de las aplicaciones de estas tecnologías en este campo. Al margen de todo lo comentado en el punto anterior, que también resultaría de utilidad, en esta línea de aplicación sanitaria debe resaltarse la posibilidad de **integrar de forma rápida y eficaz información en tiempo real sobre afectado**, así como cuantos **factores ambientales** o de otro tipo pueden tener significación, según los especialistas, para explicar patrones de afección, expansión de contagios, etc.

7.3.2 Agricultura.

Sin lugar a dudas el ámbito de la agricultura constituye, junto con medio ambiente, uno de los campos en los que la aplicación de las TIG presenta una mayor tradición. Desde los primeros mapas de cultivos y aprovechamiento, pasando por los iniciales usos de la teledetección en la extracción de estadísticas agrarias o la detección de plagas, hasta llegar a los actuales ejemplos de la agricultura de precisión (https://oa.upm.es/56889/1/Fructicultura_N_68_pags_30_39.pdf), resultan innumerables los casos de aplicaciones y herramientas TIG que se han desarrollado en este campo. De hecho, las posibilidades de creación de productos específicos que se deriven del proyecto son muy numerosas y estarían a expensas de las principales demandas que se expongan por parte de los responsables en esta materia. En cualquier caso, es evidente que uno de los conjuntos de datos que está llamado a jugar un papel muy relevante en este proyecto es el que se deriva del **SIGPAC**

(<https://www.juntadeandalucia.es/agriculturaypesca/sigpac/index.xhtml>), de entre los que cabe destacar la propia red de **recintos PAC**, que se convierten en uno de los tipo de **unidades administrativas de referencia** a la que se adscribirá buena parte de las variables residentes en el sistema. A partir de esta información, y solamente por precisar algunas aplicaciones potenciales, deben mencionarse las que se refieren a la asistencia a los **servicios de estadísticas agrarias** (<https://www.juntadeandalucia.es/organismos/agriculturaypescaaguaydesarrollorural/servicios/estadistica-cartografia/actividad/detalle/175051.htm>;
<https://www.mapa.gob.es/es/estadistica/temas/estadisticas-agrarias/agricultura/avances-superficies-producciones-agricolas/>) o el diseño y desarrollo de **modelos predictivos de producciones** a partir de la integración espacial de múltiples variables explicativas y con aplicación de técnicas de de big data y machine learning (<https://helvia.uco.es/handle/10396/22527>); https://www.mapa.gob.es/images/es/bigdatabynse_tcm30-376202.pdf).

7.3.3 Ámbito forestal.

En el caso de la gestión de los espacios y recursos forestales, también las TIG han jugado un papel relevante desde un primer momento (<http://aet.org.es/congresos/viii/alb24.pdf>) y en la actualidad siguen jugando un papel de gran relevancia, especialmente vinculado a la utilización de **datos de teledetección** y **LIDAR** (<https://zagan.unizar.es/record/61353/files/TESIS-2017-037.pdf>; https://www.miteco.gob.es/content/dam/miteco/es/biodiversidad/formacion/presentaciongeoforest_tcm30-429780.pdf), el desarrollo de **modelos predictivos de biomasa** (https://www.juntadeandalucia.es/medioambiente/portal_web/web/temas_ambientales/montes/ usos_y_aprov/jornadas_biomasa/Publicaciones/biomasa1.pdf; <https://digibuo.uniovi.es/dspace/handle/10651/45077>) o la **gestión de incendios forestales**, tanto en las fases de **elaboración de cartografía de riesgos** (<https://www.juntadeandalucia.es/medioambiente/portal/areas-tematicas/incendios-forestales/extincion-de-incendios-forestales/indice-riesgo-incendio-forestal-andalucia>), la **colaboración en emergencias** en la labores preventivas y durante las labores de extinción (<https://www.juntadeandalucia.es/medioambiente/portal/documents/20151/589228/sigdif.pdf/b7ecd aa5-717b-d7c9-15ab-63065ad87ecf?t=1434960587000>; <https://prevencionincendiosgva.es/>), o en la intervención en los momentos posteriores para la **evaluación de daños y recuperación de espacios afectados** (<file:///C:/Users/Usuario/Downloads/Dialnet-ElUsoDeLosSistemasDeInformacionGeograficaEnLosMode-4245448.pdf>; <https://docta.ucm.es/entities/publication/ca778045-7de7-401e-8fb5-292622fc7579>)

7.3.4 Ámbito de reto demográfico.

El cambio demográfico es uno de los mayores retos a los que tiene que enfrentarse la Comunidad Autónoma de Andalucía. Son muchas las razones de este cambio, ya que se trata de un fenómeno demográfico que es necesario estudiar desde la perspectiva integradora del territorio. Para ello, deben combinarse variables demográficas como el (crecimiento vegetativo negativo, saldo migratorio negativo, envejecimiento) así con variables territoriales (tipología del poblamiento, pendiente del terreno), económicas (nivel de renta, acceso a infraestructuras) o medioambientales (reducción de precipitaciones anuales, incremento de la erosión).

Además, se trata de un tipo de cambio diferente según los territorios de Andalucía. En el valle del Guadalquivir y la costa, la población crece, enfrentando el reto de gestionar la concentración de habitantes y la demanda intensiva de servicios, mientras que áreas de Sierra Morena o de la Cordillera Bética presenta problemas de despoblación, y se enfrenta al reto de gestionar la dispersión de la población y la dificultad para acceder a los servicios o las infraestructuras.

Las herramientas y soluciones tecnológicas propuestas en este proyecto abren la posibilidad de un acercamiento integrador al reto demográfico. Permitirán abordar el análisis desde distintas escalas,

con distintas aproximaciones en función del ámbito territorial concreto a estudiar y facilitarán la integración de multitud de variables de distinta resolución espacial.

8. PRESUPUESTO ESTIMATIVO

Uno de los objetivos fundamentales de someter el presente proyecto a un proceso de consulta con el mercado, consiste en recabar información sobre las capacidades y la disposición de las empresas, y agentes del conocimiento, para abordar los retos tecnológicos y de innovación que supondría su puesta en marcha, y, a partir de la evaluación de la información recibida, plantearse la viabilidad de su contratación, y las condiciones en que ésta debería realizarse.

Con ese propósito, una de las informaciones de mayor relevancia, que se espera recoger, será la de la estimación de los presupuestos necesarios para el desarrollo del proyecto, por lo que, en los cuestionarios, se solicitará una estimación de los mismos a los participantes, y, además, se hará diferenciadamente por diferentes apartados (bloques).

En el momento de lanzar la consulta la CSMAEA no se está en disposición de establecer un presupuesto cerrado para el desarrollo de un potencial servicio, sobre el que se ciñen muchas incertidumbres de índole tecnológica. No obstante, pese a no poder establecer una cifra concreta, si se estima que, por la envergadura del proyecto, las necesidades presupuestarias de una plataforma como AIDEA se sitúan en el intervalo comprendido **entre 5 y 10 millones de euros**.

En la estimación del presupuesto, se ha considerado, no únicamente el desarrollo de la solución, sino también su despliegue y mantenimiento operativo en la Agencia Digital de Andalucía durante un plazo de dos años. Para este despliegue se ha estimado una capacidad de almacenamiento en nube privada de 150 Terabytes, una capacidad en la nube pública equivalente a unos 50 Terabytes y otros 100 Terabytes para procedimientos de sincronización entre ambas.

Sin ser exhaustivos, pero en aras de facilitar el cálculo para la citada estimación, tienen que venir incluidos los gastos asociados al servicio de instalación teniendo presente que todo el equipamiento ha de ser enrackable, las licencias necesarias (sistemas operativos, software backup, software gestion, virtualización,), el equipamiento de interconexión LAN y SAN (Switches, cables, puertos y licencias necesarias), los costes inherentes al CPD (consumo eléctrico, alquiler de huellas, etc...), los equipos con características de redundancia física (controladoras, fuentes de alimentación, etc...), el suministro de cabina de almacenamiento y el suministro de Racks necesarios para toda la instalación.

Sobre la base de la capacidad de realizar el mantenimiento operativo por parte del proveedor, todo lo especificado en el párrafo anterior ha de ser gestionable de forma remota.

El despliegue implica que los equipos estarán bajo RCJA por lo que la estimación del presupuesto ha de contemplar los costes de conexión.

El despliegue se efectuará con las normas y procedimientos habilitados para ello pero, con independencia del modelo de continuidad del proyecto que se pueda adoptar, se ha de contemplar en el coste la transferencia de conocimiento que incluye, entre otros, los recursos humanos (y sus competencias) necesarios para posterior operación.

En relación con la capacidad de proceso, se han estimado unos 8 nodos de procesamiento en la nube privada, para el desarrollo de los procesos programados: 8 nodos, con dos CPUs cada uno (al menos 8 cores cada CPU) junto con, al menos, 8 GPU.

Por lo que respecta a la nube pública, ésta atenderá situaciones específicas de necesidades de grandes de computación, y estará sometida a procesos elásticos y escalables, en función de los requerimientos del análisis a realizar.

En lo que afecta al número de usuarios, éste será muy cambiante, pero a modo orientativo, se ha estimado que los gestores que puedan acceder con regularidad a la e-infraestructura, pueden estar en torno a 250.

9. INDICADORES DE IMPACTO A CONSIDERAR

9.1 Multiplicar el número de gestores que realizan análisis integrados de datos espaciales para la mejora de los servicios públicos.

No existen estudios que reflejen el nivel de uso de los sistemas de información territoriales en la administración andaluza. Sin embargo, en los trabajos preparatorios de la iniciativa Cloud_IA, se realizaron procesos de consultas y entrevistas entre todos los departamentos de las Consejerías responsables de la gestión medioambiental y agrícola, para la identificación de aplicaciones TIGs que impliquen análisis de mayor o menor complejidad con diferentes conjuntos de datos (clasificaciones, extracción de estadísticas, geoprosos, etc). Como resultado de éstas, se constató que los ejemplos de este tipo de aplicaciones eran muy puntuales, normalmente externalizados, y que se realizan, generalmente, para abordar trabajos de gran envergadura.

De la misma forma se detectaron muchos departamentos donde este tipo de aplicaciones eran inexistentes, y otros, en los que tienen identificados modelos y explotaciones que se realizan en otros organismos afines, o procedentes del mundo del sector del I+D, y que querrían aplicar, pero que no pueden hacerlo por falta de personal especializado.

Por último, se hizo patente que, en el trabajo cotidiano, existen multitud de necesidades de información que podrían ser resueltas mediante sencillas operaciones de intersección de coberturas o de geoprosos, que, existiendo datos y herramientas, no se llevan a cabo por falta de capacitación, o por el exceso de tiempo que requiere ese tipo de análisis.

El proyecto AIDEA, con las soluciones que plantea, permitirá generar herramientas y clientes de explotación personalizados, y de fácil uso, que pondrá al alcance de cualquier usuario el realizar los análisis de cualquier tipo de complejidad, integrando cualquiera de los miles de variables que estarán disponibles de forma amigable para un amplio elenco de gestores. Éstos tendrán a su alcance muchas soluciones, de mayor o menor complejidad (cálculo de indicadores, herramientas de explotación geoespacial de determinadas fuentes de datos, herramientas de modelización, etc.), lo cual redundará en impactos positivos sobre los servicios públicos.

9.2 Ampliar las capacidades de análisis y explotación de los datos espaciales.

Como se indicó con anterioridad, existen importantes limitaciones en los sistemas actuales, que imponen importantes restricciones de capacidades de computación y de manejo de grandes conjuntos de datos, al tener muy dificultada la computación paralelizada. De este hecho deriva que, solo en casos muy puntuales, e invirtiendo una gran cantidad de recursos humanos y de tiempo, se lleven a cabo análisis de esas características.

Con las capacidades de Spatial big data y de despliegue cloud que ofrece AIDEA, se abrirá la posibilidad de realizar explotaciones que impliquen enormes conjuntos de datos, los cuales, ya estarán preprocesados para facilitar su integración espacial. Igualmente, estas estructuras de datos permitirán alimentar procesos de explotación basados en técnicas de machine learning e inteligencia artificial, que, a día de hoy son anecdóticos.

Como consecuencia, se abre la puerta a aplicaciones que puedan explotar variables procedentes de grandes conjuntos de datos, como Catastro, Sig Pac o Lidar, para ámbitos extensos (cuena hidrográfica, provincia, etc), que, actualmente, son escasas o inexistentes.

Igualmente se favorecerá el desarrollo de aplicaciones multivariantes que exigen grandes capacidades de almacenamiento y de computación, basadas, o no, en las nuevas técnicas de inteligencia artificial y machine learning.

En definitiva, se pone a disposición de los usuarios nuevas capacidades que les permitan, de una manera sencilla, la realización de estudios y análisis de datos hasta ahora inaccesibles, por su volumen, o por su complejidad computacional.

Una aportación fundamental, derivada de la armonización espacial de todo tipo de variables, con independencia de su naturaleza, va a ser la facilidad para realizar análisis integrado entre parámetros socioeconómicos y físico-ambientales. Este tipo de estudios, cada vez más demandado, tiene una escasa presencia en el ámbito de la geoinformación, dada la heterogeneidad de los conjuntos de datos a analizar conjuntamente. Con AIDEA se espera reducir enormemente la complejidad de llevar a cabo estudios de coste-beneficio de infraestructuras y planes sectoriales, análisis de daños económicos de desastres naturales, mejorar evaluación de la vulnerabilidad socioeconómica según riesgos derivados de escenarios de cambio climático, etc, convirtiéndose en una potente herramienta para introducir la dimensión espacial en las interrelaciones de variables estadísticas.

9.3 Multiplicar el uso y explotación de grandes conjuntos de datos

Actualmente, por parte de muchos organismos, se producen y mantienen grandes conjuntos de datos espaciales, que, por su volumen y complejidad de manejo, son escasamente utilizados en procesos de análisis integrados para amplios territorios. Es el caso de fuentes de información tales como Catastro, SIOSE, LiDAR, DTM, SIPNA, SIGPAC, bases de datos climáticas, etc.

Con la implementación de AIDEA se facilitará que, estos conjuntos de datos, en cuya producción y mantenimiento se invierten grandes cantidades de recursos, puedan ampliar su impacto al ser utilizables en procesos de integración con otras variables de forma accesible para un gran número de usuarios.

9.4 Favorecer el uso de la Inteligencia Artificial en la explotación de datos espaciales.

En junio de 2023 se aprobó la Estrategia Andaluza de Inteligencia Artificial, la cual persigue avanzar hacia un desarrollo de las capacidades de nuestra región en esta materia. Este documento estratégico señala, en su Diagnostico, que “*el 80% del tiempo de los proyectos de IA se invierte en preparar los datos*”. Es por ello por lo que, entre las líneas estratégicas que propone, se encuentra la de facilitar la preparación de los datos para ser consumidos por herramientas de análisis que hacen uso de algoritmos de IA y machine learning.

El datalake de datos normalizados que ofrece AIDEA, en forma de vector unidimensional, junto con las herramientas de análisis y capacidades de procesamiento, se convierte así en una plataforma idónea para el desarrollo de experiencias e iniciativas que hagan uso de estas novedosas técnicas de análisis, y que podrán ser llevadas a cabo por gestores de distintos servicios públicos, por agentes del conocimiento, o por empresas del sector TIG.

9.5 Mejorar la toma de decisiones sobre aspectos de los servicios públicos vinculados a los tiempos de accesibilidad a cualquier tipo de elemento.

La accesibilidad es un criterio clave para muchas políticas públicas, tanto desde un punto de vista social (igualdad en el acceso a los servicios), económico (potencial de acceso a los mercados) como) y ambiental (presión humana sobre espacios de alto valor ambiental).

En la actualidad es posible la utilización de redes exhaustivas, con velocidades de circulación medidas a través de dispositivos móviles, para ámbitos urbanos. Sin embargo, cuando se analizan áreas extensas, con redes exhaustivas y multiplicidad de orígenes y destinos, los tiempos de computación son mucho mayores, y muchas veces inabarcables, razón por la cual no se provee este tipo información.

AIDEA plantea la generación de un módulo que proporcionará la accesibilidad de forma masiva, para todas las teselas de la Comunidad Autónoma. Es decir, todas las teselas multiescalares de 250 m. o 125m. tendrán calculadas los tiempos de accesibilidad por carretera contra todas las teselas del mismo nivel escalar.

La potencialidad de análisis integrado con multitud de otras variables en el contexto de la e-infraestructura, con carácter exhaustivo para todos los puntos de la región, constituye un elemento novedoso, que abre la puerta a multitud de soluciones útiles para la mejora de muchos servicios públicos: la provisión de servicios públicos según accesibilidad de la población a éstos, la evaluación de la movilidad, el cálculo de emisiones por transporte terrestre, los costes de transporte de biomasa y otras producciones forestales, caracterización de parcelas agrícolas según distancias a centros de venta, procesado o de distribución, y para muchos otros estudios.

AIDEA dará respuesta a esta necesidad que es transversal a muchos servicios públicos.

9.6 Favorecer el mercado de generación de soluciones basadas en datos espaciales para empresas del sector TIG.

AIDEA se plantea como una herramienta abierta, flexible y escalable, con capacidad para permitir el acceso a los datos, y soluciones, a usuarios externos, e, incluso, para introducir, de una forma controlada, nuevos desarrollos de análisis.

A ese respecto, las empresas del sector TIG, aprovechando sus prestaciones, verán favorecidos los procesos de generación de clientes y geovisores avanzados centrados en la gestión de variables de interés para un servicio público concreto, para el propio sector empresarial y para la ciudadanía.

9.7 Impulsar la I+D+i en el sector TIG

Al poner a disposición de los agentes del conocimiento un gran repositorio de datos, previamente preparados para su análisis mediante técnicas avanzadas, como machine learning e IA, se favorecerá notablemente las investigaciones y el desarrollo de innovaciones basadas en tecnologías de la

información geográfica, impulsando un sector con gran potencial de aplicación en multitud de disciplinas.

BIBLIOGRAFÍA

Anthony C. Robinson, Urška Demšar, Antoni B. Moore, Aileen Buckley, Bin Jiang, Kenneth Field, Menno-Jan Kraak, Silvana P. Camboim & Claudia R. Sluter (2017) Geospatial big data and cartography: research challenges and opportunities for making maps that matter, *International Journal of Cartography*, 3:sup1, 32-60, DOI: 10.1080/23729333.2016.1278151

Ekeanyanwu, C. , Obisakin, I. , Aduwenye, P. and Dede-Bamfo, N. (2022) Merging GIS and Machine Learning Techniques: A Paper Review. *Journal of Geoscience and Environment Protection*, 10, 61-83. doi: 10.4236/gep.2022.109004.

Haan et al. (2023). Geodata-Harvester: A Python package to jumpstart geospatial data extraction and analysis. *Journal of Open Source Software*, 8(89), 5205. <https://doi.org/10.21105/joss.05205>.

Lehto, L. and Kähkönen, J.: OGC API Features HTML-output as a Feature Dashboard, *Abstr. Int. Cartogr. Assoc.*, 3, 176, <https://doi.org/10.5194/ica-abs-3-176-2021>, 2021.

Nikparvar, B.; Thill, J.-C. Machine Learning of Spatial Data. *ISPRS Int. J. Geo-Inf.* 2021, 10, 600. <https://doi.org/10.3390/ijgi10090600>

Pérez Alcántara, J.P. Plataforma cloud para la integración espacial de geoinformación en estructuras de teselas multiescales asimétricas, su análisis y su visualización. Tesis doctoral US. <https://idus.us.es/handle/11441/116623>

Sánchez Martín, J. M., Martín Delgado, L. M., & Rengifo Gallego, J. I. (2019). La reputación online de los alojamientos rurales en Extremadura desde una óptica geoestadística. *Boletín De La Asociación De Geógrafos Españoles*, (82). <https://doi.org/10.21138/bage.2758>

Velentzas, P., Vassilakopoulos, M., Corral, A. et al. GPU-Based Algorithms for Processing the k Nearest-Neighbor Query on Spatial Data Using Partitioning and Concurrent Kernel Execution. *Int J Parallel Prog* (2023). <https://doi.org/10.1007/s10766-023-00755-8>

Werner M. Parallel Processing Strategies for Big Geospatial Data. *Front Big Data*. 2019 Dec 3; 2:44. doi: 10.3389/fdata.2019.00044